SIAM J. CONTROL OPTIM.
Vol. 36, No. 1, pp. 1–11, January 1998

ESTIMATES OF CONVERGENCE OF FULLY DISCRETE SCHEMES FOR THE ISAACS EQUATION OF PURSUIT-EVASION DIFFERENTIAL GAMES VIA MAXIMUM PRINCIPLE*

PIERPAOLO SORAVIA†

Abstract. We consider the Dirichlet problem for the Isaacs equation of pursuit-evasion differential games and prove estimates of convergence for a numerical scheme associated with it. The function that we approximate is the Hölder continuous viscosity solution of the problem, and the direct proof that we propose relies on some ideas related to the maximum principle for viscosity solutions.

Key words. pursuit-evasion games, Isaacs equations, numerical schemes, estimates of convergence, viscosity solutions

AMS subject classifications. 65N15, 49L25, 90D26, 65N06

PII. S0363012995291865

Introduction. In this paper we study the numerical approximation of the Dirichlet problem

$$\begin{cases} \mu v(x) + \min_{b\in B} \max_{a\in A}\{-f(x,a,b)\cdot Dv(x)\} - 1 = 0, & \mathbb{R}^N\backslash\mathcal{T}, \\ v(x) = 0, & \mathcal{T}, \end{cases} \tag{0.1}$$

where $\mu > 0$ and $\mathcal{T}$ is a nonempty closed set. It is known (see in various generalities Bardi and the author [6], [7], and [22]) that if the vector field f satisfies local-controllability-type assumptions on $\partial\mathcal{T}$, there is a unique Hölder continuous, bounded viscosity solution of the problem (0.1). Such a solution is nonnegative and the Kruzkov transform $v = (1 - \exp(-\mu T))/\mu$ of the capture time function T, the (lower) value of the pursuit-evasion differential game with dynamics

$$\dot{y} = f(x,a,b), \quad y(0) = x, \tag{0.2}$$

and target $\mathcal{T}$ (we refer to the mentioned papers for the actual definition of T). Therefore the pair (v,μ) can be thought of as a tool for the real problem we are interested in, which is the computation of T. From the numerical point of view, it is not convenient in general to study directly the equation for T since such function is not necessarily bounded and, as a matter of fact, may not be even finite at all points of the space. This gives us a reason for the change of variables.

As usual, for dynamic programming equations of control problems (in which case A or B is a singleton) or differential games (see, e.g., Falcone [15] for control problems, Bardi, Falcone, and the author [4], [5] for games), to recover an approximation scheme for the differential equation in (0.1) we proceed in two steps. First we approximate the directional derivative $f(x,a,b)\cdot Dv(x)$ as $\mu\tau/(1-\tau)\,(v(x+hf(x,a,b)) - v(x))$, where $\tau = \exp(-\mu h)$ and h is a time step for the underlying dynamical system (0.2)

*Received by the editors September 18, 1995; accepted for publication (in revised form) August 21, 1996.
http://www.siam.org/journals/sicon/36-1/29186.html
†Dipartimento di Matematica Pura e Applicata, via Belzoni, 7, 35131 Padova, Italy (soravia@galileo.math.unipd.it). This paper was written while the author was visiting the University of California at Santa Barbara. This research was partially supported by Consiglio Nazionale delle Ricerche of Italy.

of the differential game. This approximation looks natural when thinking that our main goal is to compute the value function T. As a result we obtain the equation

$$(0.3)\quad \begin{cases} v(x) + \tau \min_{b\in B} \max_{a\in A}\{-v(x+hf(x,a,b))\} - (1-\tau)/\mu = 0, & \mathbb{R}^N\backslash\mathcal{T}, \\ v = 0, & \mathcal{T}. \end{cases}$$

This is what we call the discrete-time approximation of (0.1) since the unique bounded solution of (0.3) is the Kruzkov transform of the value function of the discrete-time pursuit-evasion differential game in which in the dynamics (0.2) we substitute its Euler approximation with step h. The convergence of the solutions of (0.3) to the viscosity solution of (0.1) has been proved by Bardi and Falcone [3] for control problems and by Bardi and the author [8] for games.

Next we discretize the space as a countable union of simplices whose maximum diameter is k, the space step, and approximate equation (0.3) by computing it only at the vertices of the simplices. If $\mathcal{G} := \{x_i\}_{i\in\mathbb{N}}$ indicates the countable family of the vertices, we are led to consider

$$(0.4)\quad \begin{cases} v(x_i) + \tau \min_{b\in B} \max_{a\in A}\{-\sum_j \lambda_{ij}(a,b)v(x_j)\} - (1-\tau)/\mu = 0, & \mathcal{G}\backslash\mathcal{T}, \\ v(x_i) = 0, & \mathcal{T}, \end{cases}$$

where the coefficients λ_{ij} allow us to rewrite the point $\overline{x}_i = x_i + hf(x_i,a,b) = \sum_j \lambda_{ij}x_j$ in a unique way as a convex combination of the vertices of the simplex it belongs to. This is what we call the fully discrete approximation of (0.1). Equation (0.4) has a unique bounded solution which can be computed by finding the fixed point of a contraction functional, as we recall in the next section. For more details about the algorithm to actually compute the solution of (0.4), for control problems we refer to [15] and Capuzzo Dolcetta and Falcone [10], and for games we refer to [4] and the forthcoming review paper [5] for some numerical computations. In [10] the reader can also find details about an acceleration technique that can be applied to speed up the computation of the fixed point which would otherwise be rather time consuming. As one of the referees pointed out to us, another way to increase the speed of the algorithm could be to allow control and state dependent time steps. This extension would also be interesting by itself, but in order to keep technicalities at a minimum and present a rather concise proof, we preferred not to deal with such a generality in this presentation.

In this paper we show that under controllability assumptions on the vector field on $\mathcal{T}$, as long as the ratio of the steps k/h remains bounded, the solutions of (0.4) converge uniformly to the viscosity solution of (0.1). Moreover we compute explicitly the rate of convergence. When, for example, (0.1) has a Lipschitz continuous solution, as it happens under reasonable assumptions on the data (see (1.3) below), the rate is $\sqrt{h}$. In more general cases it is h^ρ, where $\rho = \min\{\beta, \alpha/2, \gamma/2\}$, γ is the Hölder exponent of the solution of (0.1), and α, β are parameters of a discrete estimate of the solutions of (0.4) at the boundary; see, however, the precise statement of the Theorem in section 1 and Remark 2. The Hölder exponent of the solution can be deduced from the controllability conditions we require. A proof of convergence of the algorithm in very general assumptions, also showing that the computations can be restricted to a bounded region, was already given in [4]. The main new contribution here concerns the estimates of convergence, remarking, however, that the different argument we give in our assumptions is easier than the one in [4] and our presentation is self-contained. Our results are new even for the minimum time problem in control theory.

Estimates of convergence of the same type were previously obtained in [15] for infinite horizon control problems, i.e., for equations in the whole space. In [15] the proof proceeds in two steps establishing first the rate of the convergence of the discrete-time approximations of (0.1) and next combining it with the rate of convergence of the fully discrete approximations to the solution of (0.3). In the case of a Dirichlet problem, the solution of (0.3) is discontinuous even if the solution of (0.1) is smooth and this makes this method difficult to implement. Instead, we directly compare the fully discrete approximations and the viscosity solution of (0.1) by means of a maximum-principle-type argument which follows the idea proposed for discrete-time approximations of infinite horizon control problems by Capuzzo Dolcetta and Ishii [11], adapted, of course, to deal with the boundary condition and the fully discrete scheme.

We finally want to mention that the approximation of Hamilton–Jacobi equations in the context of viscosity solutions was also studied by Crandall and Lions [13] and [14], Souganidis [24], Gonzalez and Rofman [16], Alziary de Roquefort [1], Barles and Souganidis [9], Pourtallier and Tidball [20], and Kocan [17]. Some results concerning the numerical approximation of discontinuous solutions of (0.1) can be found in [4]; Bardi, Bottacin, and Falcone [2]; and the paper by the author [23]. For general results concerning approximations of solutions of Hamilton–Jacobi equations with different methods, we refer to Kushner [18], Kushner and Dupuis [19], and the references therein.

1. Preliminaries and main result. We start this section with the precise assumptions we need. The function $f : \mathbb{R}^N \times A \times B \to \mathbb{R}^N$, where A, B are compact sets, is supposed to be continuous and to satisfy

$$\begin{cases} |f(x,a,b) - f(z,a,b)| \le L|x-z|, \\ |f(x,a,b)| \le L \quad \text{for all } x,z,a,b. \end{cases} \tag{1.1}$$

We are also given the closed set $\mathcal{T} \subset \mathbb{R}^N$, which is the target of the pursuit-evasion differential game with dynamics (0.2). The target may be nonsmooth and unbounded. We assume that the Dirichlet problem (0.1) has a unique continuous, bounded, nonnegative viscosity solution v. With this we mean that v satisfies the boundary condition, and for all test functions $\varphi \in C^1(\mathbb{R}^N)$ such that $v - \varphi$ attains a local maximum (resp., minimum) point at $x_0 \in \mathbb{R}^N \backslash \mathcal{T}$ we have

$$\mu v(x_0) + \min_{b\in B} \max_{a\in A} \{-f(x_0,a,b) \cdot D\varphi(x_0)\} - 1 \le 0 \text{ (resp., } \ge 0).$$

Continuous viscosity solutions exist if and only if the so-called small time local controllability of the pursuit-evasion game to $\mathcal{T}$ is satisfied. In this case, they are moreover Hölder continuous. Small time local controllability can be obtained by means of suitable assumptions on the direction of the vector field f and of its Lie brackets at the boundary of the target $\mathcal{T}$ (see, e.g., [6], [7], the paper by the author [22], and the references therein); see also (1.3) and the following discussion. More informations about the theory of viscosity solution as well as lots of references on the subject can be found in Crandall, Ishii, and Lions [12].

We now consider a discretization of the space with space step k, meaning a family of simplices $\{S_j\}_{j=1,2,\dots}$ such that $\mathbb{R}^N = \cup_j S_j$, $\text{int}(S_i) \cap \text{int}(S_j) = \emptyset$ for $i \ne j$, $\max_j\{diam(S_j)\} = k$. We denote by $\mathcal{G} = \{x_j\}_{j=1,2,\dots}$ the family of the vertices of the triangulation, we suppose that it has no accumulation points, and we remark that any

point $x \in S_i$ can be expressed, in a unique way, as a convex combination of the vertices of S_i; that is, $x = \sum_j \lambda_j x_j$, $\sum_j \lambda_j = 1$, $\lambda_j \in [0,1]$, $\lambda_j = 0$ if $x_j \notin S_i$. Given the time step h, for all a, b we indicate $\overline{x}_i = \overline{x}_i(a,b) = x_i + hf(x_i,a,b)$ and define the parameters $\lambda_{ij} = \lambda_{ij}(a,b)$ considering the unique convex representation $\overline{x}_i = \sum_j \lambda_{ij}(a,b)x_j$ with the properties above. For convenience we choose $h, k \in (0,1]$. Then we define the map $F : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^{\mathbb{N}}$ by setting

$$F_i(V) = \begin{cases} \tau \max_{b\in B} \min_{a\in A} P_i(a,b,V) + (1-\tau)/\mu & \text{if } x_i \in \mathcal{G}\backslash\mathcal{T}, \\ 0 \quad \text{if } x_i \in \mathcal{T}, \end{cases}$$

where $P_i(a,b,V) = \sum_j \lambda_{ij} V_j$. It is clear by construction that a fixed point of F is a solution of (0.4). The following result outlines the idea of the algorithm. Its easy proof is left to the reader; the detailed argument, however, can be found in the papers by Falcone [15] and Bardi, Falcone, and the author [4] and [5].

LEMMA 1. *Assume* (1.1). *The above map F is monotone with respect to the partial order $U \geq V$ if $U_i \geq V_i$ for all $i \in \mathbb{N}$ and has a restriction $F : [0, 1/\mu]^{\mathbb{N}} \to [0, 1/\mu]^{\mathbb{N}}$ which is a strict contraction.*

As we mentioned, the previous result applies to proving that the fixed point of the map F is a bounded, nonnegative solution of (0.4). Such function is defined only at grid points, so we extend it to the whole space by linear interpolation. The resulting function, which we denote by $w^{h,k}$, is our continuous approximation of the solution of (0.1). By construction $w^{h,k}$ satisfies

$$w^{h,k}(x_i) + \tau \min_{b\in B} \max_{a\in A} \{-w^{h,k}(\overline{x}_i(a,b))\} - (1-\tau)/\mu = 0, \quad x_i \in \mathcal{G}\backslash\mathcal{T},$$

and vanishes at the points of the grid in $\mathcal{T}$.

In the following we indicate with $d(x) = dist(x, \mathcal{T})$ the distance function from the target. In order to get the estimates of convergence, we need some additional conditions at the boundary. To this end we assume that there is a closed uniform neighborhood of $\mathcal{T}$, $\mathcal{T}_\delta = \{x : d(x) \leq \delta\}$ and positive constants $\alpha, \beta, \gamma \in (0,1]$ and K such that

$$\begin{cases} v(x) \leq K d^\gamma(x), \quad x \in \mathcal{T}_\delta, \\ w^{h,k}(x_i) \leq K(d^\alpha(x_i) + h^\beta(1 + (k/h)^2)), \quad x_i \in \mathcal{T}_\delta \cap \mathcal{G}. \end{cases} \tag{1.2}$$

As we assume δ independent of h, k, it is then clear that for small values of the parameters, $\mathcal{T}_\delta \cap \mathcal{G}$ is nonempty; therefore, the second inequality in (1.2) is meaningful. We can always add grid points on $\mathcal{T}$; however, the result will not depend on this fact but only on (1.2).

Remark 1. It is well known, and as a matter of fact easy to prove, that the first inequality in (1.2) is equivalent to the local γ-Hölder continuity of the capture-time function T; see, e.g., [21]. As a consequence of (1.1), when $\mu \geq L$ it is also equivalent to saying that the bounded solution v of (0.1) is γ-Hölder continuous in $\mathbb{R}^N$, as we proved in [6]. Therefore for any practical purposes, since our main goal is to compute T and we can then choose μ as we please, we will henceforth assume we are in this case. In control theory (when B is a singleton), the first inequality in (1.2) is well studied in the literature (see [7], [21], and the references therein) and sufficient conditions for it are well known. These are the so-called controllability conditions of order $[1/\gamma] - 1$ involving the vector field f and its Lie brackets up to order $[1/\gamma] - 1$ at the points of the boundary of the target $\mathcal{T}$. Sufficient conditions in the case of differential games

involving only the vector field f in a neighborhood of $\partial\mathcal{T}$ can be found in [22]. We propose the second inequality as the natural discrete version of the first, since, as we show below, it holds under the same kind of conditions. Finally note that since v, $w^{h,k}$ are bounded from above (by $1/\mu$ in fact), the properties (1.2) become global estimates by just changing the constant K.

To justify the second inequality in the assumption (1.2), we want to give an example where it is fulfilled. We show that even the second estimate in (1.2) is naturally related to the local controllability properties of the vector field f on $\partial\mathcal{T}$. It is known (see, e.g., [6] and [22] also for the extension to the nonsmooth and unbounded case) that if $\mathcal{T}$ is the closure of an open set whose boundary $\partial\mathcal{T}$ is C^2 and bounded, then the following condition on the vector field

$$\max_{b\in B}\min_{a\in A}\{f(x,a,b)\cdot n(x)\}<0,\quad x\in\partial\mathcal{T}, \tag{1.3}$$

where $n(x)$ indicates the interior unit normal vector of $\mathbb{R}^N\backslash\mathcal{T}$ at x, is sufficient for the first inequality in (1.2) to hold with $\gamma=1$ (and this is in turn equivalent to Lipschitz continuity of the solution of (0.1) if $\mu\geq L$). The next proposition shows that it is also sufficient for the second.

PROPOSITION. *Assume* (1.1) *and* (1.3). *Then there is* $C>0$ *such that for all* h,k

$$w^{h,k}(x_i)\leq C(d(x_i)+h(1+(k/h)^2))\quad \textit{for } x_i\in\mathcal{G};$$

hence the second inequality in (1.2) *is satisfied with* $\alpha=\beta=1$.

Proof. As a first step we prove the following lemma, which is a sort of discrete comparison principle for equation (0.4). □

LEMMA 2. *Assume that the sequence* $(V_i)_{i\in\mathbb{N}}$ *is such that* $V_i\in[0,1/\mu]$ *for all* i *and satisfies* $V\geq F(V)$. *Then we have* $V_i\geq w^{h,k}(x_i)$ *for all* i.

Proof of the lemma. The result follows immediately from Lemma 1 since for any fixed i, $F^n(V)_i$ is a nonincreasing sequence and $F^n(V)_i\to w^{h,k}(x_i)$, as $n\to+\infty$. Therefore $w^{h,k}(x_i)\leq F(V)_i\leq V_i$ and we can conclude. □

In the proceeding of the proof and only in this proof d is modified in $\mathcal{T}$ as the signed distance from $\partial\mathcal{T}$; i.e., $d(x)=-dist(x,\partial\mathcal{T})$ if $x\in\mathcal{T}$. We recall that if $\partial\mathcal{T}$ is of class C^2, then d is of class C^2 in a neighborhood of $\partial\mathcal{T}$. By (1.1) and (1.3) we can find $\rho,\sigma>0$, $\sigma\leq\rho^2\wedge1$, such that

$$\max_{b\in B}\min_{a\in A}\{f(x,a,b)\cdot Dd(x)\}\leq-\sigma,\quad x\in\{x:|d(x)|\leq\rho\}.$$

Let $||D^2d||_\infty=\sup_{|d(x)|\leq2\rho}|D^2d(x)|$, where the right-hand side makes sense if ρ is sufficiently small.

1. If the steps h,k are so that $\sigma/(||D^2d||_\infty+1)\leq h(L+k/h)^2$, then by choosing $C=(||D^2d||_\infty+1)/(\sigma\mu)$ we get directly that for any index i

$$w^{h,k}(x_i)\leq1/\mu\leq Ch(L+k/h)^2=Ch(L^2+2L(k/h)+(k/h)^2),$$

and immediately we obtain the conclusion since $k/h\leq1+(k/h)^2$.

2. Otherwise we can assume that the steps h,k satisfy

$$h(L+k/h)^2\leq\sigma/(||D^2d||_\infty+1). \tag{1.4}$$

Observe that a nonnegative sequence $V\in[0,1/\mu]^{\mathbb{N}}$ satisfies

$$V_i\geq F(V)_i=\tau\max_{b\in B}\min_{a\in A}\sum_j\lambda_{ij}(a,b)V_j\ +(1-\tau)/\mu, \tag{1.5}$$

for any index i such that $x_i \in \mathcal{G}\backslash\mathcal{T}$, if we can show that

$$0 \geq \max_{b\in B}\min_{a\in A}\sum_j \lambda_{ij}(V_j - V_i) \ + h\exp(\mu). \tag{1.6}$$

In fact, since it is not restrictive to assume $h \leq 1$, this follows from $\exp(x) - 1 \leq x\exp(x)$ for all $x \geq 0$, and $\tau = \exp(-\mu h)$. We want to check (1.5) for the sequence defined by the position

$$V_i = C(d(x_i) + h(L + k/h)) \wedge (1/\mu), \quad x_i \in \mathcal{G},$$

where we choose C sufficiently large so that $C\rho \geq 1/\mu$ and $C\sigma/2 \geq \exp(\mu)$. Then applying Lemma 2 we get

$$w^{h,k}(x_i) \leq C(d(x_i) + h(L + k/h)), \quad x_i \in \mathcal{G}, \tag{1.7}$$

and the conclusion. To this end, we only need to deal with indices x_i such that $0 < V_i < 1/\mu$, since (1.5) is obvious for the ith component if either $x_i \in \mathcal{T}$ or $V_i = 1/\mu$. For such indices, we will then prove (1.6). Let therefore x_i be such that $V_i \in (0, 1/\mu)$, and observe that then $x_i \in \mathcal{T}_\rho\backslash\mathcal{T}$ by the choice of C. By definition of V we have

$$V_j - V_i \leq C(d(x_j) - d(x_i)) \quad \text{for all } x_j \in \mathcal{G}. \tag{1.8}$$

We now use the regularity of the boundary $\partial\mathcal{T}$; note that by (1.1) and (1.4)

$$|x_j - x_i| \leq k + Lh (\leq \sqrt{\sigma} \leq \rho),$$

if $\lambda_{ij} \neq 0$ for some a, b, and compute

$$d(x_j) - d(x_i) \leq ||D^2 d||_\infty |x_j - x_i|^2/2 + Dd(x_i)\cdot(x_j - x_i), \quad \lambda_{ij} \neq 0. \tag{1.9}$$

By (1.8) and (1.9), the definition of $\lambda_{ij}(a, b)$, (1.3), and the assumption, we then conclude

$$\begin{aligned}\max_{b\in B}\min_{a\in A}\sum_j \lambda_{ij}(V_j - V_i) &\leq C\max_{b\in B}\min_{a\in A}\sum_j \lambda_{ij}(d(x_j) - d(x_i))\\ &\leq C(||D^2 d||_\infty (k + Lh)^2/2 + \max_{b\in B}\min_{a\in A}\sum_j \lambda_{ij}(x_j - x_i)\cdot Dd(x_i))\\ &\leq C(h\sigma/2 + h\max_{b\in B}\min_{a\in A} Dd(x_i)\cdot f(x_i, a, b)) \leq -h\exp(\mu),\end{aligned}$$

so (1.6) and, consequently, (1.5) hold for the fixed index i. □

Our main result is the following.

THEOREM. *Assume* (1.1) *and* (1.2)*, and let* $\mu \geq L$*. Then there is a constant* $C = C(K, L, \mu)$ *such that for all* h, k *we have*

$$||w^{h,k} - v||_\infty \leq Ch^\rho(1 + (k/h)^2), \tag{1.10}$$

where $\rho = \min\{\beta, \alpha/2, \gamma/2\}$ *and the parameters* α, β, γ *are defined in* (1.2)*.*

Remark 2. The constant C that appears in the estimate (1.10) can be explicitly computed in terms of K, L, and μ as can be seen from the proof of the result in the next section.

The theorem and the proposition prove that if the solution v of (0.1) is Lipschitz continuous, the target is smooth, and k/h remains bounded, then the rate of convergence is $\sqrt{h}$. If v is only Hölder continuous, we do not know the rate exactly unless we compute the parameters α, β in (1.2). In the proposition, we proved the discrete estimate in (1.2) by using (1.3) and showing that the right-hand side is a supersolution of the fully discrete equation (0.4). Then we applied a discrete comparison principle. This draws a parallel with a possible proof of the first estimate in (1.2), see, e.g., [22] and the references therein, and leads to conjecture that sufficient conditions on the vector field f and its Lie brackets to prove the first inequality in (1.2) with a certain γ also guarantee the discrete estimate with $\beta = \alpha = \gamma$. This would then give the rate of convergence for the scheme $h^{\gamma/2}$, as in infinite horizon problems without targets. However we did not attempt to prove this conjecture in more general situations than the statement of the proposition, as dealing with Lie brackets and discrete-time systems is a serious matter by itself. □

2. Proof of the theorem. In the following we indicate by $\varepsilon = h^{(2-\gamma)/2}$ and drop for convenience of notation the superscripts of $w^{h,k}$.

1. We start proving that

$$\sup_{\mathbb{R}^N}(w - v) \le Ch^\rho(1 + (k/h)^2).$$

We proceed by contradiction assuming that for any fixed constant $C > 0$, we can find h, k as small as we please and $\sigma(h, k) \in (0, 1]$ such that for any $0 < \sigma \le \sigma(h, k)$ we have

$$\sup_{\mathbb{R}^N}(w - v) \ge Ch^\rho(1 + (k/h)^2) + 2\sigma. \tag{2.1}$$

We introduce the function

$$\varphi(x, y) = w(x) - v(y) - |x - y|^2/\varepsilon, \quad (x, y) \in \mathbb{R}^{2N}.$$

By (2.1) it follows that

$$+\infty > \sup_{\mathbb{R}^{2N}} \varphi \ge Ch^\rho(1 + (k/h)^2) + 2\sigma.$$

We now choose a point (x_1, y_1) such that $\varphi(x_1, y_1) > \sup_{\mathbb{R}^{2N}} \varphi - \sigma$ and select a function $\xi \in C_c^1(\mathbb{R}^{2N})$ satisfying $0 \le \xi \le 1$, $\xi(x_1, y_1) = 1$, $|D\xi| \le 1$. If we denote $\psi = \varphi + \sigma\xi$, by construction the maximum point of ψ is attained at a point (x^o, y^o) in the support of ξ.

2. Observe that $w(x^o) > 0$. This immediately follows from

$$w(x^o) - v(y^o) + \sigma \ge \psi(x^o, y^o) \ge \psi(x_1, y_1) > \sup_{\mathbb{R}^{2N}} \varphi \ge Ch^\rho(1 + (k/h)^2) + 2\sigma. \tag{2.2}$$

We indicate the point x^o as the unique convex combination of the vertices of the simplex to which it belongs, i.e., $x^o = \sum \lambda_j x_j$, where $\sum \lambda_j = 1$, $\lambda_j \in [0, 1]$, $|x^o - x_j| \le k$ if $\lambda_j \ne 0$. Then it follows that for at least an index j such that $\lambda_j \ne 0$ we have $w(x_j) > 0$; therefore $x_j \notin \mathcal{T}$. We will need more, and as a matter of fact we prove that if $k \le \delta$, δ as in the assumption (1.2), none of such points x_j with $\lambda_j \ne 0$ can be

on $\mathcal{T}$. In fact if there is $x_i \in \mathcal{T}$, by (1.2) we obtain the following estimate

$$w(x^o) = \sum_j \lambda_j w(x_j) \le K\left(\sum_j \lambda_j d^\alpha(x_j) + h^\beta(1+(k/h)^2)\right)$$
$$\le K\left(\sum_j \lambda_j |x_j - x_i|^\alpha + h^\beta(1+(k/h)^2)\right) \le K(k^\alpha + h^\beta(1+(k/h)^2)),$$

which gives a contradiction with (2.1) when C is chosen sufficiently large, since we have $k^\alpha = (k/h)^\alpha h^\alpha \le (1+(k/h)^2)h^\alpha$, and v is nonnegative.

3. We now show that $y^o \notin \mathcal{T}$. We use the inequality $\psi(x^o, y^o) \ge \psi(x^o, x^o)$ and get

$$v(x^o) - v(y^o) + \sigma(\xi(x^o, y^o) - \xi(x^o, x^o)) \ge |x^o - y^o|^2/\varepsilon,$$

and therefore by the γ-Hölder continuity of v, see Remark 1, if we indicate with $M = M(L, \mu)$, the best γ-Hölder constant of v, we have

$$|x^o - y^o|^2/\varepsilon \le \sigma|x^o - y^o| + M|x^o - y^o|^\gamma,$$

which implies first that $|x^o - y^o| \le M_1$, where M_1 is independent of all small ε and σ. Consequently, choosing σ sufficiently small,

$$|x^o - y^o| \le (1+M)^{1/(2-\gamma)}\varepsilon^{1/(2-\gamma)}. \tag{2.3}$$

We can now conclude that $y^o \notin \mathcal{T}$; otherwise, from (1.2) and (2.3) for h, k sufficiently small so that

$$k + (1+M)^{1/(2-\gamma)}\varepsilon^{1/(2-\gamma)} \le \delta,$$

we obtain as in the final estimate in part 2 that

$$w(x^o) \le K\left(\sum_j \lambda_j d^\alpha(x_j) + h^\beta(1+(k/h)^2)\right) \le K\left(\sum_j \lambda_j |x_j - y^o|^\alpha + h^\beta(1+(k/h)^2\right)$$
$$\le K((k+|x^o - y^o|)^\alpha + h^\beta(1+(k/h)^2)) \le K(k^\alpha + (1+M)^{\alpha/(2-\gamma)}\varepsilon^{\alpha/(2-\gamma)}$$
$$+ h^\beta(1+(k/h)^2)) \le 2Kh^\rho(1+(1+M)^{\alpha/(2-\gamma)} + (k/h)^2),$$

which again gives a contradiction with (2.1) when C is sufficiently large. The last inequality follows by the definition of ε at the beginning of this section.

4. We can now use the equations for v at y^o and for w at all vertices x_j of the simplex containing x^o. By the definition of viscosity solution, equations (0.1) and (0.4), the maximality of (x^o, y^o), and the fact that w is defined in $\mathbb{R}^N$ by linear interpolation, we then obtain

$$\mu v(y^o) + \min_{b\in B}\max_{a\in A}\{-f(y^o, a, b)\cdot \sigma D_y\xi(x^o, y^o) + 2/\varepsilon\ f(y^o, a, b)\cdot(y^o - x^o)\} - 1 \ge 0,$$

$$w(x_j) + \min_{b\in B}\max_{a\in A}\{-\tau w(x_j + hf(x_j, a, b))\} - (1-\tau)/\mu = 0,$$

where we recall that $\tau = \exp(-\mu h)$. Choosing conveniently first $b_j \in B$ in the second equation and then $a_j \in A$ in the first, we then get

$$\mu v(y^o) - f(y^o, a_j, b_j) \cdot \sigma D_y \xi(x^o, y^o) + 2/\varepsilon \; f(y^o, a_j, b_j) \cdot (y^o - x^o) - 1 \geq 0, \tag{2.4}$$

$$w(x_j) - \tau w(x_j + h f(x_j, a_j, b_j)) - (1-\tau)/\mu \leq 0. \tag{2.5}$$

In the following we denote $\overline{x}_j = x_j + h f(x_j, a_j, b_j)$. By the optimality of (x^o, y^o), and therefore the fact that $\psi(x^o, y^o) \geq \psi(\overline{x}_j, y^o)$, we deduce

$$\begin{aligned} &w(\overline{x}_j) \leq w(x^o) + |\overline{x}_j - x^o|^2/\varepsilon + 2/\varepsilon \; (\overline{x}_j - x^o, x^o - y^o) + \sigma |x^o - \overline{x}_j| \\ &= w(x^o) + |\overline{x}_j - x^o|^2/\varepsilon + 2/\varepsilon \; (x_j - x^o, x^o - y^o) - 2h/\varepsilon \; f(x_j, a_j, b_j) \cdot (y^o - x^o) + \sigma |x^o - \overline{x}_j|. \end{aligned}$$

The last inequality and (2.5) then give

$$\begin{aligned} w(x_j) \leq \tau \{ & w(x^o) + |\overline{x}_j - x^o|^2/\varepsilon + 2/\varepsilon \; (x_j - x^o, x^o - y^o) \\ & - 2h/\varepsilon \; f(x_j, a_j, b_j) \cdot (y^o - x^o) + \sigma |x^o - \overline{x}_j| \} + (1-\tau)/\mu. \end{aligned}$$

We multiply by λ_j as defined in part 2 of the proof, sum on the index j, and get

$$\begin{aligned} \mu w(x^o) &\leq \mu\tau/(1-\tau) \; \sum_j \lambda_j \{ |\overline{x}_j - x^o|^2/\varepsilon + 2/\varepsilon \; (x_j - x^o, x^o - y^o) \\ &\qquad - 2h/\varepsilon \; f(x_j, a_j, b_j) \cdot (y^o - x^o) + \sigma |x^o - \overline{x}_j| \} + 1 \\ &= \mu\tau/(1-\tau) \; \sum_j \lambda_j \{ |\overline{x}_j - x^o|^2/\varepsilon - 2h/\varepsilon \; f(x_j, a_j, b_j) \\ &\qquad \cdot (y^o - x^o) + \sigma |x^o - \overline{x}_j| \} + 1, \end{aligned} \tag{2.6}$$

where the equality follows from the fact that $\sum_j \lambda_j x_j = x^o$ by definition.

We now multiply (2.4) by λ_j and sum on the index j, then add the result to (2.6) and obtain, also by the definition of τ,

$$\begin{aligned} &\mu(w(x^o) - v(y^o)) \leq \mu\tau/(1-\tau) \sum_j \lambda_j [|\overline{x}_j - x^o|^2/\varepsilon + \sigma |x^o - \overline{x}_j|] \\ &+ \sum_j \lambda_j [-(2/\varepsilon) \; \mu h \tau/(1-\tau) \; f(x_j, a_j, b_j) \cdot (y^o - x^o) - f(y^o, a_j, b_j) \cdot \sigma D_y \xi(x^o, y^o) \\ &\qquad + 2/\varepsilon \; f(y^o, a_j, b_j) \cdot (y^o - x^o)] \leq h^{-1} \sum_j \lambda_j [|\overline{x}_j - x^o|^2/\varepsilon + \sigma |x^o - \overline{x}_j|] \\ &\qquad\qquad + 2L/\varepsilon \; \sum_j \lambda_j [|x_j - y^o| |x^o - y^o| + \mu h |x^o - y^o|] + \sigma L, \end{aligned}$$

where to obtain the second inequality we added and subtracted in each bracket of the second sum the terms $(2/\varepsilon) \; f(x_j, a_j, b_j) \cdot (y^o - x^o)$ and used (1.1) and the fact that $0 \leq 1 - \mu h\tau/(1-\tau) \leq \mu h$. We now proceed with the estimate using (2.3), the fact that $|\overline{x}_j - x^o| \leq k + Lh$ if $\lambda_j \neq 0$, and the definition of ε. We then get, if we indicate $P = (1+M)^{1/(2-\gamma)}$ (and for $h \leq 1$),

$$\begin{aligned} &\mu(w(x^o) - v(y^o)) \leq (\varepsilon h)^{-1} (k + Lh)^2 + \sigma(2L + k/h) + 2LP(\mu h + k + P\varepsilon^{1/(2-\gamma)}) \\ &\varepsilon^{(\gamma-1)/(2-\gamma)} \leq h^{\gamma/2} (L + k/h)^2 + \sigma(2L + k/h) + 2LP(\mu + k/h) h^{(\gamma+1)/2} + 2LP^2 h^{\gamma/2} \\ &\qquad \leq h^{\gamma/2} [(k/h)^2 + 2L(1+P)(k/h) + L^2 + 2\mu LP + 2LP^2] + \sigma(2L + k/h). \end{aligned}$$

We finally use (2.2) and the fact that σ can be chosen arbitrarily small to get

$$Ch^\rho(1+(k/h)^2) \le \mu^{-1}h^{\gamma/2}[(k/h)^2 + 2L(1+P)(k/h) + L^2 + 2\mu LP + 2LP^2].$$

This gives a contradiction if the constant C was chosen sufficiently large, again since $k/h \le 1+(k/h)^2$.

5. To prove the other estimate we need

$$\sup_{\mathbb{R}^N}(v-w) \le Ch^\rho(1+(k/h)^2),$$

we proceed similarly, and we argue again by contradiction. We assume that for any fixed $C>0$ there are h,k as small as we want and $\sigma(h,k)\in(0,1]$ such that for all $0<\sigma\le\sigma(h,k)$ we have

$$\sup_{\mathbb{R}^N}(v-w) \ge Ch^\rho(1+(k/h)^2) + 2\sigma. \tag{2.7}$$

We follow along the lines above choosing the function $\varphi(x,y) = v(x)-w(y)-|x-y|^2/\varepsilon$ and constructing a maximum point (x^o,y^o) for ψ. As at the beginning of point 2, we prove that $v(x^o)>0$, so $x^o\notin\mathcal{T}$. Arguing as in point 3, using this time the inequality $\psi(x^o,y^o)\ge\psi(y^o,y^o)$, we show that $|x^o-y^o|\le(1+M)^{1/(2-\gamma)}\varepsilon^{1/(2-\gamma)}$. Moreover if we indicate $y^o=\sum_j\lambda_j y_j$ as the unique convex combination of the vertices of the simplex it belongs to, we can prove that none of the points y_j, $\lambda_j\neq0$, is in $\mathcal{T}$ as follows. Assume by contradiction that $y_i\in\mathcal{T}$; then we can estimate, by the first inequality in (1.2) and for h,k sufficiently small so that $k+(1+M)^{1/(2-\gamma)}\varepsilon^{1/(2-\gamma)}\le\delta$,

$$\begin{aligned} &v(x^o) \le Kd^\gamma(x^o) \le K|x^o-y_i|^\gamma \le K(|x^o-y^o|+k)^\gamma \\ &\le K((M+1)^{\gamma/(2-\gamma)}\varepsilon^{\gamma/(2-\gamma)}+k^\gamma) \le K((1+M)^{1/(2-\gamma)}h^{\gamma/2}+h^\gamma(1+(k/h)^2)), \end{aligned}$$

which provides a contradiction with (2.7) if C is sufficiently large.

We then apply part 4 of the proof with obvious modifications and the final result then follows. □

REFERENCES

[1] B. Alziary de Roquefort, *Jeux differentielles et approximation numerique de la fonctions valeur,* 2e *partie: etude numerique*, RAIRO Modél. Math. Anal. Numer., 25 (1991), pp. 535–560.

[2] M. Bardi, S. Bottacin, and M. Falcone, *Convergence of discrete schemes for discontinuous value functions of pursuit-evasion games*, in Ann. Internat. Soc. Dynam. Games, 3 (1995), pp. 273–304.

[3] M. Bardi and M. Falcone, *An approximation scheme for the minimum time function*, SIAM J. Control Optim., 28 (1990), pp. 950–965.

[4] M. Bardi, M. Falcone, and P. Soravia, *Fully discrete schemes for the value function of pursuit-evasion games*, in Advances in Dynamic Games and Applications, A. Haurie and T. Başar, eds., Birkhäuser, Switzerland, 1994.

[5] M. Bardi, M. Falcone, and P. Soravia, *Numerical methods for pursuit-evasion games via viscosity solutions*, Ann. Internat. Soc. Dynam. Games, to appear.

[6] M. Bardi and P. Soravia, *A PDE framework for games of pursuit-evasion type*, in Differential Games and Applications, Lecture Notes in Control and Inform. Sci. 119, T. Başar and P. Bernhard, eds., Springer-Verlag, New York, 1989, pp. 62–71.

[7] M. Bardi and P. Soravia, *Hamilton-Jacobi equations with a singular boundary condition on a free boundary and applications to differential games*, Trans. Amer. Math. Soc., 325 (1991), pp. 205–229.

[8] M. Bardi and P. Soravia, *Approximation of differential games of pursuit-evasion by discrete-time games*, in Differential Games. Developments in Modelling and Computations, Lecture Notes in Control and Inform. Sci. 156, R. P. Hamalainen and H. K. Ehtamo, eds., Springer-Verlag, New York, 1991.

[9] G. Barles and P. E. Souganidis, *Convergence of approximation schemes for fully nonlinear second order equations*, Asymptotic Anal., 4 (1991), pp. 271–283.

[10] I. Capuzzo Dolcetta and M. Falcone, *Discrete dynamic programming and viscosity solutions of the Bellman equation*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 6 (supplement) 1989, pp. 161–184.

[11] I. Capuzzo Dolcetta and H. Ishii, *Approximate solutions of the Bellman equation of deterministic control theory*, Appl. Math. Optim., 11 (1984), pp. 161–181.

[12] M. G. Crandall, H. Ishii, and P. L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc., 27 (1992), pp. 1–67.

[13] M. G. Crandall and P. L. Lions, *Two approximations of solutions of Hamilton-Jacobi equations*, Math. Comp., 43 (1984), pp. 1–19.

[14] M. G. Crandall and P. L. Lions, *Convergent difference schemes for nonlinear parabolic equations and mean curvature motion*, Numer. Math., 75 (1996), pp. 17–61.

[15] M. Falcone, *A numerical approach to the infinite horizon control problem of deterministic control theory*, Appl. Math. Optim., 15 (1987), pp. 1–13.

[16] R. L. V. Gonzalez and E. Rofman, *On deterministic control problems: An approximation procedure for the optimal cost, part* 1 *and* 2, SIAM J. Control Optim., 23 (1985), pp. 242–285.

[17] M. Kocan, *Approximation of viscosity solutions of elliptic partial differential equations on minimal grids*, Numer. Math., 72 (1995), pp. 73–92.

[18] H. J. Kushner, *Probability Methods for Approximations in Stochastic Control and for Elliptic Equations*, Springer-Verlag, Berlin, New York, 1977.

[19] H. J. Kushner and P. Dupuis, *Numerical Methods for Stochastic Control Problems in Continuous Time*, Springer-Verlag, Berlin, New York, 1992.

[20] O. Pourtallier and M. Tidball, *Approximation of the Value Function for a Class of Differential Games with Target*, preprint, 1994.

[21] P. Soravia, *Hölder continuity of the minimum time function with C^1-manifold targets*, J. Optim. Theory Appl., 75 (1992), pp. 401–421.

[22] P. Soravia, *Pursuit-evasion problems and viscosity solutions of Isaacs equations*, SIAM J. Control Optim., 31 (1993), pp. 604–623.

[23] P. Soravia, *Discontinuous viscosity solutions to Dirichlet problems for Hamilton-Jacobi equations with convex hamiltonians*, Comm. Partial Differential Equations, 18 (1993), pp. 1493–1514.

[24] P. E. Souganidis, *Approximation schemes for viscosity solutions of Hamilton-Jacobi equations*, J. Differential Equations, 57 (1985), pp. 1–43.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 1, pp. 12–32, January 1998

002

A GENERIC CLASSIFICATION OF TIME-OPTIMAL PLANAR STABILIZING FEEDBACKS*

A. BRESSAN† AND B. PICCOLI†

Abstract. Consider the problem of stabilization at the origin in minimum time for a planar control system affine with respect to the control. For a family of generic vector fields, a topological equivalence relation on the corresponding time-optimal feedback synthesis was introduced in a previous paper [*Dynamics of Continuous, Discrete and Impulsive Systems*, 3 (1997), pp. 335–371]. The set of equivalence classes can be put in a one-to-one correspondence with a discrete family of graphs. This provides a classification of the global structure of generic time-optimal stabilizing feedbacks in the plane, analogous to the classification of smooth dynamical systems developed by Peixoto.

Key words. time-optimal control, two-dimensional system, regular synthesis

AMS subject classifications. 93C10, 93B20

PII. S0363012995291117

1. Introduction.

Let F, G be smooth vector fields on the plane, with $F(0) = 0$, and consider the problem of reaching the origin in minimum time for the control system

$$\dot{x} = F(x) + G(x)u, \qquad |u(t)| \leq 1. \tag{1.1}$$

Under generic assumptions on F, G, it is known that the optimal control admits a regular feedback synthesis [2], [10]. Namely, for any given $\tau > 0$, on the set A_τ of points which can be steered to the origin within time τ, one can define a feedback control $u = \varphi(x)$ with the following properties:

(i) The set A_τ can be partitioned into finitely many submanifolds V_i such that the restriction of φ to each V_i is smooth.

(ii) Every trajectory of the feedback equation

$$\dot{x} = F(x) + G(x)\varphi(x) \tag{1.2}$$

starting inside A_τ reaches the origin in minimum time.

The set of manifolds V_i form a stratification of the set A_τ analogous to the analytic case; see [12], [15], [16].

For generic $F, G \in \mathcal{C}^3$, the optimal feedback $\varphi = \varphi_{F,G}$ is essentially unique. We thus regard (1.2) as a differential equation with discontinuous right-hand side, uniquely determined by the vector fields F, G. The aim of this paper is to provide a global classification of the flow generated by (1.2) in the generic case. This program can be outlined as follows:

1. Introduce an equivalence relation between couples of vector fields: $(F, G) \sim (F', G')$ determined by the topological equivalence of the corresponding flows (1.2). Roughly speaking, if $\varphi_{F,G}(x)$ and $\varphi_{F',G'}(x)$ are the corresponding time-optimal stabilizing feedbacks, the above equivalence should imply the existence of a homeomorphism, defined on a suitable subset of the plane, which maps oriented arcs of

*Received by the editors August 31, 1995; accepted for publication (in revised form) September 27, 1996. A version of this paper was presented at the 1997 Control and Design Conference, December 10–12, 1997, San Diego, CA.

http://www.siam.org/journals/sicon/36-1/29111.html

†SISSA-ISAS, via Beirut 2–4, 34014 Trieste, Italy (piccoli@sissa.it).

trajectories of

$$\dot{x} = F(x) + G(x)\varphi_{F,G}(x) \tag{1.3}$$

onto oriented arcs of trajectories of

$$\dot{x} = F'(x) + G'(x)\varphi_{F',G'}(x). \tag{1.4}$$

2. With each equivalence class associate a discrete graph in such a way that two systems are equivalent if and only if they correspond to the same graph.

3. Show that, generically, the time-optimal feedbacks are structurally stable. In other words, given a couple of generic smooth vector fields F, G, prove that $(F', G') \sim (F, G)$ whenever F' is sufficiently close to F and G' is sufficiently close to G in the $\mathcal{C}^3$ norm. A small perturbation of the fields F, G will not change the global structure of the optimal feedback flow.

4. Characterize the family of all graphs which arise in connection with an optimal feedback for some system of the form (1.1). In practice, given a suitable graph $\mathcal{G}$, this requires the construction of vector fields F, G such that the corresponding optimal feedback flow (1.2) has precisely the topological structure specified by $\mathcal{G}$.

In [2] we completed the first part of this program introducing the equivalence relation, giving a detailed description of the structure of optimal syntheses, and proving the structural stability for systems in a generic set. For general theory of syntheses and further references we refer to [17]. The definition of graph and the second part of the program are developed in this paper.

For convenience, throughout this paper we consider the entirely equivalent problem of reaching points $x \in \mathbb{R}^2$ in minimum time, starting from the origin. It is also not restrictive to assume $\tau = 1$ and to study the global feedback synthesis on the set $R = R(1)$ of points reachable from the origin within unit time.

Section 2 reviews the basic definitions and some related results from [2], [10], [11], [14]. For general theory we refer to [13]. We also recall the definition of topological equivalence between optimal feedback flows, which plays a key role.

In section 3 we give the definition of graph and describe a standard procedure to associate a graph to a stable system, i.e., to a system for which the algorithm for constructing an optimal synthesis, described in [2], succeeds. In the following section admissibility conditions are given in order to single out graphs that correspond to some system.

In section 5 we prove that two stable systems are equivalent if and only if the corresponding graphs are equivalent. Moreover, we complete the classification program exhibiting, for every admissible graph $\mathcal{G}$, a structurally stable system that is in correspondence with $\mathcal{G}$.

Finally, in the last section we outline how to generalize all results to the case of a general two-dimensional manifold.

In [11] we computed a local classification of the syntheses near frame curves and points (i.e., the singularities of the optimal feedback); see section 2 for definitions. The countable family of equivalence classes of graphs which we obtain can be regarded as a *dictionary* of all possible structures of global optimal feedbacks in connection with generic vector fields F, G. For the flows generated by these discontinuous feedbacks, the present classification is analogous to the classical work of Peixoto [8], [9] for smooth dynamical systems.

2. Basic definitions. We first review some basic notations also used in [2], [10], and [14].

The topological frontier, the closure, and the interior of a set $D \subset \mathbb{R}^2$ are denoted by $Fr(D)$, $Cl(D)$, and $Int(D)$, respectively. If D is a manifold, then ∂D denotes its relative boundary. Given a map $\gamma : [a,b] \mapsto \mathbb{R}^2$ we denote by $Dom(\gamma) \doteq [a,b]$ its domain. The restriction of γ to a subinterval $J \subset [a,b]$ is written $\gamma \restriction J$.

In connection with the control system (1.1), consider the Banach space Ξ of pairs of vector fields $\Sigma = (F,G)$, with $F, G \in \mathcal{C}^3$, $F(0) = 0$, endowed with the $\mathcal{C}^3$ norm.

We recall that the *Lie bracket* of two vector fields F, G is the vector field

$$[F,G] \doteq \nabla G \cdot F - \nabla F \cdot G.$$

For convenience, we also define the vector fields

$$X = F - G, \qquad Y = F + G. \tag{2.1}$$

For $\Sigma = (F,G)$, we write $Traj(\Sigma)$ for the set of (Carathéodory) trajectories of (1.1). A trajectory $\gamma \in Traj(\Sigma)$ is called an X-trajectory [Y-trajectory] if it corresponds to the constant control $u \equiv -1$ [$u \equiv 1$, respectively].

For a fixed $\tau \geq 0$, the *reachable set* within time τ is

$$R(\tau) = \{x : \exists \gamma \in Traj(\Sigma) \quad \text{such that} \quad \gamma(0) = 0 \in \mathbb{R}^2, \quad \gamma(t) = x \quad \text{for some } t \leq \tau\}.$$

The *minimum time function*, $T : \mathbb{R}^2 \mapsto [0, +\infty]$ is defined by

$$T(x) \doteq \inf\{\tau : x \in R(\tau)\}.$$

A *synthesis* for the control system Σ on the set $R(\tau)$ is a family of trajectories $\Gamma = \{\gamma_x : [0, b_x] \mapsto \mathbb{R}^2, \quad x \in R(\tau)\}$ satisfying the following conditions:

(a) for each $x \in R(\tau)$ one has $\gamma_x(0) = 0$, $\gamma_x(b_x) = x$;

(b) if $y = \gamma_x(t)$ for some $t \in [0, b_x]$, then $\gamma_y = \gamma_x \restriction [0,t]$.

We say that the above synthesis is *time optimal* if $b_x = T(x)$ for each x, i.e., if each trajectory γ_x steers the system from the origin to the point x in the minimum time $T(x)$.

When a trajectory γ, corresponding to a control u, satisfies the Pontryagin maximum principle (PMP) with covector field λ, then λ is called an *adjoint covector field along* (u, γ) or simply an *adjoint variable*, and we say that (γ, λ) satisfies the PMP or that γ is an *extremal trajectory*.

Consider a trajectory γ corresponding to the control u, $t_0 \in Dom(\gamma)$ and $v_0 \in \mathbb{R}^2$. We write $v(v_0, t_0; t)$ to denote the value at time t of the variational vector field along (u, γ) (see (2.4) in [2]) with the boundary condition $v(t_0) = v_0$. If $t_0, t_1 \in Dom(\gamma)$, we say that t_0 and t_1 are *conjugate* along γ if the vectors $v(G(\gamma(t_1)), t_1; t_0)$ and $G(\gamma(t_0))$ are linearly dependent.

For each $x \in \mathbb{R}^2$, one can form the 2×2 matrices whose columns are the vectors F, G, or $[F,G]$. As in [14], we consider the following scalar functions on $\mathbb{R}^2$:

$$\Delta_A(x) \doteq \det(F(x), G(x)) = F(x) \wedge G(x), \tag{2.2}$$

$$\Delta_B(x) \doteq \det(G(x), [F,G](x)) = G(x) \wedge [F,G](x), \tag{2.3}$$

where det stands for determinant and $\wedge$ denotes an exterior product. A point $x \in \mathbb{R}^2$ is called an *ordinary point* if

$$\Delta_A(x) \cdot \Delta_B(x) \neq 0. \tag{2.4}$$

On the set of ordinary points we define the scalar functions f, g as the coefficients of the linear combination

$$[F, G](x) = f(x)F(x) + g(x)G(x). \tag{2.5}$$

In [14, p. 447] it was proved that

$$f(x) = -\frac{\Delta_B(x)}{\Delta_A(x)}. \tag{2.6}$$

For a generic system (see [14]) the only time-optimal trajectories γ that are not bang–bang must verify $\Delta_B(\gamma(t)) = 0$. A $\mathcal{C}^2$ embedded one-dimensional submanifold with boundary $S \in \mathbb{R}^2$ is a regular turnpike if $\Delta_B(x) = 0$ for every $x \in S$, and it verifies some technical conditions listed in [2], [14]. For every turnpike S we can define a control φ_S such that the corresponding trajectory runs S. Such a trajectory is called a Z-trajectory or a singular trajectory.

It may happen that a same point $x \in R = R(1)$ can be reached in minimum time using different controls. An *overlap curve* is a $\mathcal{C}^2$ one-dimensional connected embedded submanifold K of $\mathbb{R}^2$, with the property that for each point of K there exist two distinct time optimal trajectories $\gamma_1, \gamma_2 : [0, T(x)] \mapsto \mathbb{R}^2$, both steering the system from the origin to x in minimum time, with the following property: for some $\varepsilon > 0$, the restrictions of γ_1, γ_2 to $[T(x) - \varepsilon,\ T(x)]$ are an X- and a Y-trajectory, respectively.

For every system $\Sigma = (F, G)$ in a suitable open dense subset $\Xi^* \subset \Xi$, the algorithm $\mathcal{A}$ described in [2] constructs a structurally stable optimal synthesis. In the first step, the algorithm constructs the two trajectories $\gamma^{\pm}$, $\gamma^{\pm}(0) = 0$, corresponding to constant control ± 1, and marks some special points along these curves from which additional special curves bifurcate. At step N, the algorithm constructs precisely those trajectories which are concatenations of N bang or singular arcs and satisfy the PMP. We say that the algorithm succeeds for Σ if it stops in a finite number of steps and some stability conditions are fulfilled; see [2]. The set $R \doteq R(1)$ of points reachable from the origin within unit time can be partitioned in a natural way into a finite number of open regions, covered by Y- or X-trajectories, separated by curves called *frame curves*. The intersections of these frame curves are called *frame points*. In connection with the above partition, there exists a piecewise smooth feedback control $u = \varphi(x)$ with the property that the (Carathéodory) solutions of

$$\dot{x} = F(x) + G(x)\varphi(x), \qquad x(0) = 0 \tag{2.7}$$

are precisely the time-optimal trajectories. Each optimal trajectory is a finite concatenation of X-, Y-, and Z-trajectories.

All frame curves and frame points were classified in [11]. In particular, only five types of frame curves can generically occur:

(F1) the trajectories γ^- and γ^+ originating from 0 and corresponding to the constant controls $u^- \equiv -1$ and $u^+ \equiv 1$, respectively;

(F2) the topological frontier of the reachable set: $Fr(R)$;

(F3) curves of points conjugate to points of other frame curves, also called *switching curves*;

(F4) regular turnpikes,

(F5) overlap curves.

To denote the frame curves we use the symbols X, Y, F, C, S, and K, respectively. Moreover we use the same notation of [11] for frame points. For example, a point

that is the intersection of a switching curve and an overlap curve is called the (C, K) frame point. See [11] for a complete description of notations.

We now recall, from [2], the equivalence relation between systems expressing the fact that their time-optimal feedback flows have similar structures. Consider two systems $\Sigma_1, \Sigma_2 \in \Xi^*$. For $i = 1, 2$ let R_i be the reachable set for Σ_i at time $t = 1$. Define

$$\mathcal{K}_i = \{x \mid x \in K \setminus \partial K,\ K \text{ is an overlap curve of } \Gamma_{\mathcal{A}}(\Sigma_i)\},$$

and set $R_i' = R_i \setminus \mathcal{K}_i,\ \ i = 1, 2$. In the following, for each $x \in R_i$, we denote by $t \mapsto \gamma_x^i(t)$ a trajectory of Σ_i which reaches x from the origin in minimum time.

DEFINITION 1 (equivalence of feedback flows). *We say that the time-optimal feedback flows for Σ_1 on R_1 and Σ_2 on R_2 are equivalent, or simply that $\Sigma_1 \sim \Sigma_2$, if there exists a homeomorphism $\Psi : R_1' \mapsto R_2'$ such that the following hold:*

(E1) *Ψ maps arcs of optimal trajectories for Σ_1 onto arcs of optimal trajectories for Σ_2. More precisely, for every $x \in R_1'$ one has $\{\Psi(\gamma_x^1(t)) : t \in Dom(\gamma_x^1)\} = \{\gamma^2_{\Psi(x)}(t) : t \in Dom(\gamma^2_{\Psi(x)})\}$.*

(E2) *Ψ induces a bijection on frame curves that are not overlap curves; i.e., for each frame curve D_1, which occurs in the construction of the optimal feedback for Σ_1 and is not a K-curve, we have that $\varphi(D_1)$ is a frame curve of the same type corresponding to Σ_2 and vice versa.*

(E3) *If A is an open region of R_1' enclosed by frame curves and entirely covered by Y- or X-trajectories, then $\Psi(A)$ is enclosed by the corresponding frame curves and is covered by Y- or X-trajectories, respectively.*

3. Graphs. In this section we introduce the definition of graph and describe a procedure to associate a graph with every system for which the algorithm $\mathcal{A}$ succeeds. The points and edges of this topological graph correspond to frame points and curves of the system. Moreover, some additional *lines* must be included in the definition of graph to describe the *history* of all trajectories that form the optimal synthesis. In Remark 4.1 we give some examples to motivate the definition of graph.

From now on, we consider only the systems of Ξ for which the algorithm $\mathcal{A}$ succeeds.

DEFINITION 2 (graph). *A graph $\mathcal{G}$ is a finite set of points of $\mathbb{R}^2$ and smooth connected one-dimensional boundary manifolds connecting the points, called edges. Moreover, inside each region enclosed by edges there are possibly some other smooth manifolds, called lines, connecting points and edges. We assume that edges and lines do not cross one another.*

Every edge can be of one of the following type: X, Y, F, S, C, K, corresponding to the types of frame curves. An edge of type X, Y, or S has an orientation and hence an initial and a terminal point. The edges of type C have a positive side, corresponding to the fact that constructed trajectories cross a frame curve of type C passing from one side to another.

Every region enclosed by edges that are not all of F type has a sign $+$ or $-$. This corresponds to the fact that a region of the reachable set in unit time R that contains no frame curve is covered by X- or by Y-trajectories. On each region we can have some curves connecting points and edges. These correspond to constructed trajectories that pass through frame points. See Remark 3.1 below.

We say that two edges E_1, E_2 are related and we write $E_1 \sim E_2$ if they have in common a point of the graph.

We now describe a canonical way of associating a graph with a system. Given a system Σ (for which $\mathcal{A}$ succeeds) we associate a graph $\mathcal{G}$ with Σ in the following way. For every frame point we construct a point of $\mathcal{G}$ having the same coordinate in $\mathbb{R}^2$. For every frame curve D, with no frame point in $D \setminus \partial D$, $\partial D = \{x_1, x_2\}$, we construct an edge E of $\mathcal{G}$ of the same type connecting the points of $\mathcal{G}$ corresponding to x_1, x_2. If D is an X, Y, or S-curve, then D has the orientation of increasing time and we endow E with the corresponding orientation. If D is of type C, then some constructed trajectories enter one side of D. We define the corresponding side of E to be positive. For every region $A \subset R$ enclosed by frame curves there is a region A' in the plane of the graph enclosed by the corresponding edges. If A is covered by Y-trajectories, we assign to A' the positive sign, otherwise we assign to A' the negative sign.

Now we pass to the construction of lines. These lines are necessary to describe the behavior of every optimal trajectory of the synthesis; see Remark 3.1. Consider a frame point x of $Cl(A)$, which is not of (K, K) type (recall the terminology of [11]), and the constructed trajectory γ_x verifying $\gamma_x(t_x) = x$ for some t_x. Assume that $\gamma_x(I) \subset A$ for some $I = [a, b] \subset Dom(\gamma)$, $t_x \in I$. Notice that it can happen $a \neq t_x \neq b$, e.g., if x is of type $(X, K)_3$. If $t_x \neq a$ and $\gamma_x(a) \in D$ frame curve, then we construct a line in A' going from a point y of the edge E, corresponding to D, to the point x' of $\mathcal{G}$ corresponding to x. If $\gamma_x(a)$ is a frame point, then we choose y to be the corresponding point of E, otherwise we choose y in $E \setminus \partial E$. If D is of C type, and $\gamma_x(a) \in D \setminus \partial D$, then we consider the last switching point z of γ_x before $\gamma_x(a)$. If D is of S type then there exists a constructed trajectory γ_1 that switches at $\gamma_x(a)$ and enters the region on the opposite side, with respect to D, of the region entered by γ_x. Indeed, a Y and an X constructed trajectory originate from every point of a turnpike. In this case, we let z be the first switching point of γ_1 after $\gamma_x(a)$. If z belongs to a frame curve D_1 then we construct a line going from a point z' of the edge E_1, corresponding to D_1, to the point y. Again if z is a frame point we let z' be the corresponding point of $\mathcal{G}$. If D_1 is a C or S frame curve then we proceed in the same way. We continue until we reach a frame curve not of C or S type. We do the same if $t_x \neq b$.

We can construct these lines in such a way that they do not cross one another. If $\mathcal{G}$ is associated to Σ in this way then we say that $\mathcal{G}$ is *canonically associated* with Σ.

Remark 3.1. Consider the system

$$\begin{cases} \dot{x}_1 = 3x_1 + u, \\ \dot{x}_2 = x_1^2 + x_1. \end{cases}$$

For every time $\tau > \ln(4)/3$ the reachable set in time τ contains two switching curves starting from γ^-. There are two frame points of type (X, C) that are not topologically equivalent. See Example 3 of [11] for an accurate description of this system Σ_3 and for the classification of (X, C) frame points. Portrayed are the reachable set of Σ_3 in Fig. 1 and its associated graph $\mathcal{G}_3$ in Fig. 2. If we do not specify a sign for every region of $\mathcal{G}_3$ then the two (X, C) frame points are not distinguishable. Hence, for some system Σ with a frame point of type $(X, C)_1$ or $(X, C)_2$, we can construct a system with the same graph, except the signs of the regions, but not equivalent to Σ. This shows the necessity of specifying a sign for every region.

Consider the system Σ_4 of Example 4 of [11]. There is a region A that is a connected component of the complement of the reachable set and is bounded. In the corresponding graph, we cannot give a sign to the region corresponding to A.

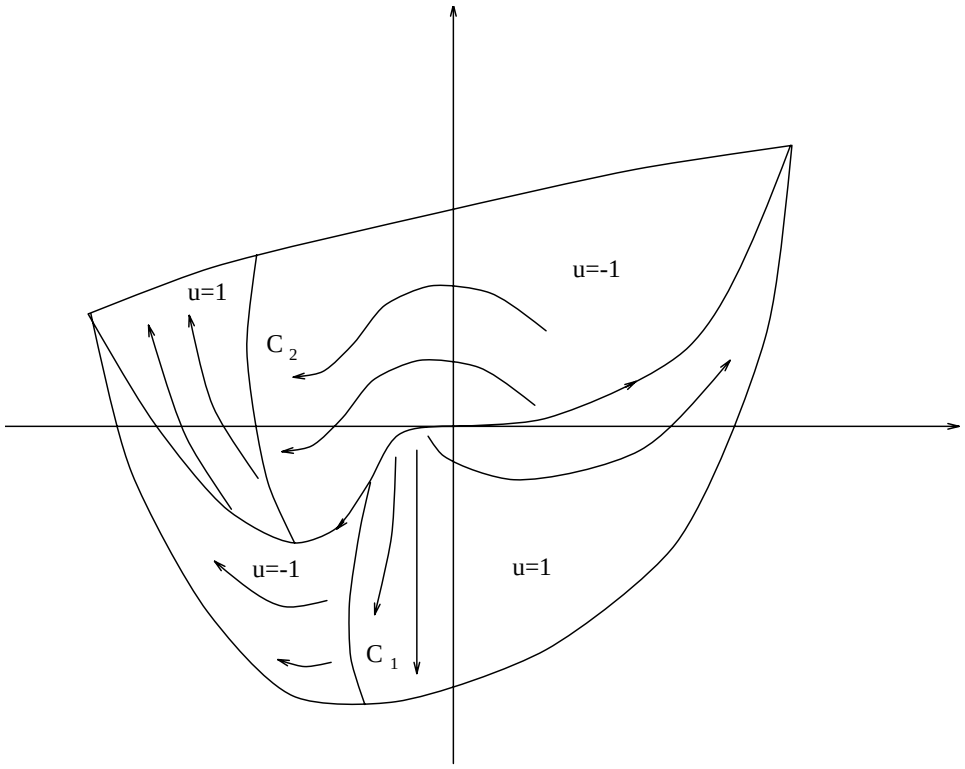


FIG. 1.

Otherwise, we would have equivalent systems corresponding to different graphs. The regions enclosed by edges all of F type correspond exactly to the *holes* of the reachable set.

Consider now the frame point x of $(C,S)_2$ type of Example 8 of [11]. If we do not specify, in the corresponding graph, a positive side for the edge corresponding to the switching curve then we do not know, from the graph, if the Y or the X trajectories enter the switching curve. Again there would exist two not equivalent systems corresponding to the same graph.

The lines divide the graphs into subregions in such a way that the trajectories, contained in the same subregion, have the same *history*, i.e., cross the same frame curves in the same order and are composed of the same sequence of elementary arcs. For example, $XYSX...$, where X, Y, and S denote, respectively, X,Y arcs and trajectories running a turnpike. If the lines are not constructed, then in some cases we cannot decide the story of every trajectory and then we cannot recognize equivalent systems. To appreciate this point, consider the following examples. First, let the syntheses of Figs. 3 and 4 correspond to some system. The associated graphs contain the same points and edges. However, the syntheses are not equivalent. Indeed, the homeomorphism Ψ should map the trajectory through the (X,S) point onto the corresponding one to satisfy (E1), but obviously in this case (E2) cannot be satisfied. To have an explicit example, consider now the system Σ_4 of the fourth example of [11]. Let γ be the constructed trajectory that passes through the (Y,S) point and then goes on as X trajectory. If we do not consider the lines, from the graph associated to

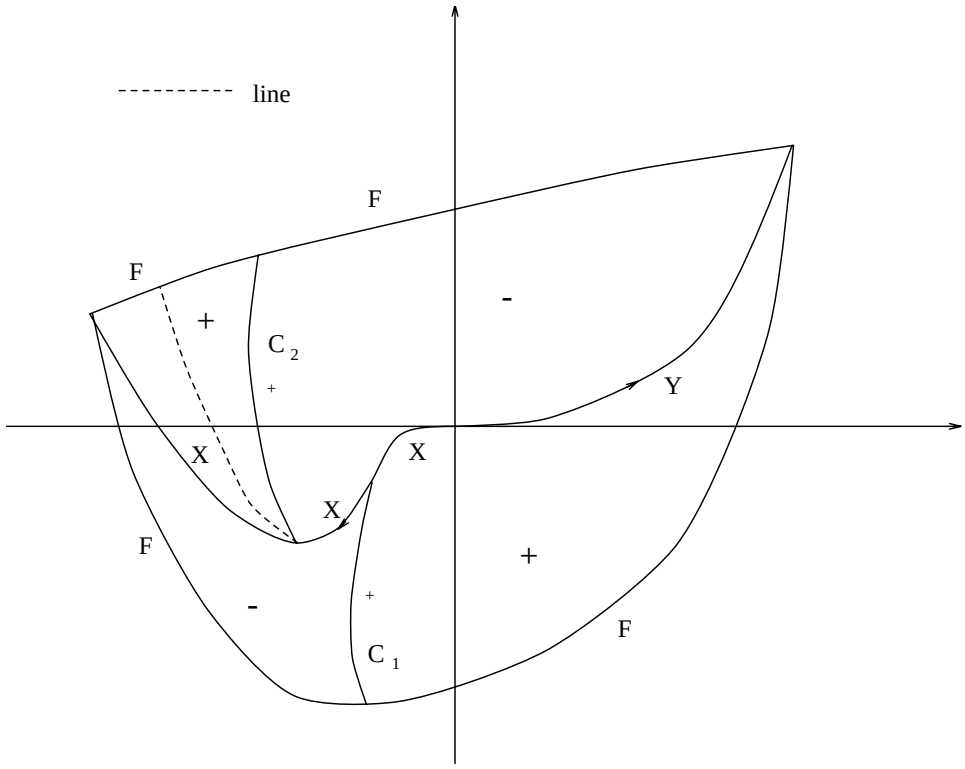


FIG. 2.

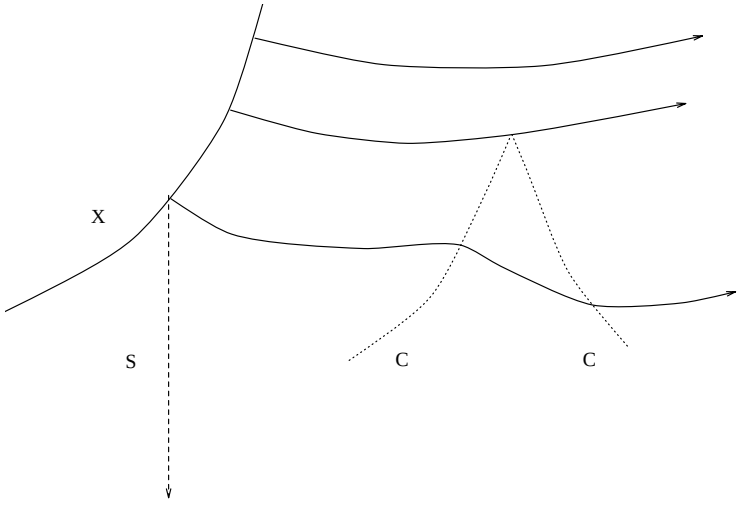


FIG. 3.

Σ_4 we cannot know if γ reaches the overlap curve or the frontier of the reachable set. Hence we cannot uniquely determine the synthesis from the graph.

4. Admissible graphs. We now give some admissibility conditions that characterize a class of graphs. This class will be proved to be the class of graphs that correspond to systems canonically.

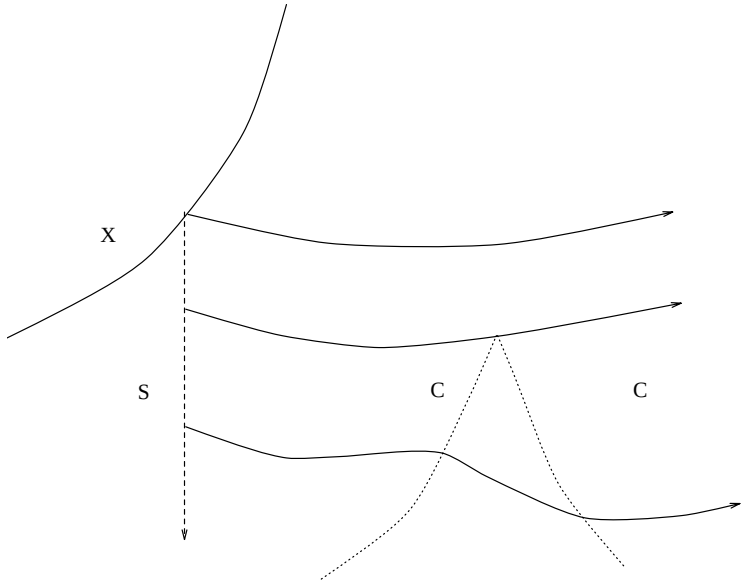


FIG. 4.

With every system of the examples of [11] we can associate a topological graph in the canonical way. We consider these examples restricted to a neighborhood of a frame point; then we obtain a set of graphs $\mathcal{E}$ whose elements are defined locally, and each one corresponds to a type of frame point. A point x' of a graph $\mathcal{G}$ is said to be *admissible* if there exists a graph $\mathcal{G}' \in \mathcal{E}$ such that $\mathcal{G}$ contains a copy of $\mathcal{G}'$ to which x' belongs. We use the same terminology for the points of $\mathcal{G}$, e.g., (X, Y) point. The first condition is

($\mathcal{G}1$) all points of $\mathcal{G}$ are admissible.

We consider graphs that contain exactly one point of the type (X, Y) and we call this point the *origin* of the graph. Assume that ($\mathcal{G}1$) holds. Let E be a Y-edge and let x be the initial point of E. If x is not the origin then there exists a Y-edge E_1 for which x is the terminal point. We consider the initial point x_1 of E_1 and do the same considerations. Since $\mathcal{G}$ is finite, proceeding by induction, we find a finite collection $E_1, \dots, E_n$ of Y-edges such that $E_i \sim E_{i+1}$, $i = 1, \dots, n-1$, and the initial point of E_n is the origin. Then, since there is only one origin, the Y-edges form a set $\{E_1, \dots, E_m\}$ such that the initial point of E_1 is the origin and for each $i = 1, \dots, m-1$ the terminal point of E_i is the initial point of E_{i+1}. We call η^+ the union of these edges. Analogously we define η^- for the X-edges. In the first step of $\mathcal{A}$ we have described all the possibilities for the sequence of frame points on a curve γ^+ of a system Σ. We say that η^+ is *admissible* if there exists a system Σ such that the curve γ^+ corresponds to η^+ canonically. That is, there is a correspondence defined for points, edges of η^+, for lines intersecting η^+, and for the regions to which η^+ belongs, that follows the rules of canonical correspondence. This happens exactly when η^+ and γ^+ have an ordered sequence of corresponding points. The second condition is

($\mathcal{G}2$) $\mathcal{G}$ has exactly one (X, Y) point, called the origin. The collections of edges $\eta^\pm$ are admissible.

Let E be a C-edge, x_1', x_2' be the endpoints of E, and A' be a region on one side of E. There exist two frame points x_1, x_2 corresponding to x_1', x_2'. Consider the correspondence between x_1' and x_1. Let D be the frame curve that corresponds to E

and $\tilde{A}_1$ the region corresponding to A'. We define A_1 to be the connected component of $\tilde{A}_1 \setminus \{x : \Delta_A(x)\Delta_B(x) = 0\}$ that contains D. Similarly we define the region A_2. We say that E is *admissible* if there exist x_1, x_2 such that the functions Δ_A, Δ_B has the same sign on A_1 and A_2.

Remark 4.1. If, for example, E connects two points of (C, S) type then A_1, A_2 are both covered by Y-trajectories or both are covered by X-trajectories. In this case, since the two vector fields must point to opposite sides of turnpikes, it follows that Δ_A has a different sign on the two regions A_1, A_2. From Theorem 3.9 of [14] we have that along C the function f (see (2.5), (2.6)) does not change sign. Hence there exists at least one curve, intersecting C, on which $\Delta_A = 0, \Delta_B = 0$. This is clearly not a generic situation and is not compatible with the condition (P_3) of section 4 of [10].

Another admissibility condition is

($\mathcal{G}3$) every C-edge is admissible.

The relation $\sim$ divides the set of F-edges into a finite number of equivalence classes. If ($\mathcal{G}1$) holds, then the union of the elements of an equivalence class forms a closed curve.

($\mathcal{G}4$) Only one closed curve that is union of the elements of an equivalence class of F-edges, encloses a region in which there are points and edges. Moreover, there are no frame curves and points outside this region.

Notice that we can have situations with more than one equivalence class of F-edges, e.g., the system in Example 4 of [11] where $R(\tau)$ has one hole.

DEFINITION 3 (entrance, exit, side). *Consider a region A' enclosed by edges of $\mathcal{G}$. If one edge E is of X type if A' is positive, of Y type if A' is negative, of C type with the negative side on A', or of S type, then we say that E is an entrance. If E is of K, F, or C type with positive side on A' then we say that E is an exit. Otherwise, we say that E is a side, i.e., if it is of Y type and A' is positive or of X type and A' is negative. The definitions are motivated by the fact that if D is a frame curve corresponding to E canonically, then through each point of D there pass a constructed trajectory that enters, resp., exits from the region corresponding to A' if and only if E is an entrance, resp., exit.*

We say that the set of lines of $\mathcal{G}$ is admissible if the following holds. Every line connects an entrance to an exit. If a point x' belongs to two entrances, resp., exits, then there is a line connecting x' with an exit, resp., entrance. Let x' be a point of one of the types $(X, C)_3, (X, K)_3, (C, C)_1, (C, S)_2, (C, K)_1, (S, K)$, and let A', B' be the two regions such that $x' \in Cl(A'), Cl(B')$. There are two lines l_1, l_2, both contained in A' or both in B', passing through x'; l_1 connects x' to an entrance and l_2 connects x' to an exit.

If x' is of type $(C, C)_2$ then there are two lines arriving at x' from different regions and at least one of them reaches another point. These are the only lines that connect two points.

If $x' \in E \setminus \partial E$, E is of C or S type and there is a line l arriving at x' from a region A' then there is a line arriving at x' from the other region B' such that $x' \in Cl(B')$.

There are no other lines. If all these conditions are satisfied then we say that the set of lines of $\mathcal{G}$ is admissible.

Remark 4.2. The conditions given for the set of lines follow directly from the canonical way of associating a graph with a system and from the description of frame points given in [11]. If we do not assume that the set of lines is admissible, then we cannot expect that there exists a system corresponding to $\mathcal{G}$.

Consider the closed curve $\tilde{F}$, union of F-edges, described in ($\mathcal{G}4$). Let U be the connected component of the complement of the union of F-edges, that is, enclosed by

$\tilde{F}$ and verifies $\tilde{F} \subset Cl(U)$. If A' is a region contained in U we define $L(A')$ to be the set of lines contained in A'. The last condition is

($\mathcal{G}5$) the set of lines of $\mathcal{G}$ is admissible. If $A' \subset U$ and A'_1 is a connected component of $A' \setminus L(A')$, then $Cl(A'_1)$ contains exactly one entrance and one exit.

We have given the conditions in ($\mathcal{G}_5$) only for regions $A' \subset U$, because if the opposite happens, then A' corresponds to an hole of R, $L(A) = \emptyset$ and $Cl(A')$ contains only exits.

If a graph $\mathcal{G}$ satisfies the conditions $(\mathcal{G}1), \dots, (\mathcal{G}5)$, then we say that $\mathcal{G}$ is *admissible*. It is easy to check that if $\mathcal{G}$ corresponds to a system Σ, then $\mathcal{G}$ is admissible. In the following we will prove the converse.

5. Classification. To ensure that the canonical way of associating a graph with a system is well defined we have to prove that two systems are equivalent if and only if the associated graphs are equivalent.

Since we have defined the equivalence between systems in *weak* form, excluding overlap curves, we have that equivalent systems may correspond to graphs having a different number of K edges. Hence we have to define an equivalence relation between graphs excluding K edges. We give below the exact definition.

Given two admissible graphs $\mathcal{G}_1, \mathcal{G}_2$, we say that they are equivalent and we write $\mathcal{G}_1 \sim \mathcal{G}_2$ if there is a correspondence ψ between edges and lines such that the following hold. We let ψ be multivalued and not injective on the set of K-edges, but it has to be a bijective function restricted to the edges not of K type. Moreover, ψ is a bijective function restricted to the set of lines. Finally the following holds:

(H1) For every edge E, not of K type, $\psi(E)$ is an edge of the same type; $E_1 \sim E_2$ if and only if $\psi(E_1) \sim \psi(E_2)$, when E_1, E_2 are not both K-edges; ψ preserves orientations and positive sides

(H2) If l is a line that connects E_1 with E_2, then $\psi(l)$ connects $\psi(E_1)$ with $\psi(E_2)$. The same holds for line connecting points. If l_1, l_2 arrive at the same point, then the same happens for $\psi(l_1), \psi(l_2)$.

(H3) If A' is a region enclosed by edges $E_1, \dots, E_n$, then the region enclosed by $\psi(E_1), \dots, \psi(E_n)$ has the same sign.

(H4) If $\mathcal{K}_1, \mathcal{K}_2$ is the set of equivalence classes of K-edges (for the relation $\sim$) of $\mathcal{G}_1, \mathcal{G}_2$, then ψ induces a bijective correspondence between $\mathcal{K}_1, \mathcal{K}_2$.

We have the following theorem.

THEOREM 5.1. *If Σ_1, Σ_2 are two systems and $\mathcal{G}_1, \mathcal{G}_2$ the corresponding graphs, then $\Sigma_1 \sim \Sigma_2$ if and only if $\mathcal{G}_1 \sim \mathcal{G}_2$.*

Proof. Assume first that $\Sigma_1 \sim \Sigma_2$ and let Ψ be as in the definition of equivalence. For simplicity we will use the symbols Γ_1, Γ_2 for $\Gamma_{\mathcal{A}}(\Sigma_1), \Gamma_{\mathcal{A}}(\Sigma_2)$, respectively.

Given a frame curve D of Γ_1 that is not a K-curve, let E_1, E_2 be the edges corresponding, respectively, to D and $\Psi(D)$. We define $\psi(E_1) = E_2$. We can proceed in the same way to define ψ on the set of lines. From (E1), (E2) it follows that (H1) and (H2) hold, and from (E3) it follows that (H3) holds.

Now, if K_1, K_2 are two K frame curves (of K type) of Γ_1, or of Γ_2, then we set $K_1 \sim K_2$ if they have a point in common. The union of the elements of an equivalence class of Γ_1 is a connected curve K. If we extend Ψ by continuity, then $\Psi(K)$ is the union of elements of an equivalence class of Γ_2. Therefore we can define ψ on K-edges in such a way that (H4) holds.

Assume now that $\mathcal{G}_1 \sim \mathcal{G}_2$. Let E_1 be an X, Y or S-edge of $\mathcal{G}_1$, let $E_2 = \psi(E_1)$, and let D_1, D_2 be the frame curves corresponding to E_1, E_2, respectively. From (H1) we have that D_1, D_2 are of the same type. Assume that $x_1, \dots, x_n$ are the points of

$D_1 \setminus \partial D_1$, ordered for increasing time, that are in relation with a frame point, not of (K, K) type, for the definition given in [2]. There are exactly n lines if D_1 is of X or Y type and $2n$ lines if D_1 is of S type, starting at x_i. From (H2) it follows that there exist $y_1, \dots, y_n \in D_2 \setminus \partial D_2$, ordered for increasing time, from which some lines of $\mathcal{G}_2$ start. Observe that, if l is a line passing through x_i and $\psi(l)$ passes through y_j, then $i = j$. Indeed, if $i \neq j$, there must be a crossing between lines, but this is not allowed by the definition of a graph. We define Ψ on D_1 in such a way that Ψ is a homeomorphism, $\Psi(D_1) = D_2$ and $\Psi(x_i) = y_i$, $i = 1, \dots, n$.

For every $y \in D_1$ consider the constructed trajectories $\gamma_y \in \Gamma_1$ for which $y = \gamma_y(b_y)$ is a switching point. If D_1 is of X or Y type, there is at most one such trajectory; if D_1 is of S type, then there are two such trajectories. If D_1 is of type X or Y and there exists γ_y, then from (H3) there exists a trajectory $\gamma_{\Psi(y)} \in \Gamma_2$ having the same property. Let $c_y > b_y$ be the first time in which γ_y reaches another frame curve and define $b_{\Psi(y)}, c_{\Psi(y)}$ similarly. We define

$$\Psi(\gamma_y(t)) \doteq \gamma_{\Psi(y)} \left(b_{\Psi(y)} + \frac{c_{\Psi(y)} - b_{\Psi(y)}}{c_y - b_y}(t - b_y) \right) \quad \forall t \in [b_y, c_y].$$

In this way we have defined Ψ also on the frame curves that are reached by the trajectories γ_y. We proceed in the same way, defining Ψ on the images of the constructed trajectories that switch at the point of these new frame curves. After a finite number of steps we define Ψ on the whole reachable set R_1 of the system Σ_1. Notice that we can have two different definitions of Ψ on the K frame curves, but Ψ restricted to R_1' (see the definition of equivalence) is well defined. Condition (E1) follows by construction. Conditions (H1), (H2) ensure that corresponding trajectories have the same history, i.e., they cross the same type of frame curves in the same order and are composed of the same elementary arcs. Finally from conditions (H1)–(H4) we have that Ψ satisfies (E1)–(E3). □

Assume now that $\mathcal{G}$ is an admissible graph. We want to find a system Σ such that $\mathcal{G}$ is associated with Σ in the canonical way. This and Theorem 5.1 show that the correspondence $\Sigma \leftrightarrow \mathcal{G}$ is a bijection between the set of equivalence classes of systems for which $\mathcal{A}$ succeeds and the set of equivalence classes of admissible graphs.

THEOREM 5.2. *If $\mathcal{G}$ is an admissible graph, then there exists a system Σ to which $\mathcal{G}$ is canonically associated.*

Proof. We construct $\Sigma = (F, G)$ defining it on a finite collection of open sets that cover $\mathcal{G}$ and then gluing together along the intersections. We proceed defining Σ and a synthesis Γ for Σ at the same time. Moreover, every trajectory $\gamma \in \Gamma$ will be endowed with an adjoint covector. At the end of the construction, we will have $\Gamma \equiv \Gamma_{\mathcal{A}}(\Sigma)$. It can happen that we will determine Σ defining two of the fields $F, G, X = F - G, Y = F + G$.

Let $\tilde{F}$ be the union of the elements of the equivalence class of F-edges described in $(\mathcal{G}4)$. Consider the connected components of the complement, in $\mathbb{R}^2$, of the union of F-edges of $\mathcal{G}$. There is only one such component R that is contained in the region enclosed by $\tilde{F}$ and such that $\tilde{F} \subset Cl(R)$. We have to construct Σ only on R.

From $(\mathcal{G}2)$, we have that there is one origin O and O will be also the origin for Σ. It is clear that, possibly translating $\mathcal{G}$, we can assume that O is the origin of $\mathbb{R}^2$. Consider a differentiable change of coordinates such that η^+ corresponds, in the new coordinates, to the line $\{(x_1, x_2) : x_2 = 0, 0 \leq x_1 \leq a\}$ for some $a > 0$. We define the field $Y = F + G$ to be the constant field $(1, 0)$ on a neighborhood N^+ of η^+ that contains only the points of $\mathcal{G}$ that are in η^+. Since η^+ is admissible there

exists a function $\tilde\theta$ such that the points $\{t_i, t_i', s_i, s_i'\}$ of (3.3), (3.4) of [2] determine the same sequence of frame points of η^+. We can define a vector field $G(x_1)$ on N^+ such that the function θ of (3.1) of [2] verifies $\theta(t) = \tilde\theta(t)$. This is easy because from the definition of Y we have that

$$\theta(t) = \theta(x_1) = \arg(G(0), G(x_1)).$$

Since the synthesis is determined by the sequence of maxima and minima of θ and not by the values at these points, we can assume that $|\theta| < \pi/2$ at every point of (3.3), (3.4) of [2]. Therefore, if $G(x_1) = (\alpha(x_1), \beta(x_1))$, then $\alpha > 0$ and

$$\nabla\Delta_B \cdot X = (1 - 2\alpha)\, \nabla\Delta_B \cdot Y.$$

Indeed, we have

$$\nabla\Delta_B = \begin{pmatrix} \alpha \frac{\partial^2 \beta}{\partial x_1^2} - \beta \frac{\partial^2 \alpha}{\partial x_1^2} \\ 0 \end{pmatrix}, \qquad X = \begin{pmatrix} 1 - 2\alpha \\ -2\beta \end{pmatrix}.$$

The choice of θ uniquely determines the direction of the vector G but not its norm. Hence, we can choose α in such a way that $\nabla\Delta_B \cdot Y$, $\nabla\Delta_B \cdot X$ have the same (resp., the opposite) sign at the points $p_i = \gamma^+(t_i), p_i' = \gamma^+(t_i')$ if at the corresponding points of η^+ there is a C-edge (resp., an S-edge). From (III) of Proposition 3.1 of [2] it follows that there is a canonical correspondence at the points p_i, p_i'.

We can again modify θ, G in such a way that $\dot\Gamma_i(0)$, $i > 1$ (see (3.23,24), (SA7) of [2]) lies in the cone determined by $Y(q_i), G(q_i)$, $q_i = \gamma^+(s_i)$. We proceed in the following way. For every y sufficiently small there exists an extremal trajectory γ_y, with second coordinate constantly equal to y after the last switching, that switches along Γ_i. Let λ_y be its associated covector. Now Y is constant; then λ_y is also constant after the last switching time of γ_y, and there exists $\zeta_1(y)$ such that $\lambda_y \cdot G(\zeta_1(y)) = 0$. Since θ is increasing near s_1, we have that $\lambda_y \cdot G(x_1)$ is a monotone function of x_1 in $[s_1 - \varepsilon, s_1 + \varepsilon]$ for some $\varepsilon > 0$ and then $\zeta_1(y)$ exists uniquely for y small. Assume we want to modify G in such a way that Γ_i is described by the points $(\zeta_2(y), y)$. Let $\xi(y, x_1)$ be a smooth function, monotone in x_1 for every y, verifying

$$\xi(y, s_i \pm \varepsilon) = s_i \pm \varepsilon, \qquad \xi(y, \zeta_2(y)) = \zeta_1(y).$$

We redefine G in such a way that if $\tilde\theta(x_1, x_2) = \arg(G(0), G(x_1, x_2))$ then $\tilde\theta(x_1, x_2) = \theta(\xi(x_2, x_1))$. From the definition of ξ and its monotonicity we have that γ_y switches at $(\zeta_2(y), y)$.

Now, choosing the module of $G(q_i)$ in a suitable way, we can assume that X, Y point to the same side, resp., to opposite side, of Γ_i if at the corresponding points of η^+ there is a C-edge, resp., a K-edge. We can repeat the same construction for $q_i' = \gamma^+(s_i')$.

Finally, possibly changing θ, G, we can assume that $\delta > 0$, resp., < 0 (see (SA8) of [2] for the definition of δ) if at the point of η^+ corresponding to q_1 there is a C-edge, resp., a K-edge. We repeat the same arguments for q_1'. Therefore from Propositions 3.1, 3.2, and 3.3 of [2] we have that η^+ corresponds to γ^+ in the canonical way. Since we have defined Y and G the system Σ is determined.

Now consider η^- and a change of coordinate as for η^+. Possibly restricting N^+, we can define X and G on a neighborhood N^- of η^- in such a way that they coincide on N^+ with the previous definitions and such that γ^- corresponds to η^- in the canonical

way. In this way we have defined Σ on $N^+ \cup N^-$, that is, a neighborhood of $\eta^+ \cup \eta^-$. We define $\Gamma = \Gamma_{\mathcal{A}}(\Sigma)$ on $N^+ \cup N^-$ and with every $\gamma \in \Gamma$ we associate the covector field constructed by $\mathcal{A}$.

Now, let x' be a point of $\mathcal{G}$ that is not in $N^+ \cup N^-$. From ($\mathcal{G}1$) there exists a frame point x, corresponding to x', that is of one of the types classified in [11]. We have shown, in [11], an example for every classified point; hence there exists a system $\Sigma(x')$, a synthesis $\Gamma(x')$ both defined on an open set $U(x')$ and a frame point $x \in \Gamma(x')$ that corresponds to x' in the canonical way. Consider an open neighborhood U' of x' that does not contain any other frame point and define a diffeomorphism $\Psi : U(x') \to U'$ in such a way that Ψ maps frame points and curves to corresponding points and edges. Moreover, Ψ maps some constructed trajectories to the corresponding lines. Using Ψ, we define Σ and Γ on U' and we associate a covector field to every $\gamma \in \Gamma$.

From ($\mathcal{G}3$) it follows that every C-edge E is admissible. However, it may happen that if x', y' are the points belonging to E, the functions Δ_A, Δ_B do not have the required signs on $U'(x'), U'(y')$. If $\Sigma(x') = (F, G)$ is one of the system of the examples of [11], we can consider the systems

$$\Sigma_1 = (F, -G), \qquad \Sigma_2 = (-F, G), \qquad \Sigma_3 = (-F, -G).$$

Let Δ_A^i, Δ_B^i be the functions Δ_A, Δ_B for Σ_i. We have that

$$\Delta_A^1, \Delta_A^2 = -\Delta_A; \qquad \Delta_A^3 = \Delta_A; \qquad \Delta_B^1 = \Delta_B; \qquad \Delta_B^2, \Delta_B^3 = -\Delta_B.$$

The systems Σ_i have the same type of synthesis of Σ (choosing the dual vectors in a suitable way). Therefore we can define $\Sigma(x'), \Sigma(y')$ in such a way that the functions Δ_A, Δ_B have the correct signs.

Next, we define Σ on neighborhoods of frame curves. Let E be a frame curve, not of X or Y type, connecting the points x', y'. We choose a differentiable change of coordinates Ψ in such a way that E corresponds to the line $\{(x_1, x_2) : x_2 = 0, 0 \leq x_1 \leq a\}$ for some $a > 0$. If E is of C, S, or K type then we define Ψ in such a way that the vector field Y (defined on $U'(x') \cup U'(y')$) corresponds to the vector field $(0, 1)$. If E is of F type and the region on one side of F is positive then again we let Y correspond to $(0, 1)$, otherwise we let X correspond to $(0, 1)$. For each type of curve we have shown an example in [11]. We choose the system $\Sigma(E)$ that gives an example of frame curve D of the same type of E and is defined on an open set $U(E)$. If E is of C type, we can choose $\Sigma(E)$ in such a way that Δ_A, Δ_B have the right sign, i.e., compatible with the systems $\Sigma(x'), \Sigma(y')$. We define a diffeomorphism $\Psi' : U(E) \to U'(E)$, where $U'(E)$ is a neighborhood of E, in such a way that Ψ' establishes a canonical correspondence between D and E, and its differential $d\Psi'$ sends either the vector field Y or X onto the vector field $(0, 1)$, following the same rules used for Ψ.

We now glue together the systems defined near points and edges. Let V_1, V_2 be two open neighborhoods of x' verifying

$$Cl(V_1) \subset V_2 \subset Cl(V_2) \subset U'(x')$$

and consider a smooth function $h_{x'}$ defined on

$$U = U'(x') \cup U'(y') \cup U'(E)$$

such that

$$h_{x'} \restriction V_1 \equiv 1, \qquad h_{x'} \restriction U \setminus V_2 \equiv 0.$$

We define $h_{y'}$ in the same way for y'. Let $(F', G'), (F'', G'')$ be the vector fields already defined on $U'(x') \cup U'(y'), U'(E)$, respectively, and define them to be zero elsewhere in U. We set

$$\tilde{F} \doteq (h_{x'} + h_{y'})F' + (1 - h_{x'} - h_{y'})F'', \qquad \tilde{G} \doteq (h_{x'} + h_{y'})G' + (1 - h_{x'} - h_{y'})G''.$$

In this way we have defined a system $\tilde{\Sigma} = (\tilde{F}, \tilde{G})$ on U. Since the syntheses corresponding to $\Sigma(x'), \Sigma(y')$, and $\Sigma(E)$ coincide on the set of intersections, Γ is well defined on U. However, if E is of C or of S type, it may happen that in the set where $h_{x'}, h_{y'} \neq 0, 1$, the functions $\tilde{\Delta}_A, \tilde{\Delta}_B$, do not have the required properties.

Consider first the case in which E is an S-edge. From E there originate Y-trajectories that enter the half plane $\{(x_1, x_2) : x_2 > 0\}$. In this case,

$$\tilde{X}_2 < 0, \qquad \tilde{X}_1 > 0, \qquad \tilde{G}_1 < 0, \qquad \tilde{\Delta}_A > 0.$$

We define a new system Σ by setting

$$Y \doteq \tilde{Y} + (0, \alpha), \qquad X \doteq \tilde{X},$$

where $|\alpha| < 1$. We have $\Delta_A = (1/2)(1 + \alpha)\tilde{X}_1 > 0$. If $\alpha(x_1, 0) \equiv 0$ then, after straightforward calculations, we obtain

$$\Delta_B(x_1, 0) = \frac{1}{2}\left(2\tilde{\Delta}_B + \frac{\partial \alpha}{\partial x_2}\tilde{G}_1\tilde{X}_2\right), \tag{5.1}$$

and then we can choose $(\partial\alpha/\partial x_2)(x_1, 0)$ in such a way that $\Delta_B(x_1, 0) \equiv 0$. Moreover,

$$\nabla\Delta_B(x_1, 0) = \nabla\tilde{\Delta}_B(x_1, 0) + \Theta_1 + \Theta_2, \qquad \Theta_1 = \begin{pmatrix} 0 \\ \frac{\partial^2 \alpha}{\partial x_2^2}\tilde{G}_1\tilde{X}_2 \end{pmatrix},$$

$$\Theta_2 = \frac{\partial \alpha}{\partial x_2}\left[\nabla(\tilde{G}_1\tilde{X}_2) + \begin{pmatrix} 0 \\ \tilde{G}_2\frac{\partial \tilde{X}_1}{x_2} - \tilde{G}_1\frac{\partial \tilde{X}_2}{x_2} - \frac{1}{2}[\tilde{X}, \tilde{Y}]_1 \end{pmatrix}\right] + \begin{pmatrix} \frac{\partial^2 \alpha}{\partial x_1 \partial x_2}\tilde{G}_1\tilde{X}_2 \\ \frac{\partial^2 \alpha}{\partial x_2 \partial x_1}\tilde{G}_1\tilde{X}_1 \end{pmatrix};$$

hence Θ_2 is determined by the previous choices but we can define α choosing

$$\frac{\partial^2 \alpha}{\partial x_2^2}(x_1, 0)$$

in such a way that $\nabla\Delta_B(x_1, 0) \neq 0$. From the compactness of E, it follows that there exists a neighborhood U' of E such that $\{x \in U' : \Delta_B(x) = 0\} = \{(x_1, x_2) : x_2 = 0\}$. Then we consider Σ restricted to U'.

Consider now the case in which E is a C-edge. Assume that from E start Y-trajectories that enter the half plane $\{(x_1, x_2) : x_2 > 0\}$ and that $\tilde{X}_1 > 0$ ($\tilde{X}_2 > 0$ follows from $\tilde{Y}_2 > 0$). Again we define $Y = \tilde{Y} + (0, \alpha), X = \tilde{X}$. If we set $\alpha(x_1, 0) = 0$, then (5.1) holds and we can choose $(\partial\alpha/\partial x_2)$ in such a way that $\Delta_B(x_1, 0) \neq 0$. Again by the compactness of E, there exists a neighborhood U' of E in which Δ_B does not vanish. We consider Σ restricted to U'.

Finally, we want to associate with every trajectory γ of Γ a covector field. If γ is contained in $V_1(x')$ or $V_1(y')$ or in $U'(E) \setminus (V_2(x') \cup V_2(y'))$ (see the definitions above), we can associate a dual variable with γ using Ψ or Ψ', because γ corresponds to a trajectory of the synthesis of $\Sigma(x')$ or $\Sigma(y')$ or $\Sigma(E)$. Otherwise assume that γ

verifies $\gamma(t_x) = x \in E \setminus \partial E$. If E is either an F- or K-edge and γ is a Y-trajectory, resp., X-trajectory, then we choose λ_γ such that $\lambda_\gamma \cdot G(x) > 0$, resp., < 0. If E is either an S- or a C-edge and γ is a Y-trajectory, resp., X-trajectory, after t_x, then we choose λ_γ in such a way that $\lambda_\gamma \cdot G(x) = 0$ and, if E is a C edge, $\lambda_\gamma \cdot [F, G](x) > 0$, resp., < 0. We associate to γ the adjoint variable that verifies $\lambda(t_x) = \lambda_\gamma$. It is clear that if $\gamma(I)$ is not a turnpike for every $I \subset Dom(\gamma)$, then (γ, λ) satisfies the PMP on some neighborhood of t_x. Assume now that $\gamma(I)$ is a turnpike, $I = [a, b]$. Let φ_S be the control defined in (2.16) of [2], and consider the system

$$\begin{cases} \dot{x} = F(x) + \varphi_S(x)G(x), \\ \dot{\lambda} = -\lambda \cdot (\nabla F(x) + \varphi_S(x)\nabla G(x)), \end{cases} \tag{5.2}$$

and the following submanifold of $\mathbb{R}^4$:

$$Z = \{(x, \lambda) : \lambda \cdot G(x) = 0\}.$$

From the definition of λ, we have $\lambda(b) \cdot G(\gamma(b)) = 0$. Since $\Delta_B(\gamma(t)) = 0$ for $t \in [a, b]$, from

$$\frac{d}{dt}(\lambda \cdot G) = \lambda \cdot [F, G],$$

we have

$$\lambda(t) \cdot G(\gamma(t)) = 0 \quad \Rightarrow \quad \frac{d}{ds}(\lambda(s) \cdot G(\gamma(s)))\Big|_{s=t} = 0.$$

By the standard theory of ODEs on closed set, we obtain the existence of a solution (x, μ) that verifies $x(b) = \gamma(b), \mu(b) = \lambda(b)$, and $(x(t), \mu(t)) \in Z$ for every $t \in [a, b]$. Since the right-hand side of (5.2) is Lipschitz continuous, there is a unique solution for every initial data. Hence $\lambda(t) \cdot G(\gamma(t)) = 0$ for every $t \in [a, b]$. We conclude that (γ, λ) satisfies the PMP.

From the compactness of E there exists a neighborhood U'' of E such that every $\gamma \in \Gamma$ restricted to U'' is extremal. We consider Σ restricted to U''.

In this way we have defined Σ, Γ on an open set that contains all frame points and curves. Now we complete the definition of Σ, Γ considering the regions enclosed by edges.

For every region $A \subset R$ let $B_i(A)$, $i = 1, \ldots, n(A)$, be the connected components of $A \setminus L(A)$, where $L(A)$ is the union of lines in A. Let $\mathcal{B}$ be the set of all $B_i(A)$, $i = 1, \ldots, n(A)$, as A ranges over the set of regions contained in R. We will define Σ on every B by induction. From $(\mathcal{G}5)$ we have that every $Cl(B)$, $B \in \mathcal{B}$, contains exactly one entrance $E(B)$. The induction hypotheses is that for every $x \in E(B)$ there exists $\gamma_x : [0, t_x] \to \mathbb{R}^2$, $\gamma_x \in \Gamma$, such that $\gamma_x(t_x) = x$; i.e., the system Σ is constructed along γ_x backward in time. We start defining Σ on the regions B for which $E(B)$ is of X or Y type. Then we consider the regions B such that on the region B', that lies on the other side of $E(B)$, the system Σ is already defined. If $E(B)$ is of S type and x is the initial point of $E(B)$, then we consider B if there is a trajectory γ_x that verifies the induction hypothesis. In a finite number of steps we define Σ on every $B \in \mathcal{B}$.

Now fix a region $B \in \mathcal{B}$ and assumes that the induction hypothesis holds. From $(\mathcal{G}5)$ we have that $Cl(B)$ contains exactly one entrance E_1 and one exit E_2. If $E_1 \sim E_2$ then B is enclosed by E_1, E_2, and either a line l or a side E_3. Otherwise, B is enclosed

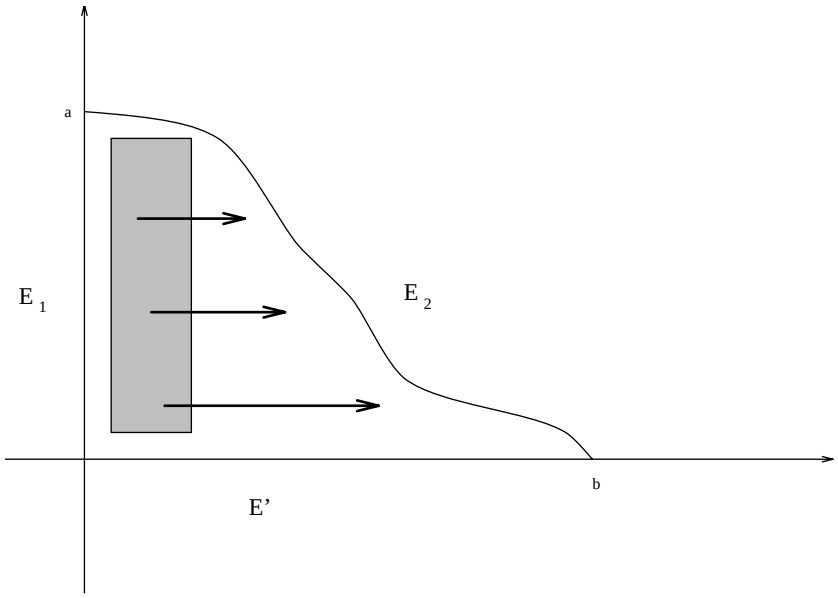


FIG. 5.

by E_1, E_2, a line l_1, and either another line l_2 or a side E_3. We define Σ on B defining Y or X, and G. Indeed, we define Σ also on a neighborhood of the lines in B if the system is not already defined near these lines. Consider the case $E_1 \sim E_2$ and assume that B is positive, being similar to the other case. Possibly using a change of coordinates, we can assume that

$$E_1 = \{(x_1, x_2) : x_1 = 0, 0 \le x_2 \le a\}, \qquad E' = \{(x_1, x_2) : x_2 = 0, 0 \le x_1 \le b\},$$

where either $E' = l$ or $E' = E_3$, and that Y is the constant vector field $(1, 0)$. We could define $Y \equiv (1, 0)$ on B and let Γ be formed by Y-trajectories, but we have to make some modifications to ensure that every $\gamma \in \Gamma$ is extremal.

Consider $\gamma_y \in \Gamma$ that verifies $\gamma_y(t_1) = (0, y)$, $\gamma_y(t_2) \in E_2$. By the induction hypothesis such a trajectory γ_y exists defined on $[0, t_2]$ for every $y \in [0, a]$. Since we have already defined Σ on a neighborhood of $E_1 \cup E_2$, there is a covector field λ_y associated with γ_y that is defined on

$$I = [t_1, t_1 + \mu_1] \cup [t_2 - \mu_2, t_2]$$

for some positive μ_1, μ_2. It can happen that $t_1 + \mu_1 = t_2 - \mu_2$, e.g., if we are near the point $E_1 \cap E_2$. We want to define Y in such a way that we can associate with γ_y a covector field, defined on $Dom(\gamma_y)$, that coincides with λ_y on I. This will ensure, choosing G in a suitable way, that every γ_y is extremal.

Consider a region

$$\Omega = [\delta_1, \delta_2] \times [\varepsilon, a - \delta_3], \qquad \delta_3 > 0, \qquad 0 < \delta_1 < \delta_2,$$

such that the following holds: $\Omega \subset A$, where A is the region containing B, and $\Omega \cap (E_1 \cup E_2) = \emptyset$. See Fig. 5 where Ω is the darkened region.

Let B' be the region on the other side of E'. If Σ is already defined on B', then $\varepsilon > 0$; otherwise $\varepsilon < 0$. Notice that if E' is a side (see the definition in section 3)

then it is of X or of Y type and the former case holds. We choose ε, δ_3 in such a way that Σ is already defined on $B \cap \{(x_1, x_2) : a - 2\delta_3 \leq x_2 \leq a\}$, and if $\varepsilon > 0$, then Σ is already defined on $B \cap \{(x_1, x_2) : 0 \leq x_2 \leq 2\varepsilon\}$. For every $y \in [\varepsilon, a - \delta_3]$, let $\gamma_y^1 \in \Gamma$ be the trajectory that verifies $\gamma_y^1(t_1(y)) = (0, y)$ and let λ_y^1 be the covector field associated with γ_y^1. We have that γ_y^1, λ_y^1 are defined on a neighborhood of $t_1(y)$. Consider the Mayer problem with final target E_2 and the cost function

$$\psi(T, x(T)) = -T + \psi_0(x(T)) \tag{5.3}$$

depending on terminal point and time, where we want to maximize ψ. For every $y \in [\varepsilon, a - \delta_3]$ let $\bar{x}(y)$ be such that $(\bar{x}(y), y) \in E_2$. There exists trajectories $\gamma_y^2 \in \Gamma$ that reach $(\bar{x}(y), y)$ with an associated covector field λ_y^2. Let $t_2(y)$ be such that $\gamma_y^2(t_2(y)) = (\bar{x}(y), y)$. Observe that γ_y^2, λ_y^2 are defined on a neighborhood of $t_2(y)$. We can define ψ in such a way that $(\gamma_y^2, \lambda_y^2)$ satisfies the PMP and the final transversality condition for the Mayer problem; see [6]. Indeed the PMP is satisfied because γ_y^2 is extremal for the time optimal problem. To satisfy the transversality condition, in view of (5.3) we need to find λ_0, λ_1 solution to

$$\lambda_0 = \max_{|\omega| \leq 1} \lambda_y^2(t_2(y)) \cdot (F + \omega G)(\bar{x}(y), y), \tag{5.4}$$

$$\lambda_y^2(t_2(y)) = \lambda_0 \nabla \psi_0(\bar{x}(y), y) + \lambda_1 n_2(\bar{x}(y), y), \tag{5.5}$$

where n_2 is a unit normal vector to E_2. Hence, (5.4) determines λ_0 and (5.5) gives a condition for λ_1, ψ_0.

Choose ν_1, ν_2, T_1, T_2 such that $\delta_1 < \nu_1 < \nu_2 < \delta_2$ and

$$T_1 > \sup\{t_1(y) : y \in [\varepsilon, a - \delta_3]\}, \qquad T_2 < \inf\{\psi_0((\bar{x}(y), y)) : y \in [\varepsilon, a - \delta_3]\}.$$

We define $Y = (\alpha, 0)$ on Ω, α continuous and positive, $\alpha \equiv 1$ on $\partial\Omega \cup [\nu_1, \nu_2] \times [\varepsilon, a - \delta_3]$, and we let $Y = (1, 0)$ outside Ω. We choose α in such a way that the following holds. For every y we have $\gamma_y^1(T_1) = (\nu_1, y)$. If $T_2(y) < t_2(y)$ is the time at which γ_y^2 reaches, backward in time, the point (ν_2, y), then

$$\psi(t_2(y) - T_2(y), (\bar{x}(y), y)) = T_2.$$

With this definition of Y we prolong $\gamma_y^{1,2}, \lambda_y^{1,2}$ defining them on the whole set B. Consider the reachable set $R(T_1)$; we have that

$$\{(\nu_1, y) : \varepsilon \leq y \leq a - \delta_3\} \subset \partial R(T_1).$$

Since $\lambda_y^1(T_1)$ has to be perpendicular to $\partial R(T_1)$, it follows that $\lambda_y^1(T_1)$ has the second component equal to zero. From Theorem 8.2 of Chapter IV of [4] we have that λ_y^2 has to be perpendicular to the level set of the function

$$\psi'(x, y) = \psi(t_2(y) - t(x, y), (\bar{x}(y), y)),$$

where $t(x, y)$ is defined by $\gamma_y^2(t(x, y)) = (x, y)$. Hence also the second component of $\lambda_y^2(T_2(y))$ has to be zero. By the PMP, since the Hamiltonian is positive (see (PMP2) of section 2 of [2]), the first components of $\lambda_y^1(T_1), \lambda_y^2(T_2(y))$ have the same sign. Since $\alpha = 1$ on $[\nu_1, \nu_2] \times [\varepsilon, a - \delta_3]$, we obtain that λ_y^1, λ_y^2 coincide up to a scalar multiple. We can now associate with every γ_y^1 the covector field λ_y^1 and define G in such a way that G is of class $\mathcal{C}^3$ and every γ_y^1 is extremal.

It may happen, however, that α is not smooth and hence Σ is not smooth. Since α is continuous there exists a sequence α_n of smooth functions converging uniformly to α. Let Σ_n, Γ_n be the system and synthesis associated with α_n. If E_2 is of K or F type, then for n large every $\gamma \in \Gamma_n$ is extremal and we are done. Indeed in this case no trajectory of Γ switches on E_2, and by compactness the same holds, if n is sufficiently large, for Γ_n. If E_2 is of C type then Σ_n has a switching curve C_n near to E_2. Since Σ has not already been defined on the region B' that lies on the other side of E_2, we can define $\Sigma = \Sigma_n$ for n sufficiently large. The only change is that we construct the system on a graph equivalent to $\mathcal{G}$, not exactly on $\mathcal{G}$.

The other case, that is, when B is enclosed by E_1, E_2 and either two lines or one line and one side, can be treated in an entirely similar manner. This concludes the construction on the regions $B \in \mathcal{B}$, and then we have defined Σ and Γ on the whole R.

We can again modify Σ on the regions $B \in \mathcal{B}$, using the same techniques described above in such a way that the following holds. If $\gamma \in \Gamma$ reaches $Fr(R)$, then it reaches $Fr(R)$ at time 1. If x belongs to an overlap curve, $\gamma_1, \gamma_2 \in \Gamma$, $\gamma_1(t_1) = x = \gamma_2(t_2)$, then $t_1 = t_2 \leq 1$, with equality holding only if $x \in Fr(R)$.

We can conclude that every $\gamma \in \Gamma$ is optimal and then $R = R(1)$, the reachable set in time 1 for Σ. It is possible to apply Theorem 3.1 of [17] or to use a dynamic programming argument. Indeed the time along the set of trajectories Γ satisfies the Hamilton–Jacobi–Bellman equation for the value function inside R and is constant on its frontier; see [1], [3]. It can happen that some points are reached by more than one trajectory of Γ. However, we can construct a synthesis from Γ, that we call again Γ, following the procedure described in section 5 of [10]. We obtain $\Gamma = \Gamma_{\mathcal{A}}(\Sigma)$. From the construction it is clear that $\mathcal{G}$ corresponds to Σ in the canonical way. □

6. Systems on two-dimensional manifolds. Since all the geometric techniques used in [2], [10], [11], [14] are local it is possible to establish analogous results for a control system defined on a general smooth two-dimensional manifold. In this case, the classification program is completed giving a new definition of graph. Indeed we have to allow reachable sets that are not diffeomorphic to a subset of the plane. By definition, a graph is a stratification (see [12]) of a connected two-dimensional manifold with the following properties.

DEFINITION 4 (graph 2). *A graph $\mathcal{G}$ is a finite collection of disjoint connected manifolds M_i such that the following holds. Each manifold M_i is two, one, or zero dimensional. For each i, $Fr(M_i)$ is the union of manifolds of $\mathcal{G}$ having strictly lower dimension. The two- (respectively, one-, zero-) dimensional manifolds correspond to regions (respectively, frame curves, frame points) of the reachable set of the system. The one-dimensional manifolds are of six types X, Y, C, S, K, or F and satisfy the same conditions of edges. The two-dimensional manifolds are not enclosed by F edges and have a sign. On each two-dimensional manifold there are possibly some marked one-dimensional submanifolds with boundary, which correspond to lines and follow the same rules. The union of the manifolds M_i is a two-dimensional connected manifold with boundary denoted by $\mathcal{T}_{\mathcal{G}}$.*

In this case, we define the relation between one-dimensional manifolds of a graph as for edges, i.e., $M_i \sim M_j$ if and only if $Fr(M_i) \cap Fr(M_j) \neq \emptyset$.

The admissibility conditions are defined in the same way as before. Given an admissible graph $\mathcal{G}$ our aim is to construct a corresponding system on a suitable manifold. For doing this, it is necessary to single out a manifold M that contains a subset diffeomorphic to $\mathcal{T}_{\mathcal{G}}$. In particular we are interested in conditions to ensure that $\mathcal{G}$ corresponds to a planar system.

The partition of $\mathcal{T}_\mathcal{G}$, naturally defined by $\mathcal{G}$, establishes a kind of triangulation, and proceeding as follows it is possible to determine its Euler–Poincaré characteristic $\chi(\mathcal{T}_\mathcal{G})$.

Recall that a finite collection T_i of boundary manifolds homeomorphic to a triangle is a triangulation provided that for every i, j, $T_i \cap T_j$ is empty, a *vertex*, or an *edge* (where by definition *vertices*, respectively, *edges*, are the images of the vertices, respectively, edges, of the corresponding triangle). If we denote by $\#T$, $\#E$, and $\#V$ the number of T_i's, edges, and vertices, respectively, then

$$\chi(\mathcal{T}_\mathcal{G}) = \#T - \#E + \#V. \tag{6.1}$$

Observe that the stratification of $\mathcal{T}_\mathcal{G}$ is not a triangulation. For every two-dimensional manifold M_i fix an internal point y and consider the frame points $x_1, \dots, x_n$ of ∂M_i. There exist n one-dimensional manifolds γ_j, $j = 1, \dots, n$, that connect x_j with y and that do not intersect each other. In this way we divide $\mathcal{T}_\mathcal{G}$ in *triangles*. It can happen that two *triangles* have two vertices in common on the boundary of $\mathcal{T}_\mathcal{G}$. However, it is easy to see that we can add an edge obtaining an admissible triangulation without changing the right-hand side of (6.1).

Now, the number of vertices is equal to the number $\#FP$ of frame points plus the number of regions $\#R$. Every frame point x produces two new edges if it either lies on the frontier of $\mathcal{T}_\mathcal{G}$ or is of type $(X,C)_3, (X,K)_3, (C,C)_1, (C,S)_2, (C,K)_1$, or (S,K) (see [11] for notations). In this case we say that x is a 2-frame point. If the frame point is of one of the remaining cases then it produces three new edges and we say that x is a 3-frame point. Let $\#FC$, $\#2FP$, and $\#3FP$ denote, respectively, the number of frame curves, 2-frame points, and 3-frame points. Then, the total number of edges is $\#FC + 2 \cdot \#2FP + 3 \cdot \#3FP$. Every region is divided in as many parts as the number of frame points on its boundary. Hence the number of triangles is $2 \cdot \#2FP + 3 \cdot \#3FP$. Finally, we obtain

$$\chi(\mathcal{T}_\mathcal{G}) = \#R - \#FC + \#FP.$$

Let $\#\mathcal{F}$ denote the number of equivalence classes of frame curves of type F. From the classification of compact two-dimensional boundary manifolds (see [5], [7]) it follows that the orientability, the Euler–Poincaré characteristic, and the number of connected components of the boundary determine one equivalence class of diffeomorphic manifolds. In particular, every reachable set of a planar system is diffeomorphic to a manifold H_k obtained from the unit closed ball B removing the interior of k distinct balls $B_1, \dots, B_k \subset Int(B)$. Moreover, $\chi(H_k) = 1 - k$ and ∂H_k has $k + 1$ connected components. Thus, we obtain the following theorem.

THEOREM 6.1. *An admissible graph $\mathcal{G}$ corresponds to a planar system if and only if $\mathcal{T}_\mathcal{G}$ is orientable and there exists $k \geq 0$ such that $\#R - \#FC + \#FP = 1 - k$ and $\#\mathcal{F} = k + 1$.*

REFERENCES

[1] M. BARDI, *A boundary value problem for the minimum-time function*, SIAM J. Control Optim., 27 (1989), pp. 776–785.

[2] A. BRESSAN AND B. PICCOLI, *Structural stability for time-optimal planar syntheses*, Dynamics of Continuous, Discrete and Impulsive Systems, 3 (1997), pp. 335–371.

[3] L.C. EVANS AND M.R. JAMES, *The Hamilton–Jacobi–Bellman equation for the time-optimal control*, SIAM J. Control Optim., 27 (1989), pp. 1477–1489.

[4] W.H. FLEMING AND R.W. RISHEL, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.

[5] M.W. Hirsch, *Differential Topology*, Springer-Verlag, New York, 1976.
[6] E.B. Lee and L. Markus, *Foundations of Optimal Control Theory*, Wiley, New York, 1967.
[7] W.S. Massey, *Algebraic Topology: An Introduction*, Springer-Verlag, New York, 1967.
[8] M.M. Peixoto, *On the classification of flows on* 2*-manifolds,* in Dynamical Systems, M.M. Peixoto, ed., Academic Press, New York, 1973, pp. 389–419.
[9] M.M. Peixoto, *Generic properties of ordinary differential equations*, in Studies in Ordinary Differential Equations, MAA Studies in Mathematics 14, J. Hale, ed., Washington, 1977, pp. 52–92.
[10] B. Piccoli, *Regular time-optimal syntheses for smooth planar systems*, Rend. Sem. Mat. Univ. Padova, 95 (1996), pp. 59–79.
[11] B. Piccoli, *Classification of generic singularities for the planar time optimal synthesis*, SIAM J. Control Optim., 34 (1996), pp. 1914–1946.
[12] H.J. Sussmann, *Analytic stratifications and control theory*, in Proc. 1978 Int. Congress of Mathematicians, Helsinki, 1980, pp. 865–871.
[13] H.J. Sussmann, *Lie brackets, real analyticity and geometric control*, in Differential Geometric Control Theory, Proceedings of the conference held at the Michigan Technological University 1982, R.W. Brockett, R.S. Millman, and H.J. Sussmann, eds., Birkhauser, Boston, 1983, pp. 1–116.
[14] H.J. Sussmann, *The structure of time-optimal trajectories for single-input systems in the plane: The* $\mathcal{C}^\infty$ *nonsingular case*, SIAM J. Control Optim., 25 (1987), pp. 433–465.
[15] H.J. Sussmann, *The structure of time-optimal trajectories for single-input systems in the plane: The general real analytic case*, SIAM J. Control Optim., 25 (1987), pp. 868–904.
[16] H.J. Sussmann, *Regular synthesis for time-optimal control of single-input real-analytic systems in the plane*, SIAM J. Control Optim., 25 (1987), pp. 1145–1162.
[17] H.J. Sussmann, *Synthesis, presynthesis, sufficient conditions for optimality and subanalytic sets*, in Nonlinear Controllability and Optimal Control, H.J. Sussmann, ed., Marcel Dekker, New York, 1990, pp. 1–20.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 1, pp. 33–81, January 1998

003

MODEL REFERENCE ADAPTIVE CONTROL OF DISTRIBUTED PARAMETER SYSTEMS*

MICHAEL BÖHM†, M. A. DEMETRIOU‡, SIMEON REICH§, AND I. G. ROSEN¶

Abstract. A model reference adaptive control law is defined for nonlinear distributed parameter systems. The reference model is assumed to be governed by a strongly coercive linear operator defined with respect to a Gelfand triple of reflexive Banach and Hilbert spaces. The resulting nonlinear closed-loop system is shown to be well posed. The tracking error is shown to converge to zero, and regularity results for the control input and the output are established. With an additional richness, or persistence of excitation assumption, the parameter error is shown to converge to zero as well. A finite-dimensional approximation theory is developed. Examples involving both first- and second-order, parabolic and hyperbolic, and linear and nonlinear systems are discussed, and numerical simulation results are presented.

Key words. model reference adaptive control, parameter convergence, persistence of excitation, distributed parameter systems, infinite-dimensional systems, finite-dimensional approximation

AMS subject classifications. 93C40, 93C20, 93B40, 93C25, 65J15, 47H15

PII. S0363012995279717

1. Introduction.

In this paper we develop a model reference adaptive control (MRAC) scheme for rather broad classes of, in general, nonlinear distributed parameter systems. By a distributed parameter system we mean one in which the state space is infinite dimensional such as occurs in the case of partial differential equations. In the context of finite-dimensional systems, MRAC is one of the standard approaches taken in designing a control law for a plant with unknown parameters. A complete description and analysis of a variety of approaches to MRAC can be found in any one of a number of standard texts on adaptive control (see, for example, [2], [14], [30], and [35]). The objective of an MRAC scheme is to determine a feedback control law which forces the state of the plant to asymptotically *track* the state of a given *reference model*. At the same time, the unknown parameters in the plant model are estimated and used to update the control law. Typically, the resulting closed loop system consisting of the plant, the reference model, and the estimator, will be nonlinear. This is true even if the underlying plant and reference models, and the estimator, are linear. The nonlinearity arises in the coupling. Consequently, the scheme requires a careful stability analysis to ensure that all signals (both input and output) remain, in some sense, bounded. It is also desirable, although not necessarily essential, that some sort of *parameter convergence* be achieved.

*Received by the editors January 11, 1995; accepted for publication (in revised form) October 6, 1996.

http://www.siam.org/journals/sicon/36-1/27971.html

†Fachbereich Mathematik, Humboldt-Universität zu Berlin, 10099, Berlin, Germany. The research of this author was supported in part by DFG.

‡Center for Research in Scientific Computation, Department of Mathematics, North Carolina State University, Raleigh, NC 27695 (mdemetri@ecs.ncsu.edu). The research of this author was supported in part by Air Force Office of Scientific Research grant AFOSR F49620-93-1-0198, NASA grant NAG-1-1600, and Air Force Office of Scientific Research grant AFOSR 91-0076.

§Department of Mathematics, Technion-Israel Institute of Technology, 32000 Haifa, Israel (sreich@techunix.technion.ac.il). The research of this author was supported in part by the Fund for the Promotion of Research at the Technion.

¶Center for Applied Mathematical Sciences, Department of Mathematics, University of Southern California, Los Angeles, CA 90089-1113 (rosen@mathc.usc.edu). The research of this author was supported in part by Air Force Office of Scientific Research grant AFOSR 91-0076.

The focus of the effort we describe here is the extension to infinite-dimensional systems of one approach to finite-dimensional MRAC. We consider nonlinear plants with the rather standard restriction that their dependence on the unknown parameters be affine. The operator describing the dynamics of the reference model is assumed to be linear and strongly V-coercive (in a Gelfand triple setting). The parameter space can be either finite or infinite dimensional, and the estimator dynamics for the unknown parameters are chosen in a fashion which renders the closed-loop error equations skew-self-adjoint. This is analogous to what is done in finite dimensions and has the effect of facilitating both tracking error and parameter convergence by forcing the time derivative of a certain energy functional to be negative semidefinite. We establish the global well-posedness of the closed-loop system via two different approaches. First we argue existence of a local solution and then its continuation by treating the closed-loop system as semilinear (i.e., a nonlinear perturbation of a linear system) with the linear component of the dynamics being the infinitesimal generator of an analytic semigroup. The second approach involves the application of an abstract version of the implicit function theorem to obtain a global solution when the initial tracking and parameter error are sufficiently small. Using an analogue of Barbălat's lemma, we establish that the tracking error approaches zero asymptotically. We also establish regularity results for both the input and output signals. In particular, we establish a boundedness result for the control signal. With the additional assumption of *persistence of excitation*, a richness condition on the plant, reference model, and input reference signal, we establish parameter convergence. The definition of persistence of excitation for infinite-dimensional systems given in section 3 below is the natural extension of the analogous definition for finite-dimensional systems as found in, for example, [29], [30], and [31]. Since the reference model and estimator are, in general, infinite dimensional, implementation requires some form of finite-dimensional approximation. Consequently, we develop an abstract finite-dimensional approximation and convergence theory. Finally, we illustrate the application of our general theory on a number of examples involving a variety of linear and nonlinear distributed parameter systems.

One drawback of our approach is that it requires (as do the analogous finite-dimensional schemes, see, for example, [30]) measurement of the full state and distributed input. Eliminating either of these restrictions represents a formidable challenge. For example, if only a partial state measurement is available, a coupled adaptive observer would be required. The corresponding analysis would be significantly more complicated than the already rather technical arguments we present here. We are currently looking at the extension of our treatment here to include partial measurements and finite-dimensional input.

Our effort here is related to our earlier treatment of adaptive identification for distributed parameter systems in [7], [8], and [9]. In fact, we employ the same estimator here to identify the unknown parameters in the plant, and the arguments used below (infinite-dimensional analogues of the finite-dimensional theory presented in [29] and [31]) to demonstrate the asymptotic convergence of the tracking and parameter error to zero, are similar to the ones employed to establish state and parameter convergence for the identification schemes. However, in the case of the identification schemes, the resulting estimator equations are linear. In the case of MRAC, the resulting closed-loop system is nonlinear. Consequently, certain aspects of the analysis, in particular, those dealing with the well-posedness of the closed-loop system and the convergence of the finite-dimensional approximation, are more delicate. Other related treatments

of on-line or adaptive identification for distributed parameter systems can be found in [1], [6], [16], [17], [18], and [37].

Recently there has been some attention given to the adaptive control of distributed parameter systems. First, with respect to approaches other than model reference, indirect adaptive control algorithms for a class of infinite-dimensional stochastic evolution equations have been developed by Duncan, Pasik-Duncan, and their coworkers in a recent series of papers [12], [13], [11], and [32]. Their approach involves the use of a least squares based estimator together with a linear quadratic control design. Parameter convergence together with a continuous dependence result (with respect to the unknown parameters) for the solutions to the operator algebraic Riccati equations yield convergence of the adaptive control law to the nonadaptive optimal LQ controller. Also, Kobayashi in [21], [22], [23], [24], and [25] has proposed a number of direct schemes based upon an input/output formulation. His approach is primarily directed toward the case of unknown input and/or output operators (i.e., the B and C operators) and places a number of restrictions on the A operator (for example, that it be self-adjoint, its eigenvalues be known, that only a finite number of modes be unstable, etc.).

In [40] and [41] a finite-dimensional approach to MRAC based upon the so-called command generator tracker is extended to infinite dimensions. The command generator tracker theory deals with the problem of a mismatch in the dimensionality of the plant and the reference model by assuming that there is an infinite-dimensional system that is input/output equivalent to the reference model. The authors establish closed-loop stability (and robustness properties) via a Lyapunov argument (which in infinite dimensions must be done with care) under a number of rather technical assumptions.

In a recent effort by Hong and Bentsman [19] the authors consider the MRAC of linear parabolic partial differential equations. The results in [19] apply only to plants and reference models which are linear parabolic partial differential equations with Dirichlet boundary conditions under the assumption that the reference signal and the plant and reference model parameters are analytic. In the present treatment, the structure of the plant and reference model are essentially independent and must satisfy only a few relatively mild abstract assumptions. In particular, we consider general nonlinear plants and require only that the reference model (but not the control system) dynamics be strongly V-coercive (in a Gelfand triple sense).

An outline of the remainder of the paper is as follows. In section 2 we define the plant, reference model, and estimator, we derive the closed-loop system, and we establish well-posedness. In section 3 we establish the convergence of the tracking error to zero, we define persistence of excitation, and we demonstrate parameter convergence. The finite-dimensional approximation and convergence theory is discussed in section 4, and examples and the results of our numerical studies are presented in section 5.

2. The MRAC Problem. Let $\{H, \langle\cdot,\cdot\rangle, |\cdot|\}$ be a Hilbert space over $\mathbf{R}$, and let $\{V, \|\cdot\|\}$ be a reflexive Banach space over $\mathbf{R}$ which is densely and continuously embedded in H. Then (see, for example, [27], [38], or [39])

$$V \hookrightarrow H \hookrightarrow V^*, \tag{2.1}$$

with the embeddings dense and continuous where V^* denotes the continuous dual of V. The notation $\langle\cdot,\cdot\rangle$ will also be used to denote the duality pairing between V^* and

V induced by the continuous and dense embeddings given in (2.1). That is, for $\varphi \in V^*$ and $\psi \in V$, $\langle \varphi, \psi \rangle$ denotes the action of the bounded linear functional φ on the vector ψ. Note that when φ is in fact an element in H (or, more precisely, can be identified with an element in H), the embeddings (2.1) imply that the value of φ *acting* on ψ is equal to the H inner product of φ and ψ. Moreover, since $\overline{H} \cong V^*$, for $\varphi \in V^*$ and $\{\varphi_n\}_{n=1}^{\infty} \subset H$ with $\lim_{n\to\infty} \varphi_n = \varphi$ in V^* we have $\lim_{n\to\infty} \langle \varphi_n, \psi \rangle = \langle \varphi, \psi \rangle$, $\psi \in V$. Consequently this minor abuse of notation is entirely justified. Let $\| \cdot \|_*$ denote the usual norm on V^*, and let $K > 0$ be such that

$$|\varphi| \leq K\|\varphi\|, \qquad \varphi \in V. \tag{2.2}$$

Let $\hat{V}^*$ be a subspace of V^*, and let $\{Q, \langle \cdot, \cdot \rangle_Q, |\cdot|_Q\}$ be a real Hilbert space.

For each $q \in Q$, let $A_1(q) : V \to V^*$ be an, in general, nonlinear operator, and for $q \in Q$, let $Dom(A_1(q)) = \{\varphi \in V : A_1(q)\varphi \in H\}$. Also, we let $A_2 : V \to V^*$ be an, in general, nonlinear operator, and we make the following standing assumptions.

(A1) (V-V^*-boundedness). There exist $\alpha_1, \alpha_2 > 0$ such that

$$|\langle A_1(q)\varphi, \psi \rangle| \leq \alpha_1 |q|_Q \|\varphi\| \|\psi\|, \ \varphi, \psi \in V, \ q \in Q, \text{ and}$$

$$|\langle A_2\varphi, \psi \rangle| \leq \alpha_2 \|\varphi\| \|\psi\|, \ \varphi, \psi \in V.$$

(A2) (Q-linearity). For each $\varphi \in V$, the map $q \to A_1(q)\varphi$ from Q into V^* is linear. For each $q \in Q$, let $A(q) : V \to V^*$ be given by

$$A(q)\varphi = A_1(q)\varphi + A_2\varphi, \qquad \varphi \in V. \tag{2.3}$$

We are interested in adaptively controlling the nonlinear plant given by

$$D_t u(t) + A(\overline{q})u(t) = f(t), \qquad \text{a.e. } t > 0, \tag{2.4}$$

$$u(0) = u_0, \tag{2.5}$$

where $\overline{q} \in Q$ is unknown, $u_0 \in H$, the operator $A(\overline{q})$ is given by (2.3) with $q = \overline{q}$, and the control input f is assumed to satisfy $f \in L_2(0, T; V^*)$ for all $T > 0$ with $f(t) \in \hat{V}^*$, a.e. $t > 0$. We assume minimally that the system (2.4), (2.5) is well-posed in at least some sense. That is, we assume that for sufficiently regular initial data, u_0, and input, f, there exists a *weak* solution. More precisely, we assume that for each $T > 0$, each $u_0 \in U_0$, U_0 a subset of H, and each $f \in L_2(0, T; V^*)$ sufficiently regular, there exists a unique V-valued function u which is V^*-absolutely continuous on $(0, T)$, $u \in C(0, T; H) \cap L_2(0, T; V)$, $D_t u \in L_2(0, T; V^*)$, and which satisfies

$$\langle D_t u(t), \varphi \rangle + \langle A(\overline{q})u(t), \varphi \rangle = \langle f(t), \varphi \rangle, \qquad \varphi \in V, \text{ a.e. } t > 0, \tag{2.6}$$

$$u(0) = u_0. \tag{2.7}$$

Theorem III.2.6 in [5] provides sufficient conditions for the existence of such a solution. Indeed, if the operator $A(\overline{q})$ is hemicontinuous (i.e., $\lim_{\lambda\to 0} \langle A(\overline{q})\{\varphi + \lambda\psi\} - A(\overline{q})\varphi, \chi \rangle = 0$, $\chi \in V$ for any $\varphi, \psi \in V$), monotone (i.e., $\langle A(\overline{q})\varphi - A(\overline{q})\psi, \varphi - \psi \rangle \geq 0$ for all $\varphi, \psi \in V$), bounded (i.e., there exists $\alpha > 0$ for which $\|A(\overline{q})\varphi\|_* \leq \alpha\{1 + \|\varphi\|\}$ for all $\varphi \in V$), and coercive (i.e., there exist $\rho > 0$ and $\sigma \in \mathbf{R}$ for which $\langle A(\overline{q})\varphi, \varphi \rangle \geq \rho\|\varphi\|^2 + \sigma$ for all $\varphi \in V$), then just such a *weak* solution exists for *all* $u_0 \in H$ and *all* $f \in L_2(0, T; V^*)$.

We are interested in designing a *model reference* adaptive controller for the plant, or system, (2.4), (2.5). That is, we wish to find a control input f in feedback form which forces the state of the unknown plant, u, to track the state of a given linear reference model,

$$\langle D_t v(t), \varphi\rangle + \langle A_0 v(t), \varphi\rangle = \langle g(t), \varphi\rangle, \qquad \varphi \in V, \text{ a.e. } t > 0, \tag{2.8}$$

$$v(0) = v_0, \tag{2.9}$$

where $v_0 \in H$, the input reference signal g is assumed to satisfy $g \in L_2(0,T;V^*)$, for all $T > 0$, with $g(t) \in \hat{V}^*$, a.e. $t > 0$, and the operator $A_0 \in \mathcal{L}(V, V^*)$ is assumed to satisfy the following conditions.

(A3) (V-V^*-boundedness). There exists $\alpha_0 > 0$ such that $|\langle A_0\varphi, \psi\rangle| \le \alpha_0\|\varphi\|\|\psi\|$, $\varphi, \psi \in V$.

(A4) (V-coercivity). There exists $\rho_0 > 0$ for which $\langle A_0\varphi, \varphi\rangle \ge \rho_0\|\varphi\|^2$, $\varphi \in V$.

(A5) ($\hat{V}^*$-range). For all $q \in Q$ we have $\mathcal{R}(A(q) - A_0) \subset \hat{V}^*$.

It is well known (see, for example, [27], [38], or [39]) that assumptions (A3) and (A4) are sufficient to conclude that the system (2.8), (2.9) admits a unique solution v satisfying $v \in C(0,T;H) \cap L_2(0,T;V)$ with $D_t v \in L_2(0,T;V^*)$ for all $T > 0$. Let $D_0 = Dom(A_0) = \{\varphi \in V : A_0\varphi \in H\}$. Then assumptions (A3) and (A4) also imply (see, for example, [33], [39], or [38]) that the operator $-A_0$ *restricted* to the subspace D_0 is the infinitesimal generator of an analytic semigroup, $\{T_0(t) : t \ge 0\}$, of bounded linear operators on H. It can also be shown (see [39]) that the operator $-A_0$ is the infinitesimal generator of an analytic semigroup on V^* and that appropriately restricted $-A_0$ generates an analytic semigroup on V (see [3]). Recalling (2.2), it follows that

$$|T_0(t)\varphi| \le e^{-\rho_0 K^{-2} t}|\varphi|, \qquad \varphi \in H, \tag{2.10}$$

and

$$\|T_0(t)\varphi\| \le M e^{-\rho_0 K^{-2} t}\|\varphi\|, \qquad \varphi \in V, \tag{2.11}$$

for some $M > 0$. The solution to the initial value problem (2.8), (2.9) is given by

$$v(t) = T_0(t)v_0 + \int_0^t T_0(t-s)g(s)ds, \qquad t \ge 0. \tag{2.12}$$

The primary motivation for the inclusion of assumption (A5) is to allow us to apply our abstract framework to second-order systems (i.e., abstract wave equations and the like). The relevance of assumption (A5) in this regard will become clearer when we discuss an example involving the control of a one-dimensional damped wave equation in section 5 below.

REMARK 2.1. *Many of the estimates contained in the arguments used to verify several of the results in this and the following section assume the existence of solutions belonging to particular regularity classes (i.e., the domains of certain operators, etc.). Thus our proofs deliver only a posteriori estimates with respect to those assumptions. A more precise argument in all of these cases would proceed as follows. The dynamical system, or initial value problem, would be approximated by a Galerkin system using smooth basis functions chosen as eigenfunctions of the relevant operator (in most cases A_0). A posteriori estimates for the approximating solutions with bounds independent*

of the number of basis functions are established using the same arguments we employ below. These estimates now serve as a priori estimates for the solutions. Weak, weak, and strong compactness properties (Aubin's lemma) of bounded subsets of time dependent functions are then used to obtain corresponding convergent subsequences and the corresponding regular solutions. Note that at the level of Galerkin solutions, with bases formed from eigenfunctions, it is immediately clear that the resulting approximate solutions are sufficiently regular to permit the estimates we make below. In particular, the Galerkin basis functions satisfy appropriate boundary conditions.*

We have the following regularity result for the reference model (2.8), (2.9).

THEOREM 2.2. *For the reference model given by* (2.8), (2.9)*, we have the following results.*

(i) *If* $g \in L_\infty(0,\infty;H)$*, then* $v \in L_\infty(0,\infty;H)$*.*
(ii) *If* $g \in L_\infty(0,\infty;V)$ *and* $v_0 \in V$*, then* $v \in L_\infty(0,\infty;V)$*.*
(iii) *If* $g \in L_2(0,\infty;V^*)$*, then* $v \in L_\infty(0,\infty;H) \cap L_2(0,\infty;V)$*.*
(iv) *If* $g \in L_2(0,\infty;H)$ *is Hölder continuous, i.e.,*

$$|g(t) - g(s)| \le C|t-s|^\rho, \qquad 0 \le t, s, < \infty, \tag{2.13}$$

for some $C > 0$ *and* $\rho \in (0,1]$*, and* $v_0 \in V$*, and if the operator* A_0 *is symmetric in the sense that*

$$\langle A_0\varphi, \psi\rangle = \langle A_0\psi, \varphi\rangle, \qquad \varphi, \psi \in V, \tag{2.14}$$

then $v \in L_\infty(0,\infty;V)$*,* $v(t) \in D_0$*, a.e.* $t > 0$*, and* $A_0 v \in L_2(0,\infty;H)$*.*

Proof. We note that (i)–(iv) are standard results for linear initial value problems. However, in order to establish some estimates for later reference, we include the following proof. Statements (i) and (ii) follow immediately from (2.10), (2.11), and (2.12). To verify (iii), for almost every $t > 0$ we have that

$$\begin{aligned} \frac{1}{2}D_t|v(t)|^2 &= \langle -A_0 v(t) + g(t), v(t)\rangle \\ &\le -\rho_0\|v(t)\|^2 + \|g(t)\|_*\|v(t)\| \\ &\le -\frac{\rho_0}{2}\,\|v(t)\|^2 + \frac{1}{2\rho_0}\,\|g(t)\|_*^2. \end{aligned} \tag{2.15}$$

Integrating both sides of the estimate (2.15) from 0 to t, it follows that

$$|v(t)|^2 + \rho_0\int_0^t \|v(s)\|^2 ds \le |v_0|^2 + \frac{1}{\rho_0}\|g\|^2_{L_2(0,\infty;V^*)}, \qquad t > 0,$$

from which the result is immediately obtained.

To verify (iv), first note that assumption (A4) and (2.14) imply that $A_0 : D_0 \subset H \to H$ is positive definite and self-adjoint. It follows that the square root of A_0, $A_0^{\frac{1}{2}}$, can be defined with $Dom(A_0^{\frac{1}{2}}) = V$ (see, for example, [39]). Moreover, for $\varphi \in V$, $\|\varphi\|_0 = |A_0^{\frac{1}{2}}\varphi|$ defines a norm on V and, by assumptions (A3) and (A4), we have that

$$\rho_0\|\varphi\|^2 \le \langle A_0\varphi, \varphi\rangle = \langle A_0^{\frac{1}{2}}\varphi, A_0^{\frac{1}{2}}\varphi\rangle = \|\varphi\|_0^2 = \langle A_0\varphi, \varphi\rangle \le \alpha_0\|\varphi\|^2 \tag{2.16}$$

for all $\varphi \in V$. The estimate (2.16) yields that the two norms $\|\cdot\|$ and $\|\cdot\|_0$ on V are equivalent.

The assumption of Hölder continuity on g, (2.13), and the fact that $\{T_0(t) : t \ge 0\}$, the semigroup of bounded linear operators on H generated by the operator $-A_0$, is

analytic, are sufficient to conclude that $A_0 v(t) \in H$ for almost all $t > 0$. It follows that $v(t) \in D_0$, a.e. $t > 0$, and recalling Remark 2.1, from (2.8), we obtain that $\langle D_t v(t), A_0 v(t)\rangle + |A_0 v(t)|^2 = \langle g(t), A_0 v(t)\rangle$, a.e. $t > 0$, and therefore that

$$\frac{1}{2} D_t \|v(t)\|_0^2 + |A_0 v(t)|^2 \le |g(t)||A_0 v(t)| \le \frac{1}{2}|g(t)|^2 + \frac{1}{2}|A_0 v(t)|^2, \quad \text{a.e. } t > 0.$$

Integrating the above estimate from 0 to t, and recalling (2.9), we find that

$$\|v(t)\|_0^2 + \int_0^t |A_0 v(s)|^2 ds \le \|v_0\|_0^2 + \int_0^t |g(s)|^2 ds \le \|v_0\|_0^2 + \|g\|^2_{L_2(0,\infty;H)}, \qquad t \ge 0,$$

from which the desired conclusion follows. □

For each $t > 0$, let $e(t) = u(t) - v(t)$. We would like to find a control input, f, such that

$$\lim_{t\to\infty} |e(t)| = 0, \tag{2.17}$$

with f remaining, in some sense, *bounded* (for example, bounded energy; $f \in L_2(0,\infty; V^*)$). If the plant (i.e., $\overline{q}$) were known, the convergence in (2.17) could be achieved by setting

$$f(t) = A(\overline{q})u(t) - A_0 u(t) + g(t), \qquad \text{a.e. } t > 0. \tag{2.18}$$

For then e would satisfy

$$\langle D_t e(t), \varphi\rangle + \langle A_0 e(t), \varphi\rangle = 0, \qquad \varphi \in V, \text{ a.e. } t > 0,$$

$$e(0) = e_0,$$

where $e_0 = u_0 - v_0 \in H$. It follows from assumption (A4) and (2.2) that $|e(t)| \le e^{-\rho_0 K^{-2} t}|e_0|$, $t \ge 0$ and, consequently, that (2.17) is satisfied. The closed-loop system is given by

$$\langle D_t u(t), \varphi\rangle + \langle A_0 u(t), \varphi\rangle = \langle g(t), \varphi\rangle, \qquad \varphi \in V, \text{ a.e. } t > 0, \tag{2.19}$$

$$u(0) = u_0. \tag{2.20}$$

THEOREM 2.3. *For the nonadaptive closed-loop system given by* (2.6), (2.7), (2.18), *or, equivalently,* (2.18)–(2.20), *we have* $f(t) \in \hat{V}^*$ *for a.e.* $t > 0$ *and the following results.*

(i) *If* $g \in L_2(0,\infty; V^*)$, *then* $u \in L_\infty(0,\infty; H) \cap L_2(0,\infty; V)$ *and, moreover,* $f \in L_2(0,\infty; V^*)$.

(ii) *If* $g \in L_\infty(0,\infty; V)$ *and* $u_0 \in V$, *then* $u \in L_\infty(0,\infty; V)$ *and* $f \in L_\infty(0,\infty; V^*)$.

(iii) *If the operator* A_0 *is symmetric in the sense of* (2.14), $u_0 \in V$, *and* $g \in L_2(0,\infty; H)$ *and satisfies* (2.13), *then* $u(t) \in D_0$, *a.e.* $t > 0$, $u \in L_\infty(0,\infty; V)$ *and* $A_0 u \in L_2(0,\infty; H)$. *If, in addition,*

(a) $g \in L_\infty(0,\infty; V^*)$, *then* $f \in L_\infty(0,\infty; V^*)$,

or

(b) *for* $\varphi \in D_0$, $A(\overline{q})\varphi \in H$, *and* $|A(\overline{q})\varphi| \le \gamma |A_0 \varphi|$, $\varphi \in D_0$, *for some* $\gamma > 0$, *then* $f \in L_2(0,\infty; H)$.

Proof. The fact that $f(t) \in \hat{V}^*$, a.e. $t > 0$, follows immediately from assumption (A5) and the assumption that $g(t) \in \hat{V}^*$, a.e. $t > 0$.

Using the same argument used in the proof of Theorem 2.2, we obtain $u \in L_\infty(0,\infty;H) \cap L_2(0,\infty;V)$. Assumptions (A1) and (A3) and the definition of the control input f given in (2.18) yield

$$\|f(t)\|_* \leq \{\alpha_1|\overline{q}|_Q + \alpha_2 + \alpha_0\}\|u(t)\| + \|g(t)\|_*, \qquad \text{a.e. } t > 0, \tag{2.21}$$

from which it follows that $f \in L_2(0,\infty;V^*)$.

The result given in (ii) follows immediately from (2.11); (2.19) and (2.20) imply that

$$u(t) = T_0(t)u_0 + \int_0^t T_0(t-s)g(s)ds, \qquad t \geq 0,$$

and (2.21).

An argument analogous to the one used in the proof of Theorem 2.2(iv) yields $u(t) \in D_0$, a.e. $t > 0$, $u \in L_\infty(0,\infty;V)$ and $A_0u \in L_2(0,\infty;H)$. The result given in (iii)(a) then follows immediately from (2.21), while the estimate $|f(t)| \leq \{\gamma + 1\}|A_0u(t)| + |g(t)|$, a.e. $t > 0$, yields the result given in (iii)(b). □

The importance of Theorem 2.3 lies in the fact that it serves as an upper bound for the results we can hope to obtain for a corresponding adaptive scheme wherein the plant $\overline{q}$ is unknown and is estimated in real time.

Since $\overline{q}$ is in fact unknown, we set

$$f(t) = A(q(t))u(t) - A_0u(t) + g(t), \qquad \text{a.e. } t > 0, \tag{2.22}$$

or

$$\langle f(t),\varphi\rangle = \langle A(q(t))u(t),\varphi\rangle - \langle A_0u(t),\varphi\rangle + \langle g(t),\varphi\rangle, \ \ \varphi \in V, \text{ a.e. } t > 0, \tag{2.23}$$

where for each $t > 0$, $q(t) \in Q$ denotes an adaptively updated *estimate* for $\overline{q}$. Once again, $f(t) \in \hat{V}^*$, a.e. $t > 0$ follows from assumption (A5) and the fact that $g(t) \in \hat{V}^*$, a.e. $t > 0$. By analogy to the finite-dimensional case, and for the purpose of forcing an appropriate energy functional which will be defined in the next section when we consider convergence, we let the adaptation law for q be given by

$$\langle D_tq(t),p\rangle_Q + \langle A_1(p)u(t),e(t)\rangle = 0, \qquad p \in Q, \text{ a.e. } t > 0, \tag{2.24}$$

$$q(0) = q_0, \tag{2.25}$$

where $q_0 \in Q$, and $e(t) = u(t) - v(t)$, $t > 0$. The closed-loop system is then given by

$$\langle D_tu(t),\varphi\rangle + \langle A_0u(t),\varphi\rangle + \langle A_1(\overline{q} - q(t))u(t),\varphi\rangle = \langle g(t),\varphi\rangle, \ \varphi \in V, \text{ a.e. } t > 0, \tag{2.26}$$

$$\langle D_tv(t),\varphi\rangle + \langle A_0v(t),\varphi\rangle = \langle g(t),\varphi\rangle, \qquad \varphi \in V, \ a.e. \ t > 0, \tag{2.27}$$

$$\langle D_tq(t),p\rangle_Q + \langle A_1(p)u(t),u(t)-v(t)\rangle = 0, \qquad p \in Q, \text{ a.e. } t > 0, \tag{2.28}$$

$$u(0) = u_0, \qquad v(0) = v_0, \qquad q(0) = q_0. \tag{2.29}$$

We are interested in showing that the nonlinear system (2.26)–(2.29) is, at least in some sense and under some set of minimally realizable assumptions, well-posed. Recalling that $u(t) = e(t) + v(t)$, and defining the parameter error r to be

$$r(t) = q(t) - \overline{q}, \qquad t > 0, \tag{2.30}$$

we consider instead the equivalent problem of establishing a well-posedness result for the nonlinear system

$$\langle D_t e(t),\varphi\rangle + \langle A_0 e(t),\varphi\rangle - \langle A_1(r(t))\{e(t)+v(t)\},\varphi\rangle = 0,\ \varphi\in V,\quad \text{a.e. } t>0, \tag{2.31}$$

$$\langle D_t v(t),\varphi\rangle + \langle A_0 v(t),\varphi\rangle = \langle g(t),\varphi\rangle,\qquad \varphi\in V,\ \text{a.e. } t>0, \tag{2.32}$$

$$\langle D_t r(t),p\rangle_Q + \langle A_1(p)\{e(t)+v(t)\},e(t)\rangle = 0,\qquad p\in Q,\ \text{a.e. } t>0, \tag{2.33}$$

$$e(0)=e_0,\qquad v(0)=v_0,\qquad r(0)=r_0, \tag{2.34}$$

where $r_0 = q_0 - \overline{q} \in Q$. In the discussion to follow, we present two approaches to demonstrating the well-posedness of the closed-loop system (2.31)–(2.34). We will first demonstrate the existence of a unique strong solution using the theory of semilinear equations with analytic semigroups. The second approach is based upon an application of an implicit function theorem. Necessarily, each of the two approaches will require its own set of additional hypotheses which must be satisfied in order for there to exist a unique solution. The nonlinear system (2.31)–(2.34) is the one we will be using to establish the tracking error and parameter convergence in the next section. It is worth noting that the *skew-self-adjoint*-like structure of the system (2.31)–(2.34) plays an essential role in the analysis to follow in sections 2.1 and 2.2. We also note that the equation for v, (2.32), could be decoupled from the rest of the system and v could be treated as an exogenous input signal. In fact, in our discussion of our convergence and approximation results in sections 3 and 4 to follow, and the implicit function theorem approach to well-posedness, it is convenient, and in some sense essential (for the arguments as given), to do just that. However, we have found that for our analytic semigroup approach to well-posedness, the arguments are most elegantly presented in the context of the complete dynamical system (2.31)–(2.34).

2.1. An analytic semigroup approach to closed-loop well-posedness. Let $X = H\times H\times Q$ be endowed with the inner product

$$\langle(\varphi_1,\psi_1,q_1),(\varphi_2,\psi_2,q_2)\rangle_X = \langle\varphi_1,\varphi_2\rangle + \langle\psi_1,\psi_2\rangle + \langle q_1,q_2\rangle_Q,\ \ (\varphi_i,\psi_i,q_i)\in X,\ i=1,2,$$

and let $|\cdot|_X$ denote the corresponding induced norm. Thus $\{X,\langle\cdot,\cdot\rangle_X,|\cdot|_X\}$ is a Hilbert space. Let $Y = V\times V\times Q$ be endowed with the norm $\|(\varphi,\psi,q)\|_Y = (\|\varphi\|^2+\|\psi\|^2+|q|_Q^2)^{\frac{1}{2}}$, $(\varphi,\psi,q)\in Y$. Then $\{Y,\|\cdot\|_Y\}$ is a reflexive Banach space which is densely and continuously embedded in X. It follows that

$$Y\hookrightarrow X\hookrightarrow Y^*, \tag{2.35}$$

with the embeddings dense and continuous. For $\lambda>0$, define the linear operator $\mathcal{A}_\lambda : Y\to Y^*$ by $\langle\mathcal{A}_\lambda(e,v,r),(\varphi,\psi,q)\rangle_{Y^*,Y} = \langle A_0 e,\varphi\rangle + \langle A_0 v,\psi\rangle + \langle\lambda r,q\rangle_Q$ for $(e,v,r),(\varphi,\psi,q)\in Y$. In the above definition, $\langle\cdot,\cdot\rangle_{Y^*,Y}$ denotes the duality pairing between Y^* and Y induced by the X inner product via the dense and continuous embeddings given in (2.35). Recalling that $D_0 = Dom(A_0) = \{\varphi\in V : A_0\varphi\in H\}$, for $\lambda>0$, define the operator $A_\lambda : Dom(A_\lambda)\subset X\to X$ by $Dom(A_\lambda) = \{(\varphi,\psi,q)\in Y : \mathcal{A}_\lambda(\varphi,\psi,q)\in X\} = D_0\times D_0\times Q$, $A_\lambda(\varphi,\psi,q) = \mathcal{A}_\lambda(\varphi,\psi,q)$, $(\varphi,\psi,q)\in Dom(A_\lambda)$. Note that $Dom(A_\lambda) = Dom(A)$ is independent of $\lambda>0$, that for $\lambda>0$, $-A_\lambda$ is the infinitesimal generator of a uniformly exponentially stable analytic semigroup, $\{T_\lambda(t) : t\geq 0\}$, on X, Y, and Y^*, and that $0\in\rho(-A_\lambda)$, the resolvent set of $-A_\lambda$.

For $\varphi \in V$, define the operator $B(\varphi) : Q \to V^*$ by

$$\langle B(\varphi)q, \psi\rangle = \langle A_1(q)\varphi, \psi\rangle, \qquad q \in Q, \ \ \psi \in V. \tag{2.36}$$

Assumptions (A1) and (A2) imply that for $\varphi \in V$, $B(\varphi) \in \mathcal{L}(Q, V^*)$ with $\|B(\varphi)\| \leq \alpha_1\|\varphi\|$. Recalling that V was assumed to be reflexive, and that Q is a Hilbert space, for $\varphi \in V$, let $B(\varphi)' \in \mathcal{L}(V, Q)$ denote the Banach space adjoint of $B(\varphi)$. That is, for $\varphi \in V$ we have

$$\langle B(\varphi)'\psi, q\rangle_Q = \langle B(\varphi)q, \psi\rangle = \langle A_1(q)\varphi, \psi\rangle, \qquad \psi \in V, \ \ q \in Q. \tag{2.37}$$

For $\lambda > 0$, define $G_\lambda : \mathbf{R}^+ \times Y \to Y^*$ by

$$\langle G_\lambda(t, \Phi), \Psi\rangle_{Y^*,Y} = \langle B(e + v)r, \varphi\rangle + \langle g(t), \psi\rangle + \langle \lambda r - B(e + v)'e, q\rangle_Q,$$

where $t \geq 0$, $\Phi = (e, v, r) \in Y$ and $\Psi = (\varphi, \psi, q) \in Y$.

We consider the system (2.31)–(2.34) written as

$$\langle D_t x(t), \Phi\rangle_{Y^*,Y} + \langle \mathcal{A}_\lambda x(t), \Phi\rangle_{Y^*,Y} = \langle G_\lambda(t, x(t)), \Phi\rangle_{Y^*,Y}, \ \Phi \in Y, \text{ a.e. } t > 0, \tag{2.38}$$

$$x(0) = x_0, \tag{2.39}$$

where $\lambda > 0$, and for each $t \geq 0$, $x(t) = (e(t), v(t), r(t))$. Under appropriate additional assumptions on the input reference signal g, the initial data e_0, v_0, and r_0, and the plant (i.e., the operator $A_1(q)$ for $q \in Q$), we establish the existence of a unique solution to the system (2.38), (2.39) by first establishing the existence of a unique local strong solution to the initial value problem in X given by

$$D_t x(t) + A_\lambda x(t) = G_\lambda(t, x(t)), \qquad \text{a.e. } t > 0, \tag{2.40}$$

$$x(0) = x_0, \tag{2.41}$$

and then showing that it is possible to continue this solution for all $t > 0$. By a strong (or classical) solution on the interval $[0, T)$ to the initial value problem (2.40), (2.41) we mean a function $x : [0, T) \to X$ which is continuous on $[0, T)$, continuously differentiable on $(0, T)$, $x(t) \in Dom(A) = Dom(A_\lambda)$ for $t \in (0, T)$, (2.40) is satisfied for $t \in (0, T)$, and (2.41) is satisfied.

To establish that the initial value problem (2.40), (2.41) is well-posed, we require the following *additional* assumptions.

(A6) (q-independent domain). The subset of V, $D_1 = Dom(A_1(q))$ is independent of $q \in Q$ and for some $\alpha \in (0, 1)$, $Dom(A_0^\alpha) \subset D_1$.

(A7) (A_0^α-boundedness). There exist $\beta_1 > 0$ such that for α as in assumption (A6), we have

$$|A_1(q)\varphi| \leq \beta_1|q|_Q|A_0^\alpha\varphi|, \qquad q \in Q, \ \ \varphi \in Dom(A_0^\alpha), \tag{2.42}$$

(A8) (A_0^α-Lipschitz). There exist $\gamma_1 > 0$ such that for α as in assumption (A6) we have

$$|A_1(q)\varphi - A_1(q)\psi| \leq \gamma_1|q|_Q|A_0^\alpha\varphi - A_0^\alpha\psi|, \quad q \in Q, \ \varphi, \psi \in Dom(A_0^\alpha), \tag{2.43}$$

(A9) (Hölder continuity). For $t \geq 0$, $g(t) \in H$, and there exist $\nu \in (0, 1]$ and $\delta > 0$ such that

$$|g(t) - g(s)| \leq \delta|t - s|^\nu, \qquad t, s \geq 0. \tag{2.44}$$

Note that assumptions (A3) and (A4) are sufficient for fractional powers of the operator A_0 to be well defined (see, for example, [33]).

THEOREM 2.4. *Suppose that assumptions* (A1)–(A9) *hold and that* $e_0, v_0 \in Dom(A_0^\alpha)$, *where* $\alpha \in (0,1)$ *is as in assumption* (A6). *Then there exists a* $T = T(x_0) > 0$ *such that the initial value problem* (2.40), (2.41) *has a unique local solution* $x \in C([0,T);X) \cap C^1((0,T);X)$.

Proof. For $\alpha \in (0,1)$, the linear operator A_0^α is closed and invertible with domain, $Dom(A_0^\alpha)$, dense in H. For the α in assumption (A6), let H_α denote the space $Dom(A_0^\alpha)$ endowed with the graph norm $\|\cdot\|_\alpha$ corresponding to A_0^α. That is, for $\varphi \in Dom(A_0^\alpha)$, $\|\varphi\|_\alpha = |\varphi| + |A_0^\alpha \varphi|$. Note that since A_0^α is closed, H_α is a Banach space, and since A_0^α is invertible, the norm $\|\cdot\|_\alpha$ is equivalent to the norm $|\cdot|_\alpha$ on $Dom(A_0^\alpha)$ given by $|\varphi|_\alpha = |A_0^\alpha \varphi|$ for $\varphi \in Dom(A_0^\alpha)$. Define the Banach space $\{X_\alpha, |\cdot|_{X_\alpha}\}$ by $X_\alpha = H_\alpha \times H_\alpha \times Q$ with $|\Phi|_{X_\alpha} = |\varphi_1|_\alpha + |\varphi_2|_\alpha + |\varphi_3|_Q$ for $\Phi = (\varphi_1, \varphi_2, \varphi_3) \in X_\alpha$.

The theorem will follow at once from Theorem 6.3.1 in [33] once we have established that for some $\lambda > 0$ and any neighborhood, $U \subset X_\alpha$, of x_0, $U = \{x \in X_\alpha : |x - x_0|_{X_\alpha} < \varepsilon\}$, there exists a constant $L = L(U,\lambda) = L(\varepsilon, x_0, \lambda) > 0$, such that

$$\begin{aligned} &|G_\lambda(t,\Phi) - G_\lambda(s,\Psi)|_X \\ &\qquad \le L\{|t-s|^\nu + |A_0^\alpha\varphi_1 - A_0^\alpha\psi_1| + |A_0^\alpha\varphi_2 - A_0^\alpha\psi_2| + |\varphi_3 - \psi_3|_Q\} \\ &\qquad = L\{|t-s|^\nu + |\varphi_1 - \psi_1|_\alpha + |\varphi_2 - \psi_2|_\alpha + |\varphi_3 - \psi_3|_Q\}, \qquad t,s > 0, \end{aligned} \tag{2.45}$$

for all $\Phi = (\varphi_1, \varphi_2, \varphi_3), \Psi = (\psi_1, \psi_2, \psi_3) \in U$. Let $\lambda > 0$ and $\Phi = (\varphi_1, \varphi_2, \varphi_3), \Psi = (\psi_1, \psi_2, \psi_3) \in U$, and consider for $t, s > 0$,

$$\begin{aligned} &|G_\lambda(t,\Phi) - G_\lambda(s,\Psi)|_X^2 \\ &\qquad = |B(\varphi_1+\varphi_2)\varphi_3 - B(\psi_1+\psi_2)\psi_3|^2 + |g(t) - g(s)|^2 \\ &\qquad + |\lambda\{\varphi_3 - \psi_3\} - \{B(\varphi_1+\varphi_2)'\varphi_1 - B(\psi_1+\psi_2)'\psi_1\}|_Q^2. \end{aligned} \tag{2.46}$$

Now, assumptions (A7) and (A8) imply that

$$\begin{aligned} |B(\varphi_1+\varphi_2)\varphi_3 - B(\psi_1+\psi_2)\psi_3| &\le |B(\varphi_1+\varphi_2)\varphi_3 - B(\varphi_1+\varphi_2)\psi_3| \\ &+ |B(\varphi_1+\varphi_2)\psi_3 - B(\psi_1+\psi_2)\psi_3| \\ &\le \beta_1|\varphi_3 - \psi_3|_Q\{|A_0^\alpha\varphi_1| + |A_0^\alpha\varphi_2|\} \\ &+ \gamma_1|\psi_3|_Q\{|A_0^\alpha\varphi_1 - A_0^\alpha\psi_1| + |A_0^\alpha\varphi_2 - A_0^\alpha\psi_2|\} \\ &= \beta_1|\varphi_3 - \psi_3|_Q\{|\varphi_1|_\alpha + |\varphi_2|_\alpha\} \\ &+ \gamma_1|\psi_3|_Q\{|\varphi_1 - \psi_1|_\alpha + |\varphi_2 - \psi_2|_\alpha\}. \end{aligned} \tag{2.47}$$

Finally, using assumptions (A7) and (A8), we obtain

$$\begin{aligned} &|\lambda\{\varphi_3 - \psi_3\} - \{B(\varphi_1+\varphi_2)'\varphi_1 - B(\psi_1+\psi_2)'\psi_1\}|_Q \\ &\le \lambda|\varphi_3 - \psi_3|_Q + \sup_{|q|_Q \le 1} |\langle\{B(\varphi_1+\varphi_2)'\varphi_1 - B(\psi_1+\psi_2)'\psi_1\}, q\rangle_Q| \\ &= \lambda|\varphi_3 - \psi_3|_Q + \sup_{|q|_Q \le 1} |\langle A_1(q)\{\varphi_1+\varphi_2\}, \varphi_1\rangle - \langle A_1(q)\{\psi_1+\psi_2\}, \psi_1\rangle| \\ &\le \lambda|\varphi_3 - \psi_3|_Q + \sup_{|q|_Q \le 1} |\langle A_1(q)\{\varphi_1+\varphi_2\} - A_1(q)\{\psi_1+\psi_2\}, \varphi_1\rangle| \\ &\qquad + \sup_{|q|_Q \le 1} |\langle A_1(q)\{\psi_1+\psi_2\}, \varphi_1 - \psi_1\rangle| \end{aligned} \tag{2.48}$$

$$
\begin{aligned}
&\leq \lambda|\varphi_3-\psi_3|_Q + \sup_{|q|_Q\leq 1}|A_1(q)\{\varphi_1+\varphi_2\}-A_1(q)\{\psi_1+\psi_2\}||\varphi_1| \\
&\qquad + \sup_{|q|_Q\leq 1}|A_1(q)\{\psi_1+\psi_2\}||\varphi_1-\psi_1| \\
&\leq \lambda|\varphi_3-\psi_3|_Q + \gamma_1|\varphi_1|\{|A_0^\alpha\varphi_1 - A_0^\alpha\psi_1| + |A_0^\alpha\varphi_2 - A_0^\alpha\psi_2|\} \\
&\qquad +\beta_1\{|A_0^\alpha\psi_1| + |A_0^\alpha\psi_2|\}|\varphi_1-\psi_1| \\
&\leq \lambda|\varphi_3-\psi_3|_Q + \gamma_1\kappa_\alpha|\varphi_1|_\alpha\{|\varphi_1-\psi_1|_\alpha + |\varphi_2-\psi_2|_\alpha\} \\
&\qquad +\beta_1\kappa_\alpha\{|\psi_1|_\alpha + |\psi_2|_\alpha\}|\varphi_1-\psi_1|_\alpha,
\end{aligned}
$$

where κ_α is such that $|\varphi| \leq \|\varphi\|_\alpha \leq \kappa_\alpha|\varphi|_\alpha$ for $\varphi \in H_\alpha$. Combining (2.46)–(2.48) and assumption (A9), we obtain (2.45), and the theorem is proved. □

In order to extend the local solution guaranteed to exist in Theorem 2.4 we require the estimate given in the following lemma.

LEMMA 2.5. *Let $x = (e, v, r)$ be the unique solution to the initial value problem* (2.40), (2.41) *guaranteed to exist on the interval $[0, T)$ by Theorem* 2.4. *It then follows that*

$$
|x(t)|_X^2 + \rho_0\int_0^t \left\{\|e(s)\|^2 + \|v(s)\|^2\right\} ds \leq |x_0|_X^2 + \frac{1}{\rho_0}\int_0^t \|g(s)\|_*^2 ds,\ 0 \leq t < T. \tag{2.49}
$$

Proof. For $s \in [0, T)$, using (2.40), we obtain

$$
\begin{aligned}
\frac{1}{2}D_t|x(s)|_X^2 &= \langle D_t x(s), x(s)\rangle_X \\
&= -\langle A_\lambda x(s), x(s)\rangle_X + \langle G_\lambda(s, x(s)), x(s)\rangle_X \\
&= -\langle A_0 e(s), e(s)\rangle - \langle A_0 v(s), v(s)\rangle + \langle g(s), v(s)\rangle \qquad (2.50) \\
&\leq -\rho_0\|e(s)\|^2 - \rho_0\|v(s)\|^2 + \|g(s)\|_*\|v(s)\| \\
&\leq -\frac{\rho_0}{2}\left\{\|e(s)\|^2 + \|v(s)\|^2\right\} + \frac{1}{2\rho_0}\|g(s)\|_*^2.
\end{aligned}
$$

Integrating both sides of (2.50) from 0 to t, and using (2.41), we obtain (2.49), and the lemma is proved. □

Note that the proof of Lemma 2.5 given above does not *explicitly* require that the additional assumptions (A6)–(A9) be satisfied.

THEOREM 2.6. *Suppose that assumptions* (A1)–(A9) *hold and that $e_0, v_0 \in Dom(A_0^\alpha)$, where $\alpha \in (0, 1)$ is as in assumption* (A6). *Then the initial value problem* (2.40), (2.41) *has a unique solution, $x = (e, v, r)$, which exists for all $t \geq 0$.*

Proof. The local solution x to the initial value problem (2.40), (2.41) guaranteed to exist by Theorem 2.4 can be continued so long as $|x(t)|_{X_\alpha}$ remains bounded. We show that this is in fact the case by using Lemma 2.5 to argue that $|x(t)|_{X_\alpha}$ remains bounded as $t \uparrow T$.

For $t \in [0, T)$ we have that

$$
x(t) = T_\lambda(t)x_0 + \int_0^t T_\lambda(t-s)G_\lambda(s, x(s))ds
$$

and therefore that

$$
A_\lambda^\alpha x(t) = A_\lambda^\alpha T_\lambda(t)x_0 + \int_0^t A_\lambda^\alpha T_\lambda(t-s)G_\lambda(s, x(s))ds.
$$

Equivalently, we have

$$A_0^\alpha e(t) = A_0^\alpha T_0(t)e_0 + \int_0^t A_0^\alpha T_0(t-s)B(e(s)+v(s))r(s)ds,$$
$$A_0^\alpha v(t) = A_0^\alpha T_0(t)v_0 + \int_0^t A_0^\alpha T_0(t-s)g(s)ds,$$

and

$$\lambda^\alpha r(t) = \lambda^\alpha e^{-\lambda t}r_0 + \int_0^t \lambda^\alpha e^{-\lambda(t-s)}\{\lambda r(s) - B(e(s)+v(s))'e(s)\}ds.$$

It follows from assumptions (A7) and (A9), (2.10), and Theorem 2.6.13 in [33] that

$$\begin{aligned}
|e(t)|_\alpha &\le e^{-\rho_0 K^{-2}t}|e_0|_\alpha + \textstyle\int_0^t M_\alpha(t-s)^{-\alpha}e^{-\rho_0 K^{-2}(t-s)}\beta_1|r(s)|_Q|e(s)|_\alpha ds \\
&\quad + \int_0^t M_\alpha(t-s)^{-\alpha}e^{-\rho_0 K^{-2}(t-s)}\beta_1|r(s)|_Q|v(s)|_\alpha ds \qquad (2.51)\\
&\le |e_0|_\alpha + M_\alpha\beta_1\int_0^t (t-s)^{-\alpha}|r(s)|_Q|x(s)|_{X_\alpha}ds, \\
|v(t)|_\alpha &\le e^{-\rho_0 K^{-2}t}|v_0|_\alpha + \int_0^t M_\alpha(t-s)^{-\alpha}e^{-\rho_0 K^{-2}(t-s)}|g(s)|ds \\
&\le e^{-\rho_0 K^{-2}t}|v_0|_\alpha + \int_0^t M_\alpha(t-s)^{-\alpha}e^{-\rho_0 K^{-2}(t-s)}|g(0)|ds \\
&\quad + \int_0^t M_\alpha(t-s)^{-\alpha}e^{-\rho_0 K^{-2}(t-s)}\left\{|g(s)-g(0)|\right\}ds \qquad (2.52)\\
&\le |v_0|_\alpha + M_\alpha\int_0^t (t-s)^{-\alpha}\{|g(0)| + \delta s^\nu\}ds \\
&\le |v_0|_\alpha + M_\alpha\left\{|g(0)| + \delta T^\nu\right\}\frac{T^{1-\alpha}}{1-\alpha},
\end{aligned}$$

and

$$\begin{aligned}
|r(t)|_Q &\le |r_0|_Q + \int_0^t e^{-\lambda(t-s)}\{\lambda|r(s)|_Q + |B(e(s)+v(s))'e(s)|_Q\}ds \\
&\le |r_0|_Q + \int_0^t \{\lambda|r(s)|_Q + \sup_{|q|_Q\le 1}|\langle A_1(q)\{e(s)+v(s)\}, e(s)\rangle|\}ds \\
&\le |r_0|_Q + \int_0^t \{\lambda|r(s)|_Q + \beta_1|e(s)|\{|A_0^\alpha e(s)| + |A_0^\alpha v(s)|\}\}ds \qquad (2.53)\\
&\le |r_0|_Q + \int_0^t \{\lambda|r(s)|_Q + \beta_1|e(s)|\{|e(s)|_\alpha + |v(s)|_\alpha\}\}ds \\
&\le |r_0|_Q + T^\alpha\int_0^t \max\{\lambda, \beta_1|e(s)|\}(t-s)^{-\alpha}|x(s)|_{X_\alpha}ds,
\end{aligned}$$

where M_α is a positive constant. Now Lemma 2.5 implies that for $s \in [0,T)$, $|x(s)|_X$ is bounded. It follows that $s \in [0,T)$, $|e(s)|$ and $|r(s)|_Q$ are bounded. Combining (2.51), (2.52), and (2.53), we obtain

$$|x(t)|_{X_\alpha} \le |x_0|_{X_\alpha} + M_\alpha\left\{|g(0)| + \delta T^\nu\right\}\frac{T^{1-\alpha}}{1-\alpha} + C\int_0^t (t-s)^{-\alpha}|x(s)|_{X_\alpha}ds, \qquad (2.54)$$

where $C > 0$. It follows from Theorem 5.6.7 in [33] that $|x(t)|_{X_\alpha} \leq C_1$ on $[0,T)$ for some $C_1 > 0$, and the theorem is proved. ∎

Theorem 2.6 yields the following regularity result for the controller f. We state it as a corollary.

COROLLARY 2.7. *Suppose that assumptions* (A1)–(A9) *hold, and that* $e_0, v_0 \in Dom(A_0^\alpha)$, *where* $\alpha \in (0,1)$ *is as in assumption* (A6). *If the operator* A_2 *is such that* $D_2 = Dom(A_2) = \{\varphi \in V : A_2\varphi \in H\} \supset Dom(A_0^\alpha)$, *where* $\alpha \in (0,1)$ *is as in assumption* (A6), *and satisfies a Lipschitz condition of the form*

$$|A_2\varphi - A_2\psi| \leq \gamma_2|\varphi - \psi|_\alpha, \qquad \varphi, \psi \in Dom(A_0^\alpha), \tag{2.55}$$

then the control law given in (2.22) *or* (2.23) *satisfies* $f(t) \in H$, $t > 0$, *and* $f \in C((0,\infty); H)$.

Proof. For $t > 0$, the controller f satisfies

$$f(t) = A(q(t))u(t) - A_0u(t) + g(t) = D_tu(t) + A(\overline{q})u(t). \tag{2.56}$$

Theorem 2.6 implies that $u(t) \in Dom(A_0)$, $t > 0$. It follows, therefore, that $u(t) \in Dom(A_0^\alpha)$, $t > 0$, and therefore that $u(t) \in D_j$, $j = 1, 2$, $t > 0$. Consequently, $A(\overline{q})u(t) \in H$, $t > 0$, and, hence, $f(t) \in H$, $t > 0$. Theorem 2.6 also implies that $D_tu \in C((0,\infty); H)$, and for $s, t > 0$, assumption (A8) together with (2.55) imply that

$$\begin{aligned} |A(\overline{q})u(t) - A(\overline{q})u(s)| &\leq |A_1(\overline{q})u(t) - A_1(\overline{q})u(s)| + |A_2u(t) - A_2u(s)| \\ &\leq \{\gamma_1|\overline{q}|_Q + \gamma_2\}|u(t) - u(s)|_\alpha. \end{aligned} \tag{2.57}$$

Inspection of the proof of Theorem 6.3.1 in [33] immediately reveals that u is continuous in H_α. It follows from (2.56) and (2.57) that $f \in C((0,\infty); H)$, which establishes the corollary. ∎

Example 2.8. We provide a simple example which satisfies assumptions (A1)–(A9). Let $H = L_2(0,1)$, and let it be endowed with the standard inner product $\langle \cdot, \cdot \rangle$ and corresponding induced norm $|\cdot|$. Let $V = H_L^1(0,1) = \{\varphi \in H^1(0,1) : \varphi(0) = 0\}$, and let it be endowed with the norm $\|\cdot\|$ given by $\|\varphi\| = \{\int_0^1 |D\varphi(x)|^2 dx\}^{\frac{1}{2}}$, $\varphi \in H_L^1(0,1)$. Then $\{V, \|\cdot\|\}$ is a reflexive Banach space and, in fact, a Hilbert space, which is densely and continuously embedded in H. We have $|\varphi| \leq \|\varphi\|$, $\varphi \in H_L^1(0,1)$. Let $\hat{V}^* = V^*$, and let $Q = \mathbf{R}^1$ with $|q|_Q = |q|$ for $q \in \mathbf{R}$. We are interested in controlling the first-order plant given by

$$\frac{\partial u}{\partial t}(t,x) + \overline{q}\frac{\partial u}{\partial x}(t,x) = f(t,x), \qquad 0 < x < 1,\ t > 0, \tag{2.58}$$

together with the boundary condition

$$u(t,0) = 0, \qquad t > 0, \tag{2.59}$$

and initial condition

$$u(0,x) = u_0(x), \qquad 0 \leq x \leq 1, \tag{2.60}$$

where $\overline{q} > 0$, $u_0 \in L_2(0,1)$, and $t \to f(t,\cdot) \in L_2(0,T;H)$ for each $T > 0$.

For each $q \in \mathbf{R}^1$, let the operator $A_1(q) : Dom(A_1(q)) \subset H \to H$ be given by $A_1(q)\varphi = qD\varphi$, $\varphi \in D_1$, where $D_1 = Dom(A_1(q)) = V$. For each $q \in Q$, let $A(q) = A_1(q)$. It follows that A_2 is the zero operator, that D_1 is independent of $q \in Q$, and that $A_1(q) : V \to V^*$. Moreover, for $q \in Q$ and $\varphi, \psi \in V$ we have that

$$|\langle A_1(q)\varphi, \psi\rangle| = |q|_Q \left| \int_0^1 D\varphi(x)\psi(x)dx \right| \leq |q|_Q|D\varphi||\psi| = |q|_Q\|\varphi\|\|\psi\| \leq |q|_Q\|\varphi\|\|\psi\|.$$

Consequently assumption (A1) is satisfied with $\alpha_1 = 1$. Assumption (A2) is trivially satisfied.

It is not difficult to show that the Hilbert space adjoint of the operator $A_1(\overline{q})$ is given by $A_1(\overline{q})^*\varphi = -\overline{q}D\varphi$, $\varphi \in D_1^*$, where $D_1^* = Dom(A_1(\overline{q})^*) = H_R^1(0,1) = \{\varphi \in H^1(0,1) : \varphi(1) = 0\}$. For $\varphi \in H_L^1(0,1)$, we have that

$$\langle A_1(\overline{q})\varphi, \varphi\rangle = \overline{q}\int_0^1 D\varphi(x)\varphi(x)dx = \frac{\overline{q}}{2}\int_0^1 D\varphi(x)^2dx = \frac{\overline{q}}{2}\varphi(1)^2 \geq 0 \tag{2.61}$$

and for $\varphi \in H_R^1(0,1)$ that

$$\langle A_1(\overline{q})^*\varphi, \varphi\rangle = -\overline{q}\int_0^1 D\varphi(x)\varphi(x)dx = -\frac{\overline{q}}{2}\int_0^1 D\varphi(x)^2dx = \frac{\overline{q}}{2}\varphi(0)^2 \geq 0. \tag{2.62}$$

It follows from (2.61) and (2.62) that (see, for example, [26, Theorem I.4.5]) the operator $-A_1(\overline{q})$ is *maximal dissipative* and therefore that it is the infinitesimal generator of a $\mathcal{C}_0$-semigroup of bounded linear operators (in fact, contractions), $\{S(t,\overline{q}) : t \geq 0\}$, on $H = L_2(0,1)$. For each $t \geq 0$, the unique mild solution, $u(t) = u(t,\cdot)$ to the system (2.58)–(2.60) is given by

$$u(t) = S(t;\overline{q})u_0 + \int_0^t S(t-s;\overline{q})f(s)ds, \tag{2.63}$$

where for each $t \geq 0$, $f(t) = f(t,\cdot) \in L_2(0,1)$. When $u_0 \in H_L^1(0,1)$, and f is strongly continuously differentiable for $t \geq 0$, the function u given by (2.63) is a strong solution (see, for example, [39, Theorem 3.2.2]). Such a solution certainly satisfies our minimal well-posedness requirement on the plant. Indeed, in addition to satisfying the differential equation and initial data, (2.4) and (2.5), we have (see, for example, [39, p. 64]) $u \in C^1([0,T];H)$, $u(t) \in Dom(A_1(\overline{q})) = V$, $t \in [0,T]$, and $A_1(\overline{q})u \in C([0,T];H)$ for all $T > 0$. It immediately follows that $u \in C(0,T;H) \cap L_2(0,T;V)$ and $D_t u \in L_2(0,T;V^*)$.

For the reference model, we consider the one-dimensional heat equation given by

$$\frac{\partial v}{\partial t}(t,x) - a_0\frac{\partial^2 v}{\partial x^2}(t,x) = g(t,x), \qquad 0 < x < 1,\ t > 0,$$

together with the boundary conditions $v(t,0) = 0$ and $\frac{\partial v}{\partial x}(t,1) = 0$, $t > 0$, and the initial conditions $v(0,x) = v_0(x)$, $0 \leq x \leq 1$, where $a_0 > 0$, $v_0 \in L_2(0,1)$, and $[t \to g(t,\cdot)] \in L_2(0,T;V^*)$ for each $T > 0$. In this case we have $A_0 \in \mathcal{L}(V,V^*)$ given by

$$\langle A_0\varphi, \psi\rangle = a_0\int_0^1 D\varphi(x)D\psi(x)dx, \qquad \varphi, \psi \in H_L^1(0,1).$$

It is immediately clear that assumptions (A3) and (A4) are satisfied with $\alpha_0 = \rho_0 = a_0$. Moreover, we have that $D_0 = Dom(A_0) = \{\varphi \in H_L^1(0,1) : \varphi \in H^2(0,1), D\varphi(1) = 0\}$ and that A_0 as an operator from V into V^* is symmetric or as an operator on H is self-adjoint. Assumption $(A5)$ is trivially satisfied with the choice of $\hat{V}^* = V^*$.

Since A_0 is symmetric, $V = Dom(A_0^{\frac{1}{2}})$ (see [39]). Consequently, we have

$$Dom(A_0^{\frac{1}{2}}) = V = Dom(A_1(q)) = D_1.$$

It follows that assumption (A6) is satisfied with $\alpha = \frac{1}{2}$. Moreover, for $\varphi \in H_L^1(0,1) = V$, we have that

$$\begin{aligned}|A_1(q)\varphi|^2 &= |qD\varphi|^2 = |q|_Q^2|D\varphi|^2 = |q|_Q^2\|\varphi\|^2 \\ &= |q|_Q^2\frac{1}{a_0}\langle A_0\varphi,\varphi\rangle = \frac{1}{a_0}|q|_Q^2\langle A_0^{\frac{1}{2}}\varphi, A_0^{\frac{1}{2}}\varphi\rangle = \frac{1}{a_0}|q|_Q^2|A_0^{\frac{1}{2}}\varphi|^2.\end{aligned}$$

It follows that assumptions (A7) and (A8) are satisfied with β_1 and γ_1 in (2.42) and (2.43), respectively, given by $\beta_1 = \gamma_1 = \frac{1}{\sqrt{a_0}}$. Thus, if $u_0, v_0 \in H_L^1(0,1)$, and g is sufficiently regular (i.e., assumption (A9) and (2.44) being satisfied for some $\delta > 0$ and $\nu \in (0,1]$), then the resulting closed-loop system will be well-posed.

2.2. Closed-loop well-posedness via an implicit function theorem. Assumptions (A6)–(A9) can be rather restrictive and may preclude the consideration of certain classes of problems of interest. In particular, assumption (A7) does not include the class of problems in which the plant and reference model dynamics are of the same *order* (i.e., $\alpha \in (0,1)$). Thus, for example, the above theory does not allow for both a plant and reference model described by a diffusion (or heat) equation. To remedy this, we propose a somewhat different approach to demonstrating the well-posedness of the closed-loop system (2.26)–(2.28). Our argument is based upon an application of the implicit function theorem (see, for example, [10]). Of course this approach requires additional assumptions as well. Indeed, in this case we can guarantee well-posedness only for initial data which is sufficiently small in norm. That is, the plant must initially be *close* to the reference model, and we require a reasonably good initial guess for the unknown parameters. Also, to simplify the presentation we make the following assumption on the linearity of the plant.

(A10) (linearity of the plant). For each $q \in Q$, $A_1(q) : V \to V^*$ is linear.

Note that assumptions (A1) and (A10) together imply that $A_1(q) \in \mathcal{L}(V, V^*)$ for each $q \in Q$. We note that assumption (A10) can be weakened quite a bit to allow for certain classes of nonlinear plants such as certain Lipschitz continuous or differentiable operators. However, the required technical assumptions would only complicate the exposition without significantly affecting its substance. Consequently, we opt for clarity and leave the generalization to the reader.

We also require the following regularity assumption on the state v of the reference model (2.8), (2.9).

(A11) (regularity of the reference model). The solution v to the system (2.8), (2.9) satisfies $v \in L_2(0,T;V)$ for all $T > 0$.

Theorem 2.2 (ii) provides sufficient conditions for assumption (A11) to be satisfied.

We consider v to be an *exogenous* signal and consider the initial value problem given by

$$\langle D_t e(t),\varphi\rangle + \langle A_0 e(t),\varphi\rangle - \langle A_1(r(t))\{e(t)+v(t)\},\varphi\rangle = 0,\ \varphi \in V,\ \text{a.e. } t > 0, \tag{2.64}$$

$$\langle D_t r(t), p\rangle_Q + \langle A_1(p)\{e(t)+v(t)\}, e(t)\rangle = 0, \qquad p \in Q,\ \text{a.e. } t > 0, \tag{2.65}$$

$$e(0) = e_0, \qquad r(0) = r_0, \tag{2.66}$$

THEOREM 2.9. *Suppose that assumptions* (A1)–(A5) *and assumptions* (A10) *and* (A11) *hold. Suppose further that* $e_0 \in V$. *Then there exists a constant* $C > 0$ *such that if*

$$\|e_0\| + |r_0|_Q < C, \tag{2.67}$$

then the initial value problem (2.64)–(2.66) *has a unique solution* (e, r) *with* $e \in L_2(0,T;V) \cap H^1(0,T;V^*)$ *and* $r \in L_\infty(0,T;Q) \cap W^{1,1}(0,T;Q)$ *for all* $T > 0$.

Proof. The proof follows from an application of the implicit function theorem (see, for example, [10]). We let $T > 0$, and we begin with the definition of the following Banach spaces. Let $X = X_1 \times X_2$ where $X_1 = V$ and $X_2 = Q$ with norm

$$\|(\varphi, q)\|_X = \|\varphi\|_{X_1} + \|q\|_{X_2} = \|\varphi\| + |q|_Q, \qquad \varphi \in V,\ q \in Q. \tag{2.68}$$

Let $Y = Y_1 \times Y_2$ where $Y_1 = L_2(0,T;V) \cap H^1_L(0,T;V^*)$ and $Y_2 = L_\infty(0,T;Q) \cap W^{1,1}_L(0,T;Q)$ with norm

$$\begin{aligned} \|(\varphi, q)\|_Y &= \|\varphi\|_{Y_1} + \|q\|_{Y_2} \\ &= \left\{\int_0^T \|\varphi(t)\|^2 dt\right\}^{\frac{1}{2}} + \left\{\int_0^T \|D_t\varphi(t)\|_*^2 dt\right\}^{\frac{1}{2}} \\ &\quad + \operatorname{ess\,sup}_{t\in[0,T]} |q(t)|_Q + \int_0^T |D_t q(t)|_Q dt, \end{aligned} \tag{2.69}$$

for $\varphi \in L_2(0,T;V) \cap H^1_L(0,T;V^*)$ and $q \in L_\infty(0,T;Q) \cap W^{1,1}_L(0,T;Q)$, and let $Z = Z_1 \times Z_2$, where $Z_1 = L_2(0,T;V^*)$ and $Z_2 = L_1(0,T;Q)$, with norm

$$\|(\varphi, q)\|_Z = \|\varphi\|_{Z_1} + \|q\|_{Z_2} = \left\{\int_0^T \|\varphi(t)\|_*^2 dt\right\}^{\frac{1}{2}} + \int_0^T |q(t)|_Q dt, \tag{2.70}$$

for $\varphi \in L_2(0,T;V^*)$ and $q \in L_1(0,T;Q)$. The subscript L in the above spaces denotes homogeneous boundary conditions at the left endpoint of the interval.

Define the function $\mathcal{F} : X \times Y \to Z$ by $\mathcal{F}(x,y) = (\mathcal{F}_1(x,y), \mathcal{F}_2(x,y))$, $x = (x_1, x_2) \in X$, $y = (y_1, y_2) \in Y$, where $\mathcal{F}_1 : X \times Y \to Z_1 = L_2(0,T;V^*)$ is given by $\mathcal{F}_1(x,y) = D_t y_1 + A_0\{y_1 + x_1\} - B(y_1 + x_1 + v)\{y_2 + x_2\}$, $x = (x_1, x_2) \in X$, $y = (y_1, y_2) \in Y$, and $\mathcal{F}_2 : X \times Y \to Z_2 = L_1(0,T;Q)$ is given by $\mathcal{F}_2(x,y) = D_t y_2 + B(y_1 + x_1 + v)'\{y_1 + x_1\}$, $x = (x_1, x_2) \in X$, $y = (y_1, y_2) \in Y$, where for $\varphi \in V$, the operator $B(\varphi) \in \mathcal{L}(Q, V^*)$ and its Banach space adjoint, $B(\varphi)' \in \mathcal{L}(V, Q)$ are given in (2.36) and (2.37), respectively.

The hypotheses of the theorem clearly imply that $\mathcal{F}(0,0) = 0$ and that $\mathcal{F} \in C(X \times Y, Z)$; that is, $\mathcal{F}$ is a continuous mapping from $X \times Y$ into Z. It is not difficult to show that $\mathcal{F}$ is continuously differentiable from $X \times Y$ into Z. Indeed, we need only to argue that (see [10, Theorem 8.9.1]) the maps $(x,y) \mapsto D_{x_i}\mathcal{F}_j$ from $X \times Y$ into $\mathcal{L}(X_i, Z_j)$, $i,j = 1,2$ and $(x,y) \mapsto D_{y_i}\mathcal{F}_j$ from $X \times Y$ into $\mathcal{L}(Y_i, Z_j)$, $i,j = 1,2$ are continuous. A straightforward calculation reveals that at $(x,y) \in X \times Y$ with $x = (x_1, x_2) \in X = X_1 \times X_2$ and $y = (y_1, y_2) \in Y = Y_1 \times Y_2$,

$$\begin{aligned}
&D_{x_1}\mathcal{F}_1\delta x_1 = A_0\delta x_1 - B(\delta x_1)\{y_2 + x_2\} \in Z_1, \qquad \delta x_1 \in X_1, \\
&D_{x_2}\mathcal{F}_1\delta x_2 = B(y_1 + x_1 + v)\delta x_2 \in Z_1, \qquad \delta x_2 \in X_2, \\
&D_{x_1}\mathcal{F}_2\delta x_1 = B(\delta x_1)'\{y_1 + x_1\} + B(y_1 + x_1 + v)'\delta x_1 \in Z_2, \qquad \delta x_1 \in X_1, \\
&D_{x_2}\mathcal{F}_2\delta x_2 = 0 \in Z_2, \qquad \delta x_2 \in X_2, \\
&D_{y_1}\mathcal{F}_1\delta y_1 = D_t\delta y_1 + A_0\delta y_1 - B(\delta y_1)\{y_2 + x_2\} \in Z_1, \qquad \delta y_1 \in Y_1, \\
&D_{y_2}\mathcal{F}_1\delta y_2 = B(y_1 + x_1 + v)\delta y_2 \in Z_1, \qquad \delta y_2 \in Y_2, \\
&D_{y_1}\mathcal{F}_2\delta y_1 = B(\delta y_1)'\{y_1 + x_1\} + B(y_1 + x_1 + v)'\delta y_1 \in Z_2, \qquad \delta y_1 \in Y_1, \\
&D_{y_2}\mathcal{F}_2\delta y_2 = D_t\delta y_2 \in Z_2, \qquad \delta y_2 \in Y_2,
\end{aligned}$$

and that the requisite continuity holds.

We show next that $D_y\mathcal{F}(0,0) = (D_y\mathcal{F}_1(0,0), D_y\mathcal{F}_2(0,0))$ is a linear homeomorphism of Y *onto* Z. We do this by demonstrating that for each $z = (z_1, z_2) \in Z$, $z_1 \in L_2(0,T;V^*)$, and $z_2 \in L_1(0,T;Q)$ there exists a unique $y = (y_1, y_2) \in Y$, $y_1 \in L_2(0,T;V) \cap H_L^1(0,T;V^*)$, and $y_2 \in L_\infty(0,T;Q) \cap W_L^{1,1}(0,T;Q)$ satisfying the linear initial value problem

$$D_t y_1 + A_0 y_1 - B(v) y_2 = z_1, \qquad t > 0, \tag{2.71}$$

$$D_t y_2 + B(v)' y_1 = z_2, \qquad t > 0, \tag{2.72}$$

$$y_1(0) = 0 \qquad \text{and} \qquad y_2(0) = 0, \tag{2.73}$$

and by providing estimates which establish the continuous dependence of y on z. If we assume that V is separable, then the argument establishing the existence of a unique solution to the system (2.71)–(2.73) is the same as the one used to prove Theorem III.1.2 in [27]. Galerkin approximation is used to define a sequence of finite-dimensional initial value problems which *approximate* the system (2.71)–(2.73). Of course each of the finite-dimensional systems admits a unique solution $y^n = (y_1^n, y_2^n)$. One then argues that these approximating solutions lie in a bounded subset of Y, that $y^n \to y$, weakly in Y, and that y is the unique solution to the initial value problem (2.71)–(2.73) (see also Remark 2.1). The key step in the proof depends upon the estimate for $\|y\|_Y$ in terms of $\|z\|_Z$ which we now derive. This estimate, which is given as an a posteriori estimate in (2.78) below, establishes the continuous dependence of y on z as well.

Taking the inner product of (2.71) with y_1 and (2.72) with y_2 and then adding, we obtain

$$\frac{1}{2}\left\{D_t|y_1|^2 + |y_2|_Q^2\right\} + \langle A_0 y_1, y_1\rangle = \langle z_1, y_1\rangle + \langle z_2, y_2\rangle_Q.$$

For any $\varepsilon > 0$, assumption (A4) implies that

$$\begin{aligned}\frac{1}{2}\left\{D_t|y_1|^2 + |y_2|_Q^2\right\} + \rho_0\|y_1\|^2 &\leq \|z_1\|_*\|y_1\| + |z_2|_Q|y_2|_Q \\ &\leq \frac{1}{2\varepsilon}\|z_1\|_*^2 + \frac{\varepsilon}{2}\|y_1\|^2 + \frac{1}{2}|z_2|_Q\left\{1 + |y_2|_Q^2\right\}.\end{aligned} \tag{2.74}$$

Choosing $\varepsilon < 2\rho_0$, setting $c_0 = 2\rho_0 - \varepsilon > 0$ and $c_1 = 1/\varepsilon$, integrating (2.74) from 0 to t, and recalling (2.73), we obtain

$$\begin{aligned}&|y_1(t)|^2 + |y_2(t)|_Q^2 + c_0\int_0^t \|y_1(s)\|^2 ds \\ &\qquad \leq c_1\int_0^t \|z_1(s)\|_*^2 ds + \int_0^t |z_2(s)|_Q ds + \int_0^t |z_2(s)|_Q|y_2(s)|_Q^2 ds \\ &\qquad \leq c_1\|z\|_Z^2 + \|z\|_Z + \int_0^t |z_2(s)|_Q|y_2(s)|_Q^2 ds.\end{aligned}$$

An application of the generalized Gronwall inequality (see [15]) yields

$$\begin{aligned}&|y_1(t)|^2 + |y_2(t)|_Q^2 + c_0\int_0^t \|y_1(s)\|^2 ds \\ &\qquad \leq \left\{c_1\|z\|_Z^2 + \|z\|_Z\right\}\left\{1 + \int_0^t |z_2(s)|_Q e^{\int_s^t |z_2(\tau)|_Q d\tau} ds\right\} \\ &\qquad \leq \left\{c_1\|z\|_Z^2 + \|z\|_Z\right\}\left\{1 + \|z\|_Z e^{\|z\|_Z}\right\}, \quad t \geq 0.\end{aligned} \tag{2.75}$$

Equation (2.71) and assumptions (A1), (A3), and (A11) yield the estimate $\|D_t y_1(t)\|_* \leq \|z_1(t)\|_* + \alpha_0\|y_1(t)\| + \alpha_1\|v(t)\||y_2(t)|_Q$, a.e. $t > 0$. Consequently there exists a constant $c_2 = c_2(\|v\|_{L_2(0,T;V)}) > 0$ such that

$$\|y_1\|^2_{H^1_L(0,T;V^*)} \leq c_2 \left\{ \|z_1\|^2_{L_2(0,T;V^*)} + \|y_1\|^2_{L_2(0,T;V)} + |y_2|^2_{L_\infty(0,T;Q)} \right\}. \tag{2.76}$$

Similarly, (2.72) yields $|D_t y_2(t)|_Q \leq |z_2(t)|_Q + \alpha_1\|v(t)\|\|y_1(t)\|$, a.e. $t > 0$, and therefore that

$$\|D_t y_2\|_{L_1(0,T;Q)} \leq \|z_2\|_{L_1(0,T;Q)} + \alpha_1\|v\|_{L_2(0,T;V)}\|y_1\|_{L_2(0,T;V)}. \tag{2.77}$$

Combining (2.75), (2.76), and (2.77), we obtain that

$$\|y\|_Y \leq h(\|z\|_Z), \tag{2.78}$$

where $h : \mathbf{R}^+ \to \mathbf{R}^+$ is continuous and monotone increasing.

The following estimates for the dependence on z on y can also be obtained. Once again, (2.71) and assumptions (A1), (A3), and (A11) imply that $\|z_1(t)\|_* \leq \|D_t y_1(t)\|_* + \alpha_0\|y_1(t)\| + \alpha_1\|v(t)\||y_2(t)|_Q$, a.e. $t > 0$, and, therefore, that

$$\|z_1\|^2_{L_2(0,T;V^*)} \leq c_2 \left\{ \|y_1\|^2_{H^1_L(0,T;V^*)} + \|y_1\|^2_{L_2(0,T;V)} + |y_2|^2_{L_\infty(0,T;Q)} \right\}. \tag{2.79}$$

Also (2.72) yields $|z_2(t)|_Q \leq |D_t y_2(t)|_Q + \alpha_1\|v(t)\|\|y_1(t)\|$, a.e. $t > 0$, and, therefore, that

$$\|z_2\|_{L_1(0,T;Q)} \leq \|D_t y_2\|_{L_1(0,T;Q)} + \alpha_1\|v\|_{L_2(0,T;V)}\|y_1\|_{L_2(0,T;V)}. \tag{2.80}$$

Combining (2.79) and (2.80), we obtain that $\|z\|_Z \leq h_0\|y\|_Y$, where $h_0 > 0$.

It follows from the implicit function theorem that there exists a $C > 0$ such that if $x_0 = (e_0, r_0) \in X$ satisfies (2.67), then there exists a unique $y = y(x_0) = (y_1(x_0), y_2(x_0)) \in Y$, which is continuously differentiable in x_0 and which satisfies $\mathcal{F}(x_0, y) = (\mathcal{F}_1(x_0, y), \mathcal{F}_2(x_0, y)) = 0$. Setting $e = y_1 + e_0$ and $r = y_2 + r_0$, we obtain the desired result. □

Under additional hypotheses a similar approach can be used to obtain a somewhat stronger result providing L_∞ estimates.

THEOREM 2.10. *Suppose that assumptions* (A1)–(A5) *and assumption* (A10) *are satisfied. Suppose further that*

(i) $D_1 = Dom(A_1(q))$ *is independent of* $q \in Q$, $D_1 \subseteq D_0$, *and there exists* $\gamma_0 > 0$ *for which* $|A_1(q)\varphi| \leq \gamma_0|q|_Q|A_0\varphi|$, $q \in Q$, *and* $\varphi \in D_1$;
(ii) A_0 *is symmetric* ($A_0 : D_0 \subset H \to H$ *is self-adjoint);*
(iii) $v \in L_\infty(0,T;V)$, $v(t) \in D_0$, *a.e.* $t > 0$, $A_0 v \in L_2(0,T;H)$; *and*
(iv) $e_0 \in D_0$.

Then there exists a constant $C > 0$ *such that if* $|A_0 e_0| + |r_0|_Q < C$, *then the initial value problem* (2.64)–(2.66) *has a unique solution* (e, r) *with* $e \in L_\infty(0,T;V) \cap H^1(0,T;H)$ *and* $r \in L_\infty(0,T;Q) \cap W^{1,1}(0,T;Q)$. *Moreover,* $e(t) \in D_0$, *a.e.* $t > 0$, *and* $A_0 e \in L_2(0,T;H)$.

Proof. Let D_0 be endowed with the graph norm. Then the proof is completely analogous to the one given above for Theorem 2.9 based upon the implicit function theorem. However, in this case we take the Banach spaces X, Y, and Z to be $X = D_0 \times Q$, $Y = \{L_2(0,T;D_0) \cap H^1_L(0,T;H)\} \times \{L_\infty(0,T;Q) \cap W^{1,1}_L(0,T;Q)\}$, and $Z = L_2(0,T;H) \times L_1(0,T;Q)$, respectively. The norms on these spaces are chosen analogously to (2.68), (2.69), and (2.70). □

It is worth noting that conditions sufficient to guarantee that hypothesis (iii) in the statement of Theorem 2.10 above holds are given in Theorem 2.2(iv).

Example 2.11. As an example of the kinds of systems to which the theory in this section applies, let $\Omega \subset \mathbf{R}^n$ be a bounded domain with Lipschitz boundary. Let $H = L_2(\Omega)$, $V = H_0^1(\Omega)$, and let Q be a closed subspace of $H^s(\Omega)^{n^2} \times H^s(\Omega)^n \times H^s(\Omega)$ with $s > n/2$. Let $V^* = V^* = H^{-1}(\Omega)$. Let $A_0 \in \mathcal{L}(V, V^*)$ be given by

$$A_0\varphi = -\sum_{i,j=1}^{n} D_j \{a_{i,j} D_i \varphi\} + \sum_{i=1}^{n} b_i D_i \varphi + c\varphi, \qquad \varphi \in V,$$

where $a_{i,j} \in L_\infty(\Omega)$, $a_{i,j}(x) = a_{j,i}(x)$, a.e. $x \in \Omega$, $i, j = 1, 2, \ldots, n$,

$$\rho_0 |\xi|^2 \le \sum_{i,j=1}^{n} a_{i,j}(x)\xi_i\xi_j \le \rho_1 |\xi|^2, \qquad \xi \in \mathbf{R}^n, \text{ a.e. } x \in \Omega,$$

for some constants $\rho_0, \rho_1 > 0$, $b_i \in W^{1,\frac{n}{2}}(\Omega)$, $i = 1, 2, \ldots, n$ with $\sum_{i=1}^n D_i b_i(x) \le 0$, a.e. $x \in \Omega$, and $c \in L^{\frac{n}{2}}(\Omega)$ with $c(x) \ge 0$, a.e. $x \in \Omega$.

For $q = (\{q_{i,j}\}, \{q_i\}, q_0) \in Q$, let $A_1(q) \in \mathcal{L}(V, V^*)$ be given by

$$A_1(q)\varphi = -\sum_{i,j=1}^{n} D_j \{q_{i,j} D_i \varphi\} + \sum_{i=1}^{n} q_i D_i \varphi + q_0 \varphi, \qquad \varphi \in V.$$

Note that these are not the most general conditions possible to guarantee that assumptions (A1)–(A5) hold for the general class of second-order elliptic plants and reference models.

3. Tracking and parameter error convergence. In this section we argue that the control objective is achieved (i.e., that the tracking error $e(t)$ converges to zero as $t \to \infty$ and that the feedback control f is, in some sense, bounded), and that under an additional richness condition on the reference model, parameter convergence is obtained (i.e., that $q(t) \to \overline{q}$ as $t \to \infty$). We require that our standing assumptions (A1)–(A5) continue to hold, and that the error equations (2.31)–(2.34) admit a unique solution (e, v, r), with $e, v \in L_2(0, T; V) \cap H^1(0, T; V^*)$ $(\subset C([0, T]; H)!)$ and $r \in H^1(0, T; Q)$ $(\subset C([0, T]; Q)!)$ for all $T > 0$.

Define $E : [0, \infty) \to \mathbf{R}^+$ by

$$E(t) = \frac{1}{2}\{|e(t)|^2 + |r(t)|_Q^2\}, \qquad t \ge 0. \tag{3.1}$$

LEMMA 3.1. *For (e, v, r) the solution to the initial value problem* (2.31)–(2.34)*, the function $E : [0, \infty) \to \mathbf{R}^+$ given by* (3.1) *is nonincreasing, and we have that*

$$E(t) + \rho_0 \int_0^t \|e(s)\|^2 ds \le \xi_0, \qquad t \ge 0, \tag{3.2}$$

where $\xi_0 = E(0) = \frac{1}{2}\{|e_0|^2 + |r_0|_Q^2\}$.

Proof. Using (2.31), (2.33), and assumption (A4), we obtain

$$\begin{aligned} D_t E(t) &= \langle D_t e(t), e(t)\rangle + \langle D_t r(t), r(t)\rangle_Q \\ &= -\langle A_0 e(t), e(t)\rangle \\ &\le -\rho_0 \|e(t)\|^2, \qquad \text{a.e. } t > 0. \end{aligned} \tag{3.3}$$

The estimate in (3.3) implies that E is nonincreasing. Integrating this expression from 0 to t, $t > 0$, we obtain the result given in (3.2). ∎

The above lemma yields the following immediate corollary.

COROLLARY 3.2. *For (e, v, r) the solution to the initial value problem* (2.31)–(2.34)*, we have $e \in L_\infty(0,\infty;H) \cap L_2(0,\infty;V)$ and $r \in L_\infty(0,\infty;Q)$. Consequently, $e \in BC([0,\infty);H)$ and $r \in BC([0,\infty);Q)$.*

LEMMA 3.3. *Let (e, v, r) be the solution to the initial value problem* (2.31)–(2.34)*. Then, if $g \in L_\infty(0,\infty;H)$, it follows that $\int_{t_1}^{t_2} \|v(t)\|^2 dt \le c_1 + c_2(t_2 - t_1)$ for all $t_2 \ge t_1 \ge 0$, where c_1 and c_2 are positive constants which do not depend on t_1 and t_2.*

Proof. It follows from (2.10) and (2.12) that if $g \in L_\infty(0,\infty;H)$ then $v \in L_\infty(0,\infty;H)$. Then, integrating (2.15) from t_1 to t_2, we obtain

$$|v(t_2)|^2 + \rho_0 \int_{t_1}^{t_2} \|v(t)\|^2 dt \le |v(t_1)|^2 + \int_{t_1}^{t_2} \|g(t)\|_*^2 dt. \tag{3.4}$$

Recalling (2.2), it follows from (3.4) that

$$\int_{t_1}^{t_2} \|v(t)\|^2 dt \le \frac{1}{\rho_0} |v|^2_{L_\infty(0,\infty;H)} + \frac{K^2}{\rho_0} |g|^2_{L_\infty(0,\infty;H)} (t_2 - t_1). \qquad \square$$

In the theorem that follows, we establish that the desired control objective is achieved. The proof we provide is in the spirit of the argument used to verify Barbălat's lemma in [34].

THEOREM 3.4. *For (e, v, r) the solution to the initial value problem* (2.31)–(2.34) *and f the adaptive feedback control law given by* (2.22) *or* (2.23)*, we have the following results.*

(i) *If $g \in L_2(0,\infty;V^*) \cup L_\infty(0,\infty;H)$, then $\lim_{t\to\infty} |e(t)| = 0$.*
(ii) *If $g \in L_2(0,\infty;V^*)$, then $u \in L_\infty(0,\infty;H) \cap L_2(0,\infty;V)$ and $f \in L_2(0,\infty;V^*)$.*
(iii) *If the operator A_0 is symmetric in the sense of* (2.14)*, $u_0, v_0 \in V$, $g \in L_2(0,\infty;H)$ and satisfies* (2.13)*, and for $\varphi \in D_0$ and $q \in Q$, we have $A(q)\varphi \in H$ and $|A_1(q)\varphi| \le \gamma_1 |q|_Q |A_0\varphi|$, for some $\gamma_1 > 0$ for which $\gamma_1 |r|_{L_\infty(0,\infty;Q)} < 1$, then $u(t) \in D_0$, a.e. $t > 0$, $u \in L_\infty(0,\infty;V)$, $A_0 u \in L_2(0,\infty;H)$. If, in addition,*
  (a) *$g \in L_\infty(0,\infty;V^*)$, then $f \in L_\infty(0,\infty;V^*)$,*

  or if, in addition,
  (b) *for $\varphi \in D_0$ we have $A_2\varphi \in H$ and $|A_2\varphi| \le \gamma_2 |A_0\varphi|$, for some $\gamma_2 > 0$, then $f \in L_2(0,\infty;H)$.*

Proof. Let $t_2 \ge t_1 \ge 0$ and note that assumptions (A1) and (A4), (2.31), and Lemma 3.1 imply that

$$\begin{aligned}
|e(t_2)|^2 - |e(t_1)|^2 &= \int_{t_1}^{t_2} \frac{d}{dt} |e(t)|^2 dt = 2 \int_{t_1}^{t_2} \langle D_t e(t), e(t) \rangle dt \\
&= -2 \int_{t_1}^{t_2} \langle A_0 e(t), e(t) \rangle dt \\
&\quad -2 \int_{t_1}^{t_2} \langle A_1(r(t))\{e(t) + v(t)\}, e(t) \rangle dt \qquad (3.5) \\
&\le -2\rho_0 \int_{t_1}^{t_2} \|e(t)\|^2 dt \\
&\quad +2\alpha_1 \int_{t_1}^{t_2} |r(t)|_Q \{\|e(t)\| + \|v(t)\|\} \|e(t)\| dt
\end{aligned}$$

$$\begin{aligned}
&\leq 2\alpha_1 \int_{t_1}^{t_2} |r(t)|_Q \{\|e(t)\| + \|v(t)\|\}\|e(t)\| dt \\
&= 2\alpha_1 \int_{t_1}^{t_2} |r(t)|_Q \|e(t)\|^2 dt + 2\alpha_1 \int_{t_1}^{t_2} |r(t)|_Q \|v(t)\|\|e(t)\| dt \\
&\leq 2\alpha_1 \sqrt{2} \xi_0^{\frac{1}{2}} \int_{t_1}^{t_2} \|e(t)\|^2 dt \\
&\quad + 2\alpha_1 \sqrt{2} \xi_0^{\frac{1}{2}} \left(\int_{t_1}^{t_2} \|v(t)\|^2 dt \right)^{\frac{1}{2}} \left(\int_{t_1}^{t_2} \|e(t)\|^2 dt \right)^{\frac{1}{2}}.
\end{aligned}$$

If $g \in L_2(0,\infty;V^*)$, then Theorem 2.2 implies that $\int_{t_1}^{t_2} \|v(t)\|^2 dt \leq \zeta_0$, for some $\zeta_0 > 0$, for all $t_2 \geq t_1 \geq 0$. It then follows from (3.5) that

$$|e(t_2)|^2 - |e(t_1)|^2 \leq \kappa_0 \int_{t_1}^{t_2} \|e(t)\|^2 dt + \kappa_1 \left(\int_{t_1}^{t_2} \|e(t)\|^2 dt \right)^{\frac{1}{2}}, \tag{3.6}$$

where $\kappa_0 = 2\sqrt{2}\alpha_1 \xi_0^{\frac{1}{2}}$ and $\kappa_1 = 2\sqrt{2}\alpha_1 \xi_0^{\frac{1}{2}} \zeta_0^{\frac{1}{2}}$. On the other hand, if $g \in L_\infty(0,\infty;H)$, then Lemma 3.3 and (3.5) imply that

$$|e(t_2)|^2 - |e(t_1)|^2 \leq \kappa_0 \int_{t_1}^{t_2} \|e(t)\|^2 dt + \kappa_0 \left(c_1 + c_2(t_2 - t_1)\right)^{\frac{1}{2}} \left(\int_{t_1}^{t_2} \|e(t)\|^2 dt \right)^{\frac{1}{2}}. \tag{3.7}$$

Now suppose that $\lim_{t\to\infty} |e(t)| \neq 0$. Then there exist $\varepsilon > 0$ and a sequence $\{t_i\}_{i=1}^{\infty}$ with $\lim_{i\to\infty} t_i = \infty$ for which

$$|e(t_i)|^2 > \varepsilon, \qquad i = 1, 2, \ldots. \tag{3.8}$$

If $g \in L_2(0,\infty;V^*)$, then (3.6) and (3.8) imply that for $\delta > 0$ and $i = 1, 2, \ldots$, we have

$$\begin{aligned}
\int_{t_i-\delta}^{t_i} |e(t)|^2 dt &= \int_{t_i-\delta}^{t_i} |e(t_i)|^2 dt - \int_{t_i-\delta}^{t_i} \{|e(t_i)|^2 - |e(t)|^2\} dt \\
&> \varepsilon\delta - \kappa_0 \int_{t_i-\delta}^{t_i} \int_t^{t_i} \|e(s)\|^2 ds dt - \kappa_1 \int_{t_i-\delta}^{t_i} \left(\int_t^{t_i} \|e(s)\|^2 ds \right)^{\frac{1}{2}} dt \\
&\geq \varepsilon\delta - \kappa_0\delta \int_{t_i-\delta}^{t_i} \|e(t)\|^2 dt - \kappa_1\delta \left(\int_{t_i-\delta}^{t_i} \|e(t)\|^2 dt \right)^{\frac{1}{2}}.
\end{aligned}$$

Recalling (2.2), it then follows that

$$\begin{aligned}
K^2 \int_{t_i-\delta}^{t_i} \|e(t)\|^2 dt &\geq \int_{t_i-\delta}^{t_i} |e(t)|^2 dt \\
&> \varepsilon\delta - \kappa_0\delta \int_{t_i-\delta}^{t_i} \|e(t)\|^2 dt - \frac{1}{2}\kappa_1^2\delta^2 - \frac{1}{2} \int_{t_i-\delta}^{t_i} \|e(t)\|^2 dt,
\end{aligned}$$

and therefore that $\left(K^2 + \kappa_0\delta + \frac{1}{2}\right) \int_{t_i-\delta}^{t_i} \|e(t)\|^2 dt > \varepsilon\delta - \frac{1}{2}\kappa_1^2\delta^2$. Choosing $\delta = \frac{\varepsilon}{\kappa_1^2}$ and replacing the sequence $\{t_i\}_{i=1}^{\infty}$ by a subsequence $\{t_{i_j}\}_{j=1}^{\infty}$ for which $t_{i_{j+1}} - t_{i_j} > \delta$, $j = 1, 2, \ldots$, we obtain

$$\int_{t_{i_j}-\delta}^{t_{i_j}} \|e(t)\|^2 dt \geq \frac{\varepsilon\delta}{2\left(K^2 + \kappa_0\delta + \frac{1}{2}\right)}, \qquad j = 1, 2, \ldots. \tag{3.9}$$

Similarly, if $g \in L_\infty(0,\infty;H)$, then (3.7) and (3.8) imply that for $\delta > 0$ and $i = 1, 2, \ldots$, we have

$$\begin{aligned}
\int_{t_i-\delta}^{t_i} |e(t)|^2 dt &= \int_{t_i-\delta}^{t_i} |e(t_i)|^2 dt - \int_{t_i-\delta}^{t_i} \{|e(t_i)|^2 - |e(t)|^2\} dt \\
&> \varepsilon\delta - \kappa_0 \int_{t_i-\delta}^{t_i} \int_t^{t_i} \|e(s)\|^2 ds dt \\
&\quad -\kappa_0 \int_{t_i-\delta}^{t_i} (c_1 + c_2(t_i - t))^{\frac{1}{2}} \left(\int_t^{t_i} \|e(s)\|^2 ds \right)^{\frac{1}{2}} dt \\
&\geq \varepsilon\delta - \kappa_0\delta \int_{t_i-\delta}^{t_i} \|e(t)\|^2 dt - \kappa_0\delta \left(c_1 + c_2\delta\right)^{\frac{1}{2}} \left(\int_{t_i-\delta}^{t_i} \|e(t)\|^2 dt \right)^{\frac{1}{2}}.
\end{aligned}$$

It follows that $\left(K^2 + \kappa_0\delta + \frac{1}{2}\right) \int_{t_i-\delta}^{t_i} \|e(t)\|^2 dt > \varepsilon\delta - \frac{1}{2}\kappa_0^2 \left(c_1 + c_2\delta\right)\delta^2$. Then, if we choose $\delta = \frac{\sqrt{c_1^2\kappa_0^2 + 4c_2\varepsilon} - c_1\kappa_0}{2c_2\kappa_0} > 0$, we again obtain (3.9). But (3.9) contradicts the fact that Lemma 3.1 implies that $\int_0^\infty \|e(t)\|^2 dt \leq \frac{\xi_0}{\rho_0} < \infty$. Consequently, $\lim_{t\to\infty} |e(t)|^2 = 0$, and therefore, $\lim_{t\to\infty} |e(t)| = 0$, which establishes (i).

Corollary 3.2 above implies that $e \in L_\infty(0,\infty;H) \cap L_2(0,\infty;V)$, and if $g \in L_2(0,\infty;V^*)$, then Theorem 2.2 implies that $v \in L_\infty(0,\infty;H) \cap L_2(0,\infty;V)$ as well. Consequently, it follows that $u = e + v \in L_\infty(0,\infty;H) \cap L_2(0,\infty;V)$. To establish that $f \in L_2(0,T;V^*)$, we note that (2.23), assumptions (A1) and (A3), and (2.30) imply that for a.e. $t > 0$

$$\begin{aligned}
\|f(t)\|_* &= \sup_{\|\varphi\|\leq 1} |\langle f(t), \varphi\rangle| \\
&= \sup_{\|\varphi\|\leq 1} |\langle A(q(t))u(t) - A_0u(t) + g(t), \varphi\rangle| \\
&= \sup_{\|\varphi\|\leq 1} |\langle A_1(q(t))u(t) + A_2u(t) - A_0u(t) + g(t), \varphi\rangle| \\
&\leq \{\alpha_1|q(t)|_Q + \alpha_2 + \alpha_0\}\|u(t)\| + \|g(t)\|_* \\
&\leq \{\alpha_1\{|r(t)|_Q + |\overline{q}|_Q\} + \alpha_2 + \alpha_0\}\|u(t)\| + \|g(t)\|_* \\
&\leq \sigma\|u(t)\| + \|g(t)\|_*,
\end{aligned}$$

where $\sigma = \alpha_1\{\sqrt{\xi_0} + |\overline{q}|_Q\} + \alpha_2 + \alpha_0$. It follows that

$$\|f(t)\|_*^2 \leq 2\sigma^2\|u(t)\|^2 + 2\|g(t)\|_*^2 \qquad \text{a.e. } t > 0. \tag{3.10}$$

The estimate (3.10) together with the fact that $g \in L_2(0,\infty;V^*)$ and $u \in L_2(0,\infty;V)$ immediately yields (ii).

To establish (iii), we first note that under the present assumptions, Theorem 2.2 implies that $v \in L_\infty(0,\infty;V)$, $v(t) \in D_0$, a.e. $t > 0$, and $A_0v \in L_2(0,\infty;H)$. Also, Corollary 3.2 implies that $r \in L_\infty(0,\infty;Q)$. Now, recalling the definition of the norm $\|\cdot\|_0$ on V from section 2, (2.31) with $\varphi = A_0e(t)$ (recall Remark 2.1) implies that

$$\langle D_te(t), A_0e(t)\rangle + |A_0e(t)|^2 = \langle A_1(r(t))\{e(t) + v(t)\}, A_0e(t)\rangle, \qquad \text{a.e. } t > 0,$$

and therefore, for any $\varepsilon > 0$, that

$$
\begin{aligned}
\tfrac{1}{2} D_t \|e(t)\|_0^2 + |A_0 e(t)|^2 &\leq |A_1(r(t))\{e(t) + v(t)\}||A_0 e(t)| \\
&\leq \gamma_1 |r(t)|_Q |A_0 e(t) + A_0 v(t)||A_0 e(t)| \\
&\leq \gamma_1 |r|_{L_\infty(0,\infty;Q)} |A_0 e(t)|^2 + \gamma_1 |r|_{L_\infty(0,\infty;Q)} \frac{\varepsilon}{2} |A_0 e(t)|^2 \\
&\quad + \gamma_1 |r|_{L_\infty(0,\infty;Q)} \frac{1}{2\varepsilon} |A_0 v(t)|^2 \qquad \text{a.e. } t > 0.
\end{aligned} \tag{3.11}
$$

Integrating (3.11) from 0 to t, recalling (2.34), (2.16), and our assumption that $\gamma_1 |r|_{L_\infty(0,\infty;Q)} < 1$, and choosing $\varepsilon > 0$ sufficiently small, we find that

$$
\begin{aligned}
&\rho_0 \|e(t)\|^2 + \sigma_0 \int_0^t |A_0 e(s)|^2 ds \\
&\qquad \leq \alpha_0 \|e_0\|^2 + \gamma_1 |r|_{L_\infty(0,\infty;Q)} \frac{1}{2\varepsilon} \int_0^t |A_0 v(s)|^2 ds, \quad t \geq 0
\end{aligned} \tag{3.12}
$$

for some $\sigma_0 > 0$. It follows from (3.12) that $e \in L_\infty(0,\infty;V)$, $e(t) \in D_0$, a.e. $t > 0$, and $A_0 e \in L_2(0,\infty;H)$. Consequently, $u = e + v \in L_\infty(0,\infty;V)$, $u(t) \in D_0$, a.e. $t > 0$, and $A_0 u \in L_2(0,\infty;H)$. This, together with (3.10) establishes the claim in (a). To establish the claim in (b), we have the estimate

$$
\begin{aligned}
|f(t)| &= \sup_{|\varphi| \leq 1} |\langle f(t), \varphi \rangle| \\
&= \sup_{|\varphi| \leq 1} |\langle A(q(t))u(t) - A_0 u(t) + g(t), \varphi \rangle| \\
&= \sup_{|\varphi| \leq 1} |\langle A_1(q(t))u(t) + A_2 u(t) - A_0 u(t) + g(t), \varphi \rangle| \\
&\leq \{\gamma_1 |q(t)|_Q + \gamma_2 + 1\} |A_0 u(t)| + |g(t)| \\
&\leq \{\gamma_1 \{|r|_{L_\infty(0,\infty;Q)} + |\overline{q}|_Q\} + \gamma_2 + 1\} |A_0 u(t)| + |g(t)|, \qquad \text{a.e. } t > 0,
\end{aligned}
$$

from which the desired result immediately follows. □

We note that the condition that $\gamma_1 |r|_{L_\infty(0,\infty;Q)} < 1$ can be satisfied with an appropriate choice of the reference model dynamics A_0, the initial estimate of the unknown parameters q_0 (i.e., that it be sufficiently close to the *true* parameters $\overline{q}$), and the initial state of the reference model v_0 (i.e., that it be sufficiently close to the initial state of the plant u_0). The last two sufficient conditions are a consequence of Lemma 3.1.

In addition to meeting the designated control objective, it is also desirable to have an adaptive control scheme provide parameter convergence as well. In order to establish that the scheme we consider here yields convergence of the parameters $q(t)$ to the *true* parameters $\overline{q}$ as $t \to \infty$, we require the following additional *richness* condition on the reference model.

DEFINITION 3.5. *The reference model* (2.8), (2.9) *or the triple* $\{A_0, g, v_0\}$ *consisting of the reference model dynamics operator* A_0*, the input reference signal,* g*, and the initial state of the reference model* v_0*, will be said to be persistently exciting, or, sufficiently rich, if there exist positive constants* τ_0*,* δ_0*, and* ε_0*, such that for each* $p \in Q$ *with* $|p|_Q = 1$ *and* $t \geq 0$ *sufficiently large, there exists* $\tilde{t} \in [t, t + \tau_0]$ *for which*

$$
\left\| \int_{\tilde{t}}^{\tilde{t}+\delta_0} A_1(p) u(\tau) d\tau \right\|_* \geq \varepsilon_0,
$$

where u *is the closed loop state of the plant as given by the system* (2.26)–(2.29).

THEOREM 3.6. *If either* $g \in L_2(0,\infty;V^*)$ *or* $g \in L_\infty(0,\infty;V)$ *and* $v_0 \in V$, *and if the reference model,* (2.8), (2.9), *is persistently exciting, then* $\lim_{t\to\infty}|r(t)|_Q = 0$.

Proof. If $g \in L_2(0,\infty;V^*)$, then Theorem 3.4 implies that $u \in L_2(0,\infty;V)$. Corollary 3.2 implies that $r \in BC([0,\infty);Q)$, and Lemma 3.1 together with Theorem 3.4 imply that $\lim_{t\to\infty}|r(t)|_Q$ exists. If we assume that $\lim_{t\to\infty}|r(t)|_Q \neq 0$, then there exists $\{t_k\}_{k=1}^{\infty}$, an increasing sequence of positive numbers for which $\lim_{k\to\infty} t_k = \infty$ and

$$|r(t_k)|_Q \geq \delta, \qquad k = 1,2,\ldots, \tag{3.13}$$

for some $\delta > 0$. If the reference model (2.8), (2.9) is persistently exciting, it then follows from assumption (A1) that for each $k = 1,2,\ldots$ and some $\tilde{t}_k \in [t_k, t_k+\tau_0]$, we have

$$\begin{aligned} 0 < \delta\varepsilon_0 &\leq |r(t_k)|_Q \left\| \int_{\tilde{t}_k}^{\tilde{t}_k+\delta_0} A_1\left(\frac{r(t_k)}{|r(t_k)|_Q}\right)u(t)dt \right\|_* \\ &= \left\| \int_{\tilde{t}_k}^{\tilde{t}_k+\delta_0} A_1(r(t_k))u(t)dt \right\|_* \\ &\leq \alpha_1 |r|_{L_\infty(0,\infty;Q)} \sqrt{\delta_0} \left\{ \int_{\tilde{t}_k}^{\tilde{t}_k+\delta_0} \|u(t)\|^2 dt \right\}^{\frac{1}{2}}. \end{aligned} \tag{3.14}$$

Letting $k \to \infty$ in (3.14), and using the fact that $u \in L_2(0,\infty;V)$ implies that

$$\lim_{k\to\infty} \int_{\tilde{t}_k}^{\tilde{t}_k+\delta_0} \|u(t)\|^2 dt = 0,$$

we obtain a contradiction.

Now suppose that $g \in L_\infty(0,\infty;V)$ and $v_0 \in V$. We first recall that Theorem 2.2 implies that $v \in L_\infty(0,\infty;V)$. Now, for $t_2 > t_1$, (2.31), assumption (A3), and (2.2) imply that

$$\begin{aligned} \left\| \int_{t_1}^{t_2} A_1(r(t))u(t)dt \right\|_* &= \left\| \int_{t_1}^{t_2} A_1(r(t))\{e(t)+v(t)\}dt \right\|_* \\ &\leq \|e(t_2)\|_* + \|e(t_1)\|_* + \int_{t_1}^{t_2} \|A_0 e(t)\|_* dt \\ &\leq K|e(t_2)| + K|e(t_1)| \\ &\quad + \alpha_0 (t_2-t_1)^{\frac{1}{2}} \left\{ \int_{t_1}^{t_2} \|e(t)\|^2 dt \right\}^{\frac{1}{2}}. \end{aligned} \tag{3.15}$$

Also, from (2.33), assumption (A1), and Lemma 3.1 it follows that

$$\begin{aligned} |r(t_2)-r(t_1)|_Q &= \sup_{|p|_Q\leq 1} |\langle r(t_2)-r(t_1), p\rangle_Q| \\ &= \sup_{|p|_Q\leq 1} \left| \left\langle \int_{t_1}^{t_2} D_t r(t)dt, p \right\rangle_Q \right| \\ &\leq \int_{t_1}^{t_2} \sup_{|p|_Q\leq 1} |\langle A_1(p)\{e(t)+v(t)\}, e(t)\rangle| \, dt \end{aligned} \tag{3.16}$$

$$
\begin{aligned}
&\leq \alpha_1 \int_{t_1}^{t_2} \{\|e(t)\| + \|v(t)\|\} \|e(t)\| dt \\
&\leq \alpha_1 \int_{t_1}^{t_2} \|e(t)\|^2 dt + \alpha_1 \|v\|_{L_\infty(0,\infty;V)} \int_{t_1}^{t_2} \|e(t)\| dt \\
&\leq \alpha_1 \int_{t_1}^{t_2} \|e(t)\|^2 dt \\
&\quad + \alpha_1 \|v\|_{L_\infty(0,\infty;V)} (t_2 - t_1)^{\frac{1}{2}} \left\{ \int_{t_1}^{t_2} \|e(t)\|^2 dt \right\}^{\frac{1}{2}}.
\end{aligned}
$$

Once again assume that $\lim_{t\to\infty} |r(t)|_Q \neq 0$, and let $\{t_k\}_{k=1}^{\infty}$ be an increasing sequence of positive numbers for which $\lim_{k\to\infty} t_k = \infty$ and for which (3.13) holds for some $\delta > 0$. Assume further that the reference model (2.8), (2.9) is persistently exciting, and for each $k = 1, 2, \ldots$, let $\tilde{t}_k \in [t_k, t_k + \tau_0]$ be such that

$$
\left\| \int_{\tilde{t}_k}^{\tilde{t}_k + \delta_0} A_1 \left(\frac{r(t_k)}{|r(t_k)|_Q} \right) u(t) dt \right\|_* \geq \varepsilon_0. \tag{3.17}
$$

Then, using (3.13), (3.15), (3.16), (3.17), and assumptions (A1) and (A2), we obtain the estimate

$$
\begin{aligned}
0 < \delta\varepsilon_0 &\leq |r(t_k)|_Q \left\| \int_{\tilde{t}_k}^{\tilde{t}_k + \delta_0} A_1 \left(\frac{r(t_k)}{|r(t_k)|_Q} \right) u(t) dt \right\|_* \\
&= \left\| \int_{\tilde{t}_k}^{\tilde{t}_k + \delta_0} A_1(r(t_k)) u(t) dt \right\|_* \\
&\leq \left\| \int_{\tilde{t}_k}^{\tilde{t}_k + \delta_0} A_1(r(t)) u(t) dt \right\|_* + \left\| \int_{\tilde{t}_k}^{\tilde{t}_k + \delta_0} A_1(r(t_k) - r(t)) \{e(t) + v(t)\} dt \right\|_* \\
&\leq K|e(\tilde{t}_k + \delta_0)| + K|e(\tilde{t}_k)| + \alpha_0 \sqrt{\delta_0 \int_{\tilde{t}_k}^{\tilde{t}_k + \delta_0} \|e(t)\|^2 dt} \\
&\quad + \alpha_1 \int_{\tilde{t}_k}^{\tilde{t}_k + \delta_0} |r(t) - r(t_k)|_Q \{\|e(t)\| + \|v(t)\|\} dt \qquad (3.18) \\
&\leq K|e(\tilde{t}_k + \delta_0)| + K|e(\tilde{t}_k)| + \alpha_0 \sqrt{\delta_0 \int_{\tilde{t}_k}^{\tilde{t}_k + \delta_0} \|e(t)\|^2 dt} \\
&\quad + \alpha_1^2 \left(\int_{\tilde{t}_k}^{\tilde{t}_k + \tau_0 + \delta_0} \|e(t)\|^2 dt + \|v\|_{L_\infty(0,\infty;V)} \sqrt{(\tau_0 + \delta_0) \int_{\tilde{t}_k}^{\tilde{t}_k + \tau_0 + \delta_0} \|e(t)\|^2 dt} \right) \\
&\quad \times \int_{\tilde{t}_k}^{\tilde{t}_k + \delta_0} \{\|e(t)\| + \|v(t)\|\} dt \\
&\leq K|e(\tilde{t}_k + \delta_0)| + K|e(\tilde{t}_k)| + \alpha_0 \sqrt{\delta_0 \int_{\tilde{t}_k}^{\tilde{t}_k + \delta_0} \|e(t)\|^2 dt}
\end{aligned}
$$

$$+\alpha_1^2\left(\int_{\tilde{t}_k}^{\tilde{t}_k+\tau_0+\delta_0}\|e(t)\|^2dt+\|v\|_{L_\infty(0,\infty;V)}\sqrt{(\tau_0+\delta_0)\int_{\tilde{t}_k}^{\tilde{t}_k+\tau_0+\delta_0}\|e(t)\|^2dt}\right)$$
$$\times\left(\sqrt{\delta_0\int_{\tilde{t}_k}^{\tilde{t}_k+\delta_0}\|e(t)\|^2dt}+\delta_0\|v(t)\|_{L_\infty(0,\infty;V)}\right).$$

Now Lemma 3.1 implies that for any $L>0$ $\lim_{t\to\infty}\int_t^{t+L}\|e(s)\|^2ds=0$. Therefore, letting $k\to\infty$ in (3.18), Lemma 3.1 and Theorem 3.4 imply that

$$\begin{aligned}
0 &< \delta\varepsilon_0\\
&\le K\lim_{k\to\infty}|e(\tilde{t}_k+\delta_0)|+K\lim_{k\to\infty}|e(\tilde{t}_k)|+\alpha_0\sqrt{\delta_0}\lim_{k\to\infty}\sqrt{\int_{t_k}^{t_k+\tau_0+\delta_0}\|e(t)\|^2dt}\\
&\quad+\alpha_1^2\left(\lim_{k\to\infty}\int_{\tilde{t}_k}^{\tilde{t}_k+\tau_0+\delta_0}\|e(t)\|^2dt\right.\\
&\qquad\left.+\|v\|_{L_\infty(0,\infty;V)}\sqrt{(\tau_0+\delta_0)\lim_{k\to\infty}\int_{\tilde{t}_k}^{\tilde{t}_k+\tau_0+\delta_0}\|e(t)\|^2dt}\right)\\
&\quad\times\left(\sqrt{\delta_0\lim_{k\to\infty}\int_{\tilde{t}_k}^{\tilde{t}_k+\delta_0}\|e(t)\|^2dt}+\delta_0\|v(t)\|_{L_\infty(0,\infty;V)}\right)\\
&= 0,
\end{aligned}$$

which is a contradiction, and the theorem is proved. □

We note that the persistence of excitation condition defined in Definition 3.5 is, in practice, difficult, if not impossible, to verify. However, this condition is analogous to a similar condition used to establish parameter convergence in an on-line identification scheme developed in [7]. In [9] a careful study and analysis of the persistence of excitation condition was carried out yielding valuable insight into how to recognize (based upon its performance) whether an adaptive scheme such as the one we treat here is either *overdamped* (i.e., the operator $-A_0$ is too stable) or *underdamped* (i.e., the persistence of excitation is insufficient). This information can then be used to tune the scheme (i.e., tune the reference model and reference input signal g) so as to achieve a balance between the tracking error convergence (i.e., $\lim_{t\to\infty}|e(t)|=0$) and parameter convergence (i.e., $\lim_{t\to\infty}|q(t)-\overline{q}|_Q=\lim_{t\to\infty}|r(t)|_Q=0$). We note also that it is possible to establish a weaker parameter convergence result in either the absence of persistence of excitation or the presence of *partial* persistence of excitation. The result and its proof are analogous to the corresponding notions in the case of a strict identification scheme (see [7] and [9]).

4. Finite-dimensional approximation. The reference model (2.8), (2.9) and the estimator, or adaptation law for q, (2.24), (2.25), reside in the memory of a computer. Moreover, they are both, in general, infinite-dimensional evolution equations. Consequently their real-time, or on-line, integration requires some form of finite-dimensional approximation. This results in an *approximating* closed-loop system. In this section we consider the finite-dimensional approximation of the reference model and the adaptation law and establish well-posedness and convergence results for the resulting approximating closed-loop systems.

For each $n = 1, 2, \ldots$, let H^n be a finite-dimensional subspace of H with $H^n \subset V$, and let Q^n be a finite-dimensional subspace of Q. Let $P^n : H \to H^n$ denote the orthogonal (with respect to the standard inner product on H) projection of H onto H^n. We then use a Galerkin approach to approximate (2.8), (2.9) and (2.24), (2.25). For each $n = 1, 2, \ldots$, we consider the approximating reference model

$$\langle D_t v^n(t), \varphi^n\rangle + \langle A_0 v^n(t), \varphi^n\rangle = \langle g(t), \varphi^n\rangle, \qquad \varphi^n \in H^n, \text{ a.e. } t > 0, \tag{4.1}$$

$$v^n(0) = v_0^n, \tag{4.2}$$

where $v^n(t), v_0^n \in H^n$, and the approximating adaptation law

$$\langle D_t q^n(t), p^n\rangle_Q + \langle A_1(p^n)u(t), u(t) - v^n(t)\rangle = 0, \qquad p^n \in Q^n, \text{ a.e. } t > 0, \tag{4.3}$$

$$q^n(0) = q_0^n, \tag{4.4}$$

where $q^n(t), q_0^n \in Q^n$. Recalling the definition of the adaptive control law given in (2.22) or (2.23), for each $n = 1, 2, \ldots$, we define an approximating adaptive feedback control law f^n by

$$f^n(t) = A(q^n(t))u(t) - A_0 u(t) + g(t), \quad \text{a.e. } t > 0, \tag{4.5}$$

or

$$\langle f^n(t), \varphi\rangle = \langle A(q^n(t))u(t), \varphi\rangle - \langle A_0 u(t), \varphi\rangle + \langle g(t), \varphi\rangle, \ \ \varphi \in V, \text{a.e. } t > 0, \tag{4.6}$$

where q^n is determined by the system (4.1)–(4.4). Combining (4.1)–(4.4) and either (4.5) or (4.6) together with the plant, (2.4), (2.5), or (2.6), (2.7), we obtain what we will refer to as the approximating closed-loop system

$$\begin{aligned}\langle D_t u^n(t), \varphi\rangle + \langle A_0 u^n(t), \varphi\rangle + \langle A_1(\overline{q} - q^n(t))u^n(t), \varphi\rangle \\ = \langle g(t), \varphi\rangle, \ \varphi \in V, \text{a.e. } t > 0,\end{aligned} \tag{4.7}$$

$$\langle D_t v^n(t), \varphi^n\rangle + \langle A_0 v^n(t), \varphi^n\rangle = \langle g(t), \varphi^n\rangle, \ \ \varphi^n \in H^n, \text{ a.e. } t > 0, \tag{4.8}$$

$$\langle D_t q^n(t), p^n\rangle_Q + \langle A_1(p^n)u^n(t), u^n(t) - v^n(t)\rangle = 0, \ \ p^n \in Q^n, \text{ a.e. } t > 0, \tag{4.9}$$

$$u^n(0) = u_0, \quad v^n(0) = v_0^n, \quad q^n(0) = q_0^n. \tag{4.10}$$

We begin by establishing a well-posedness result for the system (4.7)–(4.10). Our approach is similar to the one taken earlier in section 2 when we considered the well-posedness of the closed-loop system (2.26)–(2.29). We assume that assumptions (A1)–(A9) are satisfied, and we first note that the equation for the reference model (4.8) can be solved independently of equations (4.7) and (4.9). The solution $v^n \in C([0, \infty); H) \cap C^1((0, \infty); H)$ is given by (see, for example, [20])

$$v^n(t) = T_0^n(t)v_0^n + \int_0^t T_0^n(t - s)P^n g(s)ds, \qquad t \geq 0, \tag{4.11}$$

$\{T_0^n(t) : t \geq 0\}$ is the uniformly exponentially stable analytic semigroup of bounded linear operators on H^n generated by the Galerkin approximation, $-A_0^n \in \mathcal{L}(H^n, H^n)$, to the operator $-A_0$. That is, for each $n = 1, 2, \ldots$, $A_0^n \in \mathcal{L}(H^n, H^n)$ is the operator

defined by $A_0^n\varphi^n = \psi^n$, for $\varphi^n \in H^n$, where $\psi^n \in H^n$ is the unique element in H^n satisfying $\langle A_0\varphi^n, \chi^n\rangle = \langle \psi^n, \chi^n\rangle$, $\chi^n \in H^n$, guaranteed to exist by the Riesz representation theorem. Since H^n was assumed to be finite dimensional, we have that $T_0^n(t) = \exp(-A_0^n t) = e^{-A_0^n t}$, $t \geq 0$.

Let $\hat{X} = H \times Q$ be endowed with the inner product $\langle (\varphi, q), (\psi, p)\rangle_{\hat{X}} = \langle \varphi, \psi\rangle + \langle q, p\rangle_Q$, $(\varphi, q), (\psi, p) \in \hat{X}$, and let $|\cdot|_{\hat{X}}$ denote the corresponding induced norm. It follows that $\{\hat{X}, \langle\cdot,\cdot\rangle_{\hat{X}}, |\cdot|_{\hat{X}}\}$ is a Hilbert space. Moreover, as was done in the proof of Theorem 2.4, for the $\alpha \in (0,1)$ in assumption (A6), define the Banach space $\{\hat{X}_\alpha, |\cdot|_{\hat{X}_\alpha}\}$ by $\hat{X}_\alpha = H_\alpha \times Q$ with $|(\varphi, q)|_{\hat{X}_\alpha} = |\varphi|_\alpha + |q|_Q$ for $(\varphi, q) \in \hat{X}_\alpha$.

For $\lambda > 0$ and $\psi \in C([0,\infty); H)$ define the mapping $\hat{G}_\lambda(\cdot,\cdot;\psi) : [0,\infty) \times \hat{X}_\alpha \to \hat{X}$ by

$$\hat{G}_\lambda(t, (\varphi, q); \psi) = (g(t) - B(\varphi)\{\overline{q} - q\}, \lambda q - B(\varphi)'\{\varphi - \psi(t)\}), \tag{4.12}$$

for $t \geq 0$, $(\varphi, q) \in \hat{X}_\alpha$, where for $\varphi \in V$ the operators $B(\varphi) \in \mathcal{L}(Q, V^*)$ and its Banach space adjoint $B(\varphi)' \in \mathcal{L}(V, Q)$ are defined in (2.36) and (2.37), respectively. We note that in the above definition, since $\varphi \in H_\alpha$, assumption (A6) implies that the operator $B(\varphi)$ in fact has range in H and that the operator $B(\varphi)'$ is well defined on H. Consequently the mapping $\hat{G}_\lambda(\cdot,\cdot;\psi)$ given by (4.12) above is well defined on $[0,\infty) \times \hat{X}_\alpha$ with range in $\hat{X}$.

For $\lambda > 0$, define the operator $\hat{A}_\lambda : Dom(\hat{A}_\lambda) \subset \hat{X} \to \hat{X}$ by

$$Dom(\hat{A}_\lambda) = D_0 \times Q, \tag{4.13}$$

$$\hat{A}_\lambda(\varphi, q) = (A_0\varphi, \lambda q), \qquad (\varphi, q) \in Dom(\hat{A}_\lambda). \tag{4.14}$$

The operator $-\hat{A}_\lambda$ is the infinitesimal generator of a uniformly exponentially stable analytic semigroup, $\{\hat{T}_\lambda(t) : t \geq 0\}$, on $\hat{X}$, and $0 \in \rho(-\hat{A}_\lambda)$. For $n = 1, 2, \ldots$, let $\hat{X}^n = H \times Q^n$, and let $P_Q^n : Q \to Q^n$ denote the orthogonal (with respect to the standard inner product on Q, $\langle\cdot,\cdot\rangle_Q$) projection of Q onto Q^n. For $n = 1, 2, \ldots$, $\lambda > 0$ and $\psi \in C([0,\infty); H)$ define the mapping $\hat{G}_\lambda^n(\cdot,\cdot;\psi) : [0,\infty) \times \hat{X}_\alpha \to \hat{X}^n$ by

$$\hat{G}_\lambda^n(t, (\varphi, q); \psi) = (g(t) - B(\varphi)\{\overline{q} - q\}, \lambda P_Q^n q - P_Q^n B(\varphi)'\{\varphi - \psi(t)\}), \tag{4.15}$$

for $t \geq 0$, $(\varphi, q) \in \hat{X}_\alpha$. For $n = 1, 2, \ldots$ and $t \geq 0$, let $\hat{x}^n(t) = (u^n(t), q^n(t))$ and consider the system (4.7), (4.9), and (4.10) written in the form of an initial value problem in $\hat{X}^n$ as

$$D_t\hat{x}^n(t) + \hat{A}_\lambda\hat{x}^n(t) = \hat{G}_\lambda^n(t, \hat{x}^n(t); v^n), \qquad \text{a.e. } t > 0, \tag{4.16}$$

$$\hat{x}^n(0) = \hat{x}_0^n, \tag{4.17}$$

where $\lambda > 0$, $\hat{G}_\lambda^n$ is given by (4.15), $\hat{A}_\lambda$ is given by (4.13) and (4.14), v^n is given by (4.11), and $\hat{x}_0^n = (u_0, q_0^n) \in \hat{X}^n$.

In Theorem 4.1 to follow, we establish that the initial value problem (4.16), (4.17) has a unique local strong solution.

THEOREM 4.1. *If $u_0 \in Dom(A_0^\alpha)$, then for each $n = 1, 2, \ldots$, there exists a $T > 0$ and a unique function $\hat{x}^n \in C([0,T); \hat{X}) \cap C^1((0,T); \hat{X})$ satisfying* (4.16) *and* (4.17). *Moreover, $\hat{x}^n$ satisfies the integral equation*

$$\hat{x}^n(t) = \hat{T}_\lambda(t)\hat{x}_0^n + \int_0^t \hat{T}_\lambda(t-s)\hat{G}_\lambda^n(s, \hat{x}^n(s); v^n)ds, \qquad 0 \leq t < T. \tag{4.18}$$

Proof. For each $n = 1, 2, \ldots$, let $\hat{X}^n_\alpha = Dom(A^\alpha_0) \times Q^n$ be considered as a subspace of $\hat{X}_\alpha$, and let $\hat{U}^n \subset \hat{X}^n_\alpha$ be the neighborhood of $\hat{x}^n_0$ given by $\hat{U}^n = \{\hat{x}^n \in \hat{X}^n_\alpha : |\hat{x}^n - \hat{x}^n_0|_{\hat{X}_\alpha} < \varepsilon\}$. Let $\overline{T} > 0$ and $\lambda > 0$ be fixed. We show that there exists a constant $\hat{L}^n = \hat{L}^n(\varepsilon, \hat{x}^n_0, \lambda, \overline{T}) > 0$, such that

$$|\hat{G}^n_\lambda(t, \hat{\Phi}^n; v^n) - \hat{G}^n_\lambda(s, \hat{\Psi}^n; v^n)|_{\hat{X}} \le \hat{L}^n\{|t-s|^\nu + |\varphi - \psi|_\alpha + |q^n - p^n|_Q\}, \tag{4.19}$$

for $0 \le t, s \le \overline{T}$, and $\hat{\Phi}^n = (\varphi, q^n), \hat{\Psi}^n = (\psi, p^n) \in \hat{U}^n$. The desired result will then follow as in the proof of Theorem 6.3.1 in [33].

Let $0 \le t, s \le \overline{T}$, and let $\hat{\Phi}^n = (\varphi, q^n), \hat{\Psi}^n = (\psi, p^n) \in \hat{U}^n$. Then

$$\begin{aligned} &|\hat{G}^n_\lambda(t, \hat{\Phi}^n; v^n) - \hat{G}^n_\lambda(s, \hat{\Psi}^n; v^n)|^2_{\hat{X}} \\ &\qquad \le 2|g(t) - g(s)|^2 + 2|B(\varphi)\{\overline{q} - q^n\} - B(\psi)\{\overline{q} - p^n\}|^2 \\ &\qquad\quad + 2\lambda^2|q^n - p^n|^2_Q + 2|B(\varphi)'\{\varphi - v^n(t)\} - B(\psi)'\{\psi - v^n(s)\}|^2_Q. \end{aligned} \tag{4.20}$$

Assumptions (A7) and (A8) imply that

$$\begin{aligned} &|B(\varphi)\{\overline{q} - q^n\} - B(\psi)\{\overline{q} - p^n\}| \\ &\qquad \le |B(\varphi)\overline{q} - B(\psi)\overline{q}| + |B(\varphi)q^n - B(\varphi)p^n| + |B(\varphi)p^n - B(\psi)p^n| \\ &\qquad \le \gamma_1|\overline{q}|_Q|\varphi - \psi|_\alpha + \beta_1|\varphi|_\alpha|q^n - p^n|_Q + \gamma_1|p^n|_Q|\varphi - \psi|_\alpha. \end{aligned} \tag{4.21}$$

Now

$$\begin{aligned} &|B(\varphi)'\{\varphi - v^n(t)\} - B(\psi)'\{\psi - v^n(s)\}|_Q \\ &\qquad \le \sup_{|q|_Q \le 1} |\langle B(\varphi)'\varphi - B(\psi)'\psi, q\rangle_Q| \\ &\qquad\quad + \sup_{|q|_Q \le 1} |\langle B(\varphi)'v^n(t) - B(\psi)'v^n(s), q\rangle_Q|. \end{aligned} \tag{4.22}$$

Assumptions (A7) and (A8) imply that

$$\begin{aligned} &|\langle B(\varphi)'\varphi - B(\psi)'\psi, q\rangle_Q| \\ &\qquad \le |\langle B(\varphi)'\varphi - B(\varphi)'\psi, q\rangle_Q| + |\langle B(\varphi)'\psi - B(\psi)'\psi, q\rangle_Q| \\ &\qquad \le \kappa_\alpha\{\beta_1|\varphi|_\alpha + \gamma_1|\psi|_\alpha\}|\varphi - \psi|_\alpha|q|_Q, \end{aligned} \tag{4.23}$$

where, recalling that the space $\{H_\alpha, |\cdot|_\alpha\}$ is densely and continuously embedded in H, κ_α is such that $|\xi| \le \kappa_\alpha|\xi|_\alpha$, for $\xi \in H_\alpha$. Assumptions (A7) and (A8) also imply that

$$\begin{aligned} &|\langle B(\varphi)'v^n(t) - B(\psi)'v^n(s), q\rangle_Q| \\ &\qquad \le |\langle B(\varphi)'v^n(t) - B(\varphi)'v^n(s), q\rangle_Q| + |\langle B(\varphi)'v^n(s) - B(\psi)'v^n(s), q\rangle_Q| \\ &\qquad \le \beta_1|\varphi|_\alpha|v^n(t) - v^n(s)||q|_Q + \gamma_1|v^n(s)||\varphi - \psi|_\alpha|q|_Q \\ &\qquad \le \kappa_1|\varphi|_\alpha|t-s|^\nu|q|_Q + \kappa_2|\varphi - \psi|_\alpha|q|_Q, \end{aligned} \tag{4.24}$$

for some positive constants κ_1 and κ_2, where in the final estimate in (4.24) above, we have applied a regularity result for mild solutions to systems governed by analytic semigroups given in Theorem 4.3.1 in [33] and the fact that $v^n \in C([0, \overline{T}]; H)$ and is therefore H-bounded uniformly on $[0, \overline{T}]$.

Combining (4.20)–(4.24) and assumption (A9), we obtain (4.19), and the theorem is proved. □

It is also possible to establish a *global* existence result for the solution to the system (4.16), (4.17). However, to do this we require the following additional assumption.

(A12) The operator $A_1(\overline{q}) : V \to V^*$ is *monotone* in the sense that $\langle A_1(\overline{q})\varphi - A_1(\overline{q})\psi, \varphi - \psi\rangle \geq 0$, $\varphi, \psi \in V$.

We note that assumption (A12) is not excessively restrictive in that monotonicity can be demonstrated for relatively large classes of nonlinear operators. It corresponds physically to some form of energy dissipation in the plant. In particular, we note that the operator $A_1(\overline{q})$ appearing in the example presented in section 2.1 satisfies assumption (A12) (see (2.61)).

THEOREM 4.2. *Suppose that assumptions* (A1)–(A9) *and* (A12) *hold and that* $u_0 \in Dom(A_0^\alpha)$. *Then for each* $n = 1, 2, \ldots$, *the initial value problem* (4.16), (4.17) *has a unique solution* $\hat{x}^n = (u^n, q^n)$ *which exists for all* $t \geq 0$.

Proof. As in the proof of Theorem 2.6, we show that for each $n = 1, 2, \ldots$, the local solution $\hat{x}^n$ to the initial value problem (4.16), (4.17) shown to exist in Theorem 4.1 can be continued by arguing that $|\hat{x}^n(t)|_{\hat{X}_\alpha}$ remains bounded as $t \uparrow T$. For $t \in [0, T)$ we have $|\hat{x}^n(t)|_{\hat{X}_\alpha} = |u^n(t)|_\alpha + |q^n(t)|_Q$.

We begin by determining a bound for $|q^n(t)|_Q$ as $t \uparrow T$. From (4.7) and (4.9) and assumptions (A1), (A2), (A4), and (A12), for $t \in (0, T)$, and θ the *zero* vector in V, we obtain the estimate

$$
\begin{aligned}
\frac{1}{2}\left\{D_t|u^n(t)|^2 + D_t|q^n(t)|_Q^2\right\} &= \langle D_t u^n(t), u^n(t)\rangle + \langle D_t q^n(t), q^n(t)\rangle_Q \\
&= -\langle A_0 u^n(t), u^n(t)\rangle - \langle A_1(\overline{q})u^n(t), u^n(t)\rangle \\
&\quad +\langle g(t), u^n(t)\rangle + \langle A_1(q^n(t))u^n(t), v^n(t)\rangle \\
&= -\langle A_0 u^n(t), u^n(t)\rangle - \langle A_1(\overline{q})u^n(t) \\
&\quad -A_1(\overline{q})\theta, u^n(t)\rangle - \langle A_1(\overline{q})\theta, u^n(t)\rangle \\
&\quad +\langle g(t), u^n(t)\rangle + \langle A_1(q^n(t))u^n(t), v^n(t)\rangle \\
&\leq -\langle A_0 u^n(t), u^n(t)\rangle + |\langle A_1(\overline{q})\theta, u^n(t)\rangle| \\
&\quad +\langle g(t), u^n(t)\rangle + \langle A_1(q^n(t))u^n(t), v^n(t)\rangle \\
&\leq -\rho_0\|u^n(t)\|^2 + \|g(t)\|_*\|u^n(t)\| \\
&\quad +\alpha_1|q^n(t)|_Q\|u^n(t)\|\|v^n(t)\|.
\end{aligned} \tag{4.25}
$$

Now $v^n \in C([0, T]; H)$. But $v^n(t) \in H^n$, $t \in [0, T]$, and $H^n \subset V$ finite dimensional imply that $v^n \in C([0, T]; V)$ and, therefore, that $\|v^n(t)\|$ is bounded for $t \in [0, T]$. Consequently, for $\varepsilon > 0$, (4.25) yields

$$
\begin{aligned}
\frac{1}{2}\left\{D_t|u^n(t)|^2 + D_t|q^n(t)|_Q^2\right\} &\leq -\rho_0\|u^n(t)\|^2 + \frac{1}{2\varepsilon}\|g(t)\|_*^2 \\
&\quad +\frac{\varepsilon}{2}\|u^n(t)\|^2 + \frac{\kappa^n}{2\varepsilon}|q^n(t)|_Q^2 + \frac{\kappa^n\varepsilon}{2}\|u^n(t)\|^2
\end{aligned}
$$

or

$$
\frac{1}{2}\{D_t|u^n(t)|^2 + D_t|q^n(t)|_Q^2\} + \left\{\rho_0 - (1+\kappa^n)\frac{\varepsilon}{2}\right\}\|u^n(t)\|^2 \leq \frac{1}{2\varepsilon}\|g(t)\|_*^2 + \frac{\kappa^n}{2\varepsilon}|q^n(t)|_Q^2
$$

for some $\kappa^n > 0$. Choosing $\varepsilon = \varepsilon^n = 2\rho_0/(1+\kappa^n)$, we obtain

$$D_t|u^n(t)|^2 + D_t|q^n(t)|_Q^2 \leq \mu_1\|g(t)\|_*^2 + \mu_2^n|q^n(t)|_Q^2, \qquad 0 \leq t < T, \tag{4.26}$$

for some $\mu_1, \mu_2^n > 0$. Integrating both sides of the estimate in (4.26) from 0 to t, we find that for $0 \leq t < T$,

$$|u^n(t)|^2 + |q^n(t)|_Q^2 \leq |u_0|^2 + |q_0^n|_Q^2 + \mu_1\int_0^t \|g(s)\|_*^2 ds + \mu_2^n\int_0^t |q^n(s)|_Q^2 ds. \tag{4.27}$$

Applying the generalized Gronwall inequality (see, for example, [15]) to (4.27) above, we obtain

$$\begin{aligned} |u^n(t)|^2 + |q^n(t)|_Q^2 &\leq \zeta^n(t) + \mu_2^n\textstyle\int_0^t e^{\mu_2^n(t-s)}\zeta^n(s)ds \\ &\leq (1+\mu_2^n T e^{\mu_2^n T})\zeta^n(T) \\ &= \kappa_T^n, \qquad 0 \leq t < T, \end{aligned} \tag{4.28}$$

where for each $n = 1, 2, \ldots$, ζ^n is the monotone increasing function on $[0,T]$ given by

$$\zeta^n(t) = |u_0|^2 + |q_0^n|_Q^2 + \mu_1\int_0^t \|g(s)\|_*^2 ds, \qquad 0 \leq t \leq T.$$

Now for $t \in [0,T)$, from (4.18) we obtain

$$A_0^\alpha u^n(t) = A_0^\alpha T_0(t)u_0 + \int_0^t A_0^\alpha T_0(t-s)\{g(s) - B(u^n(s))\{\overline{q} - q^n(s)\}\}ds. \tag{4.29}$$

It follows from (4.29) and assumptions (A4) and (A7) that for $t \in [0,T)$ we have

$$\begin{aligned} |u^n(t)|_\alpha \leq\ & e^{-\rho_0 K^{-2}t}|u_0|_\alpha + \int_0^t M_\alpha(t-s)^{-\alpha}e^{-\rho_0 K^{-2}(t-s)}|g(s)|ds \\ & + \int_0^t M_\alpha(t-s)^{-\alpha}e^{-\rho_0 K^{-2}(t-s)}\beta_1\left\{|\overline{q}|_Q + |q^n(s)|_Q\right\}|u^n(s)|_\alpha ds \end{aligned}$$

for some positive constant M_α (see [33]). Assumption (A9) and (4.28) then imply that

$$\begin{aligned} |u^n(t)|_\alpha \leq\ & |u_0|_\alpha + M_\alpha\|g\|_{C([0,T];H)}\frac{T^{1-\alpha}}{1-\alpha} \\ & + M_\alpha\beta_1\{|\overline{q}|_Q + \sqrt{\kappa_T^n}\}\int_0^t (t-s)^{-\alpha}|u^n(s)|_\alpha ds \end{aligned} \tag{4.30}$$

for $0 \leq t < T$. An application of Lemma 5.6.7 in [33] to the estimate given in (4.30) above then yields the existence of a constant $\lambda_T^n > 0$ for which

$$|u^n(t)|_\alpha \leq \lambda_T^n, \qquad 0 \leq t < T. \tag{4.31}$$

Combining estimates (4.28) and (4.31), we obtain the desired result, and the theorem is proved. □

Before we present our convergence result, we discuss some computational considerations and, in particular, the matrix representations for the finite-dimensional

approximating reference model (4.1), (4.2) and adaptation law (4.3), (4.4). For each $n = 1, 2, \ldots$ let $\{\varphi_j^n\}_{j=1}^{k^n}$ be a basis for H^n and let $\{p_j^n\}_{j=1}^{\ell^n}$ be a basis for Q^n. Let

$$v^n(t) = \sum_{j=1}^{k^n} \tilde{v}_j^n(t)\varphi_j^n \qquad \text{and} \qquad q^n(t) = \sum_{j=1}^{\ell^n} \tilde{q}_j^n(t)p_j^n, \qquad t \geq 0. \tag{4.32}$$

That is, for each $t \geq 0$, let $\tilde{v}^n(t) \in \mathbf{R}^{k^n}$ and $\tilde{q}^n(t) \in \mathbf{R}^{\ell^n}$ be, respectively, the vector representations for $v^n(t) \in H^n$ and $q^n \in Q^n$ with respect to the bases $\{\varphi_j^n\}_{j=1}^{k^n}$ and $\{p_j^n\}_{j=1}^{\ell^n}$. We choose $v_0^n = P^n v_0 \in H^n$ and $q_0^n = P_Q^n q_0 \in Q^n$.

The matrix form of the approximating reference model (4.1), (4.2) then becomes

$$M^n D_t \tilde{v}^n(t) + K^n \tilde{v}^n(t) = g^n(t), \qquad t > 0, \tag{4.33}$$

$$M^n \tilde{v}^n(0) = \tilde{v}_0^n, \tag{4.34}$$

where the $k^n \times k^n$ matrices M^n and K^n are given by $[M^n]_{i,j} = \langle \varphi_j^n, \varphi_i^n \rangle$ and $[K^n]_{i,j} = \langle A_0 \varphi_j^n, \varphi_i^n \rangle$, $i, j = 1, 2, \ldots, k^n$, respectively, and $[\tilde{v}_0^n]_i = \langle v_0, \varphi_i^n \rangle$ and $[g^n(t)]_i = \langle g(t), \varphi_i^n \rangle$, $i = 1, 2, \ldots, k^n$, $t \geq 0$. Note that since $\{\varphi_j^n\}_{j=1}^{k^n}$ is a basis for H^n, the matrix M^n is nonsingular.

For $u(t) \in V$, the output of the plant, (2.4), (2.5), or (2.6), (2.7), at time $t \geq 0$, the matrix form of the approximating adaptation law (4.3), (4.4) is given by

$$M_Q^n D_t \tilde{q}^n(t) - L^n(u(t))\tilde{v}^n(t) = -h^n(u(t)), \qquad t > 0, \tag{4.35}$$

$$M_Q^n \tilde{q}^n(0) = \tilde{q}_0^n, \tag{4.36}$$

where the $\ell^n \times \ell^n$ matrix M_Q^n is given by $[M_Q^n]_{i,j} = \langle p_j^n, p_i^n \rangle_Q$, $i, j = 1, 2, \ldots, \ell^n$, for $\varphi \in V$, the $\ell^n \times k^n$ matrix $L^n(\varphi)$ and the ℓ^n-vector $h^n(\varphi)$ are given by $[L^n(\varphi)]_{i,j} = \langle A_1(p_i^n)\varphi, \varphi_j^n \rangle$ and $[h^n(\varphi)]_i = \langle A_1(p_i^n)\varphi, \varphi \rangle$, $i = 1, 2, \ldots, \ell^n$, $j = 1, 2, \ldots, k^n$, respectively, and the ℓ^n-vector $\tilde{q}_0^n$ is given by $[\tilde{q}_0^n]_i = \langle q_0, p_i^n \rangle_Q$, $i = 1, 2, \ldots, \ell^n$. Once again, since $\{p_j^n\}_{j=1}^{\ell^n}$ is a basis for Q^n, the matrix M_Q^n is nonsingular.

Combining (4.33), (4.34) and (4.35), (4.36), for $u(t) \in V$, the output of the plant, (2.4), (2.5), or (2.6), (2.7), at time $t \geq 0$, the $k^n + \ell^n$-dimensional *linear* system

$$\begin{bmatrix} D_t \tilde{v}^n(t) \\ D_t \tilde{q}^n(t) \end{bmatrix} + \begin{bmatrix} (M^n)^{-1} K^n & 0 \\ -(M_Q^n)^{-1} L^n(u(t)) & 0 \end{bmatrix} \begin{bmatrix} \tilde{v}^n(t) \\ \tilde{q}^n(t) \end{bmatrix} = \begin{bmatrix} (M^n)^{-1} g^n(t) \\ -(M_Q^n)^{-1} h^n(u(t)) \end{bmatrix}, \ t > 0,$$

$$\begin{bmatrix} \tilde{v}^n(0) \\ \tilde{q}^n(0) \end{bmatrix} = \begin{bmatrix} (M^n)^{-1} \tilde{v}_0^n \\ (M_Q^n)^{-1} \tilde{q}_0^n \end{bmatrix}$$

must be integrated to determine the state of the approximating reference model, $v^n(t)$, and the approximating parameter estimator, $q^n(t)$, at time $t > 0$. The estimate for the parameters is given by (4.32), and the control input is given by

$$f^n(t) = A\left(\sum_{j=1}^{\ell^n} \tilde{q}_j^n(t) p_j^n\right) u(t) - A_0 u(t) + g(t), \qquad t \geq 0.$$

We are now ready to turn to our convergence result. We require the following rather standard assumptions on the approximation properties of the finite-dimensional subspaces H^n and Q^n.

(A13) The subspace H^n is such that for each $n = 1, 2, \ldots$ there exists a mapping $\pi^n \in \mathcal{L}(V, V)$ for which $\pi^n \varphi \in H^n$, $\varphi \in V$, and $\lim_{n\to\infty} \|\pi^n \varphi - \varphi\| = 0$, $\varphi \in V$.

(A14) The subspace Q^n is such that $\lim_{n\to\infty} |P_Q^n q - q|_Q = 0$, $q \in Q$.

We note that assumption (A13) together with the dense and continuous embedding of V in H is sufficient to conclude that $\lim_{n\to\infty} |P^n \varphi - \varphi| = 0$, $\varphi \in H$. We note further that in many cases it is possible to choose $\pi_n = P^n$. Indeed, this is in fact the case for polynomial spline-based subspaces. Assumption (A13) can then be verified using the estimates found in, for example, [36].

The following theorem concerning the convergence of the approximating semigroups $\{T_0^n(t) : t \geq 0\}$ to the semigroup $\{T_0(t) : t \geq 0\}$ is established in [3] using the well-known Trotter–Kato theorem (see, for example, [20] and [33]).

THEOREM 4.3. *Under assumptions* (A3), (A4), *and* (A13), *for each* $T > 0$ *we have the following results.*

(i) *There exists a constant* $M_0 > 0$, *independent of* n, *for which* $\|T_0^n(t)\varphi^n\| \leq M_0\|\varphi^n\|$, $\varphi^n \in H^n$.

(ii) *For* $\varphi \in H$ *and* $t \in [0, T]$, $\lim_{n\to\infty} |T_0^n(t)P^n\varphi - T_0(t)\varphi| = 0$, *uniformly in* t *for* t *in compact subintervals of* $[0, T]$.

(iii) *For* $\varphi \in V$ *and* $t \in [0, T]$, $\lim_{n\to\infty} \|T_0^n(t)\pi^n\varphi - T_0(t)\varphi\| = 0$, *uniformly in* t *for* t *in compact subintervals of* $[0, T]$.

(iv) *For* $\varphi \in H$ *and* $t \in (0, T]$, $\lim_{n\to\infty} \|T_0^n(t)P^n\varphi - T_0(t)\varphi\| = 0$, *uniformly in* t *for* t *in compact subintervals of* $(0, T]$.

Once (i) has been established, the essence of the proof of (ii)–(iv) is demonstrating resolvent convergence in V. Let $\lambda > 0$ and $\varphi \in V$, and set $\psi = (\lambda I + A_0)^{-1}\varphi$ and $\psi^n = (\lambda I + A_0^n)^{-1}\pi^n\varphi$, $n = 1, 2, \ldots$. The triangle inequality yields

$$\|\psi - \psi^n\| \leq \|\psi - \pi^n\psi\| + \|\pi^n\psi - \psi^n\|. \tag{4.37}$$

Assumption (A13) implies that the first term on the right-hand side of the estimate in (4.37) tends to zero as $n \to \infty$. With regard to the second term, using assumptions (A3) and (A4) we obtain, for any $\varepsilon > 0$,

$$\begin{aligned}
\rho_0\|\psi^n - \pi^n\psi\|^2 &\leq \langle A_0\{\psi^n - \pi^n\psi\}, \psi^n - \pi^n\psi\rangle \\
&= \langle(\lambda I + A_0^n)\psi^n - (\lambda I + A_0)\psi, \psi^n - \pi^n\psi\rangle + \lambda\langle\psi - \pi^n\psi, \psi^n - \pi^n\psi\rangle \\
&\quad -\lambda\langle\psi^n - \pi^n\psi, \psi^n - \pi^n\psi\rangle + \langle A_0\{\psi - \pi^n\psi\}, \psi^n - \pi^n\psi\rangle \\
&\leq K\|\varphi - \pi^n\varphi\|\|\psi^n - \pi^n\psi\| + \lambda\|\psi - \pi^n\psi\|\|\psi^n - \pi^n\psi\| - \lambda|\pi^n\psi - \psi^n|^2 \\
&\quad +\alpha_0\|\psi - \pi^n\psi\|\|\psi^n - \pi^n\psi\| \\
&\leq \frac{K}{2\varepsilon}\|\varphi - \pi^n\varphi\|^2 + \frac{\lambda + \alpha_0}{2\varepsilon}\|\psi - \pi^n\psi\|^2 + \frac{1}{2}\{K + \lambda + \alpha_0\}\varepsilon\|\psi^n - \pi^n\psi\|^2.
\end{aligned}$$

Choosing $\varepsilon > 0$ sufficiently small, we find that

$$\|\psi^n - \pi^n\psi\|^2 \leq \nu_1\|\varphi - \pi^n\varphi\|^2 + \nu_2\|\psi - \pi^n\psi\|^2$$

for some constants $\nu_1, \nu_2 > 0$. Invoking assumption (A13) and recalling (4.37), we obtain

$$\lim_{n\to\infty} \|(\lambda I + A_0^n)^{-1}\pi^n\varphi - (\lambda I + A_0)^{-1}\varphi\| = 0, \qquad \varphi \in V.$$

We will require the following corollary to Theorem 4.3 above.

COROLLARY 4.4. *If $v_0 \in V$ and $v_0^n = \pi^n v_0$, then under assumptions* (A3), (A4), (A9), *and* (A13), *for each $T > 0$ we have the following results.*

(i) *For each $t \in [0,T]$, $\lim_{n\to\infty} \|v^n(t) - v(t)\| = 0$.*

(ii) *There exists a constant $\kappa_{v^n} = \kappa_{v^n}(T) > 0$, for which $\|v^n(t)\| \le \kappa_{v^n}$, $0 \le t \le T$, with $\kappa = \sup_n \kappa_{v^n} < \infty$.*

Proof. From (2.12) and (4.11) for $t \in [0,T]$, we obtain

$$\begin{aligned}\|v^n(t) - v(t)\| &\le \|T_0^n(t)\pi^n v_0 - T_0(t)v_0\| \\ &\quad + \int_0^t \|T_0^n(t-s)P^n g(s) - T_0(t-s)g(s)\|ds.\end{aligned}$$

Statement (iii) of Theorem 4.3 implies that the first term on the right-hand side of the above expression tends to zero as $n \to \infty$. Statement (iv) of Theorem 4.3 implies that the term under the integral sign tends to zero for almost every $s \in [0,T]$. Moreover, Lemma 3.6.2 in [39] implies that

$$\|T_0^n(t-s)P^n g(s) - T_0(t-s)g(s)\| \le \frac{C}{(t-s)^{\frac{1}{2}}}|g(s)|, \qquad 0 \le s < t,$$

where the constant $C > 0$ is independent of n. That the constant C is independent of n follows from the fact that the operators A_0^n are defined via Galerkin approximation and, consequently, the estimates given in Lemma 3.6.1 in [39] for the resolvent of $-A_0$ continue to hold for the resolvent of $-A_0^n$ with all constants independent of n (see also [4]). An application of the Lebesgue dominated convergence theorem then yields (i).

Statement (ii) is established analogously. Indeed, for $t \in [0,T]$, (4.11) yields

$$\begin{aligned}\|v^n(t)\| &\le \|T_0^n(t)\pi^n v_0\| + \int_0^t \|T_0^n(t-s)P^n g(s)\|ds \\ &\le M_0\kappa_\pi\|v_0\| + \int_0^t \frac{C}{(t-s)^{\frac{1}{2}}}|g(s)|ds \\ &\le M_0\kappa_\pi\|v_0\| + 2C\|g\|_{C([0,T];H)}\sqrt{T} = \kappa_{v^n}(T),\end{aligned}$$

where $\kappa_\pi > 0$ is the uniform bound on the operators $\pi^n \in \mathcal{L}(V,V)$, $n = 1,2,\ldots$, guaranteed to exist by assumption (A13) and the uniform boundedness principle. This proves the theorem. □

Using Corollary 4.4, the next corollary follows immediately by inspection.

COROLLARY 4.5. *Suppose that the assumptions* (A1)–(A9) *and* (A12)–(A14) *are satisfied, and that $u_0 \in Dom(A_0^\alpha)$. Suppose further that $v_0 \in V$, $v_0^n = \pi^n v_0$, and that $q_0^n = P_Q^n q_0$. Then the constants κ_T^n and λ_T^n defined in the proof of* Theorem 4.2 *above are in fact independent of n.*

The implication of Corollary 4.5 is that for $T > 0$ fixed, $|u^n(t)|_\alpha$ and $|q^n(t)|_Q$ are bounded uniformly in n and t for $t \in [0,T]$, where for each $n = 1,2,\ldots$ u^n and q^n satisfy (4.7)–(4.10). That is there exist constants $\kappa_T > 0$ and $\lambda_T > 0$, independent of n for which

$$|\hat{x}^n(t)|_{\hat{X}_\alpha} = |u^n(t)|_\alpha + |q^n(t)|_Q \le \lambda_T + \sqrt{\kappa_T}, \qquad 0 \le t \le T, \tag{4.38}$$

where for each $n = 1,2,\ldots$ $\hat{x}^n$, is the solution to the initial value problem (4.16), (4.17).

Our convergence result is given in Theorem 4.7 below. Its proof requires the following lemma.

LEMMA 4.6. *If $Dom(A_0^\alpha) \subset V$ for some $\alpha \in (0,1)$, then the Banach space $\{H_\alpha, |\cdot|_\alpha\}$ defined by $H_\alpha = Dom(A_0^\alpha)$ and $|\varphi|_\alpha = |A_0^\alpha \varphi|$, $\varphi \in Dom(A_0^\alpha)$ is continuously embedded in V. That is, there exists a constant $K_V > 0$ for which $\|\varphi\| \leq K_V |\varphi|_\alpha$, $\varphi \in Dom(A_0^\alpha)$.*

Proof. It can be shown (see [28, page 11]) that there exists a linear, self-adjoint, and positive operator $\Lambda : Dom(\Lambda) \subset H \to H$ for which $Dom(\Lambda) = V$ and for which the norm $\|\cdot\|_\Lambda$ on V given by $\|\varphi\|_\Lambda = \left\{|\varphi|^2 + |\Lambda\varphi|^2\right\}^{\frac{1}{2}}$, $\varphi \in V$, is equivalent to the standard norm, $\|\cdot\|$, on V. Then, for $\varphi \in Dom(A_0^\alpha)$, Corollary 2.6.11 in [33] implies that

$$\|\varphi\|^2 \leq K_1 \left\{|\varphi|^2 + |\Lambda\varphi|^2\right\} \leq K_2 \left\{|\varphi|^2 + |A_0^\alpha \varphi|^2\right\} \leq K_2 \|\varphi\|_\alpha^2 \leq K_V^2 |\varphi|_\alpha^2$$

for some constants $K_1, K_2, K_V > 0$, where the norm $\|\cdot\|_\alpha$ on H_α was defined and discussed in the proof of Theorem 2.4. □

THEOREM 4.7. *Suppose that assumptions* (A1)–(A9) *and assumptions* (A12)–(A14) *hold, that $u_0 \in Dom(A_0^\alpha)$, $\alpha \in (0,1)$ as in assumption* (A6)*, and that $v_0^n = \pi^n v_0$ and $q_0^n = P_Q^n q_0$. Then*

$$\lim_{n\to\infty} |u^n(t) - u(t)|_\alpha = \lim_{n\to\infty} |A_0^\alpha u^n(t) - A_0^\alpha u(t)| = 0 \tag{4.39}$$

and

$$\lim_{n\to\infty} |q^n(t) - q(t)|_Q = 0, \tag{4.40}$$

uniformly in t, for $t \in [0,T]$, where u^n, q^n satisfy (4.7)–(4.10)*, and u and q satisfy* (2.26)–(2.29)*. Moreover, if, in addition, the operator A_2 satisfies a Lipschitz condition of the form*

$$\|A_2\varphi - A_2\psi\|_* \leq \gamma_2 |\varphi - \psi|_\alpha, \qquad \varphi, \psi \in Dom(A_0^\alpha), \tag{4.41}$$

for some $\gamma_2 > 0$, where $\alpha \in (0,1)$ is as in assumption (A6)*, then we have*

$$\lim_{n\to\infty} f^n(t) = f(t) \tag{4.42}$$

in V^* uniformly in t, for $t \in [0,T]$, and therefore $\lim_{n\to\infty} f^n = f$ in $L_2(0,T;V^*)$, for each $T > 0$, where for each $n = 1, 2, \ldots$, f^n is given by (4.5) *or* (4.6)*, and f is given by* (2.22) *or* (2.23). Before we prove Theorem 4.7, we note that (4.39) implies that

$$\lim_{n\to\infty} |u^n(t) - u(t)| \leq \lim_{n\to\infty} \|u^n(t) - u(t)\|_\alpha \leq \kappa_\alpha \lim_{n\to\infty} |u^n(t) - u(t)|_\alpha = 0,$$

uniformly in t for $t \in [0,T]$. Moreover, (4.39) together with Lemma 4.6 also imply that

$$\lim_{n\to\infty} \|u^n(t) - u(t)\| \leq K_V \lim_{n\to\infty} |u^n(t) - u(t)|_\alpha = 0, \tag{4.43}$$

uniformly in t for $t \in [0,T]$.

Proof of Theorem 4.7. For each $t \geq 0$, let $\hat{x}(t) = (u(t), q(t))$, where u and q satisfy (2.26)–(2.29). Then for each $\lambda > 0$, $\hat{x}$ satisfies

$$\hat{x}(t) = \hat{T}_\lambda(t)\hat{x}_0 + \int_0^t \hat{T}_\lambda(t-s)\hat{G}_\lambda(s, \hat{x}(s); v)ds, \qquad t \geq 0, \tag{4.44}$$

where $\hat{G}_\lambda$ is given by (4.12), $\{\hat{T}_\lambda(t) : t \geq 0\}$ is the uniformly exponentially stable semigroup of bounded linear operators on $\hat{X}$ generated by the operator $-\hat{A}_\lambda$ defined in (4.13) and (4.14), and v satisfies (2.8), (2.9) and is given by (2.12). Subtracting (4.44) from (4.18) and taking norms in $\hat{X}_\alpha$, for each $t \geq 0$, we find that

$$
\begin{aligned}
|\hat{x}^n(t) - \hat{x}(t)|_{\hat{X}_\alpha} &\leq e^{-\lambda t}|P_Q^n q_0 - q_0|_Q \\
&\quad + \int_0^t |A_0^\alpha T_0(t-s)\{B(u^n(s))\{q^n(s) - q(s)\} \\
&\quad + B(u(s))\{\overline{q} - q(s)\}\}| \\
&\quad + e^{-\lambda(t-s)}|\lambda\{q^n(s) - q(s)\} + B(u(s))'\{u(s) - v(s)\} \\
&\quad - P_Q^n B(u^n(s))'\{u^n(s) - v^n(s)\}|_Q ds \\
&\leq |P_Q^n q_0 - q_0|_Q + \int_0^t M_\alpha (t-s)^{-\alpha} \left\{\gamma_1 |\overline{q}_Q| u^n(s) - u(s)|_\alpha \right. \\
&\quad + \gamma_1 |q^n(s)|_Q |u^n(s) - u(s)|_\alpha \\
&\quad \left. + \beta_1 |q^n(s) - q(s)|_Q |u(s)|_\alpha\right\} + \lambda |q^n(s) - q(s)|_Q \\
&\quad + \gamma_1 |u^n(s) - u(s)|_\alpha |u^n(s)| + \gamma_1 |u^n(s) - u(s)||u(s)|_\alpha \\
&\quad + \gamma_1 |u^n(s) - u(s)|_\alpha |v^n(s)| + \gamma_1 |u(s)|_\alpha |v^n(s) - v(s)| \\
&\quad + |\{I - P_Q^n\} B(u(s))'\{u(s) - v(s)\}|_Q ds,
\end{aligned} \tag{4.45}
$$

where in the above estimate we have invoked assumptions (A7) and (A8). Assumption (A14), Corollary 4.4, (2.54), and (4.38) then imply that

$$
|\hat{x}^n(t) - \hat{x}(t)|_{\hat{X}_\alpha} \leq \varepsilon^n + \kappa \int_0^t (t-s)^{-\alpha} |\hat{x}^n(s) - \hat{x}(s)|_{\hat{X}_\alpha} ds, \tag{4.46}
$$

where $\kappa > 0$ and $\lim_{n\to\infty} \varepsilon^n = 0$. The estimate given in (4.46) together with a careful inspection of the proof of Lemma 5.6.7 in [33] then yield that $|\hat{x}^n(t) - \hat{x}(t)|_{\hat{X}_\alpha} \leq \varepsilon^n K_T$, $0 \leq t \leq T$, where $K_T = \frac{T^{1-\alpha}}{1-\alpha} > 0$. This establishes (4.39) and (4.40).

We now turn our attention to establishing (4.42). For $t \in [0, T]$ and $\varphi \in V$, Assumptions (A2), (A3), (A6), (A7), and (A8), (2.2), (4.41), and Lemma 4.6 (or, equivalently, (4.43)) imply that

$$
\begin{aligned}
|\langle f^n(t) - f(t), \varphi\rangle| &= |\langle A(q^n(t))u^n(t) - A_0 u^n(t) - A(q(t))u(t) + A_0 u(t), \varphi\rangle| \\
&\leq |\langle A_1(q^n(t))u^n(t) - A_1(q(t))u^n(t), \varphi\rangle| \\
&\quad + |\langle A_1(q(t))u^n(t) - A_1(q(t))u(t), \varphi\rangle| \\
&\quad + |\langle A_2 u^n(t) - A_2 u(t), \varphi\rangle| + |\langle A_0 u^n(t) - A_0 u(t), \varphi\rangle| \\
&\leq K\beta_1 |q^n(t) - q(t)|_Q |u^n(t)|_\alpha \|\varphi\| \\
&\quad + K\gamma_1 |q(t)|_Q |u^n(t) - u(t)|_\alpha \|\varphi\| \\
&\quad + \|A_2 u^n(t) - A_2 u(t)\|_* \|\varphi\| + \alpha_0 \|u^n(t) - u(t)\| \|\varphi\| \\
&\leq K\beta_1 |q^n(t) - q(t)|_Q |u^n(t)|_\alpha \|\varphi\| \\
&\quad + K\gamma_1 |q(t)|_Q |u^n(t) - u(t)|_\alpha \|\varphi\| \\
&\quad + \gamma_2 |u^n(t) - u(t)|_\alpha \|\varphi\| + \alpha_0 K_V |u^n(t) - u(t)|_\alpha \|\varphi\|.
\end{aligned}
$$

Recalling (2.54), (4.38), (4.39), and (4.40), we obtain $\lim_{n\to\infty} \|f^n(t) - f(t)\|_* = 0$, uniformly in t for $t \in [0, T]$, and the theorem is proved. □

5. Examples and numerical results. In this section we describe and discuss four different examples which illustrate the application of the general theory developed in the previous sections. We consider the example involving a first-order hyperbolic plant and a diffusion equation reference model discussed in section 2.1, an example involving the identification of a spatially varying thermal conductivity in a heat equation plant, an example involving the identification of a damped wave equation (a parabolic regularized hyperbolic system; see [27]), and an example in which we identify the nonlinearity in a quasi-linear heat equation. All of the computations to be described below were carried out via codes written in Fortran and run on either a Sun SPARCstation 10 in the Department of Mathematics at the University of Southern California or an IBM RISCSystem 6000 at the Center for Research in Scientific Computation at North Carolina State University. The closed-loop system (2.26)–(2.29) was discretized using a spline-based Galerkin scheme. The resulting finite-dimensional system of nonlinear ordinary differential equations was integrated using either the stiff ordinary differential equation solver from the NAG Library, routine D02NBF (at USC), or a fourth-order Runge–Kutta scheme (at NCSU). All required integrals were computed numerically via a composite two-point Gauss–Legendre quadrature rule.

We note that the plants and reference models in *all* of the examples to follow satisfy our basic assumptions (A1)–(A5). In Example 5.1, assumptions (A6)–(A9), which are required by the analytic semigroup approach to well-posedness presented in section 2.1, are satisfied as well. In Examples 5.2 and 5.3, assumptions (A10) and (A11), which are required by the implicit function theorem based well-posedness theory in section 2.2, are also satisfied. As we pointed out earlier, the theory in section 2.2 can probably be extended so as to be applicable to the nonlinear plant treated in Example 5.4 as well. The basic assumptions required for the finite-dimensional approximation and convergence theory discussed in section 4, assumptions (A12)–(A14), are satisfied in all four of the examples below.

Regarding the regularity assumptions on the input reference signal g (e.g., $g \in L_2(0, \infty; V^*)$, etc.), the symmetry of the reference model operator A_0, and the persistence of excitation condition, all of which form the core of the hypotheses for our stability and convergence results in sections 2 and 3, we do not explicitly address these hypotheses in the context of the particular examples presented below. There are a number of reasons for this. First, our computational results are necessarily on a finite time interval, while our stability and convergence results are asymptotic results as $t \to \infty$. Consequently, the reconciling of our theoretical and numerical results is, in some sense, ill posed. More precisely, if $g \in L_2(0, T; V^*)$ for some $T \in (0, \infty)$, then g can be extended to a function in $L_2(0, \infty; V^*)$ by simply defining it to be identically zero on (T, ∞). Second, with regard to the persistence of excitation condition, for most systems this condition is difficult, if not impossible, to verify. However, indication of the presence or absence of persistence of excitation and the degree thereof is immediately evident from the parameter estimator trajectories (see [9]). These observations can then be used to tune the scheme, either manually or autonomously, to obtain parameter convergence and optimal performance. In fact, we did just that in carrying out the computations in the examples below. And, finally, and most significantly, we point out that as is the case with virtually all theoretical and analytical studies of the type we have presented here, the assumptions we make are the ones minimally required to allow us to establish our stability and convergence results. Whether or

not the examples we present satisfy these assumptions is essentially irrelevant. Indeed, the fact that in these cases the scheme performs satisfactorily illustrates that our results are *robust* and that our approach appears to be applicable to a far broader class of problems than those that satisfy the hypotheses of our theorems.

Example 5.1. We consider Example 2.8 discussed in section 2.1. In particular, we use this example to illustrate the approximation results obtained in section 4. Recall from section 2.1 that $H = L_2(0,1)$, $V = H_L^1(0,1) = \{\varphi \in H^1(0,1) : \varphi(0) = 0\}$, $\hat{V}^* = V^*$, and $Q = \mathbf{R}^1$. The inner product on Q was chosen to be $\langle q, p\rangle_Q = \omega\, q \cdot p$ for $q, p \in \mathbf{R}^1$. The weighting factor $\omega > 0$ serves as an *adaptive gain* which can be used to *tune* the estimator. The plant is given by

$$\frac{\partial u}{\partial t}(t,x) + \overline{q}\frac{\partial u}{\partial x}(t,x) = f(t,x), \qquad 0 < x < 1,\ t > 0,$$

$$u(t,0) = 0, \qquad t > 0,$$

$$u(0,x) = u_0(x), \qquad 0 \le x \le 1,$$

where $\overline{q} > 0$, $u_0 \in L_2(0,1)$, and $t \to f(t,\cdot) \in L_2(0,T;H)$ for each $T > 0$. The reference model is given by

$$\frac{\partial v}{\partial t}(t,x) - a_0\frac{\partial^2 v}{\partial x^2}(t,x) = g(t,x), \qquad 0 < x < 1,\ t > 0,$$

$$v(t,0) = 0, \qquad \text{and} \qquad \frac{\partial v}{\partial x}(t,1) = 0, \qquad t > 0,$$

$$v(0,x) = v_0(x), \qquad 0 \le x \le 1,$$

where $a_0 > 0$, $v_0 \in L_2(0,1)$, and $t \to g(t,\cdot) \in L_2(0,T;V^*)$ for each $T > 0$.

We approximate using linear B-splines. For $n = 1, 2, \ldots$, let $\{\varphi_j^n\}_{j=0}^n$ be the standard linear B-splines on the interval $[0,1]$ defined with respect to the uniform mesh $\{0, \frac{1}{n}, \frac{2}{n}, \ldots, 1\}$. That is, for $i = 0, 1, 2, \ldots, n$,

$$\varphi_i^n(x) = \begin{cases} 1 - |nx - i|, & x \in [\frac{i-1}{n}, \frac{i+1}{n}], \\ 0 & \text{elsewhere on } [0,1]. \end{cases} \tag{5.1}$$

Set $H^n = \operatorname{span}\{\varphi_j^n\}_{j=1}^n \subset V$. Since Q is finite dimensional, we simply set $Q^n = Q = \mathbf{R}^1$, $n = 1, 2, \ldots$. For each $n = 1, 2, \ldots$, let P^n denote the orthogonal projection of H onto H^n and setting $\pi_n = P^n$, standard approximation results for spline functions (see [36]) can be used to establish that assumption (A13) is satisfied. Thus the conclusions of Theorem 4.7 hold.

We set $\overline{q} = 1.0$, $a_0 = .1$, $\omega = .02$, and $q_0 = 0.0$. We also set $u_0(x) = 0.0$, $v_0(x) = \sin(\frac{\pi}{2}x)$, $0 \le x \le 1$, and $g(t,x) = 5\sin\left(\frac{\pi}{2}t\right)\chi_{[.215,.315]}(x)$, $0 < x < 1$, $t > 0$. We simulated the plant using a 64 linear spline based Galerkin scheme and approximated the reference model in H^n with $n = 8, 16$, and 32. In Figure 5.1a we have plotted the parameter estimator trajectories, $q^n(t)$, along with the trajectory of the *infinite-dimensional* estimator (i.e., $n = 64$), $q(t)$, for $0 \le t \le 100$. In Figure 5.1b we plot the L_2 norms of the corresponding state tracking errors, $|e^n(t)| = |u(t) - v^n(t)|$, for $0 \le t \le 100$. It is clear from the figures that the scheme performed well for n

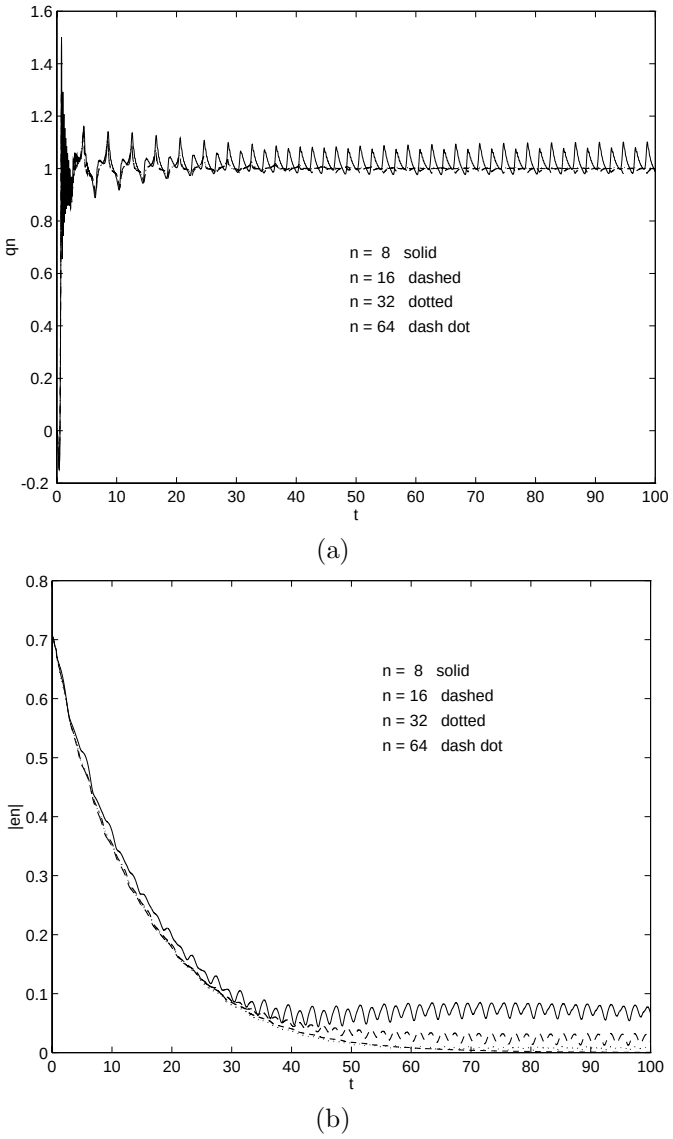


FIG. 5.1. *Results for Example* 5.1 *for various choices of* n*:* (a) *the parameter estimates versus time;* (b) *the tracking error versus time.*

as small as 8. We note that the scheme even performed reasonably well for $n = 4$, although we have not plotted these results here.

Example 5.2. In this example we consider the control of the one-dimensional heat or diffusion equation

$$\frac{\partial u}{\partial t}(t,x) = \frac{\partial}{\partial x}\left\{\overline{q}(x)\frac{\partial u}{\partial x}(t,x)\right\} + f(t,x), \quad t > 0, \quad 0 < x < 1,$$

together with the Dirichlet boundary conditions $u(t,0) = 0 = u(t,1)$, $t > 0$. We take

the reference model to be given by

$$\frac{\partial v}{\partial t}(t,x) - a_0 \frac{\partial^2 v}{\partial x^2}(t,x) = g(t,x) \qquad 0 < x < 1,\ t > 0, \tag{5.2}$$

$$v(t,0) = 0, \qquad \text{and} \qquad v(t,1) = 0, \qquad t > 0, \tag{5.3}$$

$$v(0,x) = v_0(x), \qquad 0 \le x \le 1, \tag{5.4}$$

where $a_0 > 0$, $v_0 \in L_2(0,1)$, and $t \to g(t,\cdot) \in L_2(0,T;V^*)$ for each $T > 0$.

In this case we have $H = L_2(0,1)$ and $V = H_0^1(0,1)$, each endowed with its usual inner product and corresponding induced norm. We set $\hat{V}^* = V^*$, and we let $Q = H^1(0,1)$ and take it to be endowed with the weighted inner product

$$\langle q,p\rangle_Q = \omega_1 \int_0^1 q(x)p(x)dx + \omega_2 \int_0^1 Dq(x)Dp(x)dx, \qquad p,q \in H^1(0,1),$$

where the weights ω_1 and ω_2, assumed to be positive, serve as adaptive gains or tuning parameters. For $q \in Q$, the operator $A(q) = A_1(q) \in \mathcal{L}(V,V^*)$ is given by

$$\langle A(q)\varphi,\psi\rangle = \langle A_1(q)\varphi,\psi\rangle = \int_0^1 q(x)D\varphi(x)D\psi(x)dx, \qquad \varphi,\psi \in H^1(0,1).$$

The operator $A_0 \in \mathcal{L}(V,V^*)$ is given by

$$\langle A_0\varphi,\psi\rangle = a_0 \int_0^1 D\varphi(x)D\psi(x)dx, \qquad \varphi,\psi \in H_0^1(0,1). \tag{5.5}$$

It is easily verified that assumptions (A1)–(A5) are satisfied and that the theory in section 2.2 applies.

To simulate the closed-loop system, we discretized equations (2.26)–(2.29) using a linear spline based Galerkin scheme. We approximated the plant and reference model state space H by $H^n = \text{span}\,\{\varphi_j^n\}_{j=1}^{n-1}$, where the linear B-splines φ_j^n are given by (5.1). We also used linear B-splines to discretize the parameter space Q. We set $Q^m = span\{\varphi_j^m\}_{j=0}^m$, where the linear spline basis $\{\varphi_j^m\}_{j=0}^m$ is again given by (5.1) with n replaced by m. Note that $\dim H^n = n-1$ and $\dim Q^m = m+1$. Consequently, the dimension of the approximating estimator is $n-1+m+1 = n+m$.

We set $a_0 = 0.1$, $\omega_1 = 0.1$, $\omega_2 = 0.0001$, $\overline{q}(x) = \frac{1}{10}\left\{1 - \frac{1}{2}\sin\left(2\pi\{x - \frac{1}{4}\}\right)\right\}$, $0 < x < 1$, $q_0(x) = 0.1$, $u_0(x) = 0.3(0.5 - |0.5 - x|)$, and $v_0(x) = -0.1\sin(\pi x)$ for $0 \le x \le 1$. We chose the input reference signal g to be given by

$$g(t,x) = .1\left\{\sin\left(\frac{\pi}{24}t\right) + \cos\left(\frac{\pi}{50}t\right) + \frac{1}{2}\cos\left(\frac{\pi}{30}t\right)\right\}\chi_{[.4,.6]}(x), \quad 0 < x < 1,\ t > 0.$$

The results of our numerical study with $n = 24$ and $m = 16$ are displayed in Figures 5.2a and 5.2b. In Figure 5.2a we have plotted q_0, $\overline{q}$, and the estimate for $\overline{q}$, $q(t)$, at $t = 25$. In Figure 5.2b we have plotted the L_2 norm of the state tracking error $|e(t)| = |u(t) - v(t)|$ for $0 \le t \le 50$. It is clear that the control objective has been met and that an excellent estimate for $\overline{q}$ has been obtained.

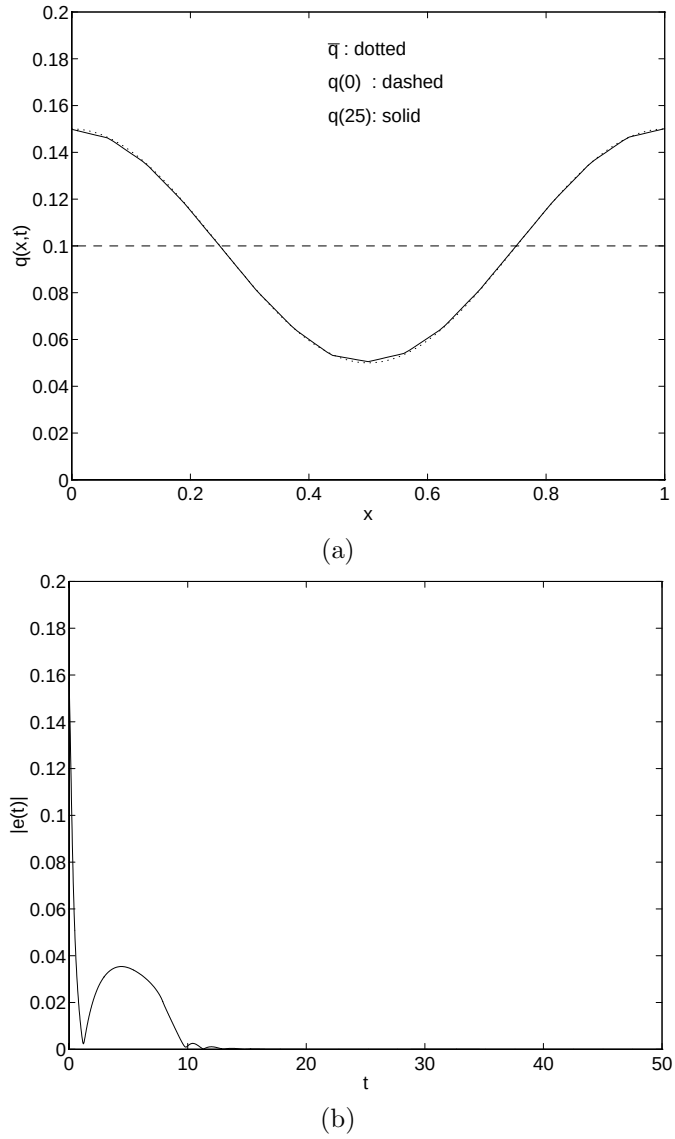


FIG. 5.2. *Results for Example* 5.2*:* (a) *the parameter estimate versus time;* (b) *the tracking error versus time.*

Example 5.3. In this example we consider the control of the one-dimensional wave equation with Kelvin–Voigt viscoelastic damping given by

$$\frac{\partial^2 u_1}{\partial t^2}(t,x) - \overline{q}_2 \frac{\partial^2}{\partial x^2}\frac{\partial u_1}{\partial t}(t,x) - \overline{q}_1 \frac{\partial^2 u_1}{\partial x^2} = f_1(t,x),\ 0 < x < 1,\ t > 0, \tag{5.6}$$

with the Dirichlet (fixed endpoint) boundary conditions

$$u_1(t,0) = 0 = u_1(t,1), \qquad t > 0, \tag{5.7}$$

and the initial conditions

$$u_1(0,x) = u_{01}(x), \qquad \text{and} \qquad \frac{\partial u_1}{\partial t}(0,x) = u_{02}(x), \qquad 0 \le x \le 1. \tag{5.8}$$

We take the reference model to be given by

$$\frac{\partial^2 v_1}{\partial t^2}(t,x) - b_0 \frac{\partial^2}{\partial x^2}\frac{\partial v_1}{\partial t}(t,x) - a_0 \frac{\partial^2 v_1}{\partial x^2} = g_1(t,x),\ 0 < x < 1,\ t > 0, \tag{5.9}$$

$$v_1(t,0) = 0 = v_1(t,1), \qquad t > 0, \tag{5.10}$$

$$v_1(0,x) = v_{01}(x), \qquad \text{and} \qquad \frac{\partial v_1}{\partial t}(0,x) = v_{02}(x), \qquad 0 \le x \le 1, \tag{5.11}$$

where $a_0, b_0 > 0$. To apply our theory in this case requires rewriting the plant and reference model as equivalent first-order systems. Let $H_1 = L_2(0,1)$ be endowed with the standard inner product and corresponding induced norm denoted by $\langle \cdot, \cdot \rangle_1$ and $|\cdot|_1$, respectively. Let $V_1 = H_0^1(0,1)$ be endowed with the inner product (and corresponding induced norm) given by $[\varphi, \psi]_1 = \langle D\varphi, D\psi \rangle_1$ for $\varphi, \psi \in V_1$. Let $H = V_1 \times H_1$ and $V = V_1 \times V_1$. We endow V with the usual product inner product, but we endow H with the inner product given by

$$\langle \varphi, \psi \rangle = \gamma \{a_0[\varphi_1, \psi_1]_1 + \langle \varphi_2, \psi_2 \rangle_1\} + \langle \varphi_1, \psi_2 \rangle_1 + \langle \varphi_2, \psi_1 \rangle_1 + b_0[\varphi_1, \psi_1]_1,$$

for $\varphi = (\varphi_1, \varphi_2), \psi = (\psi_1, \psi_2) \in H$, where $\gamma > 0$. It is not difficult to argue that if $\gamma > \max\{1, a_0^{-1}, b_0^{-1}\}$, then the norm induced by this inner product is equivalent to the standard norm on H. The inner product on H is chosen in this way so that assumption (A4) will be satisfied by the operator A_0 to be defined below. We note that this choice of an inner product affects the form of the estimator equation (2.24). Thus, in practice, γ serves as an additional tuning parameter or adaptive gain.

We let $Q = \mathbf{R}^2$ with the weighted inner product given by $\langle q, p \rangle = q^T \Omega p$, $q, p \in \mathbf{R}^2$, where Ω is the 2×2 matrix given by $\Omega = \operatorname{diag}(\omega_1, \omega_2)$, with $\omega_1, \omega_2 > 0$.

For $q = (q_1, q_2)^T \in Q$, we define the operator $A(q) \in \mathcal{L}(V, V)$ by $A(q) = A_1(q) + A_2$, where $\langle A_1(q)\varphi, \psi \rangle = q_2 \langle D\varphi_2, D\psi_2 \rangle + q_1 \langle D\varphi_1, D\psi_2 \rangle$, $\varphi = (\varphi_1, \varphi_2)$, $\psi = (\psi_1, \psi_2) \in V$, and $\langle A_2\varphi, \psi \rangle = -a_0 \langle D\varphi_2, D\psi_1 \rangle$, $\varphi = (\varphi_1, \varphi_2)$, $\psi = (\psi_1, \psi_2) \in V$. We take the operator $A_0 \in \mathcal{L}(V, V^*)$ to be given by $A_0 = A(q^*)$, where $q^* = (a_0, b_0) \in Q$. We set $f = (0, f_1)$, $g = (0, g_1)$, $u_0 = (u_{01}, u_{02})$, and $v = (v_{01}, v_{02})$. Thus we have rewritten the plant (5.6)–(5.8) in the form (2.6), (2.7) and the reference model (5.9)–(5.11) in the form (2.8), (2.9) with $u = (u_1, D_t u_1)$ and $v = (v_1, D_t v_1)$. It can be verified that assumptions (A1)–(A5) are satisfied with $\hat{V}^* = \{(0, \varphi) : \varphi \in H^{-1}(0,1)\} \subset V^* = H_0^1(0,1) \times H^{-1}(0,1)$.

To simulate the closed-loop system, we again approximate using the linear spline basis given in (5.1) and a Galerkin scheme. We set $H_1^n = \operatorname{span}\{\varphi_j^n\}_{j=1}^{n-1}$ and set $H^n = H_1^n \times H_1^n$. We took $\overline{q} = (0.0308, 0.01)$, $q^* = (a_0, b_0) = (0.0056, 0.0028)$, and $q_0 = (0.02, 0.005)$. We set $u_{01}(x) = 0.01 \sin(\pi x)$, $u_{02}(x) = 0.001 \sin(4\pi x)$, $0 \le x \le 1$, $v_{01}(x) = 0$, $v_{02}(x) = 0$, $0 \le x \le 1$, and

$$g_1(t,x) = \{4 \sin(4\pi t) + \cos(\pi t) + 2\}\, \chi_{[.215,.315]}(x),\ t > 0,\ 0 < x < 1.$$

We chose the adaptive gains to be $\omega_1 = \omega_2 = 1600/3$ and $\gamma = 100 + 1/b_0 = 457.15$. We then simulated the closed-loop system with $n = 16$. In Figure 5.3 we have plotted

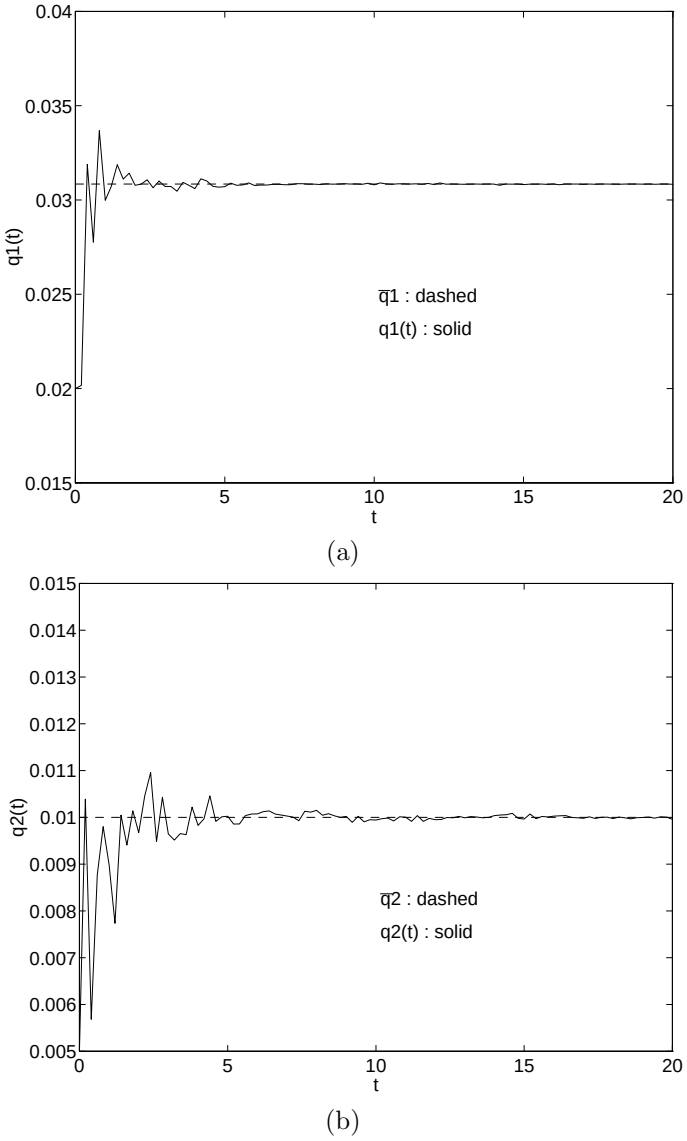


FIG. 5.3. *Results for Example* 5.3: (a) *the parameter estimate* $q_1(t)$ *versus time;* (b) *the parameter estimate* $q_2(t)$ *versus time.*

the estimate for $\overline{q}_1$ $q_1(t)$, and the estimate for $\overline{q}_2$, $q_2(t)$, for $t \in [0, 20]$. In Figure 5.4 we have plotted the V_1 norm of the displacement tracking error $\|e(t)\|_1 = \|u_1(t) - v_1(t)\|_1$ and the H_1 norm of the velocity tracking error $|D_t e(t)|_1 = |D_t u_1(t) - D_t v_1(t)|_1$ for $t \in [0, 100]$.

Example 5.4. In this example we consider the control of the one-dimensional nonlinear (strictly speaking, quasi-linear) heat equation

$$\frac{\partial u}{\partial t}(t,x) - \frac{\partial}{\partial x}\left\{\overline{q}\left(\min\left\{M, \left|\frac{\partial u}{\partial x}(t,x)\right|\right\}\right)\frac{\partial u}{\partial x}(t,x)\right\} = f(t,x), \tag{5.12}$$

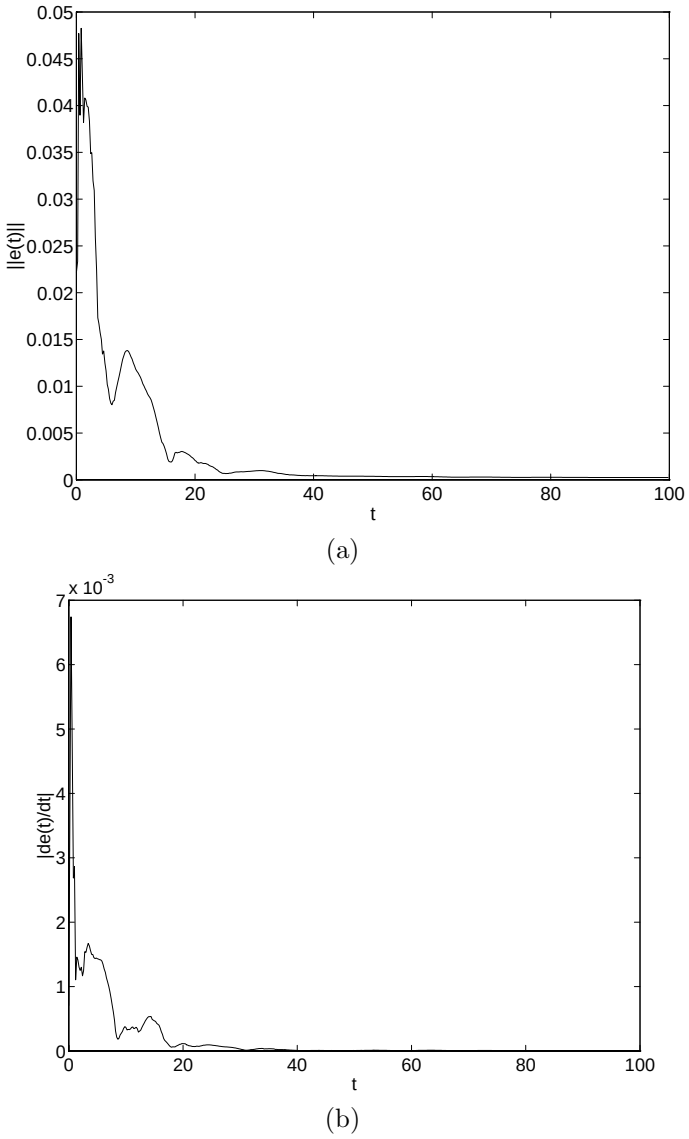


FIG. 5.4. *Results for Example* 5.3: (a) *the V_1-norm of the displacement tracking error;* (b) *the H_1-norm of the velocity tracking error.*

for $0 < x < 1$, $t > 0$, together with the Dirichlet boundary conditions

$$u(t,0) = 0 \qquad \text{and} \qquad u(t,1) = 0, \qquad t > 0, \tag{5.13}$$

and the initial conditions

$$u(0,x) = u_0(x), \qquad 0 \le x \le 1. \tag{5.14}$$

We assume that $M \in [0,\infty)$, $u_0 \in L_2(0,1)$, and $f(t,\cdot) \in L_2(0,1)$ for $t \ge 0$. We assume that the nonlinearity $\overline{q}$ is unknown and is to be identified as the system

(5.12)–(5.13) is being adaptively controlled. Once again, we take the reference model to be given by (5.2)–(5.4).

We let $H = L_2(0,1)$ be endowed with the standard inner product, we let $V = H_0^1(0,1)$ be endowed with the usual norm, and we define the Hilbert space Q as follows. Let $\hat{Q} = H^1(\mathbf{R}^+)$ and define the inner product, $\langle \cdot, \cdot \rangle_Q$, on $\hat{Q}$ by

$$\langle q, p \rangle_Q = \int_0^\infty \omega_0(\theta) q(\theta) p(\theta) d\theta + \int_0^\infty \omega_1(\theta) Dq(\theta) Dp(\theta) d\theta, \quad q, p \in \hat{Q}, \tag{5.15}$$

where $\omega_0, \omega_1 \in L_1(\mathbf{R}^+)$ are positive weighting functions. Let $|\cdot|_Q$ denote the norm induced by the inner product given in (5.15), and define the Hilbert space Q to be the completion of the inner product space $\{\hat{Q}, \langle \cdot, \cdot \rangle_Q, |\cdot|_Q\}$. For $q \in Q$, the operator $A(q) : V \to V^*$ is given by $A(q) = A_1(q)$, where $A_1(q) : V \to V^*$ is defined by

$$\langle A_1(q)\varphi, \psi \rangle = \int_0^1 q \left(\min \{M, |D\varphi(x)|\} \right) D\varphi(x) D\psi(x) dx, \qquad \varphi, \psi \in V.$$

The operator $A_0 \in \mathcal{L}(V, V^*)$ is once again given by (5.5). We set $\hat{V}^* = V^* = H^{-1}(0,1)$. It is not difficult to verify that assumptions (A1)–(A5) are satisfied.

To simulate the closed-loop system, we again approximate the plant and reference model state space H and the parameter space Q using linear B-spline functions. We approximate H by $H^n = span\{\varphi_j^n\}_{j=1}^{n-1}$, where for each $n = 2, 3, \ldots$ and $j = 1, 2, \ldots n-1$, φ_j^n is given by (5.1). For each $m = 1, 2, \ldots$ and each $r > 0$, let $\{\hat{\psi}_j^{m,r}\}_{j=0}^m$ be the standard linear B-splines on the interval $[0, r]$ defined with respect to the uniform mesh $\{0, \frac{r}{m}, \frac{2r}{m}, \ldots, r\}$. We approximate Q by $Q^{m,r} = span\{\psi_j^{m,r}\}_{j=0}^m$, where

$$\psi_j^{m,r} = \begin{cases} \hat{\psi}_j^{m,r}, & j = 0, 1, 2, \ldots, m-1, \\ \hat{\psi}_m^{m,r} + \chi_{[r,\infty)}, & j = m, \end{cases}$$

with χ_J denoting the characteristic function for the interval J. In the simulations to be described below, only q is discretized. The *true value* of $\overline{q}$, $\overline{q}(\theta) = 0.9(1 - \frac{1}{2}e^{-\frac{1}{2}\theta^2})$, $\theta \geq 0$, is used. We chose g to be given by

$$\begin{aligned} g(t,x) \quad = \quad & 1 \times 10^{-4} \{ \sin(100\pi t) + \sin(250\pi t) + \sin(450\pi t) + \sin(550\pi t) \\ & + \sin(650\pi t) + \sin(850\pi t) + \cos(150\pi t) + \cos(350\pi t) \\ & + \cos(500\pi t) + \cos(700\pi t) \} \, \chi_{[0.6, 0.8]}(x), \quad 0 < x < 1, \ t > 0, \end{aligned}$$

and set $u_0(x) = 5 \times 10^{-5}$ and $v_0(x) = -0.1 \, (0.5 - |0.5 - x|)$, $0 < x < 1$. We set $a_0 = .1$, $r = 3.5$, $M = 10.0$,

$$\omega_0(\theta) = \omega_1(\theta) = \begin{cases} 1, & 0 \leq \theta < r, \\ \frac{1}{2} e^{-20\theta}, & r < \theta < \infty, \end{cases}$$

and $q_0(\theta) = 1$, $0 < \theta < \infty$. We simulated the closed-loop system over the time interval $[0, 20]$ using $n = 32$ and $m = 24$. In Figure 5.5a we have plotted our final (i.e., at time $t = 20$) estimate for $\overline{q}$, and in Figure 5.5b we have plotted the H-norm of the tracking error, $|e(t)| = |u(t) - v(t)|$, for $t \in [0, 20]$. We note that convergence of the

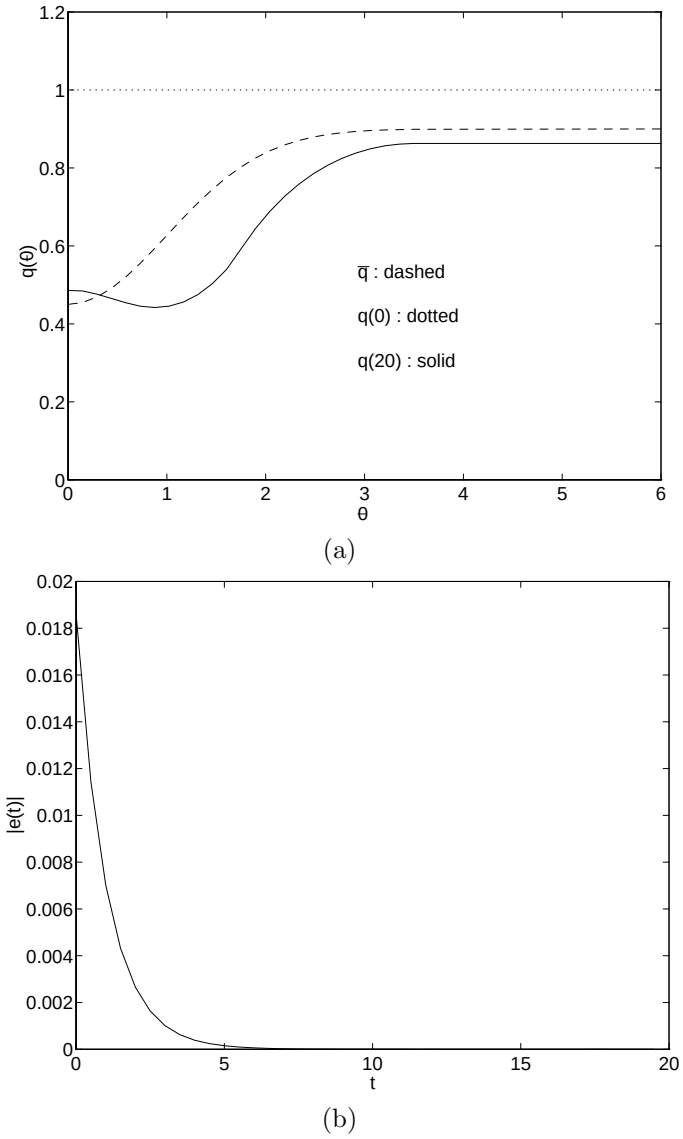


FIG. 5.5. *Results for Example* 5.4: (a) *the parameter estimate* $q(\theta)$ *versus* θ*;* (b) *the tracking error versus time.*

parameter estimates actually occurred at about $t = 5$. Our estimate for $\overline{q}$ in this example is not quite as good as the estimate obtained in Example 5.2. However, both the nonlinearity and the infinite domain of $\overline{q}$ (and therefore the need for an additional degree of approximation in the form of truncation) contribute to making this example a far more significant challenge for our scheme than the linear example discussed in Example 5.2.

Acknowledgments. The authors would like to gratefully acknowledge the referees for their careful reading of the manuscript and for their substantive comments and suggestions.

REFERENCES

[1] H. W. Alt, K. H. Hoffmann, and J. Sprekels, *A numerical procedure to solve certain identification problems*, Internat. Ser. Numer. Math, 68 (1984), pp. 11–43.

[2] K. J. Astrom and B. Wittenmark, *Adaptive Control*, Addison–Wesley, Reading, MA, 1989.

[3] H. T. Banks and K. Ito, *A unified framework for approximation and inverse problems for distributed parameter systems*, Control Theory Adv. Tech., 4 (1988), pp. 73–90.

[4] H. T. Banks and D. A. Rebnord, *Estimation of material parameters for grid structures*, J. Math. Systems, Estim. Control, 1 (1991), pp. 107–130.

[5] V. Barbu, *Nonlinear Semigroups and Differential Equations in Banach Spaces*, Noordhoff International Publishing, Leyden, The Netherlands, 1976.

[6] J. Baumeister and W. Scondo, *Asymptotic embedding methods for parameter estimation*, in Proceedings of the 26th IEEE Conference on Decision and Control, 1987, pp. 170–174.

[7] J. Baumeister, W. Scondo, M. A. Demetriou, and I. G. Rosen, *On-line parameter estimation for infinite dimensional dynamical systems*, SIAM J. Control Optim., 35 (1997), pp. 678–713.

[8] M. A. Demetriou and I. G. Rosen, *Adaptive identification of second order distributed parameter systems*, Inverse Problems, 10 (1994), pp. 261–294.

[9] M. A. Demetriou and I. G. Rosen, *On the persistence of excitation in the adaptive estimation of distributed parameter systems*, IEEE Trans. Automat. Control, 39 (1994), pp. 1117–1123.

[10] J. Dieudonné, *Foundations of Modern Analysis*, Academic Press, New York, 1960.

[11] T. E. Duncan, B. Maslowski, and B. Pasik-Duncan, *Adaptive boundary and point control of linear stochastic distributed parameter systems*, SIAM J. Control Optim., 32 (1994), pp. 648–672.

[12] T. E. Duncan and B. Pasik-Duncan, *Adaptive control of linear delay time systems*, Stochastics, 24 (1988), pp. 45–74.

[13] T. E. Duncan, B. Pasik-Duncan, and B. Goldys, *Adaptive control of linear stochastic evolution systems*, Stochastics Stochastics Rep., 36 (1991), pp. 71–90.

[14] G. C. Goodwin and K. S. Sin, *Adaptive Filtering Prediction and Control*, Prentice–Hall, Englewood Cliffs, NJ, 1984.

[15] J. K. Hale, *Ordinary Differential Equations*, Wiley-Interscience, John Wiley and Sons, New York, 1969.

[16] K. H. Hoffmann and J. Sprekels, *The method of asymptotic regularization and restricted parameter identification problems in variational inequalities*, in Free Boundary Problems: Application and Theory, IV, Proc. Intl. Colloq. Problèmes à Frontières Líbres: Applications el Théorie, Maubuisson, France, 1984, A. Bossavit, A. Damlamian, and M. Fremond, eds., Pitman Research Notes in Math., No. 121, Pitman Adv. Pub. Prog., Pitman, Boston, (1985), pp. 508–513.

[17] K. H. Hoffmann and J. Sprekels, *On the identification of coefficients of elliptic problems by asymptotic regularization*, Numer. Funct. Anal. Optim., 7 (1984-85), pp. 157–177.

[18] K. H. Hoffmann and J. Sprekels, *On the identification of parameters in general variational inequalities by asymptotic regularization*, SIAM J. Math. Anal., 17 (1986), pp. 1198–1217.

[19] K. S. Hong and J. Bentsman, *Direct adaptive control of parabolic systems: Algorithm synthesis, and convergence and stability analysis*, IEEE Trans. Automat. Control, 39 (1994), pp. 2018–2033.

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

[21] T. Kobayashi, *A digital adaptive control law for a parabolic distributed parameter system*, Systems Control Lett., 4 (1984), pp. 175–179.

[22] T. Kobayashi, *Adaptive control for infinite dimensional systems*, Internat. J. Systems Sci., 17 (1986), pp. 887–896.

[23] T. Kobayashi, *Global adaptive stabilization of infinite dimensional systems*, Systems Control Lett., 9 (1987), pp. 215–223.

[24] T. Kobayashi, *Model reference adaptive control for spectral systems*, Internat. J. Control, 46 (1987), pp. 1511–1523.

[25] T. Kobayashi, *Finite dimensional adaptive control for infinite dimensional systems*, Internat. J. Control, 48 (1988), pp. 289–302.

[26] S. G. Kreĭn, *Linear Differential Equations in Banach Space*, AMS, Providence, RI, 1971.

[27] J. L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1971.

[28] J. L. Lions and E. Magenes, *Problèmes aux Limites Non Homogènes et Applications*, Volume 1, Dunod, Paris, 1968.

[29] A. P. Morgan and K. S. Narendra, *On the stability of nonautonomous differential equations* $\dot{x} = [A + B(t)]x$, *with skew symmetric matrix* $B(t)$, SIAM J. Control Optim., 15 (1977), pp. 163–176.
[30] K. S. Narendra and A. M. Annaswamy, *Stable Adaptive Systems*, Prentice–Hall, Englewood Cliffs, NJ, 1989.
[31] K. S. Narendra and P. Kudva, *Stable adaptive schemes for system identification and control, parts* I *and* II, IEEE Trans. Systems, Man Cybernet., SMC-4 (1974), pp. 542–560.
[32] B. Pasik-Duncan, *On the consistency of a least squares identification procedure in linear evolution systems*, Stochastics Stochastics Rep., 39 (1992), pp. 83–94.
[33] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.
[34] V. M. Popov, *Hyperstability of Control Systems*, Springer-Verlag, Berlin, 1973.
[35] S. Sastry and M. Bodson, *Adaptive Control: Stability, Convergence and Robustness*, Prentice–Hall, Englewood Cliffs, NJ, 1989.
[36] M. H. Schultz, *Spline Analysis*, Prentice–Hall, Englewood Cliffs, NJ, 1973.
[37] W. Scondo, *Ein Modellabgleichsverfahren zur adaptiven Parameteridentifikation in Evolutionsgleichungen*, Ph.D. thesis, Johann Wolfgang Goethe-Universitat zu Frankfurt am Main, Frankfurt am Main, Germany, 1987.
[38] R. E. Showalter, *Hilbert Space Methods for Partial Differential Equations*, Pitman, London, 1977.
[39] H. Tanabe, *Equations of Evolution*, Pitman, London, 1979.
[40] J. T. Wen, *Direct Adaptive Control in Hilbert Space*, Ph.D. thesis, Electrical, Computer and Systems Engineering Department, Rensselaer Polytechnic Institute, Troy, NY, 1985.
[41] J. T. Wen and M. J. Balas, *Robust adaptive control in Hilbert space*, J. Math. Anal. Appl., 143 (1989), pp. 1–26.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 1, pp. 82–99, January 1998

004

AN APPROXIMATION THEORY OF SOLUTIONS TO OPERATOR RICCATI EQUATIONS FOR H^∞ CONTROL*

KAZUFUMI ITO† AND K. A. MORRIS‡

Abstract. As in the finite-dimensional case, the appropriate state feedback for the infinite-dimensional H^∞ disturbance-attenuation problem may be calculated by solving a Riccati equation. This operator Riccati equation can rarely be solved exactly. We approximate the original infinite-dimensional system by a sequence of finite-dimensional systems and consider the corresponding finite-dimensional disturbance-attenuation problems. We make the same assumptions required in approximations for the classical linear quadratic regulator problem and show that the sequence of solutions to the corresponding finite-dimensional Riccati equations converge strongly to the solution to the infinite-dimensional Riccati equation. Furthermore, the corresponding finite-dimensional feedback operators yield performance arbitrarily close to that obtained with the infinite-dimensional solution.

Key words. H^∞, approximations, partial differential equation, optimal control, infinite dimensional

AMS subject classifications. 49N10, 65P05, 93C20

PII. S0363012994274422

1. Introduction.

In this paper we discuss H^∞ control problems for the linear system in a Hilbert space X:

$$\frac{d}{dt}x(t) = Ax(t) + Bu(t) + Dv(t), \quad x(0) = x \in X, \tag{1.1}$$

where the linear closed operator A generates the C_0-semigroup $S(t)$ on X. Let W, U, and Y be separable Hilbert spaces. The signal $v(t) \in L^2(0,\infty;W)$ is a W-valued disturbance and $u(t) \in L^2(0,\infty;U)$ is the control input. We assume that the disturbance operator D and the input operator B are bounded, i.e., $D \in \mathcal{L}(W,X)$ and $B \in \mathcal{L}(U,X)$. Let $C \in \mathcal{L}(X,Y)$ be the reference output operator. For control cost $\epsilon > 0$, define the output

$$z(t) = \operatorname{col}\left(Cx(t), \sqrt{\epsilon}\, u(t)\right) \tag{1.2}$$

and the performance index

$$\rho(u,v;x) = |z|^2_{L^2(0,\infty;Y)} = \int_0^\infty |Cx(t)|^2 + \epsilon|u(t)|^2\, dt.$$

Let

$$\mathcal{U} = L^2(0,\infty;U), \quad \mathcal{W} = L^2(0,\infty;W).$$

This paper is concerned with the problem of constructing a stabilizing feedback control law $u(t) = -Kx(t)$ such that for each disturbance $w \in \mathcal{W}$, the closed-loop solution of

*Received by the editors September 19, 1994; accepted for publication (in revised form) October 8, 1996.
http://www.siam.org/journals/sicon/36-1/27442.html
†Center for Research in Scientific Computation, North Carolina State University, Raleigh, NC 27695-8205 (kito@crsc1.math.ncsu.edu).
‡Department of Applied Mathematics, University of Waterloo, Waterloo, ON N2L 3G1, Canada (kmorris@riccati.uwaterloo.ca).

(1.1) with $x(0) = 0$ satisfies

$$\rho(-Kx(t), v; 0) \leq (\gamma^2 - \delta)\, |v|^2_{L^2(0,\infty;W)} \tag{1.3}$$

for given attenuation bound $\gamma > 0$ and some $\delta > 0$. The problem described above is equivalent to so-called H^∞-disturbance attenuation: for given $\gamma > 0$ construct an exponentially stabilizing linear feedback $u(t) = -Kx(t)$ such that the attenuation bound

$$\sup_{\omega \in \mathbf{R}} \left| \begin{pmatrix} C \\ \sqrt{\epsilon} K \end{pmatrix} (i\omega I - (A - BK))^{-1} D \right| < \gamma \tag{1.4}$$

is satisfied. Such problems arise in a variety of contexts; robust stabilization is one of the most important.

DEFINITION 1.1. *If there is $\delta > 0$ such that for each $v \in \mathcal{W}$ there exists a control $u \in \mathcal{U}$ with*

$$\rho(u, v; 0) \leq (\gamma^2 - \delta)\, |v|^2_{\mathcal{W}},$$

the problem is said to be stabilizable with attenuation γ.

As in the finite-dimensional case, the H^∞ disturbance-attenuation problem is solvable if and only if the problem is stabilizable with attenuation γ (Theorem 2.2). Furthermore, in this case an appropriate state feedback may be calculated by solving an operator Riccati equation. Unfortunately, this Riccati equation can rarely be solved exactly. In this paper we approximate the original system (1.1)–(1.2) by a sequence of finite-dimensional systems and consider the corresponding finite-dimensional disturbance-attenuation problems.

The classical linear quadratic regulator (LQR) problem may be regarded as a limiting case of the H^∞ disturbance-attenuation problem, with the required disturbance attenuation $\gamma \to \infty$. The approximation theory for the linear quadratic case is fairly complete (e.g., see [GI, BK, IT1, IT2]). We make the same assumptions required in the linear quadratic case and show that a sequence of solutions to finite-dimensional Riccati equations converges strongly to the solution to the infinite-dimensional Riccati equation required to solve the H^∞ disturbance-attenuation problem. Furthermore, and more importantly, the corresponding finite-dimensional feedback control operators yield performance arbitrarily close to that obtained with the infinite-dimensional solution. A key step of the proof of these results is the game-theoretic representation ([KE, (2.10) in the proof of Theorem 2.2]) of the solution to the H^∞ Riccati equation in terms of the closed-loop solution to the standard LQR problem.

The notation that is used in this paper is standard as in [PA]. Background on linear semigroup theory may also be found in [PA]. An outline of the paper is as follows. In section 2 the solution to the H^∞ disturbance-attenuation problem in terms of the operator Riccati equation (2.1) is described, and an approximation theory of solutions to (2.1) is developed. Our theoretical results are demonstrated numerically in section 3 using a Euler–Bernoulli beam example.

2. Approximation theory. In this section we develop an approximation theory for the H^∞ disturbance-attenuation problem.

DEFINITION 2.1. (1) *The pair (A, B) is* stabilizable *if there exists a bounded linear operator $F : X \to U$ such that $A - BF$ generates an exponentially stable semigroup on X.*

(2) *The pair* (A, C) *is* detectable *if there exists a bounded linear operator* $G : Y \to X$ *such that* $A - GC$ *generates an exponentially stable semigroup on* X.

(3) *The state feedback* $K \in \mathcal{L}(X, U)$ *is* γ-admissible *if it is exponentially stabilizing and the linear feedback* $u(t) = -Kx(t)$ *is such that the attenuation bound* (1.3) *is achieved.*

A key result in the finite-dimensional and the infinite-dimensional theory is that if the problem is stabilizable with attenuation γ, then it is stabilizable by state feedback.

THEOREM 2.2 (see, e.g., [KE, Thm. 4.4]). *Assume that* (A, B) *is stabilizable and* (A, C) *is detectable. For* $\gamma > 0$ *the following are equivalent:*

- *there exists a* γ*-admissible state feedback;*
- *the system is stabilizable with disturbance attenuation* γ*;*
- *there exists a nonnegative, self-adjoint operator* Σ *on* X *satisfying the Riccati equation*

$$\left(A^*\Sigma + \Sigma A - \frac{1}{\epsilon}\Sigma BB^*\Sigma + \frac{1}{\gamma^2}\Sigma DD^*\Sigma + C^*C\right)x = 0 \tag{2.1}$$

for all $x \in dom(A)$*, and* $A - \frac{1}{\epsilon}BB^*\Sigma + \frac{1}{\gamma^2}DD^*\Sigma$ *generates an exponentially stable semigroup on* X.

Moreover, in this case a γ*-admissible state feedback is given by* $\hat{K} = \frac{1}{\epsilon}B^*\Sigma$.

An approximation theory for solutions to (2.1) which numerically approximate the feedback operator $\hat{K} = B^*\Sigma/\epsilon$ is developed below. Let X^N be a finite-dimensional subspace of X, and let P^N be the orthogonal projection of X onto X^N. The space X^N is equipped with the induced norm from X. Consider a sequence of operators $A^N \in \mathcal{L}(X^N, X^N)$, $B^N = P^N B \in \mathcal{L}(U, X^N)$, $D^N = P^N D$, and $C^N =$ the restriction of C onto X^N. The operator A^N can be extended to all of X by $A^N P^N x$.

Approximation Assumptions.

(A1) For each $x \in X$ we have

$$\text{(i)} \quad e^{A^N t}P^N x \to S(t)x,$$

$$\text{(ii)} \quad (e^{A^N t})^* P^N x \to S^*(t)x,$$

uniformly in t on bounded intervals.

(A2) (i) The family of pairs (A^N, B^N) is uniformly exponentially stabilizable; i.e., there exists a uniformly bounded sequence of operators $K^N \in \mathcal{L}(X^N, U)$ such that

$$\left|e^{(A^N - B^N K^N)t}P^N x\right|_X \le M_1\, e^{-\omega_1 t}\,|x|_X \tag{2.2}$$

for some positive constants $M_1 \ge 1$ and ω_1.

(ii) The family of pairs (A^N, C^N) is uniformly exponentially detectable; i.e., there exists a uniformly bounded sequence of operators $G^N \in \mathcal{L}(Y, X^N)$ such that

$$\left|e^{(A^N - G^N C^N)t}P^N x\right|_X \le M_2\, e^{-\omega_2 t}, \quad t \ge 0, \tag{2.3}$$

for some positive constants $M_2 \ge 1$ and ω_2.

(A3) The input operator B and disturbance operator D are compact.

Remarks. (1) Note that (A1) implies that $P^N x \to x$ for $x \in X$.

(2) Assumption (A1) and the uniform boundedess theorem imply the boundedness of $|e^{A^N t}P^N|_{\mathcal{L}(X,X)}$ uniformly in $t \in [0, 1]$ and N. Then the standard semigroup

theorem, e.g., [PA, Chapter 1, Theorem 2.2], implies that $|e^{A^N t}P^N x|_X \le M_0\, e^{\omega_0 t}|x|_X$ for some $M_0 \ge 1$ and $\omega_0 \in R$.

(3) For an important equivalent statement of (A1)(i) we note the Trotter–Kato theorem.

THEOREM 2.3 (Trotter–Kato Theorem; see, e.g., [PA, Chapter 3, Theorem 4.2]). *Assume the stability of approximation*

$$|e^{A^N t}P^N x|_X \le M_0\, e^{\omega_0 t}|x|_X \quad \textit{for some } M_0 \ge 1 \textit{ and } \omega_0 \in R.$$

Then the convergence (A1)(i)

$$e^{A^N t}P^N x \to S(t)x \quad \textit{for every } x \in X \textit{ and uniformly on bounded } t \textit{ intervals}$$

is equivalent to the consistency: for some $\lambda \in \rho(A)$

$$\left|(\lambda I - A^N)^{-1}P^N x - (\lambda I - A)^{-1}x\right|_X \to 0$$

as $N \to \infty$ *for all* $x \in X$.

(4) The convergence (A1)(ii) of the adjoint semigroup sequence $(e^{A^N t})^*$ is required for the strong convergence of the approximating feedback gain operators. A counterexample may be found in [BIP].

(5) In (A2)(i) if we let $K^N = KP^N$, then condition (2.2) becomes the preservation of exponential stability under approximation of the semigroup $T(t)$ generated by $A - BK$.

(6) Assumption (A3) is equivalent to

$$\lim |P^N D - D| = 0, \quad \lim |P^N B - B| = 0 \quad \text{as } N \to \infty$$

since X^N is finite dimensional.

Assumptions (A1)–(A2) are identical to those required to show that the solutions to the Riccati equations arising in the approximation theory for linear quadratic problem converge, e.g., [IT1]. Assumption (A3) is not required in the standard LQR problem for the existence of solutions to a family of approximating finite-dimensional Riccati equations. However, this assumption is required to ensure continuity of performance measure and to guarantee that the approximating controllers stabilize the infinite-dimensional system [IT2, MO1, MO2].

Before presenting the approximation result and its proof we state a technical lemma which plays an important role in the proof.

LEMMA 2.4 (Datko lemma; see, e.g., [SA, Theorem 6.2]). *Let* $S(t),\ t \ge 0$ *be a linear* C_0*-semigroup on a Banach space* X *satisfying the exponential bound*

$$|S(t)| \le Me^{\omega t}$$

for some constants $M \ge 1,\ \omega \ge 0$. *Moreover, let* $1 \le p < \infty$, *and suppose that there exists a constant* $c > 0$ *such that*

$$\int_0^\infty |S(t)x|_X^p\, dt \le c^p\, |x|_X^p, \quad x \in X.$$

Then, for every

$$\alpha > -\frac{1}{pc^p M^p},$$

there exists a $\gamma = \gamma(\alpha, \omega, M, c, p) \geq 1$ *such that*

$$|S(t)| \leq \gamma\, e^{\alpha t}, \quad t \geq 0.$$

Now we state our main approximation result.

THEOREM 2.5. *Assume that* (A, B) *is stabilizable,* (A, C) *is detectable, and* (A1)–(A3) *are satisfied. If the original problem is stabilizable with attenuation* γ*, then so are the approximating systems for sufficiently large* N*. For such* N*, the Riccati equation*

$$(A^N)^*\Sigma^N + \Sigma^N A^N - \frac{1}{\epsilon}\Sigma^N B^N (B^N)^*\Sigma^N + \frac{1}{\gamma^2}\Sigma^N D^N (D^N)^*\Sigma^N + (C^N)^* C^N = 0 \tag{2.4}$$

has a nonnegative, self-adjoint solution Σ^N *and* $\Sigma^N P^N x \to \Sigma\, x$ *strongly in* X *as* $N \to \infty$*. Moreover,* $\hat{K}^N = \frac{1}{\epsilon}(B^N)^*\Sigma^N$ *converges to* $\hat{K} = \frac{1}{\epsilon}B^*\Sigma$ *in norm. For* N *sufficiently large,* $\hat{K}^N$ *is* γ*-admissible for the infinite-dimensional problem.*

Proof. The proof is given in several steps. First, we give a brief description of the representation of Σ. This is used to show that for large N the approximating systems are stabilizable with attenuation γ and so for such N the finite-dimensional Riccati equation (2.4) has a solution Σ^N. We show that $\Sigma^N \to \Sigma$ and $\hat{K}^N \to \hat{K}$. Finally, we show that the approximating finite-dimensional feedback $\hat{K}^N$ is γ-admissible for the original system.

Step 1. First we briefly review the representation of Σ. Details may be found in [KE, Theorem 4.4] or [BB]. Since (A, B) is stabilizable and (A, C) is detectable, the (LQR) Riccati equation

$$\left(A^*\Pi + \Pi A - \frac{1}{\epsilon}\Pi B B^*\Pi + C^*C\right)x = 0 \quad \text{for all } x \in \mathrm{dom}(A) \tag{2.5}$$

has the unique nonnegative, self-adjoint solution Π. Let $S_c(t)$ be the exponentially stable semigroup generated by $A - \frac{1}{\epsilon}BB^*\Pi$.

Consider the quadratic differential game

$$\max_{v \in \mathcal{W}} \min_{u \in \mathcal{U}} \rho(u, v; x) - \gamma^2 |v|^2_{\mathcal{W}}$$

subject to (1.1).

Define $L \in \mathcal{L}(\mathcal{W}, L^2(0, \infty; X))$ by

$$(Lv)(t) = \int_t^\infty S_c^*(\tau - t)\Pi D v(\tau)\, d\tau. \tag{2.6}$$

For a disturbance v and $x \in X$

$$u^*(t) = -\frac{1}{\epsilon}B^*\left[\Pi x(t) + (Lv)(t)\right] \tag{2.7}$$

minimizes $\rho(u, v; x)$ over $u \in \mathcal{U}$ subject to (1.1).

For $x \in X$, $v \in \mathcal{W}$, write $r(t) = (Lv)(t)$ and define the quadratic form

$$\begin{aligned} J(v; x) &= \rho(u^*, v; x) - \gamma^2 |v|^2_{\mathcal{W}} \\ &= (x, \Pi x + 2\, r(0)) + \int_0^\infty 2\,(Dv(t), r(t)) - \frac{1}{\varepsilon}\,|B^* r(t)|^2 - \gamma^2\,|v(t)|^2\, dt. \end{aligned}$$

Defining the self-adjoint operator Q on $\mathcal{W}$ by

$$Q = \gamma^2\, I + \frac{1}{\epsilon}L^* B B^* L - D^* L - L^* D, \tag{2.8}$$

we have

$$J(v;x) = -(Qv,v)_{\mathcal{W}} + 2\,(D^*\Pi S_c(\cdot)x, v)_{\mathcal{W}} + (\Pi x, x).$$

If the system (1.1)−(1.2) is stabilizable with attenuation γ, then for some $\delta > 0$

$$J(v;0) = \rho(u^*, v; 0) - \gamma^2\,|v|^2_{\mathcal{W}} \le -\delta\,|v|^2_{\mathcal{W}}.$$

Thus $Q \ge \delta I$ and maximization of $J(v;x)$ over $v \in \mathcal{W}$ is well posed. The solution to this problem, the worst disturbance v^*, is the unique solution to

$$Qv^* - D^*\Pi S_c(\cdot)x = 0, \quad x \in X. \tag{2.9}$$

We define the self-adjoint operator Σ on X by

$$(\Sigma x, x) = \max\ J(v,x) = (D^*\Pi S_c(\cdot)x, v^*)_{\mathcal{W}} + (\Pi x, x)$$

for $x \in X$. This implies

$$\Sigma x = \Pi x + \int_0^\infty S_c^*(t)\Pi D v^*(t)\,dt. \tag{2.10}$$

It is shown in [KE, BB] that Σ satisfies the Riccati equation (2.1) and is unique within the class of nonnegative, self-adjoint solutions to (2.1) such that the closed-loop semigroup generated by $A - \frac{1}{\epsilon}BB^*\Sigma + \frac{1}{\gamma^2}DD^*\Sigma$ is exponentially stable on X.

Moreover, the optimal pair (u^*, v^*) to the differential game is of feedback form:

$$u^*(t) = -\frac{1}{\epsilon}\,B^*\Sigma x(t), \quad v^*(t) = \frac{1}{\gamma^2}\,D^*\Sigma x(t).$$

Step 2. Next, we show that the finite-dimensional Riccati equation (2.4) has a nonnegative solution by showing that the finite-dimensional system is stabilizable with attenuation γ.

Under assumptions (A1)−(A2) it is shown in [IT1] that the (LQR) Riccati equation on X^N

$$(A^N)^*\Pi + \Pi^N A^N - \frac{1}{\epsilon}\Pi^N B^N (B^N)^*\Pi^N + (C^N)^* C^N = 0 \tag{2.11}$$

has the unique nonnegative, self-adjoint solution Π^N and also that $\Pi^N P^N x \to \Pi\, x$ strongly in X as $N \to \infty$. Define $S_c^N(t) = e^{(A^N - \frac{1}{\epsilon}B^N B^{N*}\Pi^N)t}$. It is also shown in [IT1] that there exist constants $M_3 \ge 1$ and $\omega_3 > 0$ such that

$$|S_c^N(t)P^N x| \le M_3 e^{-\omega_3 t}\,|x|_X. \tag{2.12}$$

Let $\bar{K}^N = \frac{1}{\epsilon}B^{N*}\Pi^N$ and $\bar{K} = \frac{1}{\epsilon}B^*\Pi$. Then, since B is compact, $|\bar{K}^N P^N - \bar{K}| \to 0$ as $N \to \infty$. Assumption (A1) implies

$$\left|(\lambda I - (A^N - B^N\bar{K}^N))^{-1}P^N x - (\lambda I - (A - B\bar{K}))^{-1}x\right|_X \to 0$$

for all $x \in X$ and also the similar statement for the sequence of adjoint operators. It thus follows from the Trotter–Kato theorem that

$$S_c^N(t)P^N x \to S_c(t)x \quad \text{and} \quad (S_c^N)^*(t)P^N x \to S_c^*(t)x \tag{2.13}$$

for all $x \in X$, uniformly on bounded t-intervals.

Since D is compact,

$$|(S_c^N)^*(t)\Pi^N D^N - S_c^*(t)\Pi D| \to 0 \tag{2.14}$$

uniformly in any bounded t-interval. For $\tau > 0$ and $p \in [1, \infty)$,

$$\begin{aligned}\int_0^\infty |(S_c^N)^*(t)\Pi^N D^N - S_c^*(t)\Pi D|^p \, dt \\ \leq \int_0^\tau |(S_c^N)^*(t)\Pi^N D^N - S_c^*(t)\Pi D|^p \, dt \\ + \int_\tau^\infty (|S_c^N(t)|^p |\Pi^N|^p + |S_c^*(t)|^p |\Pi|^p)|D|^p \, dt.\end{aligned}$$

Since from (2.12) the second term of the right-hand side is bounded by $Me^{-\omega\tau}$ for some positive constants M and ω, it follows from (2.13) that

$$\int_0^\infty |(S_c^N)^*(t)\Pi^N D^N - S_c^*(t)\Pi D|^p \, dt \to 0 \tag{2.15}$$

as $N \to \infty$ for all $p \in [1, \infty)$.

Define the linear operators L^N and Q^N on $\mathcal{W}$ for the approximate problem that corresponds to L and Q defined in (2.6) and (2.8), respectively. It then follows from (2.15) that

$$|L^N - L| \quad \text{and} \quad |Q^N - Q| \to 0 \tag{2.16}$$

as $N \to \infty$.

Define

$$z^N(u, v; x) = \operatorname{col}(C^N x^N(t), \sqrt{\epsilon} u(t)),$$

where x^N is the state of the approximating system with control u, disturbance v, and initial condition x. Also define the finite-dimensional cost

$$\rho^N(u, v, x) = |z^N|^2_{L^2(0,\infty;Y)}.$$

As in (2.7) define the control

$$u^N(t) = -\frac{1}{\epsilon}(B^N)^*[\Pi^N x^N(t) + L^N v(t)]. \tag{2.17}$$

Then, with initial condition $x = 0$,

$$x^N(t) = \int_0^t S_c^N(t-s)\left(-\frac{1}{\epsilon} B^N (B^N)^* (L^N v)(s) + D^N v(s)\right) ds. \tag{2.18}$$

Define the linear operators $\tilde{L}^N$, $\tilde{L} \in \mathcal{L}(L^2(0,\infty;X), L^2(0,\infty,X))$ by

$$(\tilde{L}^N f)(t) = \int_0^t S_c^N(t-s)\left(\frac{1}{\epsilon} B^N (B^N)^* f(s)\right) ds$$

and

$$(\tilde{L}f)(t) = \int_0^t S_c(t-s)\left(\frac{1}{\epsilon}BB^*f(s)\right)\,ds$$

Using the same arguments as above we can show that $|\tilde{L}^N - \tilde{L}| \to 0$ as $N \to \infty$. Since from (2.7)

$$x^*(t) = \int_0^t S_c(t-s)\left(-\frac{1}{\epsilon}BB^*(Lv)(s) + Dv(s)\right)\,ds,$$

it follows from (2.16) and (2.18) that

$$|x^N - x^*|^2_{L^2(0,\infty;X)} \le \epsilon_1(N)\,|v|^2_{\mathcal{W}},$$

where $\epsilon_1(N) \to 0$ as $N \to \infty$. Since C^N is the restriction of C on X^N it follows that

$$|\rho^N(u^N, v; 0) - \rho(u^*, v; 0)| \le \epsilon(N)\,|v|^2_{\mathcal{W}},$$

where $\epsilon(N) \to 0$ as $N \to \infty$. Therefore, for sufficiently large N, the approximating problems are stabilizable with attenuation γ.

This implies that the finite-dimensional Riccati equation (2.4) has a self-adjoint solution Σ^N on X^N

$$\Sigma^N P^N x = \Pi^N P^N x + \int_0^\infty (S_c^N)^*(t)\Pi^N D^N v^N(t)\,dt, \tag{2.19}$$

where $v^N(t) \in \mathcal{W}$ is the unique solution of

$$Q^N v^N - (D^N)^*\Pi^N S_c^N(\cdot)P^N x = 0. \tag{2.20}$$

Furthermore, $A^N - \frac{1}{\epsilon}B^N(B^N)^*\Sigma^N + \frac{1}{\gamma^2}D^N(D^N)^*\Sigma^N$ generates an exponentially stable semigroup on X^N, and $\hat{K}^N = \frac{1}{\epsilon}(B^N)^*\Sigma^N$ is γ-admissible for the approximating problem.

Step 3. We now show that Σ^N converges strongly to Σ and $\hat{K}^N$ converges uniformly to $\hat{K}$.

The uniform convergence of Q^N to Q implies that Q^N is coercive with $Q^N \ge \frac{\delta}{2}$ for N sufficiently large. Also, (2.15) implies that

$$\int_0^\infty |(D^N)^*\Pi^N S_c^N(t)P^N - D^*\Pi S_c(t)|^2 dt \to 0$$

as $N \to \infty$. Therefore, the solution to (2.20) satisfies

$$|v^N|_{\mathcal{W}} \le M\,|x|_X \tag{2.21}$$

for some constant M. Note that

$$Q(v^N - v^*) = (Q - Q^N)v^N + (D^N)^*\Pi^N S_c^N(t)P^N x - D^*\Pi S_c(t)x.$$

Hence, from (2.16) and (2.21) v^N converges strongly to v^* in $\mathcal{W}$ as $N \to \infty$ for each $x \in X$.

It now follows from (2.10), (2.15), and (2.19) that

$$\Sigma^N P^N x \to \Sigma x \quad \text{for all } x \in X.$$

Since B is compact, we have that $|(B^N)^*\Sigma^N P^N - B^*\Sigma| \to 0$. That is, $\hat{K}^N$ converges uniformly to $\hat{K}$.

Step 4. To prove that for large N the approximating feedback operators $\hat{K}^N$ are γ-admissible for the system (1.1)–(1.2), first note that by the Trotter–Kato theorem the convergence of $\hat{K}^N$ to $\hat{K}$ in norm implies that the semigroup $S_{\hat{K}^N}$ generated by $A - B\hat{K}^N$ converges to the semigroup $S_{\hat{K}}$ generated by $A - B\hat{K}$, strongly in X, uniformly in bounded intervals of time. Also, there exists $\tilde{M} \geq 1$, $\tilde{\omega} > 0$ such that for sufficiently large N,

$$|S_{\hat{K}^N}(t)|_X \leq \tilde{M} e^{-\tilde{\omega} t}.$$

The output (1.2) with a disturbance v, feedback control $u(t) = -\hat{K}^N x(t)$, and initial condition $x(0) = 0$ is

$$\bar{z}^N(t) = \begin{bmatrix} C \\ -\sqrt{\epsilon}\hat{K}^N \end{bmatrix} \int_0^t S_{\hat{K}^N}(t-s) D v(s)\, ds.$$

The convergence of the semigroups and the compactness of D imply, as in equation (2.15), that for any $p \in [1, \infty)$,

$$\int_0^\infty |S_{\hat{K}^N}(t)D - S_{\hat{K}}(t)D|^p \, dt \to 0.$$

Let $\bar{z}$ indicate the output (1.2) obtained with the same disturbance v but with feedback control $-\hat{K}x(t)$. Then

$$\int_0^\infty |\bar{z}^N(t) - \bar{z}(t)|_Y^2 \, dt \leq \epsilon(N)\, |v|_{\mathcal{W}}^2,$$

where $\epsilon(N) \to 0$ as $N \to \infty$. This implies that for large N, $\hat{K}^N$ is γ-admissible for the original system. □

COROLLARY 2.6. *Under the same assumptions as in Theorem* 2.3, *we have*

$$|v^N(t) - v^*(t)|_{\mathcal{W}} = 0,$$

$$|u^N(t) - u^*(t)|_{\mathcal{U}} = 0,$$

and there exist positive constants M_4, ω_4, M_5, *and* ω_5 *such that*

$$|e^{(A^N - \frac{1}{\epsilon}B^N(B^N)^*\Sigma^N + \frac{1}{\gamma^2}D^N(D^N)^*\Sigma^N)t} P^N x| \leq M_4 e^{-\omega_4 t}\, |x|_X\,, \tag{2.22}$$

$$|e^{(A^N - B^N\hat{K}^N)t} P^N x| \leq M_5 e^{-\omega_5 t}\, |x|_X\,. \tag{2.23}$$

Proof. The convergence of the worst disturbance v^N follows from the proof of Theorem 2.3.

We show that for large N, $A^N - \frac{1}{\epsilon}B^N(B^N)^*\Sigma^N + \frac{1}{\gamma^2}D^N(D^N)^*\Sigma^N$ generates a uniformly stable semigroup $T^N(t)$. Note that

$$v^N(t) = \frac{1}{\gamma^2}(D^N)^*\Sigma^N x^N(t), \quad u^N(t) = -\frac{1}{\epsilon}(B^N)^*\Sigma^N x^N(t)$$

and

$$\int_0^\infty |C^N x^N(t)|^2 + \epsilon|u^N(t)|^2\,dt = (\Sigma^N P^N x, x) + \gamma^2|v^N(t)|^2_{\mathcal{W}}, \tag{2.24}$$

where

$$x^N(t) = T^N(t)x = e^{(A^N - \frac{1}{\epsilon}B^N(B^N)^*\Sigma^N + \frac{1}{\gamma^2}D^N(D^N)^*\Sigma^N)t}P^N x, \quad x \in X.$$

Let G^N be as in (A2)(ii) so that $e^{(A^N - G^N C^N)t}$ is uniformly exponentially stable. Then we have

$$\begin{aligned} x^N(t) &= e^{(A^N - G^N C^N)t}P^N x \\ &\quad + \int_0^t e^{(A^N - G^N C^N)(t-s)}(G^N C^N x^N(s) + B^N u^N(t) + D^N v^N(t))\,dt. \end{aligned}$$

Note that from Hölder's inequality and the Fubini theorem

$$\begin{aligned} \int_0^\infty \left|\int_0^t f(t-s)g(s)\,ds\right|^2 dt &\le \int_0^\infty \int_0^t |f(t-s)|\,ds \int_0^t |f(t-s)||g(s)|^2\,ds \\ &\le \int_0^\infty |f(\sigma)|\,d\sigma \int_0^\infty \left(\int_s^\infty |f(t-s)|\,dt\right)|g(s)|^2\,ds \\ &\le \left(\int_0^\infty |f(\sigma)|\,d\sigma\right)^2 \int_0^\infty |g(t)|^2\,dt \end{aligned}$$

for $f \in L^1(0,\infty)$ and $g \in L^2(0,\infty)$. It thus follows from (A2)(ii), (2.21), (2.24) that

$$\begin{aligned} \int_0^\infty |x^N(t)|^2\,dt &\le \frac{M_2^2}{\omega_2}|x|^2 \\ &\quad + \frac{2M_2}{\omega_2}\int_0^\infty (|G^N|^2|C^N x^N(t)|^2 + |B^N|^2|u^N|^2 + |D^N|^2|v^N(t)|^2)\,dt \\ &\le \beta|x|_X^2 \quad \text{for some } \beta > 0. \end{aligned}$$

Hence, (2.22) follows from the Datko lemma. The proof of (2.23) is identical.

Let $T(t)$ be the semigroup generated by $A - \frac{1}{\epsilon}BB^*\Sigma + \frac{1}{\gamma^2}DD^*\Sigma$ on X. The above implies that $x^N(t)$ converges in $L^2(0,\infty;X)$ to $T(t)x$ (see the proof of Theorem 2.3) and so $u^N(t) \to u^*(t)$ in $L^2(0,\infty;X)$. □

Conversely, we have the following theorem.

THEOREM 2.7. *Suppose the Riccati equation* (2.4) *has uniformly bounded nonnegative self-adjoint solutions* Σ^N *on* X^N *for* N *sufficiently large with*

$$|v^N(t)|_{\mathcal{W}} \le M|x|_X^2 \quad \textit{for some } M > 0, \tag{2.25}$$

where

$$v^N(t) = \frac{1}{\gamma^2}(D^N)^*\Sigma^N e^{(A^N - \frac{1}{\epsilon}B^N(B^N)^*\Sigma^N + \frac{1}{\gamma^2}D^N(D^N)^*\Sigma^N)t}P^N x.$$

Assume that (A1), (A2)(ii), *and* (A3) *hold. Then, the system* (1.1) *is stabilizable with attenuation* γ.

Proof. It follows from [GI] that there exist a nonnegative self-adjoint operator Σ on X and a subsequence of Σ^N such that $\Sigma^N P^N x$ converges weakly to Σx in X for $x \in X$. It follows from (A3) that $(B^N)^*\Sigma^N P^N x$ and $(D^N)^*\Sigma^N P^N x$ converge strongly to $B^*\Sigma x$ and $D^*\Sigma x$, respectively. Let

$$T^N(t) = e^{(A^N - \frac{1}{\epsilon}B^N(B^N)^*\Sigma^N + \frac{1}{\gamma^2}D^N(D^N)^*\Sigma^N)t}$$

and $T(t)$ be the semigroup generated by $A - \frac{1}{\epsilon}BB^*\Sigma + \frac{1}{\gamma^2}DD^*\Sigma$ on X. It then follows from (A1) and the Trotter–Kato theorem that

$$T^N(t)P^N x \to T(t)x \quad \text{for each } x \in X,$$

uniformly on bounded t-intervals.

Note that

$$\Sigma^N P^N x = e^{(A^N)^* t}\Sigma^N T^N(t)P^N x + \int_0^t e^{(A^N)^*(t-s)}(C^N)^* C^N T^N(t-s)P^N x\, ds,\ x \in X.$$

Taking the limit as $N \to \infty$, we obtain

$$\Sigma x = S^*(t)\Sigma T(t)x + \int_0^t S^*(t-s)C^*CT(t-s)x\, ds,$$

which implies that Σ is a solution to (2.4).

Moreover, it follows from (2.25) and the proof of Corollary 2.6 that

$$|T^N(t)P^N x| \le M_4 e^{-\omega_4 t}\,|x|_X$$

for some positive constants M_4, ω_4, provided that (A2)(ii) holds. Hence, the semigroup $T(t)$ is exponentially stable.

It follows from Theorem 2.2 that the system (1.1) is stabilizable with attenuation γ. □

The optimal disturbance attenuation problem for the infinite-dimensional system (1.1) is to find

$$\hat{\gamma} = \inf\ \gamma$$

over all γ such that (1.2) is stabilizable with attenuation γ. Let $\{\hat{\gamma}^N\}$ indicate the corresponding optimal disturbance attenuation for the approximating problems. Theorem 2.3 implies that

$$\limsup_{N\to\infty} \hat{\gamma}^N \le \hat{\gamma}. \tag{2.26}$$

However, we have a stronger result.

THEOREM 2.8. *Assume that* (A1)–(A3) *hold,* (A, B) *is stabilizable, and* (A, C) *is detectable. Then*

$$\lim_{N\to\infty} \hat{\gamma}^N = \hat{\gamma}.$$

Proof. Because of (2.26), it is sufficient to show that

$$\liminf_{N\to\infty} \hat{\gamma}^N \geq \hat{\gamma}.$$

Assume that this statement is false. Then there is an $\delta > 0$ such that for all N there is $M > N$ with $\hat{\gamma}^M < \hat{\gamma} - \delta$. In this way we can construct a subsequence $\{\hat{\gamma}^M\}$ of the sequence $\{\hat{\gamma}^N\}$ with

$$\hat{\gamma}^M < \hat{\gamma} - \delta.$$

Thus, the approximating system is stabilizable with attenuation $\hat{\gamma} - \delta/2$ and

$$\rho^M(u^M, v; 0) \leq (\hat{\gamma} - \delta/2)\,|v|^2_{\mathcal{W}},$$

where for any $v \in \mathcal{W}$, $u^M(t)$ is defined by (2.17). Moreover, we have

$$|\rho^M(u^M, v; 0) - \rho(u^*, v; 0)| \leq \epsilon(M)\,|v|^2_{\mathcal{W}},$$

where $\epsilon(M) \to 0$ as $M \to \infty$ and $u^*(t) \in \mathcal{U}$ is given by (2.7). Hence the original problem is stabilizable with attenuation $\hat{\gamma} - \delta/2$. This contradicts the optimality of $\hat{\gamma}$ and thus (2.26) holds. ∎

The above theorem implies that if a sequence of approximating problems that satisfy assumptions (A1)–(A3) are stabilizable with attenuation γ, then so is the infinite-dimensional problem. Thus, Theorem 2.7 can be regarded as a partial converse of Theorem 2.3. The difference between this theorem and Theorem 2.5 is that the assumption of uniform stabilizability (A2)(i) in Theorem 2.5 is replaced by uniform boundedness of Σ^N and v^N.

3. Example. Consider a Euler–Bernoulli beam clamped at one end and let $w(r,t)$ denote the deflection of the beam from its rigid body motion at time t and position r. The deflection can be controlled by applying a torque at the clamped end ($r = 0$). We assume that the hub inertia I_h is much larger than the beam inertia, so that, letting $\theta(t)$ indicate the rotation angle, $u(t) = I_h\ddot{\theta}(t)$ is a reasonable approximation to the applied torque. The disturbance $v(t)$ induces a uniformly distributed load $\rho d v(t)$. Use of the Kelvin–Voigt damping model leads to the following description of the beam vibrations:

$$\rho\frac{\partial^2 w}{\partial t^2} + C_v\frac{\partial w}{\partial t} + \frac{\partial^2}{\partial r^2}\left[EI\frac{\partial^2 w}{\partial r^2} + C_d I\frac{\partial^3 w}{\partial r^2 \partial t}\right] = \frac{\rho r}{I_h}u(t) + \rho\, d\, v(t), \qquad 0 < r < L.$$

The boundary conditions are

$$w(0,t) = 0, \qquad \frac{\partial w}{\partial r}(0,t) = 0,$$

$$\left[EI\frac{\partial^2 w}{\partial r^2} + C_d I\frac{\partial^3 w}{\partial r^2 \partial t}\right]_{r=L} = 0, \qquad \left[EI\frac{\partial^3 w}{\partial r^3} + C_d I\frac{\partial^4 w}{\partial r^3 \partial t}\right]_{r=L} = 0.$$

The values of the physical parameters in this example are listed in Table 1. Let $x(t) = (w(\cdot,t), \frac{\partial}{\partial t}w(\cdot,t))$, H be the closed linear subspace of $H^2(0,1)$ defined by

$$H = \left\{w \in H^2(0,1) : w(0) = \frac{dw}{dr}(0) = 0\right\},$$

TABLE 1
Physical constants.

E	$2.1 * 10^{11}\, N/m^2$
I	$1.167 \times 10^{-10}\, m^4$
ρ	$2.975\, kg/m$
C_v	$.001\, Ns/m^2$
C_d	$.01\, Ns/m^2$
L	$7\, m$
I_h	$121.9748\, kgm^2$
d	$.04\, 1/kg$

and $X = H \times L^2(0,1)$. Here $H^2(0,1)$ is the Hilbert space defined by

$$H^2(0,1) = \left\{ \phi \in C^1(0,1) : \frac{d}{dr}\phi \text{ is absolutely continuous and } \frac{d^2}{dr^2}\phi \in L^2(0,1) \right\}.$$

If the tip deflection is measured, a state-space formulation of the above partial differential equation problem is

$$\frac{d}{dt}x(t) = Ax(t) + Bu(t) + Dv(t),$$

$$y(t) = Cx(t) = w(L,t),$$

where

$$A = \begin{bmatrix} 0 & I \\ -\dfrac{EI}{\rho}\dfrac{d^4}{dr^4} & -\dfrac{C_d I}{\rho}\dfrac{d^4}{dr^4} - \dfrac{C_v}{\rho} \end{bmatrix}, \qquad B = \begin{bmatrix} 0 \\ \dfrac{r}{I_h} \end{bmatrix}, \quad D = \begin{bmatrix} 0 \\ d \end{bmatrix}$$

with domain

$$\operatorname{dom}(A) = \Bigg\{ (\phi,\psi) \in X : \psi \in H \text{ and } M = EI\,\frac{d^2}{dr^2}\phi + C_d I \frac{d^2}{dr^2}\psi \in H^2(0,1) \text{ with } M(L) = \frac{d}{dr}M(L) = 0 \Bigg\}.$$

Define $V = H \times H$. Then A can be defined by

$$\langle Ax, z\rangle_{V^* \times V} = -a(x,z) \quad \text{for } x,\ z \in V,$$

where the sesquilinear form $a(\cdot,\cdot)$ on $V \times V$ is given by

$$a((\phi_1,\psi_1),(\phi_2,\psi_2)) = -\sigma(\psi_1,\phi_2) + \sigma\left(\phi_1 + \frac{C_d}{E}\,\psi_1,\ \psi_2\right) + \left(\frac{C_v}{\rho}\,\psi_1,\ \psi_2\right)_{L^2}$$

for $(\phi_i,\psi_i) \in V,\ i = 1,2$. Here,

$$\sigma(\phi,\psi) = \int_0^L \frac{EI}{\rho}\frac{d^2}{dr^2}\phi(r)\frac{d^2}{dr^2}\psi(r)\,dr\,.$$

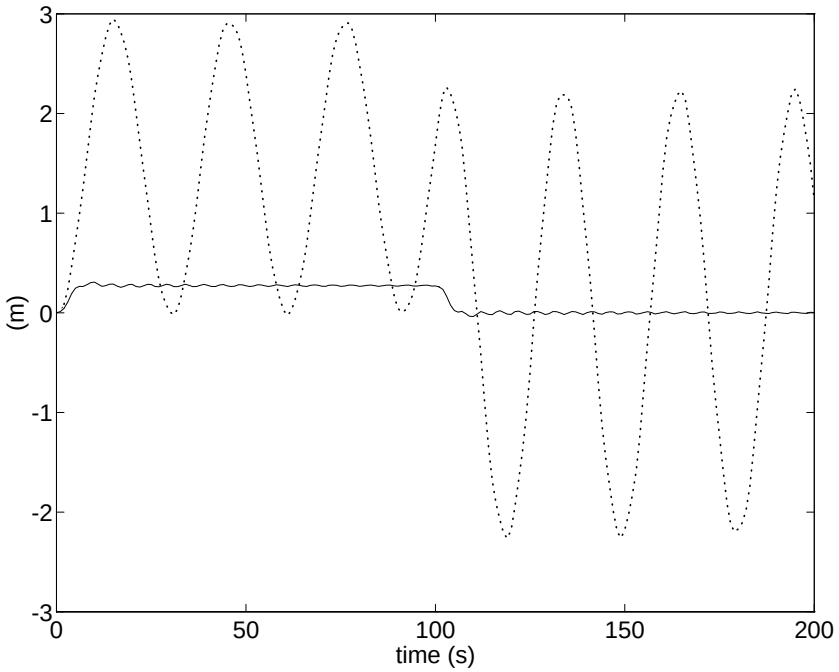


FIG. 1. *Response of open (..) and closed loop (–) to* $v(t) = 1, t \leq 100s$*: 10 elements.*

The sesquilinear form a is continuous on $V \times V$ and coercive; i.e.,

$$\operatorname{Re} a(z,z) \geq \omega \, |z|_V^2 - \beta \, |z|_X^2 \quad \text{for } z \in V$$

for appropriately chosen positive constants $\omega,\ \beta$. It thus follows that A generates an exponentially stable analytic semigroup on X [SH, section 4]. The operators B and D are clearly bounded operators from R to X. Sobolev's inequality implies that evaluation at a point is bounded on H, and so the output operator C is bounded from X to R.

Let $H^N \subset H$ be a sequence of finite-dimensional subspaces. The approximating generator A^N on $X^N = H^N \times H^N$ is defined by

$$\langle -A^N x^N, z^N \rangle = a(x^N, z^N) \quad \forall x^N,\ z^N \in X^N,$$

and $P^N,\ B^N,\ C^N$ are as defined at the beginning of section 2. This type of approximation is generally referred to as a *Galerkin* approximation. Suppose that the approximating subspaces H^N satisfy the H-approximation property: for all $\phi \in H$ there exists a sequence $\phi^N \in H^N$ with

$$\lim_{N \to \infty} |\phi^N - \phi|_H = 0. \tag{H1}$$

It is shown in [IT2, MO2] that as long as the approximating spaces satisfy the H-approximation property, assumptions (A1)–(A3) are satisfied. The standard finite-element cubic B-spline approximations [OP] do satisfy the H-approximation property, and so all the assumptions of Theorem 2.3 are satisfied by this problem.

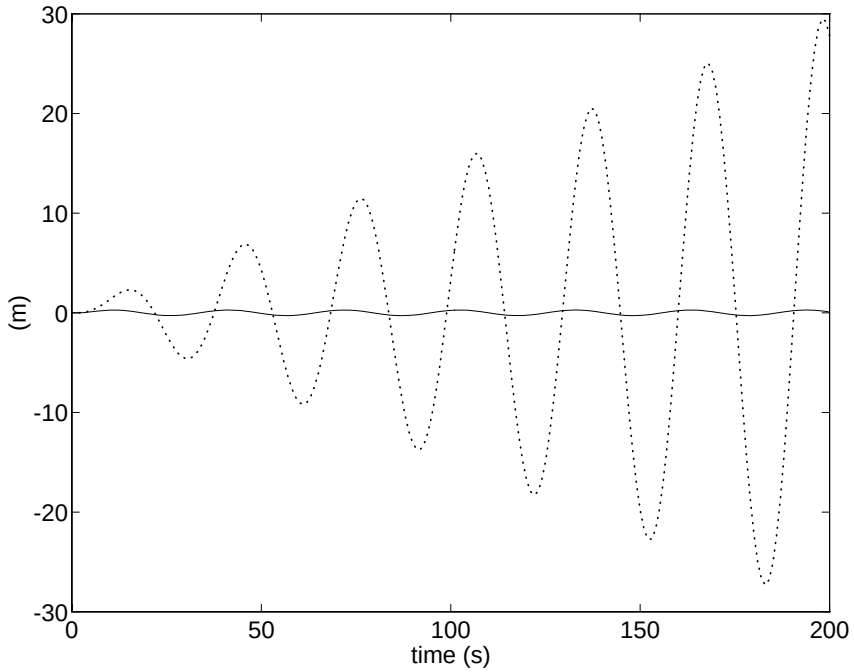


FIG. 2. *Response of open (..) and closed loop (–) to* $v(t) = \sin(\omega t)$ *:* 10 *elements.*

Our numerical calculations were carried out using a series of cubic B-spline approximations for H^N, and the corresponding series of finite-dimensional Riccati equations were solved with $\epsilon = .1$ and $\gamma = 2.3$. Figure 1 compares the open- and closed-loop responses $Cx(t) = w(L,t)$ with a temporary step disturbance for the approximation with 10 elements. The feedback controller leads to a closed loop which is able to almost entirely reject this disturbance. Figure 2 compares the open- and closed-loop responses to the periodic disturbance $\sin(\omega t)$ where ω is the first resonant frequency: $\omega = \min_i |\mathrm{Im}(\lambda_i(A^{10})|$. The resonance in the open loop is not present in the closed loop.

Since the input space $U = R$, the feedback operator $\hat{K}^N$ is a bounded linear functional on X^N and hence can be uniquely identified with an element of X^N, usually called the gain. Figure 3 displays the convergence of the feedback gains predicted by Theorem 2.3. Since X^N is a product space, the first and second components of the gains are displayed separately as displacement and velocity gains, respectively.

The sequence of operators $A^N - B^N\hat{K}^N$ is uniformly exponentially stable, and so

$$\max_{1\le i\le N} \mathrm{Re}\,\lambda_i(A^N - B^N\hat{K}^N)$$

converges to a nonzero number as $N \to \infty$, which can be verified theoretically. This convergence is displayed in Figure 4 for several different values of ϵ and $\gamma = 2.3$. Notice that as ϵ is decreased, the convergence becomes slower. A robust stability theorem provides some insight into this behavior.

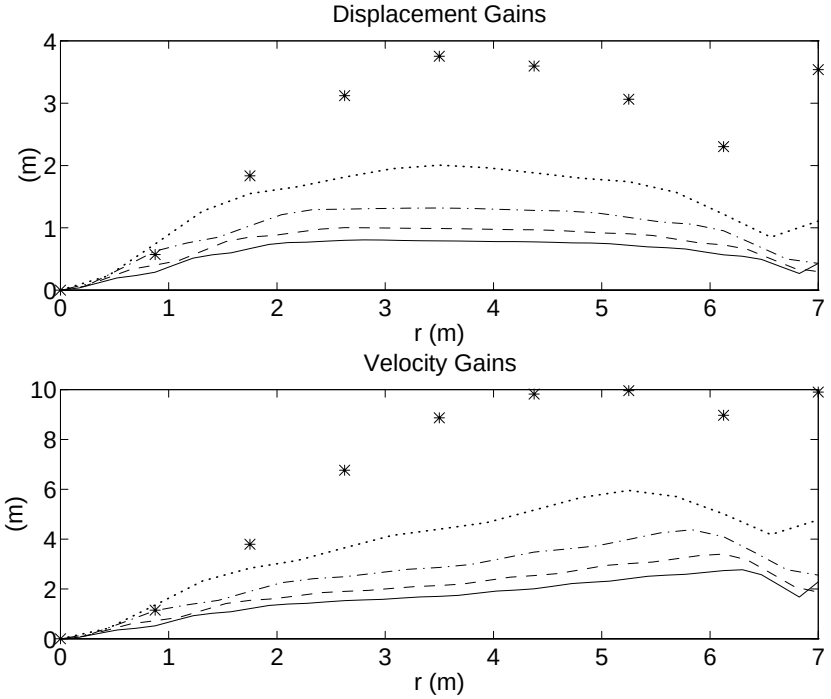


FIG. 3. *Convergence of the feedback gains:* 2 *elements* *, 4 *elements* ..., 6 *elements* _._., 8 *elements* _ _, 10 *elements,* __ .

Let the functions H, G_o, and G in the following theorem indicate transfer functions of systems with a finite number of inputs and outputs. For a given nominal plant G_o and weighting function $r \in H^\infty$, the class $\mathcal{A}(G_o, r)$ consists of all plants G that have the same number of right-hand-plane poles as G_o and that satisfy

$$|G(i\omega) - G_o(i\omega)| \; < \; |r(i\omega)|.$$

That is, $\mathcal{A}(G_o, r)$ contains the systems whose frequency response is within $r(i\omega)$ at each frequency ω of that of the nominal system G_o.

THEOREM 3.1 (see [CD]). *Suppose that a controller H stabilizes the system G_o. For any $r \in H^\infty$ the controller H stabilizes all $G \in \mathcal{A}(\, G_o, r)$ if and only if, for all ω,*

$$|H(1 + G_o H)^{-1}(i\omega)|\, |r(i\omega)| \; \leq 1. \tag{3.1}$$

For any approximation N, we have the nominal plant

$$G_o = (sI^N - A^N)^{-1}[B^N \; D^N]$$

and the controller

$$H = \begin{bmatrix} \hat{K}^N \\ 0 \end{bmatrix}.$$

Thus, we have

$$H(I + G_o H)^{-1} = \hat{K}^N (sI^N - A^N + B^N \hat{K}^N)^{-1}(sI^N - A^N)$$

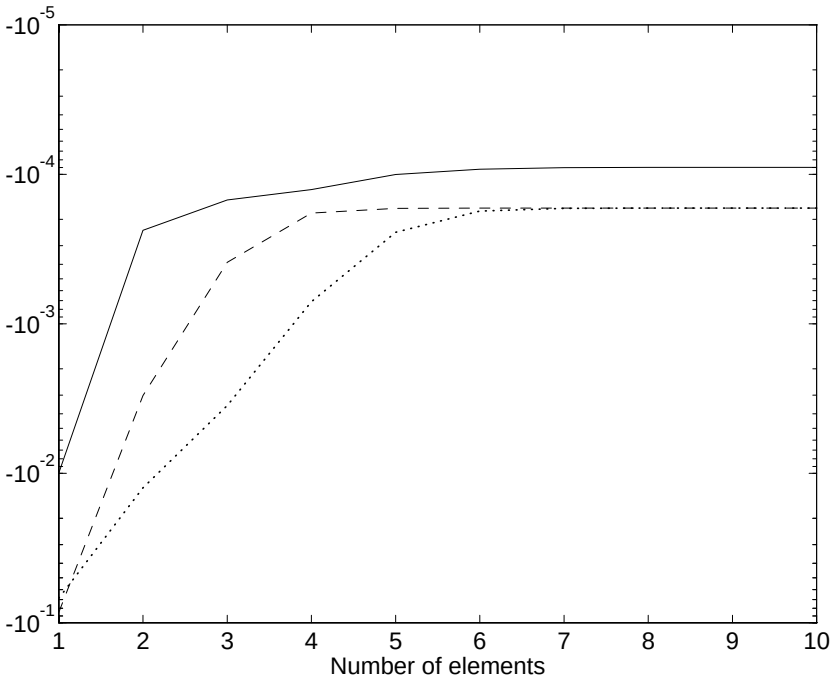


FIG. 4. *Convergence of* $\max_{1\le i\le N} \mathrm{Re}(\lambda_i)$*:* $\epsilon = .1$ *(—),* .001 *(- -),* .00001*.(...).*

and

$$H(I + G_o H)^{-1} G_o \begin{bmatrix} 0 \\ v \end{bmatrix} = \hat{K}^N (sI^N - A^N + B^N \hat{K}^N)^{-1} D^N v.$$

The second row in the attenuation bound (1.4) is commonly interpreted as a constraint on the control effort. However, it can also be interpreted as a robust stability constraint. If the inequality (1.4) is satisfied, then the robustness criterion (3.1) is satisfied with

$$r(i\omega) = \frac{\sqrt{\epsilon}}{\gamma} G_o(i\omega) \begin{bmatrix} 0 \\ I \end{bmatrix}.$$

This interpretation explains why convergence is slower for smaller ϵ. The stability margin for the approximation N is affected by the constraint that the higher-order approximations also be stabilized. Increasing ϵ while holding all other parameters such as γ constant means that the computed controller must stabilize a larger family of systems. Hence the stability margin is less sensitive to change in the approximation index N.

REFERENCES

[BB] A. BENSOUSSAN AND P. BERNHARD, *On the standard problem of H_∞-optimal control for infinite dimensional systems*, in Identification and Control in Systems Governed by Partial Differential Equations, H. T. Banks, R. H. Fabiano, and K. Ito, eds., SIAM, Philadelphia, 1993, pp. 117–140.

[BK] H. T. Banks and K. Kunisch, *The linear regulator problem for parabolic systems*, SIAM J. Control Optim., 22 (1984), pp. 684–698.

[BIP] J. A. Burns, K. Ito, and G. Propst, *On nonconvergence of adjoint semigroups for control systems with delays*, SIAM J. Control Optim., 26 (1988), pp. 1442–1454.

[CD] M. J. Chen and C. A. Desoer, *Necessary and sufficient conditions for robust stability of linear distributed feedback systems*, Internat. J. Control, 35 (1984), pp. 255–267.

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

[IT1] K. Ito, *Strong convergence and convergence rates of approximating solutions for algebraic Riccati equations in Hilbert spaces*, in Distributed Parameter Systems, F. Kappel, K. Kunisch, and W. Schappacher, eds., Springer-Verlag, Berlin, New York, 1987, pp. 151–166.

[IT2] K. Ito, *Finite-dimensional compensators for infinite-dimensional systems via Galerkin-type approximation*, SIAM J. Control Optim., 28 (1990), pp. 1251–1269.

[KE] B. Van Keulen, *H^∞-Control for Distributed Parameter Systems: A State-Space Approach*, Birkhäuser Boston, Cambridge, MA, 1993.

[MO1] K. A. Morris, *Convergence of controllers designed using state-space methods*, IEEE Trans. Automat. Control, 39 (1994), pp. 2100–2104.

[MO2] K. A. Morris, *Design of finite-dimensional controllers for infinite-dimensional systems by approximation*, J. Math. Systems, Estim. Control, 4 (1994), pp. 1–30.

[OP] N. Ottosen and H. Petersson, *Introduction to the Finite Element Method*, Prentice-Hall, Englewood Cliffs, NJ, 1992.

[PA] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.

[SA] D. Salamon, *Structure and stability of finite dimensional approximations for functional differential equations*, SIAM J. Control Optim., 23 (1985), pp. 928–951.

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

SIAM J. CONTROL OPTIM.
Vol. 36, No. 1, pp. 100–114, January 1998

005

CONVERGENCE RATE OF STOCHASTIC APPROXIMATION ALGORITHMS IN THE DEGENERATE CASE*

HAN-FU CHEN†

Abstract. Let $f(\cdot)$ be an unknown function whose root x^0 is sought by stochastic approximation (SA). Convergence rate and asymptotic normality are usually established for the nondegenerate case $f'(x^0) \neq 0$. This paper demonstrates the convergence rate of SA algorithms for the degenerate case $f'(x^0) = 0$. In comparison with the previous work, in this paper no growth rate restriction is imposed on $f(\cdot)$, no statistical property is required for the measurement noise, the general step size is considered, and the result is obtained for the multidimensional case, which is not a straightforward extension of the one-dimensional result. Although the observation noise may be either deterministic or random, the analysis is purely deterministic and elementary.

Key words. stochastic approximation, convergence rate

AMS subject classification. 62L20

PII. S0363012995281730

1. Introduction.

The topic of SA is to search the roots or extremes of an unknown function $f(\cdot) : R^l \to R^l$ which can be observed with noise. Since its pioneer work by Robbins and Monro [1], SA has obtained much attention from researchers [2, 3] and is applied in various areas, such as parameter identification, adaptive control, optimization, pattern recognition, and others [4].

In many applications not only convergence but also convergence rate of the algorithm is of interest. Intuitively, the rate of convergence depends on the derivative $f'(x^0)$ of the function at its root x^0; the rate in the nondegenerate case ($f'(x^0) \neq 0$) should be faster than it is in the degenerate case ($f'(x^0) = 0$). To be precise, the Robbins–Monro algorithm is defined by

$$x_{n+1} = x_n + a_n y_{n+1}, \tag{1.1}$$

$$y_{n+1} = f(x_n) + \epsilon_{n+1}, \tag{1.2}$$

where y_{n+1} is the observation and ϵ_{n+1} is the noise. $\{a_n\}$ is the step size and is selected to have the following properties:

$$a_n > 0, \quad a_n \xrightarrow[n \to \infty]{} 0 \quad \text{and} \quad \sum_{i=1}^{\infty} a_i = \infty.$$

Under certain conditions [1–4, 7] imposed on $f(\cdot)$ and ϵ_n, x_n defined by (1.1), (1.2) converges to the root x^0 of $f(\cdot)$, i.e.,

$$x_n \xrightarrow[n \to \infty]{} x^0, \quad f(x^0) = 0.$$

Further, in the nondegenerate multidimensional case assume

$$f(x) = H(x - x^0) + \Delta(x), \quad H < 0, \tag{1.3}$$

*Received by the editors February 17, 1995; accepted for publication (in revised form) October 9, 1996. This research was supported by the National Natural Science Foundation of China.
http://www.siam.org/journals/sicon/36-1/28173.html

†Laboratory of Systems and Control, Institute of Systems Science, Chinese Academy of Sciences, Beijing 100080, People's Republic of China (hfchen@iss03.iss.ac.cn).

$$\Delta(x) = O(\|x - x^0\|^2) \quad \text{as} \quad x \to x^0. \tag{1.4}$$

Then under some conditions on the observation noise

$$\|x_n - x^0\| = o(a_n^\delta) \quad \forall \delta \in \left(0, \frac{1}{2}\right), \tag{1.5}$$

provided $H + q\delta I$ is a stable matrix where α by assumption is defined by

$$a_{n+1}^{-1} - a_n^{-1} \xrightarrow[n \to \infty]{} q \geq 0.$$

The convergence rate in the case $f'(x^0) = 0$ was addressed in [5] for the special case where (i) $f(\cdot)$ is a scalar function, i.e., $l = 1$, and $f(x)$ grows not faster than linearly as $|x| \to \infty$; (ii) $(x - x^0)f(x) < 0 \; \forall x \neq x^0$; (iii) $f(x) = f_0|x - x^0|^{1+\gamma}\text{sign}(x - x^0)\cdot(1 + o(1))$ as $x \to x^0$, $\gamma > 0$; (iv) the conditional variance of ϵ_{n+1} given x_n is bounded, i.e., $\text{Var}(\epsilon_{n+1}|x_n) \leq \sigma^2$; (v) ϵ_{n+1} is conditionally independent of $x_0, \ldots, x_{n-1}$ given x_n; and (vi) the step size is special: $a_n = \frac{1}{n}$. In comparison with [5] this paper derives the convergence rate for the general case. To be precise, we do not impose any growth rate restriction on $f(\cdot)$; we do not require any statistical property of the noise, which is allowed to be stochastic or deterministic; we consider the general step size a_n and, finally, we give the convergence rate for both multidimensional and one-dimensional cases. The approach used here is completely different from that used in [5] and is purely deterministic. A purely deterministic approach in a discrete setting was used in [9, 10] as an alternative means for obtaining convergence results, and the approach used here is similar in flavor. We further show the power of an elementary deterministic analysis by obtaining convergence rates. It is worth noting that extension from the one-dimensional result to the multidimensional case is not straightforward. As will be seen in section 2, in the multidimensional case only the upper bound is obtained, while in the one-dimensional case it is shown that the upper bound is attainable.

2. Main results. Before describing the main results of the paper we present a convergence result, proved in [4, 6]. The algorithm considered in this paper is a modified version of (1.1), (1.2) and is defined as follows.

Let $\{M_k\}$ be a sequence of real numbers, $M_i > 0$, $M_i \uparrow \infty$ and let x^* be a fixed point in R^l. The estimate x_n is recursively given by

$$\widehat{x}_{k+1} = x_k + a_k y_{k+1}, \quad x_0 \text{ arbitrary}, \tag{2.1}$$

$$x_{k+1} = \widehat{x}_{k+1} I_{[\|\widehat{x}_{k+1}\| \leq M_{\sigma_k}]} + x^* I_{[\|\widehat{x}_{k+1}\| > M_{\sigma_k}]}, \tag{2.2}$$

$$\sigma_k = \sum_{i=0}^{k-1} I_{[\|\widehat{x}_{i+1}\| > M_{\sigma_i}]}, \tag{2.3}$$

$$y_{k+1} = f(x_k) + \epsilon_{k+1}. \tag{2.4}$$

Since M_i diverges, algorithm (2.1)–(2.4) coincides with the Robbins–Monro algorithm (1.1), (1.2) starting from some time, if we can prove that $\{x_k\}$ defined by (2.1)–(2.4) is bounded.

Let us list conditions which will be used later on.

A1. $f(\cdot)$ is an $R^l \to R^l$ measurable and locally bounded function, and $f(x) = 0$ $\forall x \in J$; i.e., J is the root set of $f(\cdot)$.

A2. $a_k > 0$, $a_k \xrightarrow[k\to\infty]{} 0$, $\sum_{i=1}^{\infty} a_i = \infty$.

A3. There is a differentiable function $v(\cdot) : R^l \to R$ such that

$$d(v(x), v(J)) > 0 \quad \text{if} \quad d(x, J) > 0$$

and

$$\sup_{\delta \le d(x,J) \le \Delta} f^{\tau}(x) v_x(x) < 0 \quad \forall \quad 0 < \delta < \Delta,$$

where $v_x(x)$ denotes the gradient of $v(x)$:

$$d(x, J) = \inf\{\|x - y\| : \forall y \in J\}, \text{ and } v(J) = \{v(x) : x \in J\}.$$

A4. As $x \to x^0$ the function $f(x)$ is expressed as

$$f(x) = H(x - x^0)\|x - x^0\|^{\gamma} + r(x), \qquad \gamma > 0, \tag{2.5}$$

where H is a stable matrix (i.e., all its eigenvalues have negative real parts) and

$$r(x) \in R^l, \quad r(x)/\|x - x^0\|^{1+\gamma} \to 0 \quad \text{as} \quad x \to x^0. \tag{2.6}$$

A5.

$$q_n \overset{\triangle}{=} a_{n+1}^{-1} - a_n^{-1}, 0 \le q_n, \limsup_{n\to\infty} q_n = q, 0 \le q < \infty, \tag{2.7}$$

$$\sum_{i=1}^{\infty} b_i = \infty, \quad \text{where} \quad b_i = \frac{a_i}{\log a_i^{-1}}. \tag{2.8}$$

PROPOSITION. *Assume* A1–A3 *hold. If there is a constant* c_0 *such that* $\|x^*\| < c_0$, $v(x^*) < \inf_{\|x\|} = c_0 v(x)$ *and if* $v(J)$ *is not dense in any interval, then* $\{x_k\}$ *defined by* (2.1)–(2.4) *converges to* J

$$\lim_{k\to\infty} d(x_k, J) = 0$$

whenever $\{\epsilon_i\}$ *satisfies the following condition:*

$$\lim_{T\to 0} \limsup_{k\to\infty} \frac{1}{T} \left\| \sum_{i=k}^{m(k,t)} a_i \epsilon_{i+1} \right\| = 0 \quad \forall t \in [0, T], \tag{2.9}$$

where

$$m(k, t) = \max\left\{ m : \sum_{i=k}^{m} a_i \le t \right\}.$$

Remark 1. An obvious condition which guarantees (2.9) is the convergence of the series

$$\sum_{i=1}^{\infty} a_i \epsilon_{i+1}.$$

Condition (2.9) is also necessary for convergence of x_n to the root of $f(x)$. This is discussed in the recent paper [8], which also shows that (2.9) is equivalent to the

standard Kushner–Clark condition [3]. However, when ϵ_k depends on $\{x_0, \ldots, x_{k-1}\}$, it is difficult to directly verify (2.9). In [4, 6] it is shown that it suffices to verify (2.9) not along the whole sequence $\{k\}$ but along the subsequence $\{n_k\}$ whenever $\{x_{n_k}\}$ converges. In [4, 6] it is also demonstrated that this verification can be done in many practically important problems.

Remark 2. If $\{x_k\}$ given by (1.1), (1.2) is a priori known to be bounded, then under conditions A1–A3 and (2.9)

$$\lim_{k\to\infty} d(x_n, J) = 0; \tag{2.10}$$

i.e., in this case the truncations introduced in (2.1)–(2.4) are not necessary.

In A4 the matrix H is stable. By the Lyapunov equality there is a positive definite matrix $P > 0$ such that

$$PH + H^\tau P = -I. \tag{2.11}$$

Denote by $\lambda_{\max}$ and $\lambda_{\min}$ the maximum and minimum eigenvalue of P, respectively, and by K the condition number $\lambda_{\max}/\lambda_{\min}$.

THEOREM. (i) *If conditions* A1–A5 *are satisfied and* x^0 *is the unique root of* $f(\cdot)$*, then for* $\{x_n\}$ *defined by* (2.1)–(2.4)

$$\limsup_{n\to\infty} (\log a_n^{-1})^{\frac{1}{\gamma}} \|x_n - x^0\| \le \sqrt{K} \left(\frac{2q\lambda_{\max}}{\gamma}\right)^{\frac{1}{\gamma}} \tag{2.12}$$

if $\{\epsilon_i\}$ *satisfies the following condition*:

$$\sum_{i=1}^{\infty} a_i (\log a_{i+1}^{-1})^{\frac{1}{\gamma}} \epsilon_{i+1} < \infty, \tag{2.13}$$

where γ *and* q *are given by* (2.5), (2.7), *respectively.*

(ii) *If, in addition,* H *is symmetric, then*

$$\limsup_{n\to\infty} (\log a_n^{-1})^{\frac{1}{\gamma}} \|x_n - x^0\| \le \left(\frac{q}{\lambda_l \gamma}\right)^{\frac{1}{\gamma}}, \tag{2.14}$$

where λ_l *is the smallest eigenvalue of* $-H$ *and* γ *and* q *are given by* (2.5), (2.7), *respectively.*

(iii) *Further, in the one-dimensional case, i.e.,* $l = 1$*, under the conditions stated in* (i) *except A3, the upper bound in* (2.14) *is attainable if* $q_n \to q > 0$.

The proof of the theorem is given in section 3.

Remark 3. From the theorem it is seen that the convergence rate of $(x_n - x^0)$ depends upon the decreasing rate of a_n. However, it is interesting to note that this dependence in the degenerate case is completely different from that in the nondegenerate case.

From (1.5) it is seen that for the nondegenerate case, if $a_n = \frac{1}{n^\alpha}$, $0 < \alpha \le 1$, then the convergence rate of $(x_n - x^0)$ is improving as α increases from 0 to 1. However, in the degenerate case the picture is different. By the theorem, $\lim_{n\to\infty} (\alpha \log n)^{\frac{1}{\gamma}} |x_n - x^0|$ equals 0 for all $\alpha \in (0, 1)$, while it may attain $(\frac{1}{|H|\gamma})^{\frac{1}{\gamma}}$ if $\alpha = 1$. This means that in contrast to the nondegenerate case, the convergence rate of $|x_n - x^0|$ for $\alpha \in (0, 1)$ is

better than that for $\alpha = 1$. This fact is verified by simulation of the following simple example:

$$f(x) = -x|x|, \quad x^0 = 0, \quad f'(x^0) = 0, \quad \gamma = 1, \quad H = -1,$$
$$\epsilon_i \equiv 0,$$
$$x_{n+1}^{(1)} = x_n^{(1)} - \frac{x_n^{(1)}|x_n^{(1)}|}{n}, \quad x_0^{(1)} = 0.5,$$
$$x_{n+1}^{(2)} = x_n^{(2)} - \frac{x_n^{(2)}|x_n^{(2)}|}{\sqrt{n}}, \quad x_0^{(2)} = 0.5.$$

The simulation shows that

$$x_n^{(1)} \log n \xrightarrow[n \to \infty]{} 1, \quad \text{while} \quad x_n^{(2)} \log n \xrightarrow[n \to \infty]{} 0,$$

which are reconciled with results stated in the theorem.

It is also worth noting that the right-hand sides of (2.14) depend upon the smallest eigenvalue λ_l of $-H$ when H is symmetric. As λ_l decreases the upper bound in (2.14) increases. In other words, the faster $f(x)$ leaves the abscissa, the faster x_n converges to x^0. This phenomenon is consistent with the convergence rate change from (1.5) for the nondegenerate case to (2.12) for the degenerate case. This is also verified by computation: if in the example considered above "$H = -1$" is replaced by "$H = -\frac{1}{2}$", i.e., if $f(x) = -\frac{1}{2}x|x|$, then the recursion with $a_n = \frac{1}{n}$ becomes $x_{n+1} = x_n - \frac{x_n|x_n|}{2n}$, $\quad x_0 = 0.5$. The computation shows

$$x_n \log n \xrightarrow[n \to \infty]{} 2,$$

which is larger than the limit of $x_n^{(1)} \log n$.

Remark 4. In the case $q > 0$, the convergence rate given in the theorem cannot be improved. However, when $q = 0$, i.e., when a slowly decreasing gain is applied, we have only established $\|x_n - x^0\| = o(\log a_n^{-1})^{-\frac{1}{r}}$. The estimate may be not sharp, but the computation shows that $x_n^{(2)} \log n$ in Remark 3 converges to zero very slowly. This means that we should not expect a much faster rate than (2.15).

3. Order of estimation error. In this section we establish the order of estimation error when the estimation algorithm (2.1)–(2.4) is applied. As a matter of fact, we intend to show that $\|z_n\| \stackrel{\triangle}{=} \|(\log a_n^{-1})^{\frac{1}{\gamma}}(x_n - x_0)\|$ is bounded. This is an intermediate step toward proving the theorem which gives either upper bound or an exact limit of $\|z_n\|$.

LEMMA 1. *If* A5 *holds, then* (2.13) *implies* (2.9).

Proof. Let (2.13) be held. Setting

$$s_n = \sum_{i=1}^{n} a_i (\log a_{i+1}^{-1})^{\frac{1}{\gamma}} \epsilon_{i+1}, \quad s_0 = 0,$$

we have

$$\begin{aligned} \sum_{i=m}^{n} a_i \epsilon_{i+1} =& \sum_{i=m}^{n} (s_i - s_{i-1})(\log a_{i+1}^{-1})^{-\frac{1}{\gamma}} \\ =& s_n (\log a_{n+1}^{-1})^{-\frac{1}{\gamma}} + \sum_{i=m}^{n-1} s_i [(\log a_{i+1}^{-1})^{-\frac{1}{\gamma}} - (\log a_{i+2}^{-1})^{-\frac{1}{\gamma}}]. \end{aligned} \tag{3.1}$$

Since s_n converges, the first term on the right-hand side of (3.1) tends to zero as $n \to \infty$, while the last term is dominated by

$$\sup_{m \le i \le n} |s_i| \sum_{i=m}^{n-1} |(\log a_{i+1}^{-1})^{-\frac{1}{\gamma}} - (\log a_{i+2}^{-1})^{-\frac{1}{\gamma}}| = \sup_{m \le i \le n} |s_i| [(\log a_{m+1}^{-1})^{-\frac{1}{\gamma}} - (\log a_{n+1}^{-1})^{-\frac{1}{\gamma}}],$$

which tends to zero as $n \to \infty$ and $m \to \infty$.

Hence, $\sum_{i=1}^{\infty} a_i \epsilon_{i+1}$ converges and (2.9) holds by Remark 1. □

LEMMA 2. *Under the conditions stated in* (i) *or in* (iii) *of the theorem,* x_k *defined by* (2.1)–(2.4) *converges to* x^0 *as* $k \to \infty$.

Proof. By Lemma 1 and the Proposition presented in section 2, under the conditions stated in (i) we see $x_k \xrightarrow[k \to \infty]{} x^0$. For the one-dimensional case stated in (i) we also have $x_k \xrightarrow[k \to \infty]{} x^0$ if we can verify A3.

Since x^0 is the unique root of $f(\cdot)$ by A1, we have by A4

$$(x - x^0) f(x) < 0 \quad \forall x \neq x^0.$$

Then the function $v(x) = (x - x^0)^2$ satisfies A3. □

Define

$$z_n = (\log a_n^{-1})^{\frac{1}{\gamma}} (x_n - x^0), \tag{3.2}$$

$$h(z) = Hz\|z\|^{\gamma} + \frac{q + \Delta}{\gamma} z, \quad z \in R^l, \quad \Delta > 0. \tag{3.3}$$

LEMMA 3 (key lemma). *Under the conditions stated in* (i) *of the theorem,* $\{z_n\}$ *is bounded if* (2.13) *holds.*

Proof. To prove boundedness of $\{z_n\}$ we first express z_n in the recursive form. For any $\Delta > 0$ and sufficiently large n by (2.7), we have $q_n \le q + \Delta$ and

$$\begin{aligned}
\left(\frac{\log a_{n+1}^{-1}}{\log a_n^{-1}} \right)^{\frac{1}{\gamma}} &= \left(\frac{\log a_n^{-1} + \log \dfrac{a_{n+1}^{-1}}{a_n^{-1}}}{\log a_n^{-1}} \right)^{\frac{1}{\gamma}} \\
&= \left(1 + \frac{\log(1 + a_n q_n)}{\log a_n^{-1}} \right)^{\frac{1}{\gamma}} \\
&= \left(1 + \frac{a_n q_n + O(a_n^2)}{\log a_n^{-1}} \right)^{\frac{1}{\gamma}} \\
&= 1 + \frac{a_n (q + \Delta + o(1))}{\gamma \log a_n^{-1}}. \qquad (3.4)
\end{aligned}$$

By Lemma 2 $\{x_n\}$ is bounded, and hence x_k is defined by the Robbins–Monro algorithm starting from some n_0. Consequently, by (3.4) for $n \ge n_0$ we derive the recursive formula for $\{z_n\}$:

$$
\begin{aligned}
z_{n+1} =& \left(1+\frac{a_n}{\gamma\log a_n^{-1}}(q+\Delta+o(1))z_n\right)\\
&+\frac{a_n}{\log a_n^{-1}}\left(1+\frac{a_n}{\gamma\log a_n^{-1}}(q+\Delta+o(1))\right)(\log a_n^{-1})^{1+\frac{1}{\gamma}}\\
&\cdot[H(x_n-x^0)\|x_n-x^0\|^\gamma+r(x_n)]+a_n(\log a_{n+1}^{-1})^{\frac{1}{\gamma}}\epsilon_{n+1}\\
=&\left(1+\frac{a_n}{\gamma\log a_n^{-1}}(q+\Delta+o(1))\right)z_n\\
&+\frac{a_n}{\log a_n^{-1}}\left(1+\frac{a_n}{\gamma\log a_n^{-1}}(q+\Delta+o(1))\right)\left[Hz_n\|z_n\|^\gamma+\frac{\|z_n\|^{1+\gamma}r(x_n)}{\|x_n-x^0\|^{1+\gamma}}\right]\\
&+a_n(\log a_{n+1}^{-1})^{\frac{1}{\gamma}}\epsilon_{n+1}
\end{aligned}
$$

$$
=z_n+b_nh_n(z_n)+a_n(\log a_{n+1}^{-1})^{\frac{1}{\gamma}}\epsilon_{n+1} \tag{3.5}
$$

$$
=z_n+b_nH_nz_n+a_n(\log a_{n+1}^{-1})^{\frac{1}{\gamma}}\epsilon_{n+1}, \tag{3.6}
$$

where

$$
\begin{aligned}
h_n(z) =&\left(Hz\|z\|^\gamma+\frac{\|z\|^{1+\gamma}r(x_n)}{\|x_n-x^0\|^{1+\gamma}}\right)\left(1+\frac{a_n}{\gamma\log a_n^{-1}}(q+\Delta+o(1))\right)\\
&+\frac{q+\Delta+o(1)}{\gamma}z=H_nz
\end{aligned} \tag{3.7}
$$

and

$$
H_n=\left[\left(H+\frac{r(x_n)}{\|x_n-x^0\|^{1+\gamma}}\cdot\frac{z_n^\tau}{\|z_n\|}\right)(1+o(1))+\frac{q+\Delta+o(1)}{\gamma\|z_n\|^\gamma}\cdot I\right]\|z_n\|^\gamma. \tag{3.8}
$$

Assume that the converse is true, i.e., assume $\{\|z_n\|\}$ is unbounded.

Let us fix a large enough constant $c>1$ such that

$$
\frac{q+\Delta}{\gamma c^\gamma}\lambda_{\max}<\frac{1}{5}. \tag{3.9}
$$

Denote by $\{z_{l_i}\}$, $i=1,2,\dots,n_c$, those of $\{z_n, n\ge n_0\}$ for which $\|z_{l_i}\|\le c$ and $\|z_i\|>c\ \forall i: i\notin\{1,\dots,n_c\}$ where n_c may be infinite. For both cases $n_c<\infty$ and $n_c=\infty$ from the unboundedness of $\{\|z_n\|\}$ we will obtain a contradiction. This implies the conclusion of the lemma.

Case 1. If $n_c<\infty$, then $\|z_i\|>c\ \forall i\ge n_c$.

We now show that by selection (3.9) for c the difference equation (3.5) in Case 1 is asymptotically stable and $z_n\xrightarrow[i\to\infty]{}0$. This implies impossibility of $\|z_i\|>c\ \forall i\ge n_c$.

Define

$$
\Phi_{n,j}=(I+b_nH_n)(I+b_{n-1}H_{n-1})\cdots(I+b_jH_j),\quad \Phi_{j,j+1}\stackrel{\triangle}{=}I, \tag{3.10}
$$

$$
\Phi_{n,j}^\tau P\Phi_{n,j}=\Phi_{n-1,j}^\tau(P+b_n(H_n^\tau P+PH_n)+b_n^2H_n^\tau PH_n)\Phi_{n-1,j}, \tag{3.11}
$$

where H_n is defined by (3.8).

From A4 and Lemma 2, notice that $r(x_n)/\|x_n - x^0\|^{1+\gamma} \xrightarrow[n\to\infty]{} 0$ and for $n \geq n_c$,

$$b_n\|H_n^\tau P H_n\| \leq c_1 b_n \|z_n\|^{2\gamma} = c_1 \frac{a_n}{\log a_n^{-1}} (\log a_n^{-1})^2 \|x_n - x^0\|^{2\gamma} \xrightarrow[n\to\infty]{} 0, \tag{3.12}$$

where c_1 is a constant. Then by (2.11), (3.9), (3.12) for sufficiently large n we have

$$(H_n^\tau P + P H_n) + b_n H_n^\tau P H_n < -\frac{1}{2}\|z_n\|^\gamma I. \tag{3.13}$$

Without loss of generality we may assume that n_0 is large enough so that (3.13) is valid for $n \geq n_0$. By (3.11), (3.13) for $j \geq n_c$ we see that

$$\Phi_{n,j}^\tau P \Phi_{n,j} \leq \Phi_{n-1,j}^\tau \left(P - \frac{1}{2} b_n \|z_n\|^\gamma I\right) \Phi_{n-1,j} \leq \left(1 - \frac{b_n\|z_n\|^\gamma}{2\lambda_{\max}}\right) \Phi_{n-1,j}^\tau P \Phi_{n-1,j},$$

where as defined in section 2 $\lambda_{\max}$ is the maximum eigenvalue of P.

This implies that

$$\begin{aligned}\Phi_{n,n_c}^\tau P \Phi_{n,n_c} \leq & (1 - \mu b_n \|z_n\|^\gamma) \Phi_{n-1,n_c}^\tau P \Phi_{n-1,n_c} \\ < & e^{-\mu b_n \|z_n\|^\gamma} \Phi_{n-1,n_c}^\tau P \Phi_{n-1,n_c} < \lambda_{\max} e^{-\mu \sum_{i=n_c}^n b_i \|z_i\|^\gamma} I,\end{aligned}$$

where $\mu = \frac{1}{2\lambda_{\max}}$.

Consequently, we have

$$\|\Phi_{n,n_c}\| < \sqrt{K} e^{-\frac{\mu}{2} \sum_{i=n_c}^n b_i \|z_i\|^\gamma}. \tag{3.14}$$

We remind the reader that $K = \lambda_{\max}/\lambda_{\min}$ and $\lambda_{\min}$ is the minimum eigenvalue of P.

From (3.6) it follows that

$$z_{n+1} = \Phi_{n,n_c} z_{n_c} + \sum_{j=n_c}^n \Phi_{n,j+1} a_j (\log a_{j+1}^{-1})^{\frac{1}{\gamma}} \epsilon_{j+1}. \tag{3.15}$$

Since $\|z_i\| > c \ \forall i \geq n_c$ and $\sum_{i=n_c}^\infty b_i = \infty$, by (3.14) the first term on the right-hand side of (3.15) tends to zero as $n \to \infty$. Let us now estimate the last term of (3.15).

Set

$$\xi_n = \sum_{j=n_c}^n a_j (\log a_{j+1}^{-1})^{\frac{1}{\gamma}} \epsilon_{j+1}.$$

By (2.13) it follows that $\xi_n \xrightarrow[n\to\infty]{} \xi < \infty$. We now have

$$\begin{aligned}&\sum_{j=n_c}^n \Phi_{n,j+1} a_j (\log a_{j+1}^{-1})^{\frac{1}{\gamma}} \epsilon_{j+1} \\ &= \sum_{j=n_c}^n \Phi_{n,j+1} (\xi_j - \xi_{j-1}) \\ &= \xi_n - \sum_{j=n_c+1}^n (\Phi_{n,j+1} - \Phi_{n,j}) \xi_{j-1} - \Phi_{n,n_c+1} \xi_{n_c-1}\end{aligned}$$

$$
\begin{aligned}
&= \xi_n - \sum_{j=n_c+1}^{n} (\Phi_{n,j+1} - \Phi_{n,j})\xi \\
&\quad - \sum_{j=n_c+1}^{n} (\Phi_{n,j+1} - \Phi_{n,j})(\xi_{j-1} - \xi) - \Phi_{n,n_{n_c+1}}\xi_{n_c-1} \\
&= (\xi_n - \xi) + \Phi_{n,n_c+1}\xi - \sum_{j=n_c+1}^{n_1} (\Phi_{n,j+1} - \Phi_{n,j})(\xi_{j-1} - \xi) \\
&\quad + \sum_{j=n_1+1}^{n} \Phi_{n,j+1} b_j H_j(\xi_{j-1} - \xi) - \Phi_{n,n_{n_c+1}}\xi_{n_c-1}. \qquad (3.16)
\end{aligned}
$$

By (3.14) it is clear that on the right-hand side of (3.16) all terms except the second-to-last one tend to zero as $n \to \infty$ for any fixed n_1. We now show that the second-to-last term of (3.16) can be made arbitrarily small by choosing n_1 sufficiently large. For any $\epsilon > 0$, take sufficiently large n_1 such that

$$
|\xi_j - \xi| < \epsilon \quad \text{and} \quad 1 \geq \frac{\mu b_j \|z_j\|^\gamma}{2} \quad \forall j \geq n_1,
$$

which is possible because

$$
b_j \|z_j\|^\gamma = \frac{a_j}{\log a_j^{-1}} \cdot [(\log a_j^{-1})^{\frac{1}{\gamma}} \|x_j - x^0\|]^\gamma = a_j \|x_j - x^0\|^\gamma \to 0. \qquad (3.17)
$$

Using (3.14), (3.17), and noticing that $\|H_n\| \leq c_2 \|z_n\|^\gamma \ \forall n \geq n_c$ for some constant $c_2 > 0$ we derive

$$
\begin{aligned}
\left\| \sum_{j=n_1+1}^{n} \Phi_{n,j+1} b_j H_j(\xi_{j-1} - \xi) \right\| &\leq \epsilon c_2 \sqrt{K} \sum_{j=n_1+1}^{n} e^{-\frac{\mu}{2}\sum_{i=j+1}^{n} b_i \|z_i\|^\gamma} b_j \|z_j\|^\gamma \\
&\leq \frac{4\epsilon c_2 \sqrt{K}}{\mu} \sum_{j=n_1+1}^{n} e^{-\frac{\mu}{2}\sum_{i=j+1}^{n} b_i \|z_i\|^\gamma} (1 - e^{-\frac{\mu}{2} b_i \|z_j\|^\gamma}) \leq \frac{4\epsilon c_2 K}{\mu},
\end{aligned}
$$

where we use the fact that $\frac{x}{2} \leq 1 - e^{-x}$ for $x \in [0, 1]$.

Consequently, the left-hand side of (3.16) tends to zero as $n \to \infty$, and hence $z_n \xrightarrow[n \to \infty]{} 0$. This contradicts $\|z_i\| > c$, $\forall i \geq n_c$. Therefore, n_c must be ∞.

Case 2. Assume $n_c = \infty$. In this case $\{z_i\}$ will come back to the ball $\{\|z\| \leq c\}$ infinitely many times and at the same time $\{z_i\}$ is unbounded. From this we can conclude that $\{\|z_i\|\}$ crosses a nonempty interval infinitely often. To be precise, let $z_{l_i}^\tau P z_{l_i} \leq \lambda_{\max} c^2$, $i = 1, \ldots, n_c$, where P is given in (2.11). Starting from any z_{l_i}, $i \in \{1, 2, \ldots, n_c\}$, there exists an $m_i > l_i$ such that $z_{m_i}^\tau P z_{m_i} > 4c^2 \lambda_{\max}^2 / \lambda_{\min}$ since $\{\|z_n\|\}$ is unbounded. Further, noticing $n_c = \infty$ we can find an integer n_{i+1} in the set $\{l_i, i = 1, 2, \ldots, n_c\}$ such that $n_{i+1} > m_i$. This procedure can be continued infinitely many times. Without loss of generality, we may assume

$$
\begin{gathered}
z_{n_i}^\tau P z_{n_i} \leq \lambda_{\max} c^2, \quad z_{m_i}^\tau P z_{m_i} \geq 4c^2 \lambda_{\max}^2 / \lambda_{\min}, \\
\lambda_{\max} c^2 < z_j^\tau P z_j < 4c^2 \lambda_{\max}^2 / \lambda_{\min}, \\
n_i < j < m_i, \quad i = 1, 2, \ldots. \qquad (3.18)
\end{gathered}
$$

This implies the crossing property of $\{\|z_i\|\}$:

$$\|z_{n_i}\| \le \sqrt{K}c, \quad \|z_{m_i}\| \ge 2c\sqrt{K}, \quad c < \|z_j\| < 2cK, \tag{3.19}$$

$$n_i < j < m_i, \quad i = 1, 2, \ldots.$$

We now show that $\sum_{j=n_i}^{m_i-1} b_j \ge T > 0$ and $\|z_s - z_{n_i}\| = O(T)$ as $T \to 0$ for all large i and s: $\sum_{j=n_i}^{s} b_j \le T$. This implies a contradiction to (3.19). We now prove this in detail.

Noticing that there are constants c_3 and c_4 such that

$$\begin{aligned} b_n\|H_n z_n\| &\le \frac{a_n}{\log a_n^{-1}}[c_3\|z_n\|^{1+\gamma} + c_4\|z_n\|] \\ &\le \frac{a_n}{\log a_n^{-1}}[c_3(\log a_n^{-1})^{\frac{\gamma+1}{\gamma}}\|x_n - x^0\|^{1+\gamma} \\ &\quad + c_4(\log a_n^{-1})^{\frac{1}{\gamma}}\|x_n - x^0\|] \xrightarrow[n\to\infty]{} 0, \end{aligned} \tag{3.20}$$

by (2.13) from (3.6) we see that

$$z_{n+1} - z_n \xrightarrow[n\to\infty]{} 0. \tag{3.21}$$

Summing up both sides of (3.6) from n_i to m_i we derive

$$z_{m_i} = z_{n_i} + \sum_{j=n_i}^{m_i-1} b_j H_j z_j + \sum_{j=n_i}^{m_i-1} a_j (\log a_{j+1}^{-1})^{\frac{1}{\gamma}} \epsilon_{j+1}. \tag{3.22}$$

From (3.22) using (3.18), (3.19), (3.20) we obtain

$$\begin{aligned} 2c\sqrt{K} &\le \sqrt{K}c + \sum_{j=n_i}^{m_i-1} b_j(c_3\|z_j\|^{1+\gamma} + c_4\|z_j\|) + \left\|\sum_{j=n_i}^{m_i-1} a_j(\log a_{j+1}^{-1})^{\frac{1}{\gamma}}\epsilon_{j+1}\right\| \\ &\le \sqrt{K}c + \sum_{j=n_i}^{m_i-1} b_j(c_3(2cK)^{1+\gamma} + c_4 2cK) + \left\|\sum_{j=n_i}^{m_i-1} a_j(\log a_{j+1}^{-1})^{\frac{1}{\gamma}}\epsilon_{j+1}\right\|. \end{aligned} \tag{3.23}$$

Since (2.13) the last term of (3.23) can be made arbitrarily small, say, less than $\epsilon\,(< \sqrt{K}c)$ if i is sufficiently large. Then from (3.23) it follows that

$$\sum_{j=n_i}^{m_i-1} b_j \ge \frac{\sqrt{K}c - \epsilon}{c_3(2cK)^{1+\gamma} + 2cc_4K} \triangleq T > 0$$

for all large enough i. This means that $l(n_i, T) \le m_i - 1$, where

$$l(n,t) = \max\left\{l : \sum_{i=n}^{l} b_i \le t\right\}, \quad b_i = \frac{a_i}{\log a_i^{-1}}. \tag{3.24}$$

Consequently, we have

$$\|z_j - z_{n_i}\| \le \left\| \sum_{s=n_i}^{j-1} b_s H_s z_s + \sum_{s=n_i}^{j-1} a_s (\log a_{s+1}^{-1})^{\frac{1}{\gamma}} \epsilon_{s+1} \right\|$$

$$\le (c_3(2cK)^{1+\gamma} + 2c_4 cK)T + o(1) \le \alpha T, \quad \alpha > 0, \tag{3.25}$$

$$\forall j \in [n_i, \dots, l(n_i, T)],$$

where α is a constant

Therefore, by Taylor's formula there exists $\widetilde{z} \in R^l$ such that

$$\|\widetilde{z} - z_{n_i}\| \le \alpha T \tag{3.26}$$

and

$$\begin{aligned} z_{l(n_i,T)}^\tau P z_{l(n_i,T)} - z_{n_i}^\tau P z_{n_i} =& \widetilde{z}^\tau P \left(\sum_{j=n_i}^{l(n_i,t)} b_j H_j z_j + \sum_{j=n_i}^{l(n_i,T)} a_j (\log a_{j+1}^{-1})^{\frac{1}{\gamma}} \epsilon_{j+1} \right) \\ =& \sum_{j=n_i}^{l(n_i,t)} b_j z_j^\tau P H_j z_j + \sum_{j=n_i}^{l(n_i,T)} b_j (\widetilde{z} - z_j)^\tau P H_j z_j \\ &+ \widetilde{z}^\tau P \sum_{j=n_i}^{l(n_i,T)} a_j (\log a_{j+1}^{-1})^{\frac{1}{\gamma}} \epsilon_{j+1}. \end{aligned} \tag{3.27}$$

Using (3.20), (3.25), and (3.26) we see that

$$\left\| \sum_{j=n_i}^{l(n_i,T)} b_j (\widetilde{z} - z_j)^\tau P H_j z_j \right\| \le 2\alpha T^2 \lambda_{\max}(c_3(2cK)^{1+\gamma} + c_4 2cK).$$

By (2.13), the last term of (3.27) can be made arbitrarily small. Hence, by (3.12), (3.13), (3.19), (3.21) we have

$$z_j^\tau P H_j z_j = \frac{1}{2} z_j^\tau (P H_j + H_j^\tau P) z_j < -\frac{1}{4} \|z_j\|^{2+\gamma} \le -\frac{1}{4} c^{2+\gamma} \quad \forall j = n_i, \dots, l(n_i, T).$$

From (3.27) we can conclude that

$$\begin{aligned} & z_{l(n_i,T)}^\tau P z_{l(n_i,T)} - z_{n_i}^\tau P z_{n_i} \\ \le & -\frac{1}{4} c^{2+\gamma} T + 2\alpha T^2 (c_3 (2cK)^{1+\gamma} + 2cc_4 K) + o(1) \le -\frac{1}{5} c^{2+\gamma} T, \end{aligned} \tag{3.28}$$

if i is large enough and T is sufficiently small. By (3.18) inequality (3.28) implies that

$$\lambda_{\max} c^2 < z_{l(n_i,T)}^\tau P z_{l(n_i,T)} \le z_{n_i}^\tau P z_{n_i} - \frac{1}{5} c^{2+\gamma} T \to \lambda_{\max} c^2 - \frac{c^{2+\gamma} T}{5},$$

which is impossible. The obtained contradiction shows that $\{z_n\}$ is bounded. □

4. Proof of the theorem. We are now in a position to prove our theorem.

Proof of the theorem. (i) For assertion (2.12) of the theorem it suffices to show

$$\limsup_{n\to\infty} z_n^\tau P z_n \le \lambda_{\max}\left(\frac{2(q+\Delta)\lambda_{\max}}{\gamma}\right)^{\frac{2}{\gamma}} \triangleq a \tag{4.1}$$

for arbitrarily small $\Delta > 0$. Let us fix $\Delta > 0$.

The idea of proof for (4.1) is that we show that $z_n^\tau P z_n$ crosses a nonempty interval infinitely often if (4.1) is not true and, at same time, $z_n^\tau P z_n$ is decreasing in a certain sense. This contains a contradiction.

By Lemma 3 $\{z_n\}$ is bounded; i.e.,

$$\|z_n\| \le \zeta < \infty \quad \forall n. \tag{4.2}$$

Hence, from (3.8) we see that

$$H_n = H\|z_n\|^\gamma + \frac{q+\Delta}{\gamma} I + o(1). \tag{4.3}$$

From (3.3), (3.6), and (4.3) it follows that

$$z_{n+1} = z_n + b_n h(z_n) + b_n o(1) + a_n(\log a_{n+1}^{-1})^{\frac{1}{\gamma}}\epsilon_{n+1}. \tag{4.4}$$

Fix any small $\epsilon > 0$ consider $z \in R^l$ for which

$$z^\tau P z \ge \lambda_{\max}\left(\frac{2(q+\Delta)\lambda_{\max}+\epsilon}{\gamma}\right)^{\frac{2}{\gamma}} \triangleq b. \tag{4.5}$$

This implies that

$$\|z\| \ge \left(\frac{2(q+\Delta)\lambda_{\max}+\epsilon}{\gamma}\right)^{\frac{1}{\gamma}}.$$

Then by (2.11) and (4.4) we have

$$\begin{aligned} & z^\tau\left((PH + H^\tau P)\|z\|^\gamma + \frac{2(q+\Delta)}{\gamma}P\right)z \\ & \le\ z^\tau\left(-\|z\|^\gamma I + \frac{2(q+\Delta)\lambda_{\max}}{\gamma} I\right) z \le -\frac{\epsilon}{\gamma}\|z\|^2. \end{aligned} \tag{4.6}$$

Assume (4.1) is not true. Then there is a small $\delta > 0$ such that

$$\limsup_{n\to\infty} z_n^\tau P z_n > a + \delta. \tag{4.7}$$

Therefore, there is a subsequence

$$z_{n_k}^\tau P z_{n_k} > a + \delta, \quad k = 1, 2, \ldots. \tag{4.8}$$

Let $\epsilon > 0$ be small enough so that

$$a + \delta > b, \tag{4.9}$$

where b is given by (4.5).

We now show that from any n_k, $k = 1, 2, \ldots$, $\{z_n\}$ will enter the ellipsoid $\{z : z^\tau P z \le b\}$. Assume the converse, i.e.,

$$\|z_i^\tau P z_i\| \ge b \quad \forall i \ge n_k \tag{4.10}$$

for some n_k.

By the boundedness of $\{z_n\}$ we have

$$\|z_j - z_n\| \le c_5 T, \quad j = n, \ldots, l(n,T) \quad \forall n, \tag{4.11}$$

and hence

$$|z_j^\tau P z_j - z_n^\tau P z_n| \le c_6 T, \quad j = n, \ldots, l(n,T) \quad \forall n, \tag{4.12}$$

where c_5 and c_6 are constants. Similar to (3.27), by (4.2), (4.11) we obtain

$$\begin{aligned} z_{l(n_k,T)}^\tau P z_{l(n_k,T)} - z_{n_k}^\tau P z_{n_k} &\le \sum_{j=n_k}^{l(n_k,T)} b_j z_j^\tau P H_j z_j + c_7 T^2 + o(1) \\ &= \sum_{j=n_k}^{l(n_k,T)} b_j z_j^\tau \left(PH \left(\|z_j\|^\gamma + \frac{q+\Delta}{\gamma} P \right) \right) z_j + c_7 T^2 + o(1). \end{aligned} \tag{4.13}$$

Using (2.11) leads to

$$z_{l(n_k,T)}^\tau P z_{l(n_k,T)} - z_{n_k}^\tau P z_{n_k} = \frac{1}{2} \sum_{j=n_k}^{l(n_k,T)} b_j z_j^\tau \left(-\|z_j\|^\gamma I + \frac{2(q+\Delta)}{\gamma} P \right) z_j + c_7 T^2 + o(1).$$

From this by (4.6), (4.10) we obtain

$$\begin{aligned} z_{l(n_k,T)}^\tau P z_{l(n_k,T)} - z_{n_k}^\tau P z_{n_k} &< -\frac{1}{2} \sum_{j=n_k}^{l(n_k,T)} b_j \frac{\epsilon}{2\gamma} \|z_j\|^2 + c_7 T^2 + o(1) \\ &< -\frac{\epsilon}{2\gamma} \left(\frac{2(q+\Delta)\lambda_{\max} + \epsilon}{\gamma} \right)^{\frac{2}{\gamma}} T + c_7 T^2 + o(1) \\ &\le -\frac{\epsilon}{3\gamma} \left(\frac{2(q+\Delta)\lambda_{\max} + \epsilon}{\gamma} \right)^{\frac{2}{\gamma}} T \end{aligned} \tag{4.14}$$

for sufficiently small T and large enough k. This means that after a finite number of steps (4.10) will not be satisfied; i.e., z_n will enter the ellipsoid $\{z : z^\tau P z \le b\}$. This together with (4.8) implies that $\{z_n^\tau P z_n\}$ will cross the interval $[b, a+\delta]$ infinitely often; i.e., there are two subsequences $\{z_{l_k}\}$ and $\{z_{m_k}\}$ such that

$$z_{l_k}^\tau P z_{l_k} \le b, \qquad z_{m_k}^\tau P z_{m_k} \ge a + \delta,$$

$$b < z_i^\tau P z_i < a + \delta \quad \forall i : l_k < i < m_k.$$

Take T sufficiently small such that $c_6 T < a + \delta - b$. Then by (4.12) we see $l(l_i, T) < m_i$ $\forall i$ and

$$z_{l(l_k,T)}^\tau P z_{l(l_k,T)} \in (b, a+\delta),$$

which combined with (4.14) leads to a contradiction:

$$0 < z_{l(l_k,T)}^{\tau} P z_{l(l_k,T)} - z_{l_k}^{\tau} P z_{l_k} \le -\frac{\epsilon}{3\gamma} \left(\frac{2(q+\Delta)\lambda_{\max} + \epsilon}{\gamma} \right)^{\frac{2}{\gamma}} T.$$

Therefore, (4.8) is impossible or (4.7) is impossible. Since δ may be arbitrarily small, the impossibility of (4.7) implies (4.1). Tending Δ to zero, from (4.1) we derive (2.12).

(ii) Now, let H be symmetric. We simply consider $\|z\|^2$ instead of $z^{\tau}Pz$ and set in (4.1) and (4.5)

$$a = \left(\frac{q+\Delta}{\lambda_l \gamma} \right)^{\frac{2}{\gamma}}, \quad b = \left(\frac{q+\Delta+\epsilon}{\lambda_l \gamma} \right)^{\frac{2}{\gamma}}.$$

The proof can be carried out along the lines of that given for the general case. For example, corresponding to (4.5), (4.6) we now have

$$\|z\|^2 \ge b, \text{ and } z^{\tau} \left(H\|z\|^{\gamma} + \frac{q+\Delta}{\gamma} I \right) z \le z^{\tau} \left(-\lambda_l \frac{q+\Delta+\epsilon}{\lambda_l \gamma} + \frac{q+\Delta}{\gamma} \right) z = -\frac{\epsilon}{\gamma} \|z\|^2,$$

respectively, while (4.13) becomes

$$\|z_{l(n_k,T)}\|^2 - \|z_{n_k}\|^2 \le -\sum_{j=n_k}^{l(n_k,T)} b_j z_j^{\tau} \left(H \left(\|z_j\|^{\gamma} + \frac{q+\Delta}{\gamma} I \right) z_j + c_7 T^2 + o(1) \right).$$

(iii) Since $q > 0$, we may set $\Delta = 0$ in (3.3) and in the proofs of Lemma 3 and part (i) of the theorem.

In the one-dimensional case H in (2.5) is a negative number, and λ_l in (2.14) equals $|H|$. The root set of $h(z)$ defined by (3.3) with $\Delta = 0$ is $J = \{0, \pm(\frac{q}{-H\gamma})^{\frac{1}{\gamma}}\}$.

It is easy to define a twice differentiable function $v(z)$ such that

$$v(z) = v(-z),\ 0 < v(z) < v(0) \quad \forall z : |z| \le \zeta,$$

$$v'(z)h(z) < 0 \quad \forall z \notin \left\{ 0, \pm \left(\frac{q}{-H\gamma} \right)^{\frac{1}{\gamma}} \right\},$$

where ζ is given in (4.2).

For $\forall t \le T$ we find that

$$\lim_{T\to 0} \limsup_{k\to\infty} \frac{1}{T} \left\| \sum_{i=k}^{l(k,t)} (b_i o(1) + a_i (\log a_{i+1}^{-1})^{\frac{1}{\gamma}} \epsilon_{i+1}) \right\|$$

$$\le \lim_{T\to 0} \limsup_{k\to\infty} \frac{1}{T} \{t \cdot o(1)\} + \lim_{T\to 0} \limsup_{k\to\infty} \frac{1}{T} \left\| \sum_{i=k}^{l(k,t)} a_i (\log a_{i+1}^{-1})^{\frac{1}{\gamma}} \epsilon_{i+1} \right\|$$

$$= 0. \tag{4.15}$$

Applying Remark 2 in section 2 to (4.4) leads to

$$\lim_{k\to\infty} d(z_k, J) = 0.$$

This is valid for any $\{\epsilon_i\}$ satisfying (4.15). In particular, if $\{\epsilon_i\}$ is such that $b_i o(1) + a_i(\log a_{i+1}^{-1})^{\frac{1}{\gamma}}\epsilon_{i+1} = 0\ \forall i \geq 1$, then (4.4) becomes the following recursion:

$$z_{n+1} = z_n + b_n h(z_n). \tag{4.16}$$

Note that $h'(0) = \frac{q}{\gamma} > 0$, and hence 0 is not stable for the equation

$$\dot{z}_t = h(z_t).$$

It is clear that 0 cannot be the limit point of (4.16).

Therefore, in this case z_n can converge either to $(\frac{q}{|H|\gamma})^{\frac{1}{\gamma}}$ or to $-(\frac{q}{|H|\gamma})^{\frac{1}{\gamma}}$. This verifies the attainability of the upper bound in (2.14). □

5. Concluding remarks. By using a deterministic analysis we have shown the pathwise convergence rate of SA when $f(x^0) = 0$ and $f'(x^0) = 0$. Some problems are still open and belong to further research. First, it might be possible to obtain more precise results. For example, as a conjecture, the limit of the left-hand side of (2.14) is one of $(\frac{q}{\lambda_i\gamma})^{\frac{1}{\gamma}}$, $i = 1, \ldots, l$, depending upon the initial value where λ_i, $i = 1, \ldots, l$ are the eigenvalues of H. Second, it is not clear what happens if $f(\cdot)$ has more complicated behavior as $x \to x^0$.

REFERENCES

[1] H. Robbins and S. Monro, *A stochastic approximation method*, Ann. Math. Statist., 22 (1951), pp. 400–407.

[2] M. B. Nevelson and R. Z. Hasminskii, *Stochastic Approximation and Recursive Estimation*, Amer. Math. Soc. Transl. Math. Monographs 47, 1976.

[3] H. J. Kushner and D. S. Clark, *Stochastic Approximation for Constrained and Unconstrained Systems*, Springer-Verlag, New York, 1978.

[4] H.-F. Chen, *Stochastic approximation and its new applications*, in Proc. 1994 Hong Kong International Workshop on New Directions of Control and Manufacturing, 1994, pp. 2–12.

[5] L. Ljung, G. Pflug, and H. Walk, *Stochastic Approximation of Random Systems*, Birkhäuser, Basel, 1992, pp. 71–76.

[6] H.-F. Chen, T. Duncan, and B. Pasik-Duncan, *On Ljung's approach to system parameter identification*, in 10th IFAC Symposium on Systems Identification, Vol. 2, preprint, M. Blanke and T. Söderström, eds., Copenhagen, 1994, pp. 667–671.

[7] H.-F. Chen, *Recursive Estimation and Control for Stochastic Systems*, John Wiley, New York, 1985.

[8] I. J. Wang, E. K. P. Chong, and S. R. Kulkarni, *Equivalent necessary and sufficient conditions on noise sequences for stochastic approximation algorithms*, Adv. in Appl. Probab., accepted for publication.

[9] S. R. Kulkarni and C. Horn, *Convergence of the Robbins–Monro algorithm under arbitrary disturbances*, in Proc. of 32nd Conf. on Decision and Control, 1993, pp. 537–538.

[10] S. R. Kulkarni and C. Horn, *Alternative approach and conditions for convergence of stochastic approximation algorithms*, in Proc. 34th Conf. on Decision and Control, New Orleans, IEEE Control Systems Society, 1995.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 1, pp. 115–136, January 1998

006

THE VALUE FUNCTION OF THE SINGULAR QUADRATIC REGULATOR PROBLEM WITH DISTRIBUTED CONTROL ACTION*

FRANCESCA BUCCI† AND LUCIANO PANDOLFI‡

Abstract. We study the regularity properties of the value function of a quadratic regulator problem for a linear distributed parameter system with distributed control action. No definiteness assumption on the cost functional is assumed. We study the regularity in time of the value function and also the space regularity in the case of a holomorphic semigroup system.

Key words. value function, quadratic regulator, distributed systems

AMS subject classification. 49J20

PII. S036301299529536X

1. Introduction. In this paper we are concerned with a general class of finite horizon linear-quadratic optimal control problems for evolution equations with distributed control and *nondefinite* cost. More precisely, we consider the following abstract differential equation over a finite interval $[\tau, T]$, $0 \le \tau < T < +\infty$:

$$\dot{x} = Ax + Bu, \qquad x(\tau) = x_0 \in X, \tag{1.1}$$

where A is the infinitesimal generator of a strongly continuous semigroup e^{At} on a Hilbert space X, B is a linear bounded operator from the control space U to X. With the dynamics (1.1), we associate the cost functional

$$J_\tau(x_0, u) = \int_\tau^T F(x(t), u(t))dt + \langle x(T), P_0\, x(T)\rangle, \tag{1.2}$$

where $x(\cdot) = x(\cdot, \tau, x_0, u)$ is the mild solution to equation (1.1) and F is the quadratic form

$$F(x, u) = \langle x, Qx\rangle + \langle x, Su\rangle + \langle Su, x\rangle + \langle u, Ru\rangle \tag{1.3}$$

(we denoted by $\langle\cdot,\cdot\rangle$ inner products in both the spaces X and U). All the operators Q, S, R, and P_0 contained in the functional (1.2) are linear bounded operators in the proper spaces, with $Q = Q^*$, $R = R^*$, $P_0 = P_0^*$. We define as usual the *value function* of the problem:

$$V(\tau, x_0) := \inf_{u \in L^2(\tau, T; U)} J_\tau(x_0, u).$$

The goal of this work is

- to characterize the property

$$V(\tau, x_0) > -\infty \quad \forall x_0 \in X, \quad \forall \tau \in [0, T]; \tag{1.4}$$

*Received by the editors November 29, 1995; accepted for publication (in revised form) October 14, 1996. This research was supported by the Italian Ministero dell'Università e della Ricerca Scientifica e Tecnologica within the program of GNAFA–CNR. The second author was also partially supported by HCM network CEC n. ERB–CHRX–CT93–0402.

http://www.siam.org/journals/sicon/36-1/29536.html

†Dipartimento di Matematica Applicata "G. Sansone," Università di Firenze, Via S. Marta 3, 50139 Firenze, Italy (fbucci@dma.unifi.it).

‡Dipartimento di Matematica, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy (lucipan@polito.it).

- to study the regularity properties of the map

$$\tau \to V(\tau, x_0)$$

on the interval $[0, T]$, when x_0 is fixed.

We shall consider also the map $x_0 \to V(\tau, x_0)$ in a special case; see section 6. It is well known that if the regulator problem is *standard*, i.e.,

$$Q \geq 0, \quad S = 0, \quad R \geq \alpha > 0, \quad P_0 \geq 0, \tag{1.5}$$

then the solution to the operator Riccati equation corresponding to problem (1.1)–(1.2) provides the synthesis of the unique optimal control. This problem is well understood, both in finite and in infinite dimensions, over a finite or infinite time horizon (compare [10], [2], [3]).

The purpose of this paper is to examine the case when (1.5) fails, with special interest in noncoercive R. We shall see that in this case the function $\tau \to V(\tau, x_0)$ has some mild regularity properties; see section 4. More regularity is obtained in the coercive case; see section 5.

The study of linear quadratic regulator (LQR) problems with nondefinite cost is related to a large variety of problems. Among them, we recall the study of *dissipative* systems (see [20]), the analysis of the stability of feedback systems [14], the analysis of second variations of nonlinear optimization problems (see [5], [15]). When game theory is studied for linear systems then the quadratic form (1.3) is nonpositive. In particular, the suboptimal H^∞-problem can be recast in this setting [1]. Finally, very recently, singular control theory has been used to obtain new results on *regular* control problems for some class of boundary control systems: systems with input delays first [16], and later systems described by wave- or plate-like equations with high internal damping [9].

We recall that the existing results for finite-dimensional systems over an infinite time interval ([19], [21]; see also [4]) were extended to distributed systems in [22], [23], [12], [13], [8]. If $T < +\infty$, the only work we know in an infinite-dimensional context, in which a nonpositive cost functional is studied, is [6]. This paper considers even time-varying systems but under the restriction $R = I$.

2. A simple example. The interest of the results presented in this paper is justified by the possible applications that we already quoted, for instance to H^∞-control theory over a finite time interval, or to the analysis of the second variation of general cost functionals. However, the following example may help the reader to understand our problem. The example is a bit artificial, since we want to present a very simple one. Nevertheless it is suggested by nontrivial problems in network theory.

A delay line in its simpler form is described by an input-output relation,

$$v(t, x) = v(t + x) = \int_{-1}^{0} u(t + s + x)\, d\eta(s), \tag{2.1}$$

where $t > 0$, $x \in (-1, 0)$ and the integral is a Stieltjes integral.

For simplicity we assume that the input $u(\cdot)$ is continuous, a condition that can be very much relaxed.

The simplest case described by (2.1) is

$$v(t, x) = u(t + x - 1) \tag{2.2}$$

and corresponds to a jump function η, with jump at -1. If the system is started at $t = 0$ then the input (2.1) is read only for $t > 0$, so that the output $v(\cdot)$ from (2.1) is given by

$$v(t,x) = \begin{cases} \phi(x+t), & -1 < x+t < 0, \\ \int_{-1}^{0} u(t+s+x)\, d\eta(s) & \text{otherwise.} \end{cases}$$

The function ϕ describes the "initial state" of the system (quite often it will be $\phi = 0$). In the case of equation (2.2) we have in particular

$$v(t,x) = \begin{cases} u(t+x-1), & t+x > 1, \\ \phi(t+x-1), & t+x < 1. \end{cases}$$

Notice that if $\phi(\cdot)$ and $u(\cdot)$ are regular then $v(t,x) = v(t+x)$ solves the first-order hyperbolic equation

$$v_t = v_x, \qquad v(0,x) = \phi(x), \quad v(t,0) = u(t-1).$$

The function v can be interpreted as a delayed potential at the output of the network produced by the potential $u(\cdot)$ at the input. If the delay line is connected to a resistive load, it produces a current $i(t) = \frac{1}{R} v(t-1)$, and the energy dissipated by the load in time T is given by

$$-\int_0^T i(t)v(t-1)\, dt = -\int_0^T \frac{1}{R}|v(t-1)|^2\, dt.$$

Since

$$v(t,-1) = \begin{cases} \phi(t-1), & 0 < t < 1, \\ u(t-1), & t > 1, \end{cases}$$

then

$$-\int_0^T i(t)v(t-1)\, dt = \begin{cases} -\int_0^T \frac{1}{R}|\phi(t-1)|^2\, dt & \text{if } T < 1, \\ -\int_0^1 \frac{1}{R}|\phi(t-1)|^2\, dt - \int_1^T \frac{1}{R}|u(t-1)|^2\, dt & \text{if } T > 1. \end{cases}$$

The energy that the load can dissipate is at most

$$\inf_{u(\cdot)} -\int_0^T i(t)v(t-1)\, dt.$$

We see from this that the load dissipates a finite amount of energy $V(\phi)$ if $T < 1$, described by the quadratic functional

$$V(\phi) = -\int_0^T \frac{1}{R}|\phi(t-1)|^2\, dt. \tag{2.3}$$

Otherwise, the load can dissipate as much energy as we want.

Hence it makes sense to study the energy function $E(T)$:

$$E(T) = \inf_{u \in L^2(0,T)} -\int_0^T i(t)v(t-1)\, dt.$$

In this example the function $E(T)$ is finite only if $T < 1$, and in this case $E(T)$ is the quadratic functional (2.3).

In this paper we consider an analogous problem in more generality: we study the dependence on the interval $[\tau, T]$ of the "energy" dissipated by a certain linear time invariant system.

3. Preliminary results. We recall that the solution to (1.1) is

$$x(t) = e^{A(t-\tau)}x_0 + (L_\tau u)(t), \tag{3.1}$$

with

$$(L_\tau u)(t) = \int_\tau^t e^{A(t-s)}Bu(s)\,ds,$$

$$: \text{ continuous } L^2(\tau,T;U) \to L^2(\tau,T;X). \tag{3.2}$$

Note that $t \to (L_\tau u)(t)$ is an X-valued continuous function.

The adjoint L_τ^* of L_τ: $\langle L_\tau u, f\rangle_{L^2(\tau,T;X)} = \langle u, L_\tau^* f\rangle_{L^2(\tau,T;U)}$ is given by

$$(L_\tau^* f)(t) = B^* \int_t^T e^{A^*(s-t)}\, f(s)\,ds,$$

$$: \text{ continuous } L^2(\tau,T;X) \to L^2(\tau,T;U).$$

Introduce also the bounded operator from U to X:

$$L_{\tau,T}u = \int_\tau^T e^{A(T-s)}Bu(s)\,ds$$

(which describes the map (3.1) from the input u to the solution of (1.1) at time $t = T$, with initial time τ and $x_0 = 0$). The adjoint of $L_{\tau,T}$ is the map given by

$$(L_{\tau,T}^* y)(t) = B^*\, e^{A^*(T-t)}y.$$

Using (3.1), one can easily show the following lemma.

LEMMA 3.1. *The cost functional* (1.2) *can be rewritten as*

$$J_\tau(x_0,u) = \langle \mathcal{M}_\tau x_0, x_0\rangle + 2\,Re\,\langle \mathcal{N}_\tau x_0, u\rangle + \langle \mathcal{R}_\tau u, u\rangle, \tag{3.3}$$

with $\mathcal{M}_\tau \in \mathcal{L}(X)$, $\mathcal{N}_\tau \in \mathcal{L}(X, L^2(\tau,T;U))$, *and* $\mathcal{R}_\tau \in \mathcal{L}(L^2(\tau,T;U))$, $\mathcal{M}_\tau$ *and* $\mathcal{R}_\tau$ *self-adjoint, defined as follows:*

$$\mathcal{M}_\tau x = e^{A^*(T-\tau)}\,P_0\,e^{A(T-\tau)}x + \int_\tau^T e^{A^*(t-\tau)}Q\,e^{A(t-\tau)}x\,dt, \tag{3.4}$$

$$(\mathcal{N}_\tau x)(t) = (L_\tau^* Q\, e^{A(\cdot-\tau)}x)(t) + S^* e^{A(t-\tau)}x + (L_{\tau,T}^* P_0\, e^{A(T-\tau)}x)(t), \tag{3.5}$$

$$(\mathcal{R}_\tau u)(t) = (L_\tau^* Q\,(L_\tau u))(t) + S^*(L_\tau u)(t) + (L_\tau^* S u)(t)$$

$$+ Ru(t) + (L_{\tau,T}^* P_0\, L_{\tau,T} u)(t). \tag{3.6}$$

We first state a lemma, which will be useful later.

LEMMA 3.2. *If there exists* τ_0 *and a constant* γ *such that*

$$\mathcal{R}_{\tau_0} \geq \gamma\, I, \tag{3.7}$$

then $\mathcal{R}_\tau \geq \gamma\, I$ *for any* $\tau > \tau_0$.

Proof. It is sufficient to notice that if $\tau > \tau_0$ we can write

$$\langle \mathcal{R}_\tau u, u\rangle_{L^2(\tau,T;U)} = \langle \mathcal{R}_{\tau_0} v, v\rangle_{L^2(\tau_0,T;U)},$$

where $v(\cdot)$ is given by $v(t) = 0$ if $\tau_0 \le t < \tau$ and $u(t)$ when $t \ge \tau$. Hence, from (3.7) it follows that $\mathcal{R}_\tau \ge \gamma\, I$ for any $\tau \in [\tau_0, T]$. □

We shall use the following general result pertaining to continuous quadratic forms in Hilbert spaces, whose proof is given for the sake of completeness.

LEMMA 3.3. *Let X and U be two Hilbert spaces, and consider*

$$f(x,u) = 2Re\langle \mathcal{N}x, u\rangle + \langle \mathcal{R}u, u\rangle$$

with $\mathcal{N} \in \mathcal{L}(X,U)$, $\mathcal{R} \in \mathcal{L}(U)$, $\mathcal{R} = \mathcal{R}^*$.

1. *If there exists $x \in X$ such that*

$$V(x) := \inf_{u\in U} f(x,u) > -\infty,$$

then $\mathcal{R} \ge 0$.

2. *The infimum of $f(x,\cdot)$ is attained if and only if the equation*

$$\mathcal{R}u = -\mathcal{N}x \tag{3.8}$$

is solvable, and in this case any solution u of (3.8) gives a minimum.

3. *If for each $x \in X$ there exists a* unique *u_x such that*

$$f(x,u_x) = \min_u f(x,u),$$

then $\mathcal{R}$ is invertible (the inverse $\mathcal{R}^{-1}$ may not be bounded) and $u_x = -\mathcal{R}^{-1}\mathcal{N}x$ so that the transformation $x \to u_x$ is linear and continuous from X to U.

4. *Let us assume that $V(x) > -\infty$ for each $x \in X$. Then there exists a linear bounded operator $P \in \mathcal{L}(X)$ such that*

$$V(x) = \langle x, Px\rangle \qquad \forall x \in X\,. \tag{3.9}$$

Proof. If there exists v such that $\langle \mathcal{R}v, v\rangle < 0$ then $f(x,\lambda v) \to -\infty$ as $\lambda \to +\infty$. This proves Lemma 3.3(1). The second item is well known [23, Lemma 2.3]. To prove the third item we use item 2: the minimum u_x is characterized by (3.8). This equation is uniquely solvable for every x by assumption. Hence, $\ker \mathcal{R} = \{0\}$ and $\operatorname{im} \mathcal{N} \subseteq \operatorname{im} \mathcal{R}$. Consequently, $u_x = -\mathcal{R}^{-1}\mathcal{N}x$ where $\mathcal{R}^{-1}$ acts from the closure of the image of $\mathcal{R}$. Hence, $\mathcal{R}^{-1}\mathcal{N}$ is bounded since $\mathcal{R}^{-1}$ is closed and $\mathcal{N}$ is bounded.

The proof of the fourth item follows an approach in [7]. If $\mathcal{R}$ is coercive, then it is boundedly invertible, so that $f(x,\cdot)$ admits a unique minimum, namely, $u^+ = -\mathcal{R}^{-1}\mathcal{N}x$, and

$$V(x) = f(x,u^+) = -\langle x, \mathcal{N}^*\mathcal{R}^{-1}\mathcal{N}x\rangle.$$

Hence, (3.9) holds true and we have obtained an explicit expression for P, i.e.,

$$P = -\mathcal{N}^*\mathcal{R}^{-1}\mathcal{N}.$$

If we simply have $\mathcal{R} \ge 0$, we consider the function

$$f_n(x,u) = f(x,u) + \frac{1}{n}|u|^2.$$

Now $\mathcal{R}_n = \mathcal{R} + \frac{1}{n} I$ is coercive; hence

$$V_n(x) = \min_u f_n(x,u) = \langle x, P_n x \rangle,$$

with $P_n \in \mathcal{L}(X)$. By construction

$$n \to \langle x, P_n x \rangle$$

is a decreasing numerical sequence for any $x \in X$, and

$$V(x) \le \langle x, P_n x \rangle \le \langle x, P_1 x \rangle; \tag{3.10}$$

hence there exists $P \in \mathcal{L}(X)$ such that

$$\langle x, Px \rangle = \lim_{n \to +\infty} \langle x, P_n x \rangle = \inf_n \langle x, P_n x \rangle \ge V(x) \quad \forall x \in X.$$

To conclude, it remains to show that $V(x)$ *coincides* with $\langle x, Px \rangle$ for any $x \in X$. Assume by contradiction that $V(x) < \langle x, Px \rangle$ for a given $x \in X$, and let $\alpha > 0$ such that

$$\langle x, Px \rangle = V(x) + \alpha\,.$$

Since $V(x) = \inf_u f(x,u)$ there exists $\overline{u} \in U$ such that

$$f(x, \overline{u}) < V(x) + \frac{\alpha}{2}. \tag{3.11}$$

Correspondingly, there exists an integer $n_0 \in \mathbb{N}$ such that

$$0 \le f_{n_0}(x, \overline{u}) - f(x, \overline{u}) = \frac{1}{n_0} |\overline{u}|^2 < \frac{\alpha}{2}. \tag{3.12}$$

From (3.11) and (3.12) it follows that

$$V(x) \le \langle x, P_{n_0} x \rangle \le f_{n_0}(x, \overline{u}) < V(x) + \alpha,$$

which is a contradiction; compare (3.10). □

The above lemma and (3.3) imply a first necessary condition for finiteness of the value function.

LEMMA 3.4. *If there exists x_0 such that $V(\tau, x_0) > -\infty$, then*

$$J_\tau(0,u) = \langle \mathcal{R}_\tau u, u \rangle \ge 0 \quad \forall u \in L^2(\tau, T; U). \tag{3.13}$$

This observation is now used to obtain a necessary condition of more practical interest, which is well known in the finite-dimensional case. The symbol I denotes the identity operator acting on a space which will be clear from the context.

PROPOSITION 3.5. *If there exists $\tau_0 \in [0, T)$ and a constant $\gamma \ge 0$ such that $\mathcal{R}_{\tau_0} \ge \gamma\, I$, then $R \ge \gamma I$.*

Consequently,

$$\text{if there exists } x_0 \text{ and } \tau_0 \text{ such that } V(\tau_0, x_0) > -\infty, \text{ then } R \ge 0. \tag{3.14}$$

Proof. We first consider the case $\gamma = 0$; hence by assumption $\mathcal{R}_{\tau_0} \ge 0$. By contradiction, suppose that there exists a control $u_0 \in U$ and a constant $\alpha > 0$ such that $\langle R u_0, u_0 \rangle = -\alpha$. Given a small $\epsilon > 0$, choose a control u as follows:

$$u(t) = \begin{cases} 0, & \tau_0 \le t < T - \epsilon, \\ u_0, & T - \epsilon \le t \le T, \end{cases}$$

and compute

$$\begin{aligned}\langle \mathcal{R}_{\tau_0} u, u\rangle &= \int_{T-\epsilon}^{T} \left\langle Q \int_{T-\epsilon}^{t} e^{A(t-s)} B u_0 \, ds, \int_{T-\epsilon}^{t} e^{A(t-s)} B u_0 \, ds \right\rangle dt \\ &+ \epsilon \langle R u_0, u_0\rangle + 2\, Re \int_{T-\epsilon}^{T} \left\langle \int_{T-\epsilon}^{t} e^{A(t-s)} B u_0 \, ds, S u_0 \right\rangle dt \\ &+ \int_{T-\epsilon}^{T} \left\langle P_0 \int_{T-\epsilon}^{T} e^{A(T-s)} B u_0 \, ds, \int_{T-\epsilon}^{T} e^{A(T-s)} B u_0 \, ds \right\rangle dt \\ &= -\epsilon\alpha + o(\epsilon) + o(\epsilon^2) \quad \text{as } \epsilon \text{ tends to zero.} \end{aligned} \tag{3.15}$$

Since ϵ can be taken arbitrarily small, (3.15) yields $\langle \mathcal{R}_{\tau_0} u, u\rangle < 0$, and this contradicts the assumption.

Assume instead $\mathcal{R}_{\tau_0} \geq \gamma I > 0$. By choosing $u(t) = 0$ for $t \in [\tau_0, T-\epsilon[$, $u(t) = u_0 \in U$ arbitrary when $t \in [T-\epsilon, T]$, a direct computation yields

$$\gamma\epsilon \|u_0\|^2 \leq \epsilon \langle R u_0, u_0\rangle + o(\epsilon)\|u_0\|^2,$$

which implies $\langle R u_0, u_0\rangle \geq \gamma \|u_0\|^2$ for any $u_0 \in U$.

Finally, if $V(\tau_0, x_0) > -\infty$ for some $\tau_0 \in [0, T)$ and $x_0 \in X$, then from Lemma 3.4 it follows that $\mathcal{R}_\tau$ is a nonnegative operator for $\tau \geq \tau_0$. Therefore, from the previous part of the proof, $R \geq 0$. ∎

We now show that the value function satisfies Bellman's optimality principle, which is known, in the context of linear-quadratic problems, as linear operator inequality (LOI) or dissipation inequality (DI).

We begin with the following lemma.

LEMMA 3.6. *If for some number τ and some $x_0 \in X$ we have $V(\tau, x_0) > -\infty$, then we have also $V(t, x(t)) > -\infty$ for each $t \in [\tau, T]$. Here, $x(t)$ denotes the value at time t of the function given by* (3.1) *for any fixed control $u(\cdot)$ on $[\tau, t]$.*

Proof. Let $t \in (\tau, T)$. Then

$$\begin{aligned} J_\tau(x_0, u) &= \int_\tau^t F(x(s), u(s))\, ds + \int_t^T F(x(s), u(s))\, ds \\ &+ \langle x(T), P_0 x(T)\rangle = \int_\tau^t F(x(s), u(s)) ds + J_t(x(t), u), \end{aligned}$$

where $x(\cdot) = x(\cdot, \tau, x_0, u)$ for any $u \in L^2(\tau, T; U)$. Now take a control $v \equiv 0$ on $[\tau, t)$; then

$$J_\tau(x_0, u+v) = \int_\tau^t F(x(s), u(s))\, ds + J_t(x(t), u+v)$$

and

$$\inf_v J_\tau(x_0, u+v) = \int_\tau^t F(x(s), u(s))\, ds + \inf_v J_t(x(t), u+v). \tag{3.16}$$

The conclusion immediately follows since in fact $\inf_v J_t(x(t), u+v) = V(t, x(t))$. ∎

THEOREM 3.7. *Let $\tau \in [0,T]$ and $x_0 \in X$ be given. Let V be the value function of problem* (1.1), (1.2) *and assume that $V(\tau, x_0) > -\infty$. Then*

$$\int_\tau^t F(x(s), u(s))\, ds \,+\, V(t, x(t)) - V(\tau, x_0) \geq 0 \tag{3.17}$$

for any $u(\cdot) \in L^2(\tau, T; U)$ and any $t \in (\tau, T)$, with $x(\cdot) = x(\cdot, \tau, x_0, u)$. Moreover, the equality holds true if and only if the control u in (3.17) *is optimal.*

Proof. We return to the conclusion of Lemma 3.6 and observe again that

$$\inf_v J_t(x(t), u+v) = \inf_u J_t(x(t), u) \,=\, V(t, x(t)), \tag{3.18}$$

while

$$\inf_v J_\tau(x_0, u+v) \geq V(\tau, x_0); \tag{3.19}$$

hence plugging (3.18) into (3.16) and taking into account (3.19), we get

$$V(\tau, x_0) \leq \int_\tau^t F(x(s), u(s)) ds \,+\, V(t, x(t)), \tag{3.20}$$

which is nothing but (3.17). Thus, if for a given initial datum x_0 there exists an optimal control $u^+(\cdot, \tau, x_0)$ minimizing $J_\tau(x_0, u)$, then we can rewrite (3.16) and (3.19) with $u = u^+(\cdot, \tau, x_0)$, and (3.19) is in fact an equality. Therefore, (3.20) becomes an equality as well. For these arguments compare also [11].

Vice versa, assume that (3.17) is satisfied for any control $u \in L^2(\tau, T; U)$ and it is an equality for a given u^*. Then, passing to the limit, as $t \to T^-$, in (3.17) with $u = u^*$ and $x = x^* = x(\cdot\,, \tau, x_0, u^*)$ and assuming for the moment that

$$\lim_{t \to T^-} V(t, x^*(t)) = \langle x^*(T), P_0 x^*(T) \rangle, \tag{3.21}$$

we readily get

$$V(\tau, x_0) = \int_\tau^T F(x^*(s), u^*(s))\, ds \,+ \langle x^*(T), P_0 x^*(T) \rangle;$$

that is,

$$V(\tau, x_0) = J_\tau(x_0, u^*);$$

hence by definition u^* is optimal.

To conclude, it remains to show that if (x^*, u^*) satisfies

$$\int_\tau^t F(x^*(s), u^*(s))\, ds \,+\, V(t, x^*(t)) - V(\tau, x_0) \,=\, 0, \tag{3.22}$$

then (3.21) holds true. From (3.22) it follows that there exists

$$\lim_{t \to T^-} V(t, x^*(t)) = V(\tau, x_0) - \int_\tau^T F(x^*(s), u^*(s))\, ds,$$

and by the very definition of the value function it follows that

$$\lim_{t \to T^-} V(t, x^*(t)) \leq \langle x^*(T), P_0 x^*(T) \rangle.$$

To see this rewrite the above limit as

$$\lim_{t\to T^-} V(t,x^*(t)) = V(\tau,x_0) - \int_\tau^T F(x^*(s),u^*(s))\,ds - \langle x^*(T),P_0x^*(T)\rangle + \langle x^*(T),P_0x^*(T)\rangle.$$

By contradiction, assume now that

$$\lim_{t\to T^-} V(t,x^*(t)) = \langle x^*(T),P_0x^*(T)\rangle - \gamma,$$

where γ is a suitable positive constant. Then there exists $\delta>0$ such that

$$V(t,x^*(t)) < \langle x^*(T),P_0x^*(T)\rangle - \frac{\gamma}{2} \tag{3.23}$$

for any $t\in(T-\delta,T)$. Recall now that

$$x^*(T) = x(T,\tau,x_0,u^*) = x(T,t,x^*(t),u^*{}_{|s\geq t});$$

hence we can rewrite

$$\langle x^*(T),P_0x^*(T)\rangle = \underbrace{\langle e^{A(T-t)}x^*(t),P_0\,e^{A(T-t)}x^*(t)\rangle}_{A_1} + \underbrace{2\,Re\int_t^T \langle u^*(s),B^*e^{A^*(T-s)}P_0\,e^{A(T-t)}x^*(t)\rangle\,ds}_{A_2} + \underbrace{\int_t^T \langle u^*(s),B^*e^{(T-s)A^*}P_0\,L_{t,T}u^*\rangle\,ds}_{A_3}.$$

Take a possibly smaller δ, in order to get

$$A_2 + A_3 < \frac{\gamma}{4}, \tag{3.24}$$

so that (3.23) yields

$$V(t,x^*(t)) < A_1 - \frac{\gamma}{4}. \tag{3.25}$$

Finally, let δ such that

$$\left|\int_t^T \langle Qe^{A(s-t)}x^*(t),e^{A(s-t)}x^*(t)\rangle ds\right| < \frac{\gamma}{8}. \tag{3.26}$$

Now fix $t\in(T-\delta,T)$ so that (3.24) and (3.26) hold true. From (3.25) it follows that there exists a control $v=v_t\in L^2(t,T;U)$ such that

$$J_t(x^*(t),v_t) < A_1 - \frac{\gamma}{4};$$

that is, by means of (3.3),

$$\langle \mathcal{M}_t x^*(t), x^*(t)\rangle + 2\,Re\,\langle \mathcal{N}_t x^*(t), v_t\rangle + \langle \mathcal{R}_t v_t, v_t\rangle \,<\, A_1 - \frac{\gamma}{4},$$

with $\mathcal{M}_t$, $\mathcal{N}_t$, $\mathcal{R}_t$ defined in (3.4), (3.5), and (3.6), respectively. We know that $\langle \mathcal{M}_t x^*(t), x^*(t)\rangle$ is $A_1 + \int_t^T \langle\, e^{A^*(s-t)} Q\, e^{A(s-t)} x^*(t), x^*(t)\rangle\, ds$. Thus we cancel the term A_1, we take into account (3.26), and we obtain

$$2\,Re\,\langle \mathcal{N}_t x^*(t), v_t\rangle + \langle \mathcal{R}_t v_t, v_t\rangle \,<\, -\frac{\gamma}{8}\,. \tag{3.27}$$

In particular this implies that $v_t \neq 0$. Notice now that

$$|\langle \mathcal{N}_t x^*(t), v_t\rangle| \le \epsilon\,(T-t)\cdot |v_t(\cdot)|_{L^2(t,T;U)}, \quad |\langle \mathcal{R}_{t,T} v_t, v_t\rangle| \le \text{const}\cdot |v_t(\cdot)|^2_{L^2(t,T;U)},$$

and therefore

$$\liminf_{t\to T^-} |v_t(\cdot)|_{L^2(t,T;U)} = \beta > 0 \qquad (\text{possibly } \beta = +\infty).$$

Hence there exists a sequence t_n such that

$$\frac{1}{|v_{t_n}(\cdot)|^2}\,\langle \mathcal{N}_{t_n} x^*(t), v_{t_n}\rangle \to 0,$$

so that we see from (3.27) for n large

$$\frac{1}{|v_{t_n}(\cdot)|^2}\,\langle \mathcal{R}_{t_n} v_{t_n}, v_{t_n}\rangle \,<\, 0.$$

In other words $J_{t_n}(0, v_{t_n}) < 0$ and this is a contradiction since by assumption $J_\tau(0,u)$ is nonnegative for any $u \in L^2(\tau, T; U)$. □

The next proposition is an immediate consequence of Lemmas 3.1 and 3.3. We omit the proof.

PROPOSITION 3.8. *Let $\tau \in [0,T]$. If*

$$V(\tau, x_0) \,>\, -\infty \quad \forall x_0 \in X, \tag{3.28}$$

there exists a self-adjoint operator $W(\cdot) \in \mathcal{L}(X)$ such that $W(T) = P_0$ and

$$V(\tau, x_0) = \langle x_0, W(\tau) x_0\rangle. \tag{3.29}$$

4. Time regularity of the value function: The noncoercive case. In this section we investigate the regularity properties of $V(\tau, x_0)$ with respect to the initial time τ.

We note that several regularity results are known for the value function even of nonlinear systems, and with more general cost but under special boundedness properties, which are not satisfied in the present case, compare [11, Chapter 6].

Our first result is Lemma 4.1.

LEMMA 4.1. *Let $\tau_0 \in [0,T]$ and x_0 be such that $V(\tau_0, x_0)$ is finite. Then $\tau \to V(\tau, x_0)$ is upper semicontinuous at τ_0.*

Proof. Fix $x_0 \in X$, and let $\tau_0 \in [0,T]$. In order to show that

$$\limsup_{\tau\to\tau_0} V(\tau, x_0) \le V(\tau_0, x_0),$$

we shall show that for any real number $\alpha > V(\tau_0, x_0)$ we have $\alpha > V(\tau, x_0)$ if $|\tau - \tau_0|$ is taken small enough. We first consider the case when $\tau > \tau_0$. Let u be an admissible control such that

$$J_{\tau_0}(x_0, u) < \alpha, \tag{4.1}$$

and define

$$x_\tau(t) = e^{A(t-\tau)}x_0 + \int_\tau^t e^{A(t-s)}Bu(s)\,ds.$$

It is readily verified that

1. $\lim_{\tau\to\tau_0} x_\tau(T) = x_{\tau_0}(T)$,
2. $\lim_{\tau\to\tau_0} \int_\tau^T F(x_\tau(s), u(s))\,ds = \int_{\tau_0}^T F(x_{\tau_0}(s), u(s))\,ds$

so that if $|\tau - \tau_0|$ is small enough,

$$V(\tau, x_0) \le J_\tau(x_0, u) < \alpha.$$

Finally, if $\tau < \tau_0$, choose once more $u \in L^2(\tau_0, T; U)$ in such a way that (4.1) holds true. It is now sufficient to repeat the same arguments used before, after replacing u with $\hat{u}$ defined as follows:

$$\hat{u}(t) := \begin{cases} 0, & t < \tau_0, \\ u(t), & t \ge \tau_0. \end{cases}$$

The proof is complete. □

As to lower semicontinuity, the following result holds true.

LEMMA 4.2. *Let x_0 be such that $\tau \to V(\tau, x_0)$ is finite on $[0, T]$. Then*

- *the map $\tau \to V(\tau, x_0)$ is lower semicontinuous at τ_0 provided that for each element τ_n of a sequence $\{\tau_n\}$ which tends monotonically to τ_0 there exists a control $u_{\tau_n} \in L^2(\tau_n, T; U)$ such that*

$$\begin{aligned} &\text{(i)} \quad V(\tau_n, x_0) \le J_{\tau_n}(x_0, u_{\tau_n}) \le V(\tau_n, x_0) + \tfrac{1}{n}\,; \\ &\text{(ii)} \quad \text{there exists } \gamma_0 > 0 \text{ for which } |u_{\tau_n}(\cdot)|_{L^2(\tau_n, T; U)} \le \gamma_0. \end{aligned} \tag{4.2}$$

Proof. Let $\tau_0 \in [0, T]$ be given, and consider a sequence $\{\tau_n\}_{n\in\mathbb{N}}$ such that $\tau_n \downarrow \tau_0$. Introduce the inputs

$$\hat{u}_{\tau_n}(t) := \begin{cases} 0, & \tau_0 < t < \tau_n, \\ u_{\tau_n}(t), & t \ge \tau_n, \end{cases}$$

and define

$$x_{\tau_n}(t) := x(t; \tau_n, x_0, u_{\tau_n}), \quad \hat{x}_{\tau_n}(t) := x(t; \tau_0, x_0, \hat{u}_{\tau_n}).$$

Notice that $x_{\tau_n}(t) - \hat{x}_{\tau_n}(t) \to 0$, as $n \to \infty$, for any t, and that its norm is uniformly bounded in $L^2(\tau_0, T; U)$; hence

$$\lim_{n\to\infty} [J_{\tau_0}(x_0, \hat{u}_{\tau_n}) - J_{\tau_n}(x_0, u_{\tau_n})] = 0.$$

Therefore,

$$\liminf_{n\to\infty} J_{\tau_0}(x_0, \hat{u}_{\tau_n}) = \liminf_{n\to\infty} J_{\tau_n}(x_0, u_{\tau_n}) = \liminf_{n\to\infty} V(\tau_n, x_0),$$

where the last equality is due to (i). On the other hand (ii) implies the existence of an admissible control $v \in L^2(\tau_0, T; U)$ such that

$$\hat{u}_{\tau_n} \rightharpoonup v$$

as $n \to \infty$. Now the map

$$u(\cdot) \to J_{\tau_0}(x_0, u)$$

is convex continuous, hence weakly lower semicontinuous, so that

$$V(\tau_0, x_0) \le J_{\tau_0}(x_0, v) \le \liminf_{n\to\infty} J_{\tau_0}(x_0, \hat{u}_{\tau_n}) = \liminf_{n\to\infty} V(\tau_n, x_0). \tag{4.3}$$

To conclude the proof, we need to consider a sequence $\{r_n\}_{n\in\mathbb{N}}$ such that $r_n \uparrow \tau_0$. In this case, we introduce

$$\tilde{u}_{r_n}(t) := \begin{cases} u_{r_n}(t), & t \ge \tau_0, \\ 0 & \text{otherwise.} \end{cases}$$

Again from (ii) it follows that there exists an input $v \in L^2(\tau_0, T; U)$ such that $\tilde{u}_{r_n} \rightharpoonup v$ in $L^2(\tau_0, T; U)$. A similar argument gives

$$V(\tau_0, x_0) \le \liminf_{n\to\infty} V(r_n, x_0),$$

which finally yields

$$V(\tau_0, x_0) \le \liminf_{\tau\to\tau_0^-} V(\tau, x_0). \qquad \square$$

Consequently, we can deduce Theorem 4.3.

THEOREM 4.3. *Under the same assumptions as Lemma* 4.2*, the map* $\tau \to V(\tau, x_0)$ *is continuous for any* $\tau \in [0, T]$.

In the case that an optimal control exists for each τ near τ_0, Lemma 4.2 takes a simpler form. We state this form under the assumption that an optimal control exists for each τ.

COROLLARY 4.4. *Let* $x_0 \in X$ *be fixed. Assume that*

(i) *for any* $\tau \in [0, T]$ *there exists an optimal control* $u_\tau^+ \in L^2(\tau, T; U)$*;*

(ii) *there exists a constant* $\gamma > 0$*, independent of* τ*, such that*

$$|u_\tau^+|_{L^2(\tau,T:U)} \le \gamma \quad \forall \tau \in [0, T]. \tag{4.4}$$

Under these conditions, the map $\tau \to V(\tau, x_0)$ *is continuous.*

We note explicitly that if there exists an optimal control u^* for $J_\tau(x_0, v)$ then for each $\tau' > \tau$ there exists an optimal control for $J_{\tau'}(x(\tau'; \tau, x_0, u^*), v)$.

It has some interest to see that if the operator A generates a strongly continuous *group* then we can prove more in Theorem 4.5.

THEOREM 4.5. *Let us assume that for each* $\tau \in [0, T)$ *and each* $x_0 \in X$ *there exists a unique optimal control* $u^+(\cdot, \tau, x_0)$ *which minimizes* $J_\tau(x_0, u)$*. If* e^{At} *is a strongly continuous* group *then the value function is continuous from the right.*

Proof. We prove continuity from the right at a fixed $\tau_0 \in [0,T)$. We know from Lemma 3.3(2) that $x_0 \to u_\tau^+(\cdot,x_0)$ is linear and continuous from X to $L^2(\tau,T;U)$ for each $\tau \in [0,T)$.

Now we consider points $\tau > \tau_0$. We show that for each fixed $\tau > \tau_0$ there exists $x_1 = x_1(x_0)$ such that

$$u^+(\cdot,\tau_0,x_1(x_0))_{|[\tau,T]} = u^+(\cdot,\tau,x_0)\,. \tag{4.5}$$

It is sufficient to see for this that there exists a solution x_1 of

$$x^+(\tau,\tau_0,x_1) = e^{A(\tau-\tau_0)}x_1 + \int_{\tau_0}^{\tau} e^{A(\tau-s)}Bu^+(s,\tau_0,x_1)\,ds = x_0\,. \tag{4.6}$$

If this is true, unicity of the optimal control shows that (4.5) holds.

We noted above that $\|u^+(\cdot,\tau_0,x_1)\|_{L^2(\tau_0,T)} \le M\|x_1\|$ so that the norm of the operator

$$\mathcal{T}x_1 = \int_{\tau_0}^{\tau} e^{A(\tau-s)}Bu^+(s,\tau_0,x_1)\,ds$$

can be estimated as follows: $\|\mathcal{T}x_1\| \le (\tau-\tau_0)Mk\|x_1\|$.

We write (4.6) in the form

$$x_1 + e^{-A(\tau-\tau_0)}\mathcal{T}x_1 = e^{-A(\tau-\tau_0)}x_0\,. \tag{4.7}$$

If $\tau-\tau_0$ is sufficiently small, $\|e^{-A(\tau-\tau_0)}\mathcal{T}\|$ is less then 1; hence (4.7) can be continuously solved for x_1 and gives a linear continuous transformation $x_1 = x_1(x_0)$, which, of course, depends upon τ. The element x_1,

$$x_1 = x_1(x_0) = [I + e^{-A(\tau-\tau_0)}\mathcal{T}]^{-1}e^{-A(\tau-\tau_0)}x_0$$

is continuous with respect to x_0 and also with respect to τ if τ is close to τ_0. In particular, $\tau \to x_1(x_0)$ is bounded in a neighborhood of τ_0. Therefore,

$$\begin{aligned}\|u_\tau^+(\cdot,x_0)\| &= \|u_{\tau_0}^+(\cdot,x_1)_{|_{[\tau,T]}}\|_{L^2(\tau,T)}\\ &\le \|u_{\tau_0}^+(\cdot,x_1)\|_{L^2(\tau_0,T)} \le c\cdot\|x_1(x_0)\| \le \gamma\|x_0\|\,.\end{aligned}$$

Right continuity follows from Lemma 4.2. □

The previous theorem presents a case in which the quite involved condition of Lemma 4.2 is satisfied. The next example shows that the condition in that lemma cannot be avoided if we are to obtain continuity of the value function.

We note first that the value function is not continuous in general, even for finite-dimensional systems: if the cost is $|x(T)|^2$ and the system is controllable then the value function has a jump at T. The following example shows that the value function may be discontinuous even at points $\tau < T$.

Example 4.6. Consider the *delay system* given by

$$\begin{cases} \dot{x} = y(t-1), \\ \dot{y} = u(t), \end{cases} \tag{4.8}$$

with initial datum $\phi_0 = \mathrm{col}[x_0, y_0(\cdot)] \in \mathbb{R}\times L^2(-1,0)$. The quadratic functional is

$$J_\tau(\phi_0,u) = \int_\tau^2 |x(t)|^2\,dt + 3|x(2)|^2\,.$$

Take $\phi_0 = \text{col}(1,0)$. When $\tau \in [1,2]$, then $y(t-1) = 0$; hence $x(t) = 1$ on $[\tau, 2]$. Consequently,

$$J_\tau(\phi_0, u) = (2-\tau) + 3 \quad \forall u\,, \forall \tau \in [1,2]\,.$$

In particular $J_1(\phi_0, u) = 4$ and

$$V(1, \phi_0) = 4.$$

On the other hand, if $\tau \in [0,1[$, then $y(t-1) \neq 0$ when $t > \tau + 1$, and it can be arbitrarily fixed, by means of suitable choices of the control u, within the class of $W^{1,2}$ functions which are zero at $t = \tau + 1$. This set is dense in $L^2(1+\tau, 2)$; hence suitable functions y can be found in order to drive $x(t)$ to zero in time $\epsilon > 0$, namely, from $1+\tau$ to $1+\tau+\epsilon$, while remaining uniformly bounded. Therefore, we have that $x(t) = 1$ in $(\tau, 1+\tau)$, and

$$\int_{1+\tau}^{1+\tau+\epsilon} x^2(t)dt \to 0$$

as $\epsilon \to 0$. In conclusion, if $\tau < 1$, $V(\tau, \phi_0) = 1$ and the value function is *not* continuous at $\tau = 1$.

REMARK 4.7. The previous example shows that in the statement of Lemma 4.2—which concerns lower semicontinuity of $V(\tau, x_0)$—assumption (ii) *cannot* be dispensed with. In fact that assumption holds in the previous example for $\tau \to 1^+$ but not for $\tau \to 1^-$.

5. Time regularity of the value function: The coercive case. Let $\hat\tau \in [0,T]$ be given, and consider the operator $\mathcal{R}_{\hat\tau}$ as defined in (3.6). Throughout this section we shall assume that $\mathcal{R}_{\hat\tau}$ is *coercive*, i.e.,

$$\exists\, \gamma > 0 : \quad \mathcal{R}_{\hat\tau} \geq \gamma. \tag{5.1}$$

Our current goal is to show that under assumption (5.1) the value function $V(\tau, x_0)$ displays better regularity properties with respect to τ. We start by showing that the map

$$\tau \to V(\tau, x_0)$$

is *continuous* for any $\tau \in [\hat\tau, T]$, with $x_0 \in X$ fixed.

We recall that from (5.1), by virtue of Lemma 3.2, it follows that $\mathcal{R}_\tau \geq \gamma$ for any $\tau \in [\hat\tau, T]$, and by continuity also on an interval $(\tau', T] \supset [\hat\tau, T]$. Hence there exists a constant γ' such that

$$\|\mathcal{R}_\tau^{-1}\| \leq \frac{1}{\gamma'} \qquad \forall \tau \in (\tau', T] \supseteq [\hat\tau, T]. \tag{5.2}$$

Moreover (5.1) implies that for any initial time $\tau \in [\hat\tau, T]$ there exists a unique optimal control $u_\tau^+(t) = u^+(\cdot\,; \tau, x_0)$ ($u_\tau^+(\cdot)$ for short), explicitly given in terms of the initial state by

$$u^+(t; \tau, x_0) = -\left(\mathcal{R}_\tau^{-1}(\mathcal{N}_\tau x_0)(\cdot)\right)(t) \tag{5.3}$$

(compare item 3 of Lemma 3.3); and from (5.2) it follows

$$|u_\tau^+(\cdot)|_{L^2(\tau,T;U)} \leq k_{\hat\tau}|x_0|, \quad \text{with } k_{\hat\tau} \text{ independent of } \tau. \tag{5.4}$$

The following theorem provides a simple explicit expression of the value function in terms of the optimal pair which will be useful in the next section.

THEOREM 5.1. *Let $\mathcal{R}_\tau$ be coercive, and let $(x^+(\cdot,\tau,x_0),u^+(\cdot,\tau,x_0))$ the optimal pair for problem* (1.1)–(1.2). *Then*

$$W(\tau)x_0 = e^{A^*(T-\tau)}P_0x^+(T,\tau,x_0) + \int_\tau^T e^{A^*(t-\tau)}\left(Qx^+(t,\tau,x_0)+Su^+(t,\tau,x_0)\right)dt. \tag{5.5}$$

Proof. Since the infimum of the cost is attained at $u^+(\cdot,\tau,x_0)$ (u_τ^+ for short), plugging (5.3) into (3.3) we easily obtain

$$W(\tau)x_0 = \mathcal{M}_\tau x_0 + \mathcal{N}_\tau^* u_\tau^+. \tag{5.6}$$

The adjoint operator $\mathcal{N}_\tau^* : L^2(\tau,T;U)\to X$ maps any $L^2(\tau,T)$-function v in

$$\mathcal{N}_\tau^* v = e^{A^*(T-\tau)}P_0L_{\tau,T}v + \int_\tau^T e^{A^*(t-\tau)}\left((Q\,L_\tau+S)v\right)(t)\,dt;$$

hence (5.5) follows from (5.6) by a direct computation. □

As a consequence of Corollary 4.4, we first have Theorem 5.2.

THEOREM 5.2. *Let $x\in X$ be given. Assume that* (5.1) *is satisfied. Then $\tau\to V(\tau,x)$ is continuous on $[\hat\tau,T]$.*

Actually we are able to show that the value function satisfies a further regularity property. Before we state a preliminary result, see Lemma 5.3.

LEMMA 5.3. *Assume that $\mathcal{R}_{\hat\tau}$ is coercive. If $w(\cdot)$ is a continuous function, then the function*

$$s\to\phi(s) := (\mathcal{R}_\tau^{-1}w)(s) \tag{5.7}$$

is continuous for any $\tau\ge\hat\tau$.

In particular, if $\mathcal{R}_\tau$ is coercive then the optimal control is continuous.

Proof. Since $\mathcal{R}_{\hat\tau}$ is coercive, R is coercive, so that we can assume that $R=I$. Moreover, for any $\tau>\hat\tau$, $\mathcal{R}_\tau$ is coercive, hence invertible.

Let $\phi(t) := (\mathcal{R}_\tau^{-1}w)(t)$, with $w(\cdot)$ continuous: we know that $\phi(\cdot)$ is at least a U-valued L^2 function. But

$$\begin{aligned}\phi(t) = w(t) &- B^*\int_t^T e^{A^*(s-t)}Q\int_\tau^s e^{A(s-r)}\,B\phi(r)\,dr\,ds\\ &- S^*\int_\tau^t e^{A(t-s)}\,B\phi(s)\,ds - B^*\int_t^T e^{A^*(s-t)}S\phi(s)\,ds\\ &- B^*\,e^{A^*(T-t)}\,P_0\int_\tau^T e^{A(T-s)}B\phi(s)\,ds,\end{aligned}$$

and the right-hand side is apparently a U-valued *continuous* function.

The second statement follows from (5.3) since $(\mathcal{N}_\tau x_0)(\cdot)$ is a continuous function; compare (3.5). □

THEOREM 5.4. *Let $x\in D(A)$ be given. Assume that* (5.1) *is satisfied. Then the map $\tau\to V(\tau,x)$ is differentiable in $[\hat\tau,T]$.*

Proof. Let $x_0 \in D(A)$ and let $u_\tau^+ = u^+(\cdot, \tau, x_0)$ the unique optimal control of problem (1.1)–(1.2), $\tau \geq \hat{\tau}$. As in (5.6)

$$V(\tau, x_0) = \langle x_0, W(\tau)x_0 \rangle = \langle \mathcal{M}_\tau x_0, x_0 \rangle + \langle \mathcal{N}_\tau x_0, u_\tau^+ \rangle \,, \tag{5.8}$$

with $\mathcal{M}_\tau$ and $\mathcal{N}_\tau$ given by (3.4), (3.5), respectively.

From the very definition of $\mathcal{M}_\tau$ it readily follows that the derivative $\frac{\partial}{\partial \tau}\langle \mathcal{M}_\tau x_0, x_0 \rangle$ exists for any $x_0 \in D(A)$. In order to show that the second summand in (5.8), namely,

$$\int_\tau^T (\mathcal{N}_\tau x_0)(t) \cdot \overline{u_\tau^+(t)}\, dt, \tag{5.9}$$

is differentiable with respect to τ, we first observe that the factor $(\mathcal{N}_\tau x_0)(\cdot)$ is differentiable, with

$$\frac{\partial}{\partial \tau}(\mathcal{N}_\tau x_0)(t) = -(\mathcal{N}_\tau(Ax_0))(t). \tag{5.10}$$

Moreover, again from (3.5) it follows that (5.10) is a continuous function.

We next want to show that for each $t > \tau$ the U-valued function $\tau \to u_\tau^+(t)$ admits the first derivative with respect to τ and that this is continuous. Fix τ_0 and first consider the case $\tau > \tau_0$. Introduce the operator $\hat{\mathcal{N}}_\tau \in \mathcal{L}(X, L^2(\tau_0, T; U))$ defined as follows:

$$(\hat{\mathcal{N}}_\tau x_0)(t) = \begin{cases} (\mathcal{N}_\tau x_0)(t), & t \in [\tau, T], \\ (\mathcal{N}_{\tau_0} x_0)(t), & t \in [\tau_0, \tau[. \end{cases}$$

By construction

$$\hat{\mathcal{N}}_\tau x_0\big|_{t \geq \tau} \equiv \mathcal{N}_\tau x_0,$$

and for instance

$$\mathcal{R}_\tau^{-1}(\mathcal{N}_\tau x_0) = \mathcal{R}_\tau^{-1}(\hat{\mathcal{N}}_\tau x_0).$$

Moreover, we take into account (5.10) and we see that

$$\lim_{\tau \to \tau_0^+} \frac{(\hat{\mathcal{N}}_\tau x_0)(t) - (\mathcal{N}_{\tau_0} x_0)(t)}{\tau - \tau_0} = -\mathcal{N}_{\tau_0} A x_0 \quad \forall x_0 \in D(A). \tag{5.11}$$

In fact it is sufficient to observe that

$$(\hat{\mathcal{N}}_\tau x_0)(t) - (\mathcal{N}_{\tau_0} x_0)(t) = \begin{cases} 0, & t \in [\tau_0, \tau\,[, \\ (\mathcal{N}_\tau x_0)(t) - (\mathcal{N}_{\tau_0} x_0)(t), & t \in [\tau, T]. \end{cases}$$

Now we compute, via (5.3),

$$\frac{1}{\tau - \tau_0}(u_\tau^+ - u_{\tau_0}^+)(t)$$

$$= \frac{1}{\tau - \tau_0}\left[\left(\mathcal{R}_{\tau_0}^{-1}(\mathcal{N}_{\tau_0} x_0)(\cdot)\right)(t) - \left(\mathcal{R}_\tau^{-1}(\mathcal{N}_\tau x_0)(\cdot)\right)(t)\right]$$

$$= -\mathcal{R}_{\tau_0}^{-1}\left[\frac{(\hat{\mathcal{N}}_\tau x_0 - \mathcal{N}_{\tau_0} x_0)(\cdot)}{\tau - \tau_0}\right](t) + \mathcal{R}_\tau^{-1}\frac{[\mathcal{R}_{\tau_0} - \mathcal{R}_\tau]}{\tau - \tau_0}\mathcal{R}_{\tau_0}^{-1}(\hat{\mathcal{N}}_\tau x_0)(t). \tag{5.12}$$

The first summand in (5.12) tends to

$$(5.13)\quad -\left[\mathcal{R}_{\tau_0}^{-1}\left(\frac{\partial}{\partial\tau_0}(\mathcal{N}_{\tau_0}x_0)(\cdot)\right)\right](t)=\left(\mathcal{R}_{\tau_0}^{-1}(\mathcal{N}_{\tau_0}Ax_0)(\cdot)\right)(t)=-u_{\tau_0}^{+}(t,Ax_0),$$

when $\tau\to\tau_0^+$, due to (5.11).

As for the second summand, it can be rewritten in the following way:

$$\begin{aligned}
&\mathcal{R}_\tau^{-1}\frac{[\mathcal{R}_{\tau_0}-\mathcal{R}_\tau]}{\tau-\tau_0}\mathcal{R}_{\tau_0}^{-1}(\hat{\mathcal{N}}_\tau x_0)(t)\\
&=\mathcal{R}_\tau^{-1}B^*\int_t^T e^{A^*(s-t)}Q\underbrace{\frac{1}{\tau-\tau_0}\left[\int_{\tau_0}^{\tau}e^{A(s-r)}B(\mathcal{R}_{\tau_0}^{-1}\hat{\mathcal{N}}_\tau x_0)(r)dr\right]}_{a(\tau,s)}ds\\
&+\mathcal{R}_\tau^{-1}S^*\frac{1}{\tau-\tau_0}\int_{\tau_0}^{\tau}e^{A(t-s)}B(\mathcal{R}_{\tau_0}^{-1}\hat{\mathcal{N}}_\tau x_0)(s)ds\\
&+\mathcal{R}_\tau^{-1}B^*\,e^{A^*(T-t)}P_0\frac{1}{\tau-\tau_0}\int_{\tau_0}^{\tau}e^{A(T-s)}B(\mathcal{R}_{\tau_0}^{-1}\hat{\mathcal{N}}_\tau x_0)(s)ds\\
&=\boxed{1}+\boxed{2}+\boxed{3}.
\end{aligned}$$

We rewrite, in turn,

$$\begin{aligned}
a(\tau,s)=&\underbrace{\frac{1}{\tau-\tau_0}\int_{\tau_0}^{\tau}e^{A(s-r)}B(\mathcal{R}_{\tau_0}^{-1}\mathcal{N}_{\tau_0}x_0)(r)dr}_{b(\tau,s)}\\
&+\underbrace{\int_{\tau_0}^{\tau}e^{A(s-r)}B\left(\mathcal{R}_{\tau_0}^{-1}\left(\frac{\hat{\mathcal{N}}_\tau x_0-\mathcal{N}_{\tau_0}x_0}{\tau-\tau_0}\right)\right)(r)dr}_{c(\tau,s)}.
\end{aligned}$$

Observe now that as a consequence of Lemma 5.3 we have

$$\lim_{\tau\to\tau_0^+}b(\tau,s)=e^{A(s-\tau_0)}B(\mathcal{R}_{\tau_0}^{-1}\mathcal{N}_{\tau_0}x_0)(\tau_0)=-e^{A(s-\tau_0)}Bu_{\tau_0}^{+}(\tau_0,x_0),$$

while $\lim_{\tau\to\tau_0^+}c(\tau,s)=0$; hence

$$a(s):=\lim_{\tau\to\tau_0^+}a(\tau,s)=-e^{A(s-\tau_0)}Bu_{\tau_0}^{+}(\tau_0,x_0).$$

Finally, since $(\tau,s)\to a(\tau,s)$ is bounded, we can conclude that $\boxed{1}$ converges to

$$-\mathcal{R}_{\tau_0}^{-1}B^*\int_t^T e^{A^*(s-t)}Qe^{A(s-\tau_0)}B\,u_{\tau_0}^{+}(\tau_0,x_0)ds$$

as τ tends to τ_0^+. The convergence of the terms $\boxed{2}$ and $\boxed{3}$ can be proved even more easily.

If $\tau < \tau_0$ we define instead

$$(\hat{\mathcal{N}}_{\tau_0} x_0)(t) = \begin{cases} (\mathcal{N}_{\tau_0} x_0)(t), & t \in [\tau_0, T], \\ (\mathcal{N}_\tau x_0)(t), & t \in [\tau, \tau_0[, \end{cases}$$

and rewrite the term $\mathcal{R}_\tau^{-1}\mathcal{N}_\tau - \mathcal{R}_{\tau_0}^{-1}\mathcal{N}_{\tau_0}$ as

$$\mathcal{R}_\tau^{-1}(\mathcal{N}_\tau - \hat{\mathcal{N}}_{\tau_0}) + (\mathcal{R}_\tau^{-1} - \mathcal{R}_{\tau_0}^{-1})\hat{\mathcal{N}}_{\tau_0}$$

$$= \mathcal{R}_\tau^{-1}(\mathcal{N}_\tau - \hat{\mathcal{N}}_{\tau_0}) + \mathcal{R}_{\tau_0}^{-1}(\mathcal{R}_{\tau_0} - \mathcal{R}_\tau)\mathcal{R}_\tau^{-1}\hat{\mathcal{N}}_{\tau_0}.$$

The rest of the proof is completely similar.

Therefore we have proved that for each τ there exists $\frac{\partial}{\partial\tau} u_\tau^+(t)$ and that

$$\frac{\partial}{\partial\tau} u_\tau^+(t) = u_\tau^+(t, Ax_0) - \mathcal{R}_\tau^{-1} B^* \int_t^T e^{A^*(s-t)} Q e^{A(s-\tau)} B u_\tau^+(\tau, x_0)\, ds$$

$$-\mathcal{R}_\tau^{-1} S^* e^{A(t-\tau)} B u_\tau^+(\tau, x_0) - \mathcal{R}_\tau^{-1} B^* e^{A^*(T-t)} P_0 e^{A(T-\tau)} B u_\tau^+(\tau, x_0).$$

In conclusion we saw that the function $(\mathcal{N}_\tau x_0)(t)\,\overline{u_\tau^+(t)}$ is differentiable with respect to τ, and moreover its derivative is a continuous function in $[\hat{\tau}, T] \times [\hat{\tau}, T]$. Therefore, (5.9) is differentiable, and

$$\frac{\partial}{\partial\tau} \int_\tau^T (\mathcal{N}_\tau x_0)(t) \cdot \overline{u_\tau^+(t)}\, dt$$

$$= -(\mathcal{N}_\tau x_0)(\tau)\overline{u_\tau^+(\tau)} - \int_\tau^T (\mathcal{N}_\tau A x_0)(t)\overline{u_\tau^+(t)}\, dt + \int_\tau^T (\mathcal{N}_\tau x_0)(t)\overline{\frac{\partial}{\partial\tau} u_\tau^+(t)}\, dt. \qquad \square$$

We are now able to deduce a *differential form* of the DI.

PROPOSITION 5.5. *Assume that* (5.1) *holds true. Then there exists a self-adjoint operator* $W(\cdot) \in \mathcal{L}(X)$ *such that*

(i) $W(T) = P_0$*;*

(ii) $W(\cdot)$ *satisfies*

$$\frac{d}{d\tau}\langle a, W(\tau)\, a\rangle + 2Re\langle Aa + Bv, W(\tau)a\rangle + F(a, v) \geq 0 \tag{5.14}$$

for any $(a, v) \in D(A) \times U$ *for any* $\tau \in [\hat{\tau}, T]$.

Proof. We fix $a \in D(A)$, $v \in U$, and take a control $u(\cdot) \in C^1([\tau, T]; U)$ such that $u(\tau) = v$. We define $x(t) = x(t, \tau, a, u)$. It is well known (see, for instance, [2]) that in this case x is a *strict* solution to (1.1); that is, $x \in C^1([\tau, T]; X) \cap C([\tau, T]; D(A))$ and it satisfies (1.1) on $[\tau, T]$.

We write the DI (3.17) for $(x(t), u(t))$, $t \in [\tau, T]$; namely,

$$\int_\tau^t F(x(s), u(s))ds \,+\, \langle x(t), W(t)x(t)\rangle - \langle x(\tau), W(\tau)x(\tau)\rangle \geq 0. \tag{5.15}$$

If we divide in (5.15) by $t - \tau$ and let $t \to \tau$, we have

$$\frac{d}{ds}\langle x(\tau), W(s)x(\tau)\rangle_{|s=\tau} + 2\mathrm{Re}\langle Ax(\tau) + Bu(\tau), W(\tau)x(\tau)\rangle + F(x(\tau), u(\tau)) \geq 0.$$

To conclude, substitute $x(\tau) = a$ and $u(\tau) = v$. $\square$

We proved that if we replace an optimal pair in the left-hand side of the DI in integral form, then we get an equality. Hence we get an equality also in the differential form (5.14). In particular, we fix $a \in \operatorname{dom} A$ and we see that $u^+(\tau,\tau,a)$ is a minimum of the left-hand side of inequality (5.14). Hence we find that $u = u^+(\tau,\tau,a)$ satisfies

$$Ru + S^*a + B^*W(\tau)a = 0\,.$$

Since $\mathcal{R}$ is coercive then R is coercive too and we see that the optimal control has the well-known feedback form

$$u = u^+(\tau,\tau,a) = -R^{-1}[S^* + B^*W(\tau)]\,a$$

(if $a \in D(A)$ and, by continuity, for each $a \in X$, see item 3 of Lemma 3.3). Moreover, as $u^+(t,\tau,a) = u^+(t,t,x^+(t,\tau,a))$, the previous equality gives the feedback form of the optimal control on the interval $[0,T]$. We replace this expression for the unique optimal control in the left-hand side of (5.14) and find a quadratic differential equation for $W(\tau)$ which is the usual Riccati equation.

Of course, the Riccati equation can be written provided that R^{-1} is a bounded operator. But, an example in [6] shows that if $\mathcal{R}$ is not coercive then the minimum of the cost may exist and be unique, in spite of the fact that the corresponding Riccati equation is not solvable on $[\tau,T]$.

6. Space regularity of the value function. This section is devoted to the study of some space regularity properties of the value function in the case that the optimal control problem is driven by an abstract equation of *parabolic* type. See [17] for analogous arguments. More precisely, we shall make the following assumption.

*H*1. A is the generator of an analytic semigroup e^{tA} on X.

It is well known (see for instance [18]) that in this case there exists an $\omega \in \mathbb{R}$ such that the fractional powers $(\omega I - A)^\alpha$ are well defined for any $\alpha \in (0,1)$, and moreover there exist constants M_α, β such that the following estimates hold true

$$\|t^\alpha(\omega I - A)^\alpha\, e^{At}\|_{\mathcal{L}(X)} \le M_\alpha e^{\beta t}, \quad t > 0. \tag{6.1}$$

For the sake of simplicity we assume that the semigroup is exponentially stable, i.e., that we can choose $\omega = 0$.

We associate the following output to system (1.1):

$$y = Cx + Du,$$

where y belongs to a third Hilbert space Y and $C \in L(X,Y)$, $D \in L(U,Y)$. We assume that the cost penalizes the output y, i.e., that the quadratic functional F in (1.3) is given by

$$F(x,u) = \|y\|_Y^2 + \langle u, R_1 u\rangle$$

so that $Q = C^*C$, $S = C^*D$, $R = R_1 + D^*D$. (A special and important case is $D = 0$.) We make the following assumption.

*H*2. $R_1 \ge 0, \quad P_0 \ge 0$.

We now use similar arguments as in Lemma 3.3. Introduce a regularized optimal control problem with cost given by

$$J_{\tau,n}(x_0,u) = J_\tau(x_0,u) + \frac{1}{n}\,\|u\|^2_{L^2(\tau,T;U)}, \quad n \in \mathbb{N} \tag{6.2}$$

and observe that since the operator $\mathcal{R}_n = \mathcal{R} + \frac{1}{n}I$ is coercive for each n, then there exists a unique optimal control u_n^+ and

$$V_n(\tau, x_0) := \inf_u J_{\tau,n}(x_0, u) = J_{\tau,n}(x_0, u_n^+) = \langle x_0, W_n(\tau)x_0\rangle.$$

Arguing as in the proof of statement 4 in Lemma 3.3 we know that

$$W_n(\tau)x_0 \to W(\tau)x_0 \qquad \forall x_0 \in X.$$

Let $x_n^+(\cdot) = x_n(\cdot, \tau, x_0, u_n^+)$ and $y_n^+(\cdot) = Cx_n^+(\cdot) + Du_n^+(\cdot)$. Then we have the following lemma.

LEMMA 6.1. *Let $\gamma_0 \geq 0$ such that $(-A^*)^{\gamma_0}C^* \in \mathcal{L}(X)$, and assume that there exists a number $\gamma \in (0, \gamma_0 + \frac{1}{2})$ such that*

$$(-A^*)^{\gamma} P_0^{1/2} \in \mathcal{L}(X). \tag{6.3}$$

Then there exists a constant c such that

$$\|y_n^+(\cdot)\|^2_{L^2(\tau,T)} + \|P_0^{1/2}x_n^+(T)\|_X^2 \leq c\,\|(-A)^{-\gamma}x_0\|^2 \quad \forall n \in \mathbb{N}. \tag{6.4}$$

Proof. The estimate is easily obtained as follows (note that $0 \in \rho(A)$ since we assumed $\omega = 0$):

$$\int_\tau^T \|y_n^+(t)\|^2\,dt + \|P_0^{1/2}x_n^+(T)\|^2 \leq J_{\tau,n}(x_0, u_n^+) \leq J_1(x_0, u_1^+)$$

$$\leq J_1(x_0, 0) = \int_\tau^T \|Ce^{A(t-\tau)}x_0\|^2\,dt + \|P_0^{1/2}e^{A(T-\tau)}x_0\|^2$$

$$\leq \|C(-A)^{\gamma_0}\|^2 \cdot M^2\,\frac{T^{1-2(\gamma-\gamma_0)}}{1-2(\gamma-\gamma_0)}\,\|(-A)^{-\gamma}x_0\|^2$$

$$+ \|P_0^{1/2}(-A)^{\gamma}\|^2\,\|(-A)^{-\gamma}x_0\|^2. \qquad \square$$

REMARK 6.2. We stress that since

$$c = \max\left(\|C(-A)^{\gamma_0}\|^2 \cdot M^2\,\frac{T^{1-2(\gamma-\gamma_0)}}{1-2(\gamma-\gamma_0)}, \|P_0^{1/2}(-A)^{\gamma}\|^2\right),$$

the estimate (6.4) is uniform with respect to n and τ.

LEMMA 6.3. *Under the same assumptions of Lemma* 6.1 *there exists a constant k such that*

$$\|(-A^*)^{\gamma}W_n(\tau)(-A)^{\gamma}\| \leq k \quad \forall n \in \mathbb{N}. \tag{6.5}$$

Proof. Let $\xi_0 \in X$. We recall that since by construction the operator $\mathcal{R}_{\tau,n}$ relative to $J_n(\xi_0, u)$ is coercive for each fixed n, then the regularized control problem admits a unique optimal pair $(x_n^+(\cdot, \xi_0), u_n^+(\cdot, \xi_0))$, and Theorem 5.1 yields

$$W_n(\tau)\xi_0 = e^{A^*(T-\tau)}P_0x_n^+(T, \xi_0) + \int_\tau^T e^{A^*(t-\tau)}C^*y_n^+(t, \xi_0)\,dt.$$

The regularity assumptions on C and P_0 imply that $W_n(\tau)\xi_0 \in D((-A^*)^\gamma)$ and

$$(-A^*)^\gamma W_n(\tau)\xi_0 = e^{A^*(T-\tau)}[(-A^*)^\gamma P_0^{1/2}][P_0^{1/2}x_n^+(T,\xi_0)]$$
$$+(-A^*)^{\gamma-\gamma_0}\int_\tau^T e^{A^*(s-\tau)}[(-A^*)^{\gamma_0}C^*][y_n^+(s,\xi_0)]\,ds.$$

Now, as a consequence of (6.4) there exists k such that

$$\|(-A^*)^\gamma W_n(\tau)\xi_0\| \le k\,\|(-A)^{-\gamma}\xi_0\|$$

uniformly in n. The conclusion follows immediately by choosing $\xi_0 = (-A)^\gamma x_0$ with $x_0 \in D((-A)^\gamma)$. □

Consequently we have the following theorem.

THEOREM 6.4. *Under the same assumptions of Lemma* 6.1 *the operator*

$$(-A^*)^\gamma W(\tau)(-A)^\gamma$$

admits a bounded extension to X *for any* $\gamma < \gamma_0 + \frac{1}{2}$.

REFERENCES

[1] T. Başar and P. Bernhard, *H^∞-Optimal Control and Related Minimax Design Problems. A Dynamic Game Approach*, Birkhäuser, Boston, 1991.

[2] A. Bensoussan, G. Da Prato, M. C. Delfour, and S. K. Mitter, *Representation and Control of Infinite Dimensional Systems*, Vol. I, Birkhäuser, Boston, 1992.

[3] A. Bensoussan, G. Da Prato, M. C. Delfour, and S. K. Mitter, *Representation and Control of Infinite Dimensional Systems*, Vol. II, Birkhäuser, Boston, 1993.

[4] S. Bittanti, A. J. Laub, and J. C. Willems, eds., *The Riccati Equation*, Springer-Verlag, Berlin, New York, 1991.

[5] D. J. Clements and B. D. O. Anderson, *Singular Optimal Control: The Linear-Quadratic Problem*, Lecture Notes in Control and Inform. Sci. 5, Springer-Verlag, New York, 1978.

[6] B. Jacob, *Linear quadratic optimal control of time-varying systems with indefinite costs on Hilbert spaces: The finite horizon problem*, J. Math. Systems Estim. Control, 5 (1995), pp. 1–28.

[7] E. A. Jonckheere and L. M. Silverman, *Spectral theory of the linear quadratic optimal control problem: Discrete-time single-input case*, IEEE Trans. Circuits and Systems CAS–25, 1978, pp. 810–825.

[8] B. van Keulen, *Equivalent conditions for the solvability of the nonstandard LQ-problem for Pritchard–Salamon systems*, SIAM J. Control Optim., 33 (1995), pp. 1326–1356.

[9] I. Lasiecka, L. Pandolfi, and R. Triggiani, *A singular control approach to highly damped second-order abstract equations and applications*, in Control of Partial Differential Equations, E. Casas, ed., Marcel Dekker, New York, 1995, pp. 157–169.

[10] I. Lasiecka and R. Triggiani, *Differential and Algebraic Riccati Equations with Application to Boundary/Point Control Problems: Continuous Theory and Approximation Theory*, Lecture Notes in Control and Inform. Sci. 164, Springer-Verlag, New York, 1991.

[11] X. Li and J. Yong, *Optimal Control Theory for Infinite Dimensional Systems*, Birkhäuser, Basel, 1995.

[12] A. L. Likhtarnikov and V. A. Yakubovich, *The frequency theorem for equations of evolutionary type*, Siberian Math. J., 17 (1976), pp. 790–803.

[13] J.-Cl. Louis and D. Wexler, *The Hilbert space regulator problem and operator Riccati equation under stabilizability*, Ann. Soc. Sci. Bruxelles, Ser. I, 105 (1991), pp. 137–165.

[14] A. I. Lur'e, *Some Nonlinear Problems in the Theory of Automatic Control*, Her Majesty's Stationery Office, London, 1957.

[15] B. P. Molinari, *Nonnegativity of a quadratic functional*, SIAM J. Control Optim., 13 (1975), pp. 792–806.

[16] L. Pandolfi, *The standard regulator problem for systems with input delays: An approach through singular control theory*, Appl. Math. Optim., 31 (1995), pp. 119–136.

[17] L. Pandolfi, *Singular Regulator Problem for Holomorphic Semigroup Systems*, Rapporto interno 28, Dipartimento di Matematica, Politecnico di Torino, 1994.
[18] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, Berlin, 1983.
[19] J. C. Willems, *Least squares stationary optimal control and the algebraic Riccati Equation*, IEEE Trans. Automat. Control, AC–16 (1971), pp. 621–634; a correction in *On the existence of a nonpositive solution to the Riccati Equation*, ibidem, AC–19 (1974), pp. 592–593.
[20] J. C. Willems, *Dissipative dynamical systems, Parts* I *and* II, Arch. Rational Mech. Anal., 45 (1972), pp. 321–393.
[21] V. A. Yakubovich, *The frequency theorem in control theory*, Siberian Math. J., 14 (1973), pp. 384–419.
[22] V. A. Yakubovich, *A frequency theorem for the case in which the state and control spaces are Hilbert spaces with an application to some problems in the synthesis of optimal controls,* I, Siberian Math J., 15 (1974), pp. 457–476.
[23] V. A. Yakubovich, *A frequency theorem for the case in which the state and control spaces are Hilbert spaces with an application to some problems in the synthesis of optimal controls,* II, Siberian Math J., 16 (1975), pp. 828–845.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 1, pp. 137–147, January 1998

007

MINIMAL (MAX,+) REALIZATION OF CONVEX SEQUENCES*

STÉPHANE GAUBERT†, PETER BUTKOVIČ‡, AND RAYMOND CUNINGHAME-GREEN‡

Abstract. We show that the minimal dimension of a linear realization over the (max,+) semiring of a convex sequence is equal to the minimal size of a decomposition of the sequence as a supremum of discrete affine maps. The minimal-dimensional realization of any convex realizable sequence can thus be found in linear time. The result is based on a bound in terms of minors of the Hankel matrix.

Key words. max plus algebra, minimal realization, rational sequences, discrete-event systems

AMS subject classifications. 93B20, 15A15

PII. S036301299528534X

1. Introduction.

A classical problem consists in studying infinite sequences $h_0, h_1, \dots$ with values in a semiring $(\mathcal{S}, \oplus, \otimes)$, generated by a finite device. The simplest and most studied class is probably that of *realizable* or *recognizable* sequences, obtained as the scalar output of finite-dimensional recurrent $\mathcal{S}$-linear systems:

$$x_0 = b, \quad x_{k+1} = Ax_k, \quad h_k = cx_k, \quad k = 0, 1, \dots, \tag{1.1}$$

where $A \in \mathcal{S}^{n\times n}, b, x_0, x_1, \dots \in \mathcal{S}^{n\times 1}, c \in \mathcal{S}^{1\times n}$ for some positive integer n, and where concatenation denotes the matrix product as usual.[1] Equivalently,

$$h_k = cA^k b, \quad k = 0, 1, \dots. \tag{1.2}$$

The triple (A, b, c) is called a linear *realization* or *representation* of the sequence h, and n is called the dimension of the realization (A, b, c). By the Kleene–Schützenberger theorem [3], realizable sequences coincide with *rational sequences* (sequences of coefficients of rational series). The theory of these sequences is much developed in the case of fields (particularly $\mathcal{S} = \mathbb{R}$ or $\mathbb{C}$). The case of realizable sequences over the semiring of nonnegative reals $(\mathbb{R}^+, +, \times)$ is also much studied in connection with probability measures and Markov chains [13, 22]. Here, we are concerned with realizable sequences over the "(max,+)" semiring $\mathbb{R}_{\max} \stackrel{\text{def}}{=} (\mathbb{R} \cup \{-\infty\}, \oplus, \otimes)$, with max as addition ($a \oplus b \stackrel{\text{def}}{=} \max(a, b)$) and + as product ($a \otimes b \stackrel{\text{def}}{=} a + b$). The interest in realizable sequences over $\mathbb{R}_{\max}$ arises from at least two fields.

a) In discrete-event systems theory, it is known that an important subclass of man-made systems with synchronization features (manufacturing systems, transportations networks, etc.) can be modeled by input-output variants of the dynamics (1.1), namely,

$$x_{k+1} = Ax_k \oplus bu_{k+1}, \quad y_k = cx_k, \qquad k \in \mathbb{Z}, \tag{1.3}$$

where $u_k, y_k \in \mathbb{R} \cup \{-\infty\}$. In the case of manufacturing systems, typically, the input u_k represents the availability date of the kth unit of raw material, and y_k represents

*Received by the editors April 28, 1995; accepted for publication (in revised form) October 17, 1996.

http://www.siam.org/journals/sicon/36-1/28534.html

†INRIA, Domaine de Voluceau, BP 105, 78153 Le Chesnay cédex, France (Stephane.Gaubert@inria.fr).

‡School of Mathematics and Statistics, University of Birmingham, Birmingham B15 2TT, UK (p.butkovic@bham.ac.uk, r.a.cuninghame-green@bham.ac.uk).

[1] E.g., $(Ax_k)_i = \oplus_{j=1}^n A_{ij} \otimes (x_k)_j$, $cx_k = \oplus_{j=1}^n c_j \otimes (x_k)_j$, etc.

the availability date of the kth finished part, while the vector x_k represents the dates of completion of internal events. It is not difficult to see that the *minimal* output y of (1.3) corresponding to the *earliest* behavior of the system is given by the sup-convolution

$$y_k = \bigoplus_{p\in\mathbb{N}} h_p \otimes u_{k-p} = \sup_{p\in\mathbb{N}}[h_p + u_{k-p}], \tag{1.4}$$

so that the realizable sequence h determines the input-output relation of (1.3). As in the case of conventional linear systems, the sequence $h_0, h_1, \ldots$ is called the *impulse response* of the system, for it coincides with the output y associated with the *impulse* input u: $u_k = 0$ if $k = 0$, $u_k = -\infty$ otherwise. See [1] for a complete presentation.

b) In dynamic programming, the simplest stationary deterministic Markovian decision problem with finite state space $Q = \{1, \ldots, n\}$, transition reward $A : Q\times Q \to \mathbb{R} \cup \{-\infty\}$, initial reward $c : Q \to \mathbb{R} \cup \{-\infty\}$, final reward $b : Q \to \mathbb{R} \cup \{-\infty\}$, and horizon k, writes

$$\max_{q_0,\ldots,q_k} \left[c(q_0) + \sum_{i=1}^{k} A(q_{i-1}, q_i) + b(q_k) \right]. \tag{1.5}$$

Identifying A, b, c with matrices of appropriate sizes, it is immediately seen that the optimal reward in horizon k, given by (1.5), coincides with $h_k = cA^k b$.

In this paper, we are concerned with the *minimal realization problem*, which, given a sequence h, consists in finding a (linear) realization (A, b, c) with minimal dimension. For instance, in the Markov decision context, the minimal realization problem asks whether or not there exists another decision problem (A', b', c') of type (1.5), with state space Q' of strictly smaller cardinality but the same reward history $h_0, h_1, \ldots$ as (A, b, c). In the context of discrete-event systems, this is a natural problem, which consists in finding a minimal internal realization of the system (1.4), known only by its input-output relation $u \to y$. This has often interesting practical interpretations: loosely speaking, the nonminimality of the triple (A, b, c) arises from the existence of nontrivial temporal relations between the different physical events in the system. Particularly, nonminimalities occur when some component of the system (a particular machine or process) which plays a physical role in the production process is invisible from the performance evaluation point of view, i.e., when the normal functioning of this process will never delay the output dates due to the existence of margins. This striking phenomenon is illustrated on the cover page of the book [1], to which the reader is referred for more motivation.

Unlike in the field's case, the minimal realization problem over $\mathbb{R}_{\max}$ is not solved by rank arguments. It is indeed very much analogous to that of the nonnegative realization (over the usual algebra) [14] for which only partial solutions are known. We refer the reader to Olsder [29, 28], Cuninghame-Green [6], Qi and Chen [31], Gaubert [15, Chap. VI], De Schutter and De Moor [10, 11, 12, 9] for existing results (realization procedures, bounds, heuristics, reduction of the partial realization problem to an extended linear complementarity problem). See also [1, sections 1.3 and 9.2.3]. In the present paper, we characterize the minimal dimension of a realization for the subclass of *convex* realizable sequences, extending a result given by Cuninghame-Green and Butkovič [7] for the strictly convex case. The proof requires the minor bound given by Gaubert in [15, Chapter VI; 16], together with a classical *majorization* result [26].

It is worth noting that the convexity assumption, although restrictive, is natural with respect to a subclass of discrete-event systems. Input-output systems (1.4) with

affine realizable impulse response, $h_k = \alpha + \beta \times k$, can be interpreted as (delayed) *flow limiters* [1, section 6.2.2], i.e., as periodic systems, with minimal interevent delay β and transfer delay α. When building complex discrete-event systems from simple ones, one uses in particular the synchronization (or parallel composition) operation [1], which corresponds to the pointwise max of the impulse responses. Since a convex map is the upper envelope of its tangent lines, it is not difficult to see that realizable convex responses correspond exactly to parallel composition of finitely many delayed flow limiters.

Finally, we refer the reader interested in an overview of the theory and applications of the (max,+) semiring to [5, 19, 1, 27, 21]. General results on semirings can be found in [18].

2. Statement of the result. A sequence $h_0, h_1, \ldots \in \mathbb{R}$ is *convex* if

$$k \geq 1 \Rightarrow h_{k+1} - h_k \geq h_k - h_{k-1}.$$

It follows from the well-known periodicity property of (max,+) realizable sequences (see [5], [1, Theorem 3.112], [17, Theorem 7]) that a realizable convex sequence is eventually periodic of period one; that is, there exists $N \in \mathbb{N}$ and $\lambda \in \mathbb{R}$ such that

$$k \geq N \Rightarrow \qquad h_{k+1} = \lambda + h_k. \tag{2.1}$$

In what follows, N will always stand for the least natural number satisfying (2.1) and will be called the *length of the transient* of the sequence. A *max-polynomial* [8] is the function

$$p(x) = \max_{1 \leq i \leq r} (\alpha_i + \beta_i x), \tag{2.2}$$

with $\alpha_i \in \mathbb{R}, \beta_i \in \mathbb{R}$. The name "polynomial" refers to the notation of the semiring $\mathbb{R}_{\max}$: $a \oplus b = \max(a, b)$, $a \otimes b = a + b$, and, for $n \in \mathbb{N}$, $a^n = \underbrace{a \otimes \cdots \otimes a}_{n \text{ times}}$ $(= n \times a)$.

Then, (2.2) becomes

$$p(x) = \bigoplus_{i=1}^{r} \alpha_i \otimes x^{\beta_i}. \tag{2.3}$$

Note that we extend the exponent notation and use x^{β_i} for $x \times \beta_i$, even when $\beta_i \notin \mathbb{N}$: unlike in the conventional case, maxpolynomials may have real (nonintegral) exponents. We say that p is a *polynomial realization* of h if for all nonnegative integers k, $p(k) = h_k$. We denote by $\mathsf{mpr}(h)$ the minimal number of monomials of a polynomial realization of h (i.e., the minimal value of r). By convention, $\mathsf{mpr}(h) = +\infty$ when h does not admit a polynomial realization. Denote by $\mathsf{mlr}(h)$ the minimal dimension of a linear realization (with $\mathsf{mlr}(h) = +\infty$ if h is not realizable). The main result of this paper is the following characterization which solves the minimal realization problem for convex sequences.

THEOREM 2.1. *For every convex sequence h there holds*

$$\mathsf{mpr}(h) = \mathsf{mlr}(h). \tag{2.4}$$

Note that the theorem states in particular that the existence of a polynomial realization is equivalent to that of a linear one, for convex sequences.

Efficient computation of $\mathsf{mpr}(h)$ is not difficult: given a convex sequence h with the length of the transient N, the algorithm given in the Appendix provides a minimal polynomial realization in time $O(N)$.

The inequality $\mathsf{mlr}(h) \leq \mathsf{mpr}(h)$ is immediate: if p is a polynomial realization of h of type (2.3), then $h_k = c\,\mathrm{diag}(\beta)^k\, b$, where b denotes the $r \times 1$ matrix with entries $0, c$ the $1 \times r$ matrix with entries $\alpha_1, \ldots, \alpha_r$, and $\mathrm{diag}(\beta)$ the $r \times r$ matrix with diagonal entries $\beta_1, \ldots, \beta_r$ and off-diagonal elements equal to $-\infty$. The proof of the reverse inequality will use a bound in terms of determinants and Hankel matrices, which we introduce next.

3. Minor bound for the minimal dimension of realization. Recall that the *Hankel matrix* [13, 3] associated with the sequence $h_0, h_1, \ldots$ is the $\mathbb{N} \times \mathbb{N}$-matrix

$$\mathcal{H} = (\mathcal{H}_{ij}), \mathcal{H}_{ij} = h_{i+j} \text{ for all } i, j = 0, 1, \ldots .$$

A classical result for the minimal realization problem over fields states that the minimal dimension of any realization of a sequence h is equal to the *rank* of its Hankel matrix, which can be defined equivalently as the cardinality of a basis of the vector space generated by the rows (or columns) of $\mathcal{H}$ or as the maximal size of a square submatrix of $\mathcal{H}$ with nonzero determinant.

Over a general (commutative) semiring $\mathcal{S}$, several nonequivalent rank notions exist,[2] which do not characterize the minimal dimension of realization but only provide bounds. Here, we will need the rank notion originating from determinants and bideterminants over semirings.

Given a positive integer n, let $\mathfrak{S}_n^+$, $\mathfrak{S}_n^-$, respectively, denote the sets of even and odd permutations on $\{1, \ldots, n\}$ (we use the concepts of even and odd permutations in the conventional sense [4]). The *positive* and *negative* determinants of an $n \times n$ matrix A with entries from a (commutative) semiring $\mathcal{S}$ are defined as follows:

$$\det{}^+ A = \bigoplus_{\sigma \in \mathfrak{S}_n^+} \bigotimes_{i=1}^{n} A_{i\sigma(i)},$$

$$\det{}^- A = \bigoplus_{\sigma \in \mathfrak{S}_n^-} \bigotimes_{i=1}^{n} A_{i\sigma(i)}.$$

The *bideterminant* [20] or, equivalently, the determinant in the *symmetrized* semiring $\mathcal{S}^2$ [30, 15] of a square matrix A is the ordered pair

$$\det A \stackrel{\text{def}}{=} (\det{}^+ A, \det{}^- A).$$

We say that the determinant is *balanced* if $\det^+ A = \det^- A$, otherwise it is *unbalanced*. Let $I, J \subseteq \mathbb{N}$ denote (possibly infinite) sets of row and column indices. Given an $I \times J$ matrix A and two subsets $I' = \{i_1, \ldots, i_r\} \subseteq I, J' = \{j_1, \ldots, j_s\} \subseteq J$, with $i_1 < \cdots < i_r$, $j_1 < \cdots < j_s$, we denote by $A[I'|J']$ or $A[i_1, \ldots, i_r | j_1, \ldots, j_s]$ the $r \times s$ submatrix $(A_{i_m j_t})_{1 \leq m \leq r, 1 \leq t \leq s}$. Let $\| X \|$ denote the cardinality of a set X. The *minor rank* $\mathrm{rk}_m A$ of a matrix A is the supremum of the order of the finite square submatrices of A with unbalanced determinant

$$\mathrm{rk}_m A = \sup \left\{ r > 0 \,;\; \begin{array}{c} \exists I' \subseteq I,\ \exists J' \subseteq J,\ \| I' \| = \| J' \| = r \\ \text{and } \det^+ A[I'|J'] \neq \det^- A[I'|J'] \end{array} \right\},$$

[2] Row rank, column rank, Schein rank are distinct standard notions for Boolean matrices [24], which can also be defined in general semirings. Moreover, in the $\mathbb{R}_{\max}$ case, other rank notions have been used in relation to the uniqueness of solutions of linear systems [5], [15, Chapter 3, section 10].

and $\mathrm{rk}_m A = 0$ if no submatrix of A with unbalanced determinant exists. The following result taken from [15, 16] is a semiring weak version of a well-known result over fields [13, 3].

THEOREM 3.1 (minor bound). *The dimension of a linear realization of a sequence h is not less than the minor rank of its Hankel matrix.*

Hence,

$$\mathrm{rk}_m \mathcal{H} \leq \mathsf{mlr}(h). \tag{3.1}$$

This result holds in an arbitrary commutative semiring $\mathcal{S}$ (and not only in $\mathbb{R}_{\max}$). It is purely combinatorial in nature. We will prove it as a consequence of the following semiring version [15, 16] of the classical Binet–Cauchy identity [25, section 2.4.14].

LEMMA 3.2 (Binet–Cauchy formula). *Let $\mathcal{S}$ be an arbitrary commutative semiring. Let $A \in \mathcal{S}^{n\times r}$, $B \in \mathcal{S}^{r\times p}$. For all subsets $I \subseteq \{1,\dots,n\}$, $J \subseteq \{1,\dots,p\}$ of k elements, we have*

$$\begin{aligned}\det{}^+(AB)[I|J] \oplus \bigoplus_{K'} \big(\det{}^+ A[I|K'] \det{}^- B[K'|J] \oplus \det{}^- A[I|K'] \det{}^+ B[K'|J]\big) \\ = \det{}^-(AB)[I|J] \oplus \bigoplus_{K} \big(\det{}^+ A[I|K] \det{}^+ B[K|J] \oplus \det{}^- A[I|K] \det{}^- B[K|J]\big);\end{aligned} \tag{3.2}$$

where the sums are taken over all the k-element subsets $K, K' \subseteq \{1,\dots,r\}$. By convention, these two sums are equal to the zero element of $\mathcal{S}$ if $k > r$.

More generally, a folklore "transfer principle" [15, Chapter 1] asserts that usual ring identities admit semiring analogues whenever written without minus sign. Such semiring analogues can be obtained by direct combinatorial means (Zeilberger [33] proves the case $n = p = r$ of (3.2); the general case can be proved along the same lines). They can also be deduced formally from their classical ring versions, following an algebraic argument due to Reutenauer and Straubing [32], which we reproduce here for the sake of completeness. Note also that a different Binet–Cauchy identity in the (max,+) semiring (valid for permanents) has been given by Bapat [2].

Proof of Lemma 3.2. Let $X = \{a'_{ij}, b'_{kl}; 1 \leq i \leq n; 1 \leq j,k \leq r; 1 \leq l \leq p\}$ denote a family of distinct commuting indeterminates, and consider the semiring of (formal) commutative polynomials with coefficients in $\mathbb{N}$ and indeterminates $x \in X$: $\mathcal{S}' = \mathbb{N}[X] \subseteq \mathbb{Z}[X]$. Introduce the two matrices $A' = (a'_{ij}), B' = (b'_{kl})$ with entries in $\mathcal{S}'$. We first note that the identity (3.2) holds for A' and B'. Indeed, using the invertibility of the addition of $\mathbb{Z}[X]$, the identity (3.2) for A' and B' is equivalent to the conventional Binet–Cauchy identity which is known to be valid in the *ring* $\mathbb{Z}[X]$:

$$\det A'B'[I|J] = \sum_K \det A'[I|K] \det B'[K|J]. \tag{3.3}$$

Now, there is a unique morphism of semirings $\varphi : \mathcal{S}' \to \mathcal{S}$, such that $\forall i,j,k,l,\ \varphi(a'_{ij}) = a_{ij}, \varphi(b'_{kl}) = b_{kl}$. This morphism transforms the identity (3.2), applied to the matrices A', B' with entries in $\mathcal{S}'$, to the required identity for the matrices A, B with entries in $\mathcal{S}$. □

Proof of Theorem 3.1. Consider a linear realization (A, b, c) of h of dimension r. Let $\mathcal{O} = [c, cA, cA^2, \dots]^T$ and $\mathcal{C} = [b, Ab, A^2b, \dots]$, and consider two finite subsets $I, J \subseteq \mathbb{N}$. Applying the Binet–Cauchy identity to the finite size factorization $\mathcal{H}[I|J] = \mathcal{O}[I|1\dots r]\mathcal{C}[1\dots r|J]$ following from $\mathcal{H} = \mathcal{O}\mathcal{C}$, we get $\det^+(\mathcal{H}[I|J]) = \det^-(\mathcal{H}[I|J])$ if $|I| = |J| > r$. Hence, $\mathrm{rk}_m \mathcal{H} \leq r$. □

4. Preliminary majorization results. The proof of Theorem 2.1 will use standard convexity results that we recall here. The following famous result, due to Hardy, Littlewood, and Pólya, states the equivalence of two possible definitions of the relation of *majorization*.

THEOREM 4.1 (see [23, section 2.20], [26]). *Let $\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_k$ and $\alpha'_1 \leq \alpha'_2 \leq \cdots \leq \alpha'_k$ be real nonnegative numbers. The following two statements are equivalent:*

1. *$\alpha_1 + \cdots + \alpha_k = \alpha'_1 + \cdots + \alpha'_k$ and*
   $\alpha_\nu + \cdots + \alpha_k \geq \alpha'_\nu + \cdots + \alpha'_k$ for all ν, $2 \leq \nu \leq k$.
2. *There exists a doubly stochastic[3] matrix P such that $\alpha' = P\alpha$ where $\alpha = (\alpha_1, \ldots, \alpha_k)^T$, $\alpha' = (\alpha'_1, \ldots, \alpha'_k)^T$.*

We write $\alpha' \prec \alpha$ and say that α' is majorized *by α when these two equivalent statements hold.*

The following majorization inequality is standard [23, 26]. We single out the strict-inequality case for further use.

THEOREM 4.2. *Let g be a convex function $\mathbb{R} \to \mathbb{R}$.*

1. *If $\alpha' \prec \alpha$, then*

$$\sum_i g(\alpha'_i) \leq \sum_i g(\alpha_i). \tag{4.1}$$

2. *Take $P = (P_{jm})$ such that $\alpha' = P\alpha$ as in Theorem 4.1. If for some j the restriction of g to the set $\{\alpha_m \mid P_{jm} \neq 0\}$ does not coincide with an affine function, then the strict inequality holds in (4.1).*

Proof. Classically, (4.1) is obtained by summing up the convexity inequalities

$$\forall j, \;\; g(\alpha'_j) \leq \sum_m P_{jm} g(\alpha_m). \tag{4.2}$$

The strict inequality in (4.1) follows from the strict inequality in (4.2), as soon as j satisfies condition 4.2 of the theorem. ∎

5. Proof of Theorem 2.1. Suppose that the sequence $h_0, h_1, \ldots$ is convex. To prove Theorem 2.1, it remains to show that $\mathsf{mlr}(h) \geq \mathsf{mpr}(h)$. We will assume that $\mathsf{mlr}(h) < \infty$ (otherwise, there is nothing to prove). Then, the existence of a polynomial realization follows readily from the convexity of h together with (2.1). Let p be such a polynomial realization satisfying

$$\forall x \in \mathbb{R}, \;\; p(x) = \max_{1 \leq i \leq r} \ell_i(x) = \max_{1 \leq i \leq r} (\alpha_i + \beta_i x), \tag{5.1}$$

with

$$\beta_1 < \beta_2 < \cdots < \beta_r \quad \text{and} \quad r = \mathsf{mpr}(h). \tag{5.2}$$

We set

$$\begin{array}{lll} I_0 = [x_0, z_0] & \stackrel{\text{def}}{=} & \{x \in \mathbb{N} \mid \; p(x) = \ell_1(x)\} \qquad\qquad \text{and} \\ I_i = [x_i, z_i] & \stackrel{\text{def}}{=} & \{x \in \mathbb{N} \mid \; p(x) = \ell_{i+1}(x) > \ell_i(x)\} \quad \text{for} \;\; i = 1, \ldots, r-1, \end{array} \tag{5.3}$$

[3] "Doubly stochastic" refers, as usual, to a matrix with nonnegative entries in which both the row and column sums are equal to 1.

with $x_i, z_i \in \mathbb{N}$, except the last value $z_{r-1} = \infty$. (The fact that I_i is nonempty follows from the minimality of r. The fact that it is an interval is immediate due to the convexity of $\max_i \ell_i$.) Note that $\alpha_1 > \alpha_i$ for $i = 2, \dots, r$ (otherwise, using (5.2), we get $\ell_1(k) \le \ell_i(k) \forall k \ge 0$, which contradicts the minimality or r). Hence, $p(0) = \alpha_1 = \ell_1(0)$, and thus $x_0 = 0$. To summarize,

$$0 = x_0 \le z_0 < x_1 \le z_1 \cdots \le z_{r-2} < x_{r-1}. \tag{5.4}$$

The following elementary lemma states the existence of a minimal polynomial realization in which each line passes through at least two consecutive points.

LEMMA 5.1. *There exists a minimal polynomial realization* (5.1) *such that*

$$x_{i+1} \ge x_i + 2 \quad \text{for } i = 0, \dots, r-2. \tag{5.5}$$

The proof of the lemmas is at the end of this part.

COROLLARY 5.2. *If p is the polynomial realization in Lemma* 5.1*, then p does not coincide with an affine function on $[x_i, x_{i+1}]$ for $i = 0, \dots, r-2$.* □

Without loss of generality, we suppose that the polynomial realization p satisfies the conditions of Lemma 5.1 and Corollary 5.2. We set

$$u_i \stackrel{\text{def}}{=} x_i - i \qquad \text{for } i = 0, \dots, r-1. \tag{5.6}$$

We get from (5.5) that

$$u_0 < u_1 < \cdots < u_{r-1}.$$

LEMMA 5.3. *Let $M = \mathcal{H}[0, 1, \dots, r-1 | u_0, u_1, \dots, u_{r-1}]$. We have*

$$\det{}^+ M = \bigotimes_{i=0}^{r-1} p(x_i) > \det{}^- M. \tag{5.7}$$

Thus, $\mathcal{H}$ contains an $r \times r$ submatrix with unbalanced minor, i.e., $\mathrm{rk}_m \mathcal{H} \ge r$, which together with the minor bound gives $\mathsf{mlr}(h) \ge \mathrm{rk}_m \mathcal{H} \ge r = \mathsf{mpr}(h)$, and Theorem 2.1 follows.

Proof of Lemma 5.1. We show that there exists a minimal polynomial realization of the form (5.1) with

$$z_i \ge x_i + 1 \quad \text{for } i = 0, \dots, r-1. \tag{5.8}$$

We start from an arbitrary minimal realization (5.1). Let $i_0 = \min\{i \mid x_i = z_i\}$. By replacing ℓ_{i_0+1} with the affine map ℓ passing through the two points $(x_{i_0}, h_{x_{i_0}}), (x_{i_0} + 1, h_{x_{i_0}+1})$, we obtain a new decomposition of the form (5.1) with $i_0 < i_0' = \min\{j \mid x_j' = z_j'\}$, where x_j', z_j' are points defined by (5.3) for the new polynomial realization. Indeed, $x_i' = x_i, z_i' = z_i$ for all $i < i_0$ and $x_{i_0}' = x_{i_0}, z_{i_0}' \ge x_{i_0} + 1 > x_{i_0} = x_{i_0}'$. Note that $\ell \ne \ell_{i_0+2}$; otherwise ℓ_{i_0+1} could be removed from p and the arising function would still be a polynomial realization of h which contradicts the minimality of p. After a finite number of such replacements, we get $z_i > x_i$ for all i, so that (5.5) becomes satisfied. □

Proof of Lemma 5.3. From the definition of M we have

$$M = (h_{i+u_j}) = (p(i + u_j))_{i,j=0,\dots,r-1}.$$

We prove that for all permutations σ of $0,\dots,r-1$ such that $\sigma \neq \mathrm{Id}$,

$$\bigotimes_{i=0}^{r-1} p(i+u_i) > \bigotimes_{i=0}^{r-1} p(i+u_{\sigma(i)}). \tag{5.9}$$

Clearly, it is sufficient to show that for all elementary cycles $c=(i_1,\dots,i_k)$ $(k\geq 2)$,

$$\begin{aligned} &p(i_1+u_{i_1})\otimes p(i_2+u_{i_2})\otimes\cdots\otimes p(i_k+u_{i_k}) \\ &> \ p(i_1+u_{i_2})\otimes p(i_2+u_{i_3})\otimes\cdots\otimes p(i_k+u_{i_1}) \end{aligned} \tag{5.10}$$

or with the conventional notation

$$\begin{aligned} &p(i_1+u_{i_1})+p(i_2+u_{i_2})+\cdots+p(i_k+u_{i_k}) \\ &> \ p(i_1+u_{i_2})+p(i_2+u_{i_3})+\cdots+p(i_k+u_{i_1}). \end{aligned} \tag{5.11}$$

Let $\alpha_1,\dots,\alpha_k$, with $\alpha_1<\cdots<\alpha_k$, denote the sequence obtained by reordering the $x_l=i_l+u_{i_l}$ (since $x_t<x_s \Leftrightarrow t<s \Leftrightarrow u_t<u_s$, i_l and u_{i_l} are ordered in the same way), and let $\alpha'_1,\dots,\alpha'_k$ with $\alpha'_1\leq\cdots\leq\alpha'_k$ denote the sequence obtained by reordering the $i_l+u_{i_{l+1}}$.

We claim that $\alpha'\prec\alpha$. Indeed, condition 1 of Theorem 4.1 is satisfied, because $\alpha_\nu+\cdots+\alpha_k$ is equal to the sum of the $k-\nu+1$ highest possible values of i_l and u_{i_l}; hence, it is greater than $\alpha'_\nu+\cdots+\alpha'_k$ which is also the sum of $k-\nu+1$ values of i_l and u_{i_l}.

Moreover, take P such that $\alpha'=P\alpha$ as in Theorem 4.1. Since P is doubly stochastic, there is at least one $j\in\{1,2,\dots,k\}$ such that $P_{jk}\neq 0$. Since $\alpha'_j\leq\alpha'_k<\alpha_k$, we have $P_{jk}\neq 1$; thus there is at least one $m\in\{1,2,\dots,k-1\}$ such that $P_{jm}\neq 0$. It remains to apply Corollary 5.2 together with the strict inequality case in Theorem 4.2 to get (5.11). □

Example 1. The function $p(x)=\max(0,-3+x,-8+2x,-22+4x)$ is a polynomial realization of the sequence $h=0,0,0,0,1,2,4,6,10,14,\dots$. From (5.3) we have

$$x_0=0,\quad x_1=4,\quad x_2=6,\quad x_3=8,$$

$$u_0=0,\quad u_1=3,\quad u_2=4,\quad u_3=5,$$

$$\det\mathcal{H}[0,1,2,3|0,3,4,5]=\det\begin{bmatrix}0&0&1&2\\0&1&2&4\\0&2&4&6\\0&4&6&10\end{bmatrix}=(15,14),$$

which is unbalanced. Hence, $\mathsf{mlr}(h)=4$ and a minimal linear realization of h is (A,b,c), where $A=\mathrm{diag}(0,1,2,4)$, $c=(0,-3,-8,-22)$, $b=(0,0,0,0)^T$. Note that this minimal realization is not unique.

6. Appendix. Now we develop a method for finding a polynomial realization of the minimal dimension.

Suppose that $h_0,h_1,\dots$ is a convex sequence satisfying (2.1). Consider the points $P_j=[j,h_j], j=0,1,\dots,M$, in the plane with Cartesian coordinate system. If $M=N+1$, then p is a polynomial realization of h iff $p(j)=h_j$ for $j=0,\dots,M$,

so that we may restrict our investigation to a method for finding a polynomial realization of the finite sequence $P_0, \ldots, P_M$. It is evident that to every polynomial realization (which in general may contain redundant monomials) we can assign another polynomial realization of no greater dimension in which every line (monomial) passes through at least two points of the set $\{P_0, \ldots, P_M\}$. Indeed, in what follows we consider only polynomial realizations which possess this property.

A subset T of the set

$$S = \{[j, h_j] \mid \ j = 0, 1, \ldots, M\}$$

is called *aligned* if $\| T \| \geq 3$ and there exists a line q such that

1. $T \subseteq q$,
2. $(S - T) \cap q = \emptyset$.

Note that an aligned subset of S may not exist and that two different aligned subsets are either disjoint or have exactly one common point.

Let T be a fixed aligned set of points lying on a line q. Since $\| T \| \geq 3$, $[j - 1, h_{j-1}], [j, h_j], [j + 1, h_{j+1}] \in q$ for some j.

Let $s(t)$ represent a line of an arbitrary polynomial realization which passes through $[j, h_j]$. Since

$$h_{j-1} \geq s(j - 1), \qquad h_{j+1} \geq s(j + 1),$$

and $[j - 1, h_{j-1}], [j, h_j], [j + 1, h_{j+1}]$ are collinear, we have that q coincides with s. Hence every polynomial realization contains each line passing through all points of an aligned subset, and thus in the construction of the minimal polynomial realization we must always include these lines. This concerns also the two special lines: one covering $[0, h_0], [1, h_1]$ and the other passing through $[M - 1, h_{M-1}], [M, h_M]$, which by trivial reasons must be involved too.

The point $[j, h_j]$ $(0 < j < M)$ is called a *breaking point* if $[j - 1, h_{j-1}], [j, h_j], [j + 1, h_{j+1}]$ are not collinear. Clearly, the first and last points of every aligned set are breaking points (except $[0, h_0], [M, h_M]$).

Consider a fixed set

$$B = \{[r, h_r], [r + 1, h_{r+1}], \ldots, [s, h_s]\}$$

of consecutive breaking points which is maximal; i.e., both $[r, h_r]$ belongs to an aligned set or $r = 1$ and $[s, h_s]$ belongs to an aligned set or $s = M - 1$. Hence both $[r, h_r]$ and $[s, h_s]$ can be assumed to belong already to a line of the realization. Clearly, a line cannot pass through more than two consecutive breaking points. If B consists of k points, then the minimal number of lines joining pairs of consecutive points which contain all the points in B (except for the extreme points $[r, h_r], [s, h_s]$) is

$$\left\lceil \frac{k}{2} \right\rceil - 1.$$

A self-evident strategy to achieve this lower bound is to take alternatively every other line consecutively joining the pairs of adjacent points starting by the line passing through

$$[r + 1, h_{r+1}], \quad [r + 2, h_{r+2}].$$

The foregoing discussion justifies the following algorithm, which starts from $[0, h_0]$. (Note that $\ell \equiv (P, Q)$ reads, "line ℓ determined by the points P and Q.")

Algorithm. MINIMAL POLYNOMIAL REALIZATION (MPR)

Input: Finite sequence $h_0, \ldots, h_M$ of real numbers, with $M \geq 3$.

Output: Polynomial realization of h of minimal dimension represented by the lines $\ell_1, \ell_2, \ldots$; "no" if h is not convex.

(1) $r := 1$, $s := 0$, $j := 0$, $P_i = [i, h_i]$, $(i = 0, 1, \ldots, M)$.
(2) Accept $\ell_r \equiv (P_j, P_{j+1})$.
(3) $r := r + 1$.
(4) Let j $(j > s)$ be the least index for which ℓ_{r-1} does not pass through P_j (if no such exists then stop).
(5) If P_j is below ℓ_{r-1} then stop ("no").
(6) If $j = M$ then go to (8).
(7) $s := j$ and go to (2).
(8) Accept $\ell_r \equiv (P_{j-1}, P_j)$, stop.

Example 2. Input: $h = 1, 0, -1, -1, 0, 2, 5, 8$. $(M = 7)$

$P_0 = (0, 1)$, $P_1 = (1, 0)$, $P_2 = (2, -1)$, $P_3 = (3, -1)$, $P_4 = (4, 0)$, $P_5 = (5, 2)$, $P_6 = (6, 5)$, $P_7 = (7, 8)$.

The algorithm MPR produces the following:

$r = 1, s = 0$		
accept $\ell_1 \equiv (P_0, P_1)$	accept $\ell_2 \equiv (P_3, P_4)$	accept $\ell_3 \equiv (P_5, P_6)$
$r = 2$	$r = 3$	$r = 4$
$j = 3$	$j = 5$	
$s = 3$	$s = 5$	
		stop

$p_1(t) = -t + 1$ (line ℓ_1)
$p_2(t) = t - 4$ (line ℓ_2)
$p_3(t) = 3t - 13$ (line ℓ_3)
$A = \mathrm{diag}(-1, 1, 3)$, $c = (1, -4, -13)$, $b = (0, 0, 0)^T$.

REFERENCES

[1] F. Baccelli, G. Cohen, G. Olsder, and J. Quadrat, *Synchronization and Linearity*, John Wiley, New York, 1992.

[2] R. B. Bapat, *Permanents, max algebra and optimal assignment*, Linear Algebra Appl., 226/228 (1995), pp. 73–86.

[3] J. Berstel and C. Reutenauer, *Rational Series and Their Languages*, Springer-Verlag, New York, 1988.

[4] G. Birkhoff and S. MacLane, *A Survey of Modern Algebra*, Macmillan, New York, 1965.

[5] R. Cuninghame-Green, *Minimax Algebra*, Lecture Notes in Econom. and Math. Systems 166, Springer-Verlag, New York, 1979.

[6] R. Cuninghame-Green, *Algebraic realization of discrete dynamic systems*, in Proc. of the 1991 IFAC Workshop on Discrete Event System Theory and Applications in Manufacturing and Social Phenomena, Shenyang, China, June 1991.

[7] R. Cuninghame-Green and P. Butkovič, *Discrete-event dynamic systems: The strictly convex case*, Ann. Oper. Res., 57 (1995), pp. 45–63.

[8] R. Cuninghame-Green and P. Meijer, *An algebra for piecewise-linear minimax problems*, Discrete Appl. Math., 2 (1980), pp. 267–294.

[9] B. De Schutter, *Max-Algebraic System Theory for Discrete-Event Systems*, Ph.D. thesis, Katholieke Univ. Leuven, Feb. 1996.

[10] B. De Schutter and B. De Moor, *The characteristic equation and minimal state space realization of siso systems in the max algebra*, in Proc. of the 11th Conf. on Anal. and Opt. of Systems: Discrete-Event Systems, Lect. Notes. in Control and Inform. Sci. 199, Springer-Verlag, New York, 1994.

[11] B. De Schutter and B. De Moor, *Minimal realization in the max algebra is an extended linear complementarity problem*, Sys. Control Lett., 25 (1995), pp. 103–111.

[12] B. De Schutter and B. De Moor, *Minimal state space realization of mimo systems in the max algebra*, in Proc. of the 3rd European Control Conference, Roma, Italy, Sept. 1995, pp. 411–416.
[13] M. Fliess, *Matrices de Hankel*, J. Math. Pures. Appl., 53 (1974), pp. 197–222.
[14] M. Fliess, *Séries rationnelles positives et processus stochastiques*, Ann. Inst. H. Poincaré, XI (1975), pp. 1–21.
[15] S. Gaubert, *Théorie des systèmes linéaires dans les dioïdes*, thèse, École des Mines de Paris, 1992.
[16] S. Gaubert, *On Rational Series in One Variable over Certain Dioids*, Rapport de Recherche 2162, INRIA, Le Chesnay, France, Jan. 1994.
[17] S. Gaubert, *Rational series over dioids and discrete-event systems*, in Proc. of the 11th Conf. on Anal. and Opt. of Systems: Discrete-Event Systems, Lect. Notes. in Control and Inform. Sci. 199, Springer-Verlag, New York, 1994.
[18] J. S. Golan, *The Theory of Semirings with Applications in Mathematics and Theoretical Computer Science*, Pitman Monogr. Surveys Pure Appl. Math. 54, Longman Scientific and Technical Publishers, Harlow, U.K., 1992.
[19] M. Gondran and M. Minoux, *Graphes et algorithmes*, Eyrolles, Paris, 1979. Engl. transl. *Graphs and Algorithms*, John Wiley, New York, 1984.
[20] M. Gondran and M. Minoux, *Linear algebra in dioids: A survey of recent results*, Ann. Discrete Math., 19 (1984), pp. 147–164.
[21] J. Gunawardena, ed., *Idempotency*, Publications of the Newton Institute, Cambridge University Press, London, 1996.
[22] G. Hansel and D. Perrin, *Mesures de probabilités rationnelles*, in Mots, M. Lothaire, ed., Hermes, Paris, 1990.
[23] G. Hardy, J. Littlewood, and G. Pólya, *Inequalities*, Cambridge University Press, London, 1934, reprinted 1973.
[24] K. Kim, *Boolean Matrix Theory and Applications*, Marcel Dekker, New York, 1982.
[25] M. Marcus and H. Minc, *A Survey of Matrix Theory and Matrix Inequalities*, Allyn and Bacon, Boston, 1964.
[26] A. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications*, Academic Press, New York, 1979.
[27] V. Maslov and S. Samborskiĭ, eds., *Idempotent analysis*, Adv. in Soviet Math. 13, AMS, Providence, RI, 1992.
[28] G. Olsder, *On the characteristic equation and minimal realizations for discrete-event dynamic systems*, in Analysis and Optimization of Systems, A. Bensoussan and J. Lions, eds., in Lecture Notes in Control and Inform. Sci. 83, Springer-Verlag, New York, 1986, pp. 189–201.
[29] G. Olsder, *Some results on the minimal realization of discrete-event systems*, in 25th IEEE Conf. on Decision and Control, Athens, Greece, 1986.
[30] M. Plus, *Linear systems in* $(\max,+)$*-algebra*, in Proceedings of the 29th Conference on Decision and Control, Honolulu, Dec. 1990.
[31] X. Qi and W. Chen, *The minimal realization of discrete-event systems*, in Proc. of the 1991 IFAC Workshop on Discrete Event System Theory and Applications in Manufacturing and Social Phenomena, Shenyang, China, June 1991, pp. 29–33.
[32] C. Reutenauer and H. Straubing, *Inversion of matrices over a commutative semiring*, J. Algebra, 88 (1984), pp. 350–360.
[33] D. Zeilberger, *A combinatorial approach to matrix algebra*, Discrete Math., 56 (1985), pp. 61–72.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 1, pp. 148–163, January 1998

008

STABILIZATION OF ELASTIC PLATES WITH DYNAMICAL BOUNDARY CONTROL*

BOPENG RAO†

Abstract. We consider a hybrid system composed of a plate equation and two ordinary differential equations. We prove that the system is strongly but not uniformly stable. By a new approach, we show that the smooth solution has a rational decay rate. Finally we establish the uniform energy decay rate for a simplified hybrid system.

Key words. hybrid system, compact boundary feedback, lack of uniform stability, rational energy decay rate

AMS subject classifications. 35B40, 35M10, 93C15, 93C20, 93D15

PII. S0363012996300975

1. Introduction.

The purpose of this work (the results of which were announced in Rao [18]) is to investigate the stabilization of a thin elastic plate, which is clamped on one part of the edge and rimmed along the other part with a flange that has mass and moment of inertia of the boundary. Let $\Omega \subset \mathbb{R}^2$ denote a bounded domain, with smooth boundary Γ consisting of two disjoint pieces: Γ_0, the clamped part, and Γ_1, the rimmed part. Following the linear elasticity theory (see Lagnese and Lions [6]), the vibration y of the plate is governed by the plate equation associated with two dynamical boundary conditions:

$$\begin{cases} y'' + \Delta^2 y = 0 & \text{in } \Omega\times]0,+\infty[, \\ y = \partial_\nu y = 0 & \text{on } \Gamma_0\times]0,+\infty[, \\ J\partial_\nu y'' + \Delta y + (1-\mu)B_1 y = m & \text{on } \Gamma_1\times]0,+\infty[, \\ \rho y'' - \partial_\nu \Delta y - (1-\mu)\partial_\tau B_2 y = f & \text{on } \Gamma_1\times]0,+\infty[, \end{cases} \tag{1.1}$$

where $\nu = (\nu_1, \nu_2)$ is the unit outer normal vector and $\tau = (-\nu_2, \nu_1)$ is the unit tangent vector. The inertial properties of the boundary are supported along the rimmed part Γ_1 whereon the boundary feedback controls m, f are applied. $0 < \mu < 1/2$ is the Poisson coefficient. $\rho > 0$ is the linear boundary density; $J > 0$ is the bending moment of inertia of the boundary. B_1, B_2 are the usual boundary operators associated with the plate equation:

$$B_1 y = 2\nu_1\nu_2 \frac{\partial^2 y}{\partial x_1 \partial x_2} - \nu_1^2 \frac{\partial^2 y}{\partial x_2^2} - \nu_2^2 \frac{\partial^2 y}{\partial x_1^2}, \tag{1.2}$$

$$B_2 y = (\nu_1^2 - \nu_2^2)\frac{\partial^2 y}{\partial x_1 \partial x_2} + \nu_1\nu_2\Big(\frac{\partial^2 y}{\partial x_1^2} - \frac{\partial^2 y}{\partial x_2^2}\Big). \tag{1.3}$$

In the present work, we use the following boundary feedback controls:

$$m = -L\partial_\nu y', \qquad f = -Ly', \tag{1.4}$$

*Received by the editors March 25, 1996; accepted for publication (in revised form) October 30, 1996.

http://www.siam.org/journals/sicon/36-1/30097.hmtl

†Institut de Recherche Mathématique Avancée, Université de Louis Pasteur de Strasbourg, 7 Rue René-Descartes, 67084 Strasbourg cedex, France (rao@math.u-strasbg.fr).

where L is the canonical isomorphism from $H^{-s}(\Gamma_1)$ onto $H^s(\Gamma_1)$. If we consider the usual initial data, $y(0) \in H^2(\Omega), y_t(0) \in L^2(\Omega)$, then the regularity of the weak solution is insufficient for defining the dynamical terms $J\partial_\nu y''$ and $\rho y''$ on the boundary Γ_1. Following an idea of Slemrod [21], the dynamical boundary conditions have to be treated as ordinary differential equations in the time variable. Therefore, indeed we have a system made up of one partial differential equation and two ordinary differential equations, called a hybrid system.

For these kinds of problems, a classical method developed in Littman and Markus [11] is based on the spectrum theory. Roughly speaking, if the spectrum of the system approaches asymptotically the imaginary axis, then the system loses the uniform energy decay rate. Further, if the eigenvectors of the system form a Riesz basis, then the smooth solution has a rational decay rate. Because of the essential difficulty intervening in the determination of the spectrum of the system, this method is obviously limited to one-dimensional problems (Lee and You [8], Littman and Markus [11, 12]).

In [17], we considered the well-known SCOLE model (Euler–Bernoulli beam equation associated with dynamical boundary conditions). We introduced a new method, which is based on a result of compact perturbation due to Russell [20]. We established the uniform stability in the case of hinged beam with the usual boundary feedback controls, as well as in the case of a clamped beam with the boundary feedback controls of high order. In the case of a clamped beam with the usual boundary feedback controls, we proved that the system loses the uniform energy decay rate. Unlike the usual method, this method does not require any knowledge of the spectrum of the system. For other applications, we refer to Rao [19] for the Rayleigh beam equation and Komornik and Rao [4] for compactly coupled waves equations.

Now we outline briefly the content of this work. In section 2, we formulate the system (1.1)–(1.4) into an evolutionary equation,

$$u' + Au + Bu = 0, \tag{1.5}$$

where A is a maximal monotone operator and B is the canonical isomorphism from $H^{-s}(\Gamma_1) \times H^{-s}(\Gamma_1)$ onto $H^s(\Gamma_1) \times H^s(\Gamma_1)$. Therefore, we can give a semigroup approach of the system (1.1)–(1.4). In section 3, we prove the strong stability of the system (1.5) for all $s \geq 0$. Therefore, we improve a recent result of Markus and You [13], which asserts the strong stability of the system (1.5) in the case $s = 0$. Further, we show that the system (1.5) actually loses the uniform energy decay rate if $s > 0$. Notice that in this case the control operator B is compact. In section 4, we establish the rational energy decay rate for the smooth solution. To this end, we introduce a new multiplier method which is based on a nonlinear technique. Once again, this method does not require any knowledge of the spectrum of the system. In section 5, we consider a simplified plate model in which we have taken $J = 0$. We prove that the usual multiplier method can be applied for obtaining the uniform energy decay rate of the system.

2. Well-posedness. Throughout this paper, we assume that $\Omega \subset \mathbb{R}^2$ is a bounded domain with smooth boundary $\Gamma = \Gamma_0 \cup \Gamma_1$. We assume that Γ_0 and Γ_1 are of class C^2 and $\bar{\Gamma}_0 \cap \bar{\Gamma}_1 = \{\emptyset\}$.

Let us denote by L the canonical isomorphism from $H^{-s}(\Gamma_1)$ onto $H^s(\Gamma_1)$:

$$\langle L\eta, \zeta\rangle_{H^s(\Gamma_1)\times H^{-s}(\Gamma_1)} = \langle \eta, \zeta\rangle_{H^{-s}(\Gamma_1)}, \qquad s \geq 0. \tag{2.1}$$

If $s = 0$, then L is the identity of $L^2(\Gamma_1)$; we find again the usual boundary feedback controls used in Markus and You [13].

Now let y be a smooth solution of the closed-loop hybrid system

$$\begin{cases} y'' + \Delta^2 y = 0 & \text{in } \Omega\times]0,+\infty[, \\ y = \partial_\nu y = 0 & \text{on } \Gamma_0\times]0,+\infty[, \\ J\partial_\nu y'' + \Delta y + (1-\mu)B_1 y + L\partial_\nu y' = 0 & \text{on } \Gamma_1\times]0,+\infty[, \\ \rho y'' - \partial_\nu \Delta y - (1-\mu)\partial_\tau B_2 y + L y' = 0 & \text{on } \Gamma_1\times]0,+\infty[. \end{cases} \tag{2.2}$$

Setting

$$z = y', \qquad \xi = \partial_\nu y'|_{\Gamma_1}, \qquad \eta = y'|_{\Gamma_1}, \tag{2.3}$$

then we transform the hybrid system (2.2) into a system of first order

$$\begin{pmatrix} y \\ z \\ \xi \\ \eta \end{pmatrix}' + \begin{pmatrix} -z \\ \Delta^2 y \\ \frac{1}{J}\big(\Delta y + (1-\mu)B_1 y\big) \\ -\frac{1}{\rho}\big(\partial_\nu \Delta y + (1-\mu)\partial_\tau B_2 y\big) \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \frac{1}{J}L\xi \\ \frac{1}{\rho}L\eta \end{pmatrix} = 0. \tag{2.4}$$

Denoting the columns by u', Au, and Bu, we get an abstract evolutionary equation

$$u' + Au + Bu = 0, \qquad u(0) = u_0 \in H. \tag{2.5}$$

According to the formulation (2.4), we are led to introduce the following energy space:

$$H = H^2_{\Gamma_0}(\Omega) \times L^2(\Omega) \times L^2(\Gamma_1) \times L^2(\Gamma_1), \tag{2.6}$$

where we have put

$$H^2_{\Gamma_0}(\Omega) = \{y \in H^2(\Omega) \ : \quad y = \partial_\nu y = 0 \quad \text{on } \Gamma_0\}. \tag{2.7}$$

For $u = (y, z, \xi, \eta), \widetilde{u} = (\widetilde{y}, \widetilde{z}, \widetilde{\xi}, \widetilde{\eta}) \in H$, we define the inner product by

$$(u, \widetilde{u})_H = \int_\Omega \big(a(y,\widetilde{y}) + z\widetilde{z}\big)dx + J\int_{\Gamma_1} \xi\widetilde{\xi} d\Gamma + \rho\int_{\Gamma_1} \eta\widetilde{\eta} d\Gamma, \tag{2.8}$$

where the bilinear form $a(\cdot,\cdot)$ is defined by

$$\begin{aligned} a(y,z) \ :=& \frac{\partial^2 y}{\partial x_1^2}\frac{\partial^2 z}{\partial x_1^2} + \frac{\partial^2 y}{\partial x_2^2}\frac{\partial^2 z}{\partial x_2^2} \\ &+\mu\Big(\frac{\partial^2 y}{\partial x_1^2}\frac{\partial^2 z}{\partial x_2^2} + \frac{\partial^2 y}{\partial x_2^2}\frac{\partial^2 z}{\partial x_1^2}\Big) + 2(1-\mu)\frac{\partial^2 y}{\partial x_1 \partial x_2}\frac{\partial^2 z}{\partial x_1 \partial x_2} \ . \end{aligned} \tag{2.9}$$

We next introduce the linear bounded operator B and the linear unbounded operator A as follows:

$$Bu = \begin{pmatrix} 0 \\ 0 \\ \frac{1}{\rho}L\xi \\ \frac{1}{J}L\eta \end{pmatrix}, \qquad Au = \begin{pmatrix} -z \\ \Delta^2 y \\ \frac{1}{\rho}\big(\Delta y + (1-\mu)B_1 y\big) \\ -\frac{1}{J}\big(\partial_\nu \Delta y + (1-\mu)\partial_\tau B_2 y\big) \end{pmatrix}. \tag{2.10}$$

The domain $D(A)$ is defined by

$$D(A) = \begin{pmatrix} u = (y, z, \xi, \eta) \in W \times H^2_{\Gamma_0}(\Omega) \times L^2(\Gamma_1) \times L^2(\Gamma_1) \\ \text{such that} \quad \xi = \partial_\nu z|_{\Gamma_1} \quad \text{and} \quad \eta = z|_{\Gamma_1} \end{pmatrix}, \tag{2.11}$$

where the subspace W is defined as

$$W = \begin{pmatrix} y \in H^2_{\Gamma_0}(\Omega), \quad \Delta^2 y \in L^2(\Omega) \\ \Delta y + (1-\mu) B_1 y \in L^2(\Gamma_1) = v_1 \\ \partial_\nu \Delta y + (1-\mu)\partial_\tau B_2 y = v_2 \in L^2(\Gamma_1) \end{pmatrix}. \tag{2.12}$$

The trace functions in (2.12) are defined by means of the following Green formula:

$$\int_\Omega \Delta^2 y \phi dx = \int_\Omega a(y, \phi) dx + \int_{\Gamma_1} v_2 \phi d\Gamma - \int_{\Gamma_1} v_1 \phi d\Gamma \qquad \forall \phi \in H^2_{\Gamma_0}(\Omega). \tag{2.13}$$

LEMMA 2.1. *The operator B is self-adjoint nonnegative definite and compact if $s > 0$.*

Proof. Let $u = (y, z, \xi, \eta) \in H$. Then using (2.8) and (2.10) we have

$$(Bu, u)_H = (L\xi, \xi)_{L^2(\Gamma_1)} + (L\eta, \eta)_{L^2(\Gamma_1)}. \tag{2.14}$$

Since $\xi \in L^2(\Gamma_1), L\xi \in H^s(\Gamma_1)$, we have

$$(L\xi, \xi)_{L^2(\Gamma_1)} = \langle L\xi, \xi \rangle_{H^s(\Gamma_1) \times H^{-s}(\Gamma_1)} = \|\xi\|^2_{H^{-s}(\Gamma_1)}.$$

It follows from (2.14) that

$$(Bu, u)_H = \|\xi\|^2_{H^{-s}(\Gamma_1)} + \|\eta\|^2_{H^{-s}(\Gamma_1)} \geq 0. \tag{2.15}$$

Since L is self-adjoint and compact for $s > 0$, so is B. The proof is complete. □

LEMMA 2.2. *The operator A is maximal monotone and skew adjoint. Moreover the resolvent $(I + A)^{-1}$ is compact in H.*

Proof. Let $u = (y, z, \xi, \eta) \in D(A)$. Using (2.8) and (2.10) we have

$$(Au, u)_H = \int_\Omega a(-z, y) dx + \int_\Omega \Delta^2 yz dx$$

$$- \int_{\Gamma_1} (\partial_\nu \Delta y + (1-\mu)\partial_\tau B_2 y) z d\Gamma_1 + \int_{\Gamma_1} (\Delta y + (1-\mu) B_1 y) \partial_\nu z d\Gamma.$$

Since $y \in W$ and $z \in H^2_{\Gamma_0}(\Omega)$, it follows from (2.13) that

$$(Au, u)_H = 0 \qquad \forall u \in D(A). \tag{2.16}$$

Now given $u_0 \in H$, we solve the equation $u + Au = u_0$. This means that

$$\begin{cases} y - z = y_0, \\ z + \Delta^2 y = z_0, \\ J\partial_\nu z + \Delta y + (1-\mu) B_1 y = J\xi_0, \\ \rho z - \partial_\nu \Delta y - (1-\mu)\partial_\tau B_2 y = \rho \eta_0. \end{cases} \tag{2.17}$$

Then eliminating z and using $\xi = \partial_\nu z|_{\Gamma_1}, \eta = z|_{\Gamma_1}$, we find that y satisfies the system

$$\begin{cases} y + \Delta^2 y = y_0 + z_0 \in L^2(\Omega), \\ y = \partial_\nu y = 0 \qquad \text{on } \Gamma_0, \\ J\partial_\nu y + \Delta y + (1-\mu)B_1 y = J(\xi_0 + \partial_\nu y_0) \qquad \text{on } \Gamma_1, \\ \rho y - \partial_\nu \Delta y - (1-\mu)\partial_\tau B_2 y = \rho(\eta_0 + y_0) \qquad \text{on } \Gamma_1. \end{cases} \tag{2.18}$$

Using the Green formula (2.13), we prove that the system (2.18) is equivalent to the following variational equation:

$$\begin{aligned} &\int_\Omega \big(y\phi + a(y,\phi)\big)dx + \rho\int_{\Gamma_1} y\phi d\Gamma + J\int_{\Gamma_1} \partial_\nu y \partial_\nu \phi d\Gamma \\ &= \int_\Omega \big(y_0 + z_0\big)\phi dx + \rho\int_{\Gamma_1} (\eta_0 + y_0)\phi + J\int_{\Gamma_1} (\xi_0 + \partial_\nu y_0)\partial_\nu \phi d\Gamma \end{aligned} \tag{2.19}$$

for all $\phi \in H^2_{\Gamma_0}(\Omega)$. Thanks to the Lax–Milgram theorem, equation (2.19) admits a unique solution $y \in H^2_{\Gamma_0}(\Omega)$. Then defining

$$z = y - y_0 \in H^2_{\Gamma_0}(\Omega), \quad \xi = \partial_\nu z|_{\Gamma_1}, \quad \eta = z|_{\Gamma_1},$$

we see that $u = (y, z, \xi, \eta) \in D(A)$ and solves the equation $u + Au = u_0$. Therefore, we conclude that A is maximal monotone in the energy space H.

On the other hand, thanks to the elliptic theory (Lions and Magenes [9]), we see that $y \in H^{5/2}(\Omega)$. It follows that

$$\|(y, z, \eta, \xi)\|_{H^{5/2}(\Omega)\times H^2_{\Gamma_0}(\Omega)\times H^{3/2}(\Gamma_1)\times H^{1/2}(\Gamma_1)} \le C\|u_0\|_H. \tag{2.20}$$

In particular, we obtain the compactness of the resolvent $(I + A)^{-1}$.

Finally, since A is antisymmetric and maximal monotone with compact resolvent, A is therefore skew adjoint in H (see Brezis [1]). The proof is thus complete. □

THEOREM 2.3. *Assume that $s \ge 0$. Then*

(i) *for any $u_0 \in D(A)$, equation (2.5) admits a unique strong solution $u(t) = \big(y(t), z(t), \xi(t), \eta(t)\big)$ such that*

$$y(t) \in C^0\big(\mathbb{R}^+; H^{5/2}(\Omega)\big) \cap C^1\big(\mathbb{R}^+; H^2_{\Gamma_0}(\Omega)\big) \cap C^2\big(\mathbb{R}^+; L^2(\Omega)\big), \tag{2.21}$$

$$y|_{\Gamma_1} \in C^2\big(\mathbb{R}^+; L^2(\Gamma_1)\big), \qquad \partial_\nu y|_{\Gamma_1} \in C^2\big(\mathbb{R}^+; L^2(\Gamma_1)\big). \tag{2.22}$$

(ii) *for any $u_0 \in H$, equation (2.5) admits a unique weak solution $u(t) = \big(y(t), z(t), \eta(t), \xi(t)\big)$ such that*

$$y(t) \in C^0\big(\mathbb{R}^+; H^2_{\Gamma_0}(\Omega)\big) \cap C^1\big(\mathbb{R}^+; L^2(\Omega)\big), \tag{2.23}$$

$$y|_{\Gamma_1} \in C^1\big(\mathbb{R}^+; L^2(\Gamma_1)\big), \qquad \partial_\nu y|_{\Gamma_1} \in C^1\big(\mathbb{R}^+; L^2(\Gamma_1)\big). \tag{2.24}$$

Proof. Since A is maximal monotone and B is nonnegative definite, the operator $A + B$ with the domain $D(A + B) = D(A)$ is maximal monotone. Applying the Hille–Yosida theorem (see Brezis [1] and Pazy [14]), we know that for any $u_0 \in D(A)$ equation (2.5) admits a unique strong solution:

$$u(t) = \big(y(t), z(t), \xi(t), \eta(t)\big) \in C^0\big(\mathbb{R}^+; D(A)\big) \cap C^1\big(\mathbb{R}^+; H\big). \tag{2.25}$$

On the other hand, thanks to (2.20), we have $D(A) \subset H^{5/2}(\Omega) \times H^2_{\Gamma_0}(\Omega) \times H^{3/2}(\Gamma_1) \times H^{1/2}(\Gamma_1)$. It follows that

$$\begin{cases} y(t) \in C^0\big(\mathbb{R}^+; H^{5/2}(\Omega)\big) \cap C^1\big(\mathbb{R}^+; H^2_{\Gamma_0}(\Omega)\big), \\ y_t(t) \in C^0\big(\mathbb{R}^+; H^2_{\Gamma_0}(\Omega)\big) \cap C^1\big(\mathbb{R}^+; L^2(\Omega)\big), \\ y(t)|_{\Gamma_1} \in C^0\big(\mathbb{R}^+; H^{3/2}(\Gamma_1)\big) \cap C^1\big(\mathbb{R}^+; L^2(\Gamma_1)\big), \\ \partial_\nu y(t)|_{\Gamma_1} \in C^0\big(\mathbb{R}^+; H^{1/2}(\Gamma_1)\big) \cap C^1\big(\mathbb{R}^+; L^2(\Gamma_1)\big). \end{cases} \tag{2.26}$$

This gives (2.21)–(2.22).

Now let $u_0 \in H$; then equation (2.5) admits a unique weak solution given by $u(t) = S_{A+B}(t)u_0$, where $S_{A+B}(t)$ is the semigroup of contractions generated by the operator $-(A+B)$ on the energy space H. Moreover we have

$$u(t) = \big(y(t), z(t), \xi(t), \eta(t)\big) \in C^0\big(\mathbb{R}^+; H\big). \tag{2.27}$$

Interpreting (2.27) gives (2.23)–(2.24). The proof is complete. □

Notice that for a general function y possessing only the smoothness property (2.21), the trace functions $y''|_{\Gamma_1}$ and $\partial_\nu y''|_{\Gamma_1}$ don't make sense. Here they are defined by means of the equations

$$\begin{cases} y'' = \frac{1}{\rho}\big(\partial_\nu \Delta y + (1-\mu)\partial_\tau B_2 y\big) & \text{on } \Gamma_1, \\ \partial_\nu y'' = -\frac{1}{J}\big(\Delta y + (1-\mu) B_1 y\big) & \text{on } \Gamma_1. \end{cases} \tag{2.28}$$

From (2.21), we see that the right-hand sides of these equations are continuous functions.

3. Strong stability and lack of uniform energy decay rate. Let $u = (y, z, \xi, \eta)$ be a solution of equation (2.5). We define the associated energy by

$$E(t) = \frac{1}{2}\Big\{ \int_\Omega \big(|y'|^2 + a(y,y)\big)dx + \int_{\Gamma_1} \big(\rho|y'|^2 + J|\partial_\nu y'|^2\big)d\Gamma \Big\}. \tag{3.1}$$

Then using (2.15)–(2.16) we have

$$\frac{d}{dt}E(t) = -\|y'\|^2_{H^{-s}(\Gamma_1)} - \|\partial_\nu y'\|^2_{H^{-s}(\Gamma_1)} \le 0. \tag{3.2}$$

Therefore, $E(t)$ is a nonincreasing function. Moreover, we have the following strong stability result.

THEOREM 3.1. *For any initial data $u_0 \in H$, the energy $E(t)$ of system (2.5) satisfies*

$$\lim_{t\to+\infty} E(t) = 0. \tag{3.3}$$

Proof. Since B is self-adjoint and nonnegative definite, there exists a linear operator $\widetilde{B}$ such that $B = \widetilde{B}\widetilde{B}^*$. Then indeed we have

$$(Bu, u)_H = (\widetilde{B}^* u, \widetilde{B}^* u)_H, \tag{3.4}$$

which, together with (2.15), implies that

$$Bu = 0 \Leftrightarrow \widetilde{B}^* u = 0. \tag{3.5}$$

Applying LaSalle's invariance principle (see Slemrod [21]), it is sufficient to prove that the equation $\widetilde{B}^* S_A(t)u_0 = 0$ has only trivial solution $u_0 = 0$, where $S_A(t)$ denotes the semigroup generated by the operator A on the space H. Thanks to (3.5), this is equivalent to proving that the equation $BS_A(t)u_0 = 0$ has only trivial solution $u_0 = 0$. Assume that $u_0 \in D(A)$. Then indeed writing $S_A(t)u_0 = \big(y(t), z(t), \eta(t), \xi(t)\big)$, we see that y satisfies the system

$$\begin{cases} y'' + \Delta^2 y = 0 & \text{in } \Omega\times]0,+\infty[, \\ J\partial_\nu y'' + \Delta y + (1-\mu)B_1 y = 0 & \text{on } \Gamma_1\times]0,+\infty[, \\ \rho y'' - \partial_\nu \Delta y - (1-\mu)\partial_\tau B_2 y = 0 & \text{on } \Gamma_1\times]0,+\infty[, \end{cases} \tag{3.6}$$

with the supplementary conditions

$$y' = \partial_\nu y' = 0 \qquad \text{on } \Gamma_1\times]0,+\infty[. \tag{3.7}$$

A straightforward computation shows that z satisfies the following conditions (see Lasiecka [7]):

$$\begin{cases} z'' + \Delta^2 z = 0 & \text{in } \Omega\times]0,+\infty[, \\ z = \partial_\nu z = \Delta z = \partial_\nu \Delta z = 0 & \text{on } \Gamma_1\times]0,+\infty[. \end{cases} \tag{3.8}$$

Applying the Holmgren theorem (see Lions [10]), it follows that $z = 0$. This, together with (3.6)–(3.7), implies that y satisfies the following conditions:

$$\begin{cases} \Delta^2 y = 0 & \text{in } \Omega\times]0,+\infty[, \\ y = \partial_\nu y = 0 & \text{on } \Gamma_0\times]0,+\infty[, \\ \Delta y + (1-\mu)B_1 y = 0 & \text{on } \Gamma_1\times]0,+\infty[, \\ \partial_\nu \Delta y + (1-\mu)\partial_\tau B_2 y = 0 & \text{on } \Gamma_1\times]0,+\infty[. \end{cases} \tag{3.9}$$

Multiplying the first equation by y and using the Green formula (2.13), we obtain

$$\int_\Omega a(y,y)dx = 0 \Rightarrow y = 0.$$

This yields that $u_0 = 0$. We complete the case $u_0 \in H$ by a standard argument of density of $D(A)$ in H and the continuity of the semigroup $S_A(t)$. The proof is thus complete. □

LEMMA 3.2. *Let $A = -A^*$ be the infinitesimal generator of a C_0 group, and let B be a compact operator in the Hilbert space H. Then the group $S_{A+B}(t)$, generated by the operator $-(A+B)$, has no uniform energy decay rate for $t > 0$.*

Proof. We first notice that $S_{A-B^*}(t)$ is also a group in H. Since $(A+B)^* = -(A-B^*)$, we have

$$\|S_{A+B}(t)\| = \|S_{-(A-B^*)}(t)\| = \|S_{A-B^*}(-t)\| \qquad \forall t \in \mathbb{R}. \tag{3.10}$$

Since B and B^* are compact, thanks to a result of compact perturbation due to Russell [20], we know that for any $t > 0$ the following two conditions can't be simultaneously held

$$\|S_{A+B}(t)\| < 1 \quad \text{and} \quad \|S_{A-B^*}(-t)\| < 1. \tag{3.11}$$

It follows from (3.10) and (3.11) that $\|S_{A+B}(t)\| \geq 1$ for any $t > 0$. The proof is complete. □

Remark. Lemma 3.2 has been formulated in another form in Gibson [2]. But the present version seems to adapt well to the first-order evolutionary equations.

THEOREM 3.3. *Assume that $s > 0$. Then the energy $E(t)$ of system (2.5) has no uniform decay rate.*

Proof. Since A is maximal monotone and skew adjoint (Lemma 2.2), and since B is compact (Lemma 2.1), then applying Lemma 3.2 we conclude that the group $S_{A+B}(t)$ has no uniform decay rate for $t > 0$. □

Remark. If $s = 0$, then L becomes the identity of $L^2(\Gamma_1)$, and the control operator B is not compact. Therefore Lemma 3.2 does not apply directly to the system (2.5). Let us consider an example where Ω is the unit disc and Γ_1 is the whole circle. Considering the radial solutions of the system (2.5), then the control space $L^2(\Gamma_1) \times L^2(\Gamma_1)$ will be reduced into $\mathbb{R}^2$ and the control operator B becomes again compact. Consequently, system (2.5) actually loses the uniform energy decay rate, even in this geometrically favorable case (see Lagnese [5]). Of course, the uniform stability of system (2.5) remains an open problem in the general case.

4. Rational energy decay rate. In this section, we will establish the rational energy decay rate for the smooth solution of system (2.5). We first recall the following classical result (Komornik [3] and Lagnese [5]).

LEMMA 4.1. *Let $E : \mathbb{R}^+ \to \mathbb{R}^+$ be a nonincreasing function. Assume that there exists a positive constant M such that*

$$\int_T^\infty E^2(t)dt \leq ME(0)E(T) \qquad \forall T > 0. \tag{4.1}$$

Then we have

$$E(t) \leq E(0)\frac{2M}{M+t} \qquad \forall t \geq 0. \tag{4.2}$$

In sections 4 and 5, we assume that there exists a point $x_0 \in \mathbb{R}^2$ such that, setting $m = x - x_0$, we have

$$\Gamma_0 = \{x \in \Gamma \ : \quad (m \cdot \nu) \leq 0\}, \qquad \Gamma_1 = \{x \in \Gamma \ : \quad (m \cdot \nu) > 0\}. \tag{4.3}$$

LEMMA 4.2. *Let y satisfy the following conditions:*

$$\begin{cases} y \in H^2_{\Gamma_0}(\Omega), \qquad \Delta^2 y \in L^2(\Omega), \\ \Delta y + (1-\mu)B_1 y = v_1 \in L^2(\Gamma_1), \\ \partial_\nu \Delta y + (1-\mu)\partial_\tau B_2 y = v_2 \in L^2(\Gamma_1). \end{cases} \tag{4.4}$$

Then we have

$$-\int_\Omega m \cdot \nabla y \Delta^2 y dx \leq -\frac{1}{2}\int_\Omega a(y,y)dx + C_0 \int_{\Gamma_1} (|v_1|^2 + |v_2|^2)d\Gamma, \tag{4.5}$$

where C_0 is a positive constant depending only on the domain Ω.

Proof. We start with $v_1 \in H^{3/2}(\Gamma_1)$ and $v_2 \in H^{1/2}(\Gamma_1)$. In that case, since $y \in H^4(\Omega)$ we have the following Green formula (see Lagnese [5]):

$$\int_\Omega \Delta^2 y\ (m\cdot\nabla y)dx \tag{4.6}$$

$$= \int_\Omega a(y,y)dx + \int_\Gamma v_2(m\cdot\nabla y)d\Gamma - \int_\Gamma v_1\partial_\nu(m\cdot\nabla y)d\Gamma$$

$$+\frac{1}{2}\int_\Gamma (m\cdot\nu)\Big\{\Big(\frac{\partial^2 y}{\partial x_1^2}\Big)^2 + \Big(\frac{\partial^2 y}{\partial x_2^2}\Big)^2 + 2\mu\frac{\partial^2 y}{\partial x_1^2}\frac{\partial^2 y}{\partial x_2^2} + 2(1-\mu)\Big(\frac{\partial^2 y}{\partial x_1\partial x_2}\Big)^2\Big\}d\Gamma.$$

Since $y = \partial_\nu y = 0$ on Γ_0, it follows that

$$\nabla y = 0, \qquad B_1 y = 0, \qquad \partial_\nu(m\cdot\nabla y) = (m\cdot\nu)\Delta y, \tag{4.7}$$

$$\Big(\frac{\partial^2 y}{\partial x_1^2}\Big)^2 + \Big(\frac{\partial^2 y}{\partial x_2^2}\Big)^2 + 2\mu\frac{\partial^2 y}{\partial x_1^2}\frac{\partial^2 y}{\partial x_2^2} + 2(1-\mu)\Big(\frac{\partial^2 y}{\partial x_1\partial x_2}\Big)^2 = (\Delta y)^2. \tag{4.8}$$

On the other hand, we have

$$\Big(\frac{\partial^2 y}{\partial x_1^2}\Big)^2 + \Big(\frac{\partial^2 y}{\partial x_2^2}\Big)^2 + 2\mu\frac{\partial^2 y}{\partial x_1^2}\frac{\partial^2 y}{\partial x_2^2} + 2(1-\mu)\Big(\frac{\partial^2 y}{\partial x_1\partial x_2}\Big)^2$$

$$\geq (1-\mu)\left\{\Big(\frac{\partial^2 y}{\partial x_1^2}\Big)^2 + \Big(\frac{\partial^2 y}{\partial x_2^2}\Big)^2 + 2\Big(\frac{\partial^2 y}{\partial x_1\partial x_2}\Big)^2\right\}. \tag{4.9}$$

Inserting (4.7)–(4.9) into (4.6) gives

$$\int_\Omega \Delta^2 y(m\cdot\nabla y)dx \tag{4.10}$$

$$\geq \int_\Omega a(y,y)dx + \int_{\Gamma_1} v_2(m\cdot\nabla y)d\Gamma - \int_{\Gamma_1} v_1\partial_\nu(m\cdot\nabla y)d\Gamma$$

$$+\frac{\delta}{2}(1-\mu)\int_{\Gamma_1}\Big\{\Big(\frac{\partial^2 y}{\partial x_1^2}\Big)^2 + \Big(\frac{\partial^2 y}{\partial x_2^2}\Big)^2 + 2\Big(\frac{\partial^2 y}{\partial x_1\partial x_2}\Big)^2\Big\}d\Gamma,$$

where δ is positive constant such that $(m\cdot\nu)\geq\delta$ for all $x\in\Gamma_1$.

Now a direct computation gives

$$2\big|\partial_\nu(m\cdot\nabla y)\big|^2 \leq |\partial_\nu y|^2 + R^2\left\{\Big(\frac{\partial^2 y}{\partial x_1^2}\Big)^2 + \Big(\frac{\partial^2 y}{\partial x_2^2}\Big)^2 + 2\Big(\frac{\partial^2 y}{\partial x_1\partial x_2}\Big)^2\right\}, \tag{4.11}$$

where $R = \|m\|_{L^\infty(\Gamma_1)}$ is the diameter of Ω. For any $\lambda>0$, it follows that

$$\int_{\Gamma_1} v_1\partial_\nu(m\cdot\nabla y)d\Gamma \geq -\lambda\int_{\Gamma_1}|v_1|^2 d\Gamma - \frac{1}{8\lambda}\int_{\Gamma_1}|\partial_\nu y|^2 d\Gamma \tag{4.12}$$

$$-\frac{R^2}{8\lambda}\int_{\Gamma_1}\left\{\Big(\frac{\partial^2 y}{\partial x_1^2}\Big)^2 + \Big(\frac{\partial^2 y}{\partial x_2^2}\Big)^2 + 2\Big(\frac{\partial^2 y}{\partial x_1\partial x_2}\Big)^2\right\}d\Gamma,$$

$$\int_{\Gamma_1} v_2(m\cdot\nabla y)d\Gamma \geq -\lambda\int_{\Gamma_1}|v_2|^2 d\Gamma - \frac{R^2}{4\lambda}\int_{\Gamma_1}|\nabla y|^2 d\Gamma. \tag{4.13}$$

Inserting (4.12)–(4.13) into (4.10), we obtain

$$\int_\Omega \Delta^2 y(m\cdot\nabla y)dx \geq \int_\Omega a(y,y)dx \tag{4.14}$$
$$-\lambda\int_{\Gamma_1}(|v_1|^2+|v_2|^2)d\Gamma - \frac{1}{4\lambda}\int_{\Gamma_1}\big(|\partial_\nu y|^2 + R^2|\nabla y|^2\big)d\Gamma$$

provided

$$\lambda \geq \frac{R^2}{4\delta(1-\mu)}. \tag{4.15}$$

We obtain (4.3) by taking $\lambda > 0$ sufficiently large in (4.14):

$$\int_{\Gamma_1}\big(|\partial_\nu y|^2 + R^2|\nabla y|^2\big)d\Gamma \leq \lambda\int_\Omega a(y,y)dx \qquad \forall y\in H^2_{\Gamma_0}(\Omega). \tag{4.16}$$

The case $v_1, v_2 \in L^2(\Gamma_1)$ can be easily completed by a standard argument of density (see Lemma 3.1 in Rao [16]). The proof is complete. □

THEOREM 4.3. *Assume that $s = 0$. Then given any $u_0 \in D(A)$, there exists a constant $M > 0$ depending only on u_0 such that the following rational energy decay rate holds:*

$$E(t) \leq E(0)\frac{2M}{M+t} \qquad \forall t > 0 \tag{4.17}$$

for all smooth solution u of system (2.5).

Proof. Let $0 \leq T < S < +\infty$, we multiply the plate equation by $E(t)(m\cdot\nabla y)$ and integrate over $\Omega\times[T,S]$:

$$\int_T^S\int_\Omega E(t)(m\cdot\nabla y)y''dxdt = -\int_T^S\int_\Omega E(t)(m\cdot\nabla y)\Delta^2 y dxdt. \tag{4.18}$$

In the left-hand side of (4.18), one integration by parts gives

$$\int_T^S\int_\Omega E(t)(m\cdot\nabla y)y''dxdt \tag{4.19}$$
$$= \int_\Omega\Big[E(t)(m\cdot\nabla y)y'dx\Big]_T^S - \int_T^S\int_\Omega E'(t)(m\cdot\nabla y)y'dxdt$$
$$+\int_T^S\int_\Omega E(t)|y'|^2dxdt - \frac{1}{2}\int_T^S\int_{\Gamma_1}E(t)(m\cdot\nu)|y'|^2d\Gamma dt.$$

Next using the Cauchy–Schwarz inequality we have

$$\Big|\int_\Omega (m\cdot\nabla y)y'dx\Big| \leq C_1E(t). \tag{4.20}$$

Then it follows that

$$\int_\Omega\Big[E(t)(m\cdot\nabla y)y'dx\Big]_T^S - \int_T^S\int_\Omega E'(t)(m\cdot\nabla y)y'dxdt \tag{4.21}$$
$$\geq -C_1(E^2(T)+E^2(S)) + C_1\int_T^S E'(t)E(t)dt \geq -2C_1E^2(T).$$

Inserting (4.21) into (4.19) gives

$$\int_T^S \int_\Omega E(t)(m\cdot\nabla y)y''dxdt \tag{4.22}$$

$$\geq \int_T^S \int_\Omega E(t)|y'|^2dxdt - \frac{R}{2}\int_T^S \int_{\Gamma_1} E(t)|y'|^2d\Gamma dt - 2C_1E^2(T).$$

Because of (2.21)–(2.22), y satisfies conditions (4.4). Then applying Lemma 4.2 in the right-hand side of (4.18) gives

$$-\int_T^S \int_\Omega E(t)(m\cdot\nabla y)dxdt \leq -\frac{1}{2}\int_T^S \int_\Omega E(t)a(y,y)\Delta^2 ydxdt \tag{4.23}$$

$$+C_0\int_T^S \int_{\Gamma_1} E(t)\big(|\rho y''+y'|^2+|J\partial_\nu y''+\partial_\nu y'|^2\big)d\Gamma dt.$$

Inserting (4.22)–(4.23) into (4.18) gives that

$$\int_T^S E^2(t)dt \leq 2C_1E^2(T) \tag{4.24}$$

$$+C_2\int_T^S \int_{\Gamma_1} E(t)\big(|y'|^2+|\partial_\nu y'|^2+|y''|^2+|\partial_\nu y''|^2\big)d\Gamma dt,$$

where we have put

$$C_2 = 2C_0(1+\rho^2+J^2)+\frac{R}{2}. \tag{4.25}$$

Since the energy $E(t)$ is nonincreasing, it follows that

$$\int_T^S E^2(t)dt \leq 2C_1E^2(T) \tag{4.26}$$

$$+C_2E(T)\int_T^S \int_{\Gamma_1} \big(|y'|^2+|\partial_\nu y'|^2+|y''|^2+|\partial_\nu y''|^2\big)d\Gamma dt.$$

On the other hand, from (3.2) we deduce that

$$\int_T^S \int_{\Gamma_1} \big(|y'|^2+|\partial_\nu y'|^2\big)d\Gamma dt \leq E(T). \tag{4.27}$$

Differentiating the system (2.5) with respect to the variable t gives

$$\int_T^S \int_{\Gamma_1} \big(|y''|^2+|\partial_\nu y''|^2\big)d\Gamma dt \leq E_1(T), \tag{4.28}$$

where the energy of high-order $E_1(t)$ is defined by

$$E_1(t) = \frac{1}{2}\|u'(t)\|^2 = \frac{1}{2}\|(A+B)u(t)\|^2. \tag{4.29}$$

Use of (4.27)–(4.28) in (4.26) gives

$$\int_T^S E^2(t)dt \le ME(T)E(0), \tag{4.30}$$

where we have put

$$M = 2C_1 + C_2 + C_2\frac{\|(A+B)u_0\|^2}{\|u_0\|^2}. \tag{4.31}$$

Finally, thanks to Lemma 4.1 we deduce the rational energy decay rate (4.17) from (4.30). The proof is thus complete. □

Remark. For most linear problems we use the classical multiplier $m \cdot \nabla y$. The idea of the proof of Theorem 4.3 consists in taking the multiplier $E(t)m \cdot \nabla y$ which is usually used in nonlinear problems. This method is very simple and can be easily applied to other problems.

There is another natural approach based on the spectral theory (Littman and Markus [11]). Roughly speaking, let $\lambda_n = \alpha_n + i\beta_n$ be the eigenvalues of the operator $A + B$. Assume that (i) $\alpha_n \ge 1/n^p$ $(p > 0)$ and (ii) the associated eigenvectors ϕ_n form a Riesz basis; then the trajectory $S_{A+B}(t)u_0$ has a rational decay rate for smooth initial data u_0. This method was applied to one-dimensional problems such as the Euler–Bernoulli beam model (Littman and Markus [11]) and string/mass model (Lee and You [8]) with very careful calculation of the eigenvalues. It seems to be impossible to justify conditions (i) and (ii) for the plate model of general shape.

5. Uniform stability of a simplified model. In this section we consider a simplified model in which the bending moment of inertia of the boundary J is neglected. Therefore, we obtain the following hybrid system:

$$\begin{cases} y'' + \Delta^2 y = 0 & \text{in } \Omega\times]0,+\infty[, \\ y = \partial_\nu y = 0 & \text{on } \Gamma_0\times]0,+\infty[, \\ \Delta y + (1-\mu)B_1 y = -\partial_\nu y' & \text{on } \Gamma_1\times]0,+\infty[, \\ \rho y'' - \partial_\nu \Delta y - (1-\mu)\partial_\tau B_2 y = -y' & \text{on } \Gamma_1\times]0,+\infty[. \end{cases} \tag{5.1}$$

Let us first introduce the energy space H,

$$H = H^2_{\Gamma_0}(\Omega) \times L^2(\Omega) \times L^2(\Gamma_1), \tag{5.2}$$

and the linear unbounded operator A,

$$Au = \begin{pmatrix} -z \\ \Delta^2 y \\ -\frac{1}{\rho}\big(\partial_\nu \Delta y + (1-\mu)\partial_\tau B_2 y - z\big) \end{pmatrix}. \tag{5.3}$$

The domain of the operator A is defined as

$$D(A) = \begin{pmatrix} u = (y,z,\eta) \in W \times H^2_{\Gamma_0}(\Omega) \times L^2(\Gamma_1) \\ \eta = z|_{\Gamma_1}, \quad \Delta y + (1-\mu)B_1 y + \partial_\nu z = 0 \quad \text{on } \Gamma_1 \end{pmatrix}, \tag{5.4}$$

where we have put

$$W = \begin{pmatrix} y \in H^2_{\Gamma_0}(\Omega), \quad \Delta^2 y \in L^2(\Omega) \\ \Delta y + (1-\mu)B_1 y \in H^{1/2}(\Gamma_1) \\ \partial_\nu \Delta y + (1-\mu)\partial_\tau B_2 y \in L^2(\Gamma_1) \end{pmatrix} . \tag{5.5}$$

Now let y be a smooth solution of the hybrid system (5.1). Setting

$$z = y', \qquad \eta = y'|_{\Gamma_1},$$

we transform the hybrid system (5.1) into an abstract evolutionary equation:

$$u' + Au = 0, \qquad u(0) = u_0 \in H. \tag{5.6}$$

Given $u_0 \in D(A)$, using the definition (5.3)–(5.5) and the Green formula (2.13), a straightforward computation gives

$$(Au, u)_H = \int_{\Gamma_1} \big(|z|^2 + |\partial_\nu z|^2\big) d\Gamma \geq 0 \ . \tag{5.7}$$

Moreover by an argument analogous to that of Lemma 2.2, we can easily verify the rank condition $R(I+A) = H$. This implies that the operator A is maximal monotone. Therefore, equation (5.6) is well posed in the sense of a semigroup of contractions. Similarly we have the following result.

THEOREM 5.1. (i) *For any $u_0 \in D(A)$, equation* (5.6) *admits a unique strong solution $u(t) = \big(y(t), z(t), \eta(t)\big) \in D(A)$ such that*

$$y(t) \in C^0\big(\mathbb{R}^+; H^{7/2}(\Omega)\big) \cap C^1\big(\mathbb{R}^+; H^2_{\Gamma_0}(\Omega)\big) \cap C^2\big(\mathbb{R}^+; L^2(\Omega)\big), \tag{5.8}$$

$$y|_{\Gamma_1} \in C^2\big(\mathbb{R}^+; L^2(\Gamma_1)\big). \tag{5.9}$$

(ii) *For any $u_0 \in H$, equation* (5.6) *admits a unique weak solution $u(t) = S_A(t)u_0 = \big(y(t), z(t), \eta(t)\big) \in H$ such that*

$$y(t) \in C^0\big(\mathbb{R}^+; H^2_{\Gamma_0}(\Omega)\big) \cap C^1\big(\mathbb{R}^+; L^2(\Omega)\big), \tag{5.10}$$

$$y|_{\Gamma_1} \in C^1\big(\mathbb{R}^+; L^2(\Gamma_1)\big), \tag{5.11}$$

where $S_A(t)$ is the semigroup of contractions generated by $-A$. □

Now let $u(t) = (y(t), z(t), \eta(t))$ be a solution of equation (5.6); we define the associated energy by setting

$$E(t) = \frac{1}{2}\Big\{ \int_\Omega \big(|y'|^2 + a(y,y)\big) dx + \rho \int_{\Gamma_1} |y'|^2 d\Gamma \Big\}. \tag{5.12}$$

Then using (5.7) we have

$$\frac{d}{dt} E(t) = -\int_{\Gamma_1} \big(|y'|^2 + |\partial_\nu y'|^2\big) d\Gamma \leq 0.$$

Hence for any $T > 0$ we have

$$\int_0^T \int_{\Gamma_1} \big(|y'|^2 + |\partial_\nu y'|^2\big) d\Gamma dt \leq E(0). \tag{5.13}$$

By a slight modification, we can easily establish the following analogue of Lemma 4.2.

LEMMA 5.2. *Let y satisfy the following conditions:*

$$\begin{cases} y \in H^2_{\Gamma_0}(\Omega), \qquad \Delta^2 y \in L^2(\Omega), \\ \Delta y + (1-\mu)B_1 y = v_1, \\ \partial_\nu \Delta y + (1-\mu)\partial_\tau B_2 y = v_2 + \widehat{v}_2 \end{cases} \tag{5.14}$$

with $v_1, v_2, \widehat{v}_2 \in L^2(\Gamma_1)$. Then the following estimate holds:

$$-\int_\Omega m \cdot \nabla y \Delta^2 y dx \tag{5.15}$$

$$\leq -\frac{1}{2}\int_\Omega a(y,y)dx - \int_{\Gamma_1} \widehat{v}_2 (m \cdot \nabla y) d\Gamma + C_0 \int_{\Gamma_1} (|v_1|^2 + |v_2|^2) d\Gamma,$$

where C_0 is a positive constant depending only on Ω. □

THEOREM 5.3. *There exist two positive constants M and ω such that*

$$E(t) \leq M E(0) e^{-\omega t} \qquad \forall t > 0 \tag{5.16}$$

for any solution u of equation (5.6).

Proof. Because of the density of $D(A)$ into H and the continuity of the energy $E(t)$ with respect to the initial data u_0, it is sufficient to consider the smooth initial data $u_0 \in D(A)$. Then indeed, multiplying the plate equation (5.1) by $(m \cdot \nabla y)$ and integrating over $\Omega \times [0,T]$, we obtain

$$\int_0^T \int_\Omega (m \cdot \nabla y) y'' dx dt = -\int_0^T \int_\Omega (m \cdot \nabla y) \Delta^2 y dx dt. \tag{5.17}$$

In the left-hand side, one integration by parts gives

$$\int_0^T \int_\Omega (m \cdot \nabla y) y'' dx dt = \int_\Omega \left[(m \cdot \nabla y) y' dx \right]_0^T \tag{5.18}$$

$$+ \int_0^T \int_\Omega |y'|^2 dx dt - \frac{1}{2}\int_0^T \int_{\Gamma_1} (m \cdot \nu) |y'|^2 d\Gamma dt.$$

Next using the Cauchy–Schwarz inequality we have

$$\int_\Omega \left[(m \cdot \nabla y) y' dx \right]_0^T \geq -C_1 E(0), \tag{5.19}$$

where $C_1 > 0$ is a constant depending only on Ω. Inserting (5.19) into (5.18) and taking (5.13) into account gives

$$\int_0^T \int_\Omega (m \cdot \nabla y) y'' dx dt \geq \int_0^T \int_\Omega |y'|^2 dx dt - C_1 E(0). \tag{5.20}$$

Using (5.8)–(5.9), we verify easily that $y, v_1 = -\partial_\nu y', v_2 = y'$, and $\widehat{v}_2 = \rho y''$ satisfy conditions (5.14). Applying Lemma 5.2 in the right-hand side of (5.17), we

obtain that

$$-\int_0^T\int_\Omega (m\cdot\nabla y)\Delta^2 y dxdt \le -\frac{1}{2}\int_0^T\int_\Omega a(y,y)dxdt \tag{5.21}$$
$$-\rho\int_0^T\int_{\Gamma_1}(m\cdot\nabla y)y''d\Gamma dt + C_0\int_0^T\int_{\Gamma_1}\big(|y'|^2+|\partial_\nu y'|^2\big)d\Gamma dt.$$

One integration by parts gives

$$-\int_0^T\int_{\Gamma_1}(m\cdot\nabla y)y''d\Gamma dt \tag{5.22}$$
$$=-\int_{\Gamma_1}\big[(m\cdot\nabla y)y'd\Gamma\big]_0^T+\int_0^T\int_{\Gamma_1}(m\cdot\nabla y')y'd\Gamma dt.$$

Using the Cauchy–Schwarz inequality we have

$$-\int_{\Gamma_1}\big[(m\cdot\nabla y)y'\big]_0^T d\Gamma \le CE(0). \tag{5.23}$$

On the other hand, a straightforward computation gives

$$\int_{\Gamma_1}(m\cdot\nabla y')y'd\Gamma \tag{5.24}$$
$$=\int_{\Gamma_1}\Big((m\cdot\nu)y'\partial_\nu y'-\frac{1}{2}\partial_\tau(m\cdot\tau)|y'|^2\Big)d\Gamma$$
$$\le (R+K)\int_{\Gamma_1}(|y'|^2+|\partial_\nu y'|^2)d\Gamma,$$

where we have put

$$R=\|m\|_{L^\infty(\Gamma)},\qquad K=\|\partial_\tau(m\cdot\tau)\|_{L^\infty(\Gamma)}. \tag{5.25}$$

Inserting (5.22)–(5.24) into (5.21) and taking (5.13) into account, we get

$$\int_0^T\int_\Omega(m\cdot\nabla y)\Delta^2 y dxdt \le -\frac{1}{2}\int_0^T\int_\Omega a(y,y)dxdt+CE(0), \tag{5.26}$$

where $C>0$ is a constant depending only on Ω.

Finally inserting (5.20) and (5.26) into (5.17), we obtain the integral inequality

$$\int_0^T E(t)dt \le CE(0)\qquad \forall T>0. \tag{5.27}$$

Thanks to the well-known result of Datko (Pazy [14]), this implies the uniform energy decay rate (5.16). The proof is thus complete. □

Remark. We have shown that the hybrid system (5.1) can be uniformly stabilized by means of the usual boundary feedback controls. But if we take $\rho=0$ and $J>0$, then contrary to the previous case the uniform stability of the corresponding hybrid system remains an *open* problem. See also [15] for other uniformly stable hybrid systems.

REFERENCES

[1] H. Brezis, *Analyse Fonctionelle, Théorie et Applications*, Masson, Paris, 1983.

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

[3] V. Komornik, *Exact Controllability and Stabilization, The Multiplier Method*, Masson, Paris, 1994.

[4] V. Komornik and B. Rao, *Boundary stabilization of compactly coupled wave equations*, C. R. Acad. Sci. Paris Sér. I Math., 320 (1995), pp. 833–838

[5] J. E. Lagnese, *Boundary Stabilization of Thin Plates*, SIAM, Philadelphia, PA, 1989.

[6] J. E. Lagnese and J. L. Lions, *Modelling, Analysis and Control of Thin Plates*, Masson, Paris, 1988.

[7] I. Lasiecka, *Asymptotic behavior of solutions to plate equations with nonlinear dissipation occurring through shear forces and bending moments*, Appl. Math. Optim., 21 (1990), pp. 167–189.

[8] E. B. Lee and Y. C. You, *Stabilization of a hybrid (string/point mass) system*, in Proc. Fifth Internat. Conf. on System Engineering, Dayton, OH, 1987.

[9] J. L. Lions and E. Magenes, *Problèmes aux limites non homogènes*, Vol. I, Dunod, Paris, 1968.

[10] J. L. Lions, *Contrôlabilité exacte, perturbations et stabilisation de systèmes distribués*, Vol. I, Masson, Paris, 1988.

[11] W. Littman and L. Markus, *Some recent results on control and stabilization of flexible structures*, in Proc. COMCON Workshop, Montpellier, 1987.

[12] W. Littman and L. Markus, *Stabilization of a hybrid system of elasticity by feedback boundary damping*, Ann. Mat. Pura Appl., 152 (1988), pp. 281–330.

[13] L. Markus and Y. C. You, *Dynamical boundary control for elastical plates of general shape*, SIAM J. Control Optim., 31 (1993), pp. 983–992.

[14] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.

[15] B. Rao, *Decay estimates of solutions for a hybrid system of flexible structures*, European J. Appl. Math., 4 (1993), pp. 303–319.

[16] B. Rao, *Stabilization of Kirchhoff plate equation in star-shaped domain by nonlinear boundary feedback*, Nonlinear Anal., Theory, Meth. Appl., 20 (1993), pp. 605–626.

[17] B. Rao, *Uniform stabilization of a hybrid system of elasticity*, SIAM J. Control Optim., 33 (1995), pp. 440–454.

[18] B. Rao, *Stabilisation d'une équation de plaque par contrôle frontière dynamique*, C. R. Acad. Sci. Paris Sér. I Math., 321 (1995), pp. 1449–1454.

[19] B. Rao, *A compact perturbation method for the boundary stabilization of the Rayleigh beam equation*, Appl. Math. Optim., 33 (1966), pp. 253–263.

[20] D. L. Russell, *Decay rates for weakly damped systems in Hilbert space obtained with control-theoretic methods*, J. Differential Equations, 19 (1975), pp. 344–370.

[21] M. Slemrod, *Feedback stabilization of a linear system in Hilbert space with an a priori bounded control*, Math. Control Signals Systems, 2 (1989), pp. 265–285.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 1, pp. 164–179, January 1998

009

CONFIGURATION FLATNESS OF LAGRANGIAN SYSTEMS UNDERACTUATED BY ONE CONTROL*

MURUHAN RATHINAM† AND RICHARD M. MURRAY‡

Abstract. Lagrangian control systems that are differentially flat with flat outputs that depend only on configuration variables are said to be configuration flat. We provide a complete characterization of configuration flatness for systems with n degrees of freedom and $n-1$ controls whose range of control forces only depends on configuration and whose Lagrangian has the form of kinetic energy minus potential. The method presented allows us to determine if such a system is configuration flat and, if so, provides a constructive method for finding all possible configuration flat outputs. Our characterization relates configuration flatness to Riemannian geometry. We illustrate the method with two examples.

Key words. differential flatness, nonlinear control, Lagrangian mechanics

AMS subject classifications. 93C10, 93B29, 58F05, 58B21

PII. S0363012996300987

1. Introduction.

Roughly speaking, an underdetermined system of ODEs

$$F^k(t, x^1, \ldots, x^N, \dot{x}^1, \ldots, \dot{x}^N) = 0, \qquad k = 1, \ldots, n < N,$$

is differentially flat if there is a smooth locally one-to-one correspondence between solutions $x(t)$ of the system and arbitrary functions $y(t)$ of the form

$$x(t) = g(t, y(t), \ldots, y^{(l)}(t)),$$
$$y(t) = h(t, x(t), \ldots, x^{(q)}(t)),$$

where $(y^1, \ldots, y^p) \in \mathbb{R}^p$ and $p = N - n$. Here g, h are smooth maps, $y^{(k)}$ is the kth derivative of y, and l, q are integers. The variables y^j are referred to as flat outputs. The special class of systems given by

$$\dot{x}^i = f^i(t, x^1, \ldots, x^n, u^1, \ldots, u^p), \quad i = 1, \ldots, n,$$

is more familiar to control theorists and the flat outputs depend on states, inputs, and derivatives of inputs

$$y^j = h^j(t, x, u, u^{(1)}, \ldots, u^{(q)}), \quad j = 1, \ldots, p.$$

For a detailed discussion of differential flatness see Fliess et al. [3, 4], Martin [9], Pomet [12], van Nieuwstadt, Rathinam, and Murray [21], and Rathinam and Sluis [13].

*Received by the editors March 25, 1996; accepted for publication (in revised form) October 30, 1996. A preliminary version of this article has appeared in the *Proceedings of the* 35*th IEEE Control and Decision Conference*, Kobe, Japan, 1996.

http://www.siam.org/journals/sicon/36-1/30098.html

†Applied Mathematics, California Institute of Technology, Mail Code 217-50, Pasadena, CA 91125 (muruhan@ama.caltech.edu). The research of this author was supported in part by NSF grant CMS-9502224.

‡Mechanical Engineering, California Institute of Technology, Mail Code 104-44, Pasadena, CA 91125 (murray@indra.caltech.edu). The research of this author was supported in part by NSF grant CMS-9502224 and AFOSR grant F49620-95-1-0419.

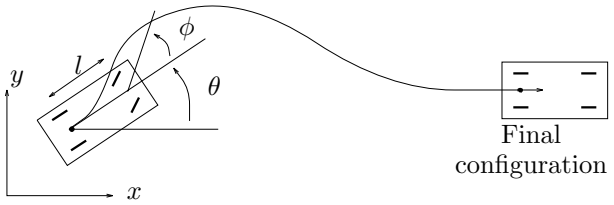


FIG. 1.1. *Path planning for kinematic car.*

The importance of flatness to control applications lies in the fact that it provides a systematic and relatively simple way to generate solution trajectories between two given states. One uses the maps g and h to transform between original system space (states as well as inputs) and the smaller-dimensional flat output space. See van Nieuwstadt and Murray [20] and Murray, Rathinam, and Sluis [11] for more details.

For example consider the "kinematic car" shown in Figure 1.1. Ignoring dynamics we assume the velocity of the midpoint between the rear wheels and the steering velocity are directly controlled. Then the system is differentially flat with the coordinates of the midpoint between rear wheels providing the two flat outputs (see Rouchon et al. [14] and Tilbury, Murray, and Sastry [19]). Given any trajectory for this point one can determine the entire motion of the car: the tangent to the trajectory determines the orientation of the car and the curvature (second derivative) determines the orientation of the front wheels. Hence all feasible paths of the vehicle can be parametrized in terms of the trajectories of the flat output point. A given set of initial and final configurations of the car then determine two end points and first- and second-order derivatives at these end points for feasible trajectories of the flat output point. One could choose any trajectory for the flat output point that satisfies these end conditions and obtain a feasible trajectory for the car that passes through the given initial and final conditions. In this example, flat outputs are rather obvious. This is not the case with many other examples, and one needs a theoretical tool to provide a systematic way of finding them if they exist.

In the case of single input systems a complete characterization of differential flatness is available; see, e.g., Shadwick [16]. In that case, flatness is the same as static feedback linearizability. See also [2]. In the framework of exterior differential systems, checking for flatness of a single input system reduces to calculating "derived systems" and checking certain rank and integrability conditions. See van Nieuwstadt, Rathinam, and Murray [21], Sluis [17], and Sluis and Tilbury [18]. For multi-input systems no complete theory exists.

Many interesting examples of mechanical systems are differentially flat, and in most known examples flat outputs have been found that depend only on the configuration variables but not on their derivatives. We refer to such flat outputs as "configuration flat outputs" and systems possessing such outputs as "configuration flat." For instance, the above example of the kinematic car is configuration flat. All Lagrangian systems that are fully actuated (number of controls equals number of degrees of freedom) are configuration flat with all the configuration variables as flat outputs. See [11] for a catalogue of other examples. The reasons for studying configuration flatness are as follows. First, it is a simpler case than the general case of differential flatness and is possibly the first thing to study if one were to be able to relate the mechanical structure with differential flatness. For instance, configuration controllability of mechanical systems has already been studied and related to the

mechanical structure (see Lewis and Murray [8]). Second, the smaller the number of derivatives of configuration variables the flat outputs depend upon the simpler the numerical implementation of the transformations involved in trajectory generation. In this paper we completely characterize configuration flatness for a special class of mechanical systems. The class under consideration involves systems whose dynamics are described by Lagrangian mechanics with a Lagrangian function of the form "kinetic energy minus potential." Also, the number of independent controls is assumed to be one less than the number of degrees of freedom (the simplest case next to fully actuated systems) and the possible range of control forces only depends on the configuration and not on the velocity. We describe an algorithm for deciding if such a system is configuration flat and if it is so, we describe a procedure for finding all possible configuration flat outputs. We do not consider systems with nonholonomic constraints. The kinematic car example hence does not fall into the class of systems under our consideration.

The paper is organized as follows. Section 2 introduces some concepts from Lagrangian control systems theory and also provides a definition of configuration flatness. Section 3 introduces some concepts from Riemannian geometry that are necessary for our theory and also states and proves the main proposition and outlines an algorithm for coordinate calculations to check configuration flatness. Section 4 explores how system symmetries relate to symmetries of the flat outputs. Finally, section 5 gives two examples to illustrate the theory.

2. Lagrangian control systems and configuration flatness. Consider a Lagrangian system with configuration manifold Q of dimension n and a Lagrangian $L : TQ \to \mathbb{R}$. When no external (generalized) forces are applied, the motion of this system satisfies the Euler–Lagrange equations, written in coordinates $(q^1, \dots, q^n)$ as

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}^i}\right) - \frac{\partial L}{\partial q^i} = 0, \quad i = 1, \dots, n. \tag{2.1}$$

In a control situation external control forces are applied and it is natural to think of forces as covectors on the manifold Q. In other words, for a configuration $q \in Q$ the total external force acting on the system can be represented by an element of T_q^*Q. This is because forces naturally pair with velocities, which can be thought of as elements of T_qQ, to give instantaneous power. The possible range of control forces lies in a subspace of T_q^*Q which may depend on position q as well as velocity v_q. In other words the control forces can be described by a horizontal-valued codistribution $\bar{P} \subset T^*(TQ)$, and $p = \dim \bar{P}$ is the number of independent controls. For an interesting and wide class of systems this subspace depends only on configuration q and hence can be described by a codistribution $P \subset T^*Q$ of dimension p. For the rest of the discussion we shall consider only this case. All feasible paths (solutions) of such a system are characterized by the following underdetermined system of ODEs in coordinates $(q^1, \dots, q^n)$:

$$a_k^i \left(\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}^i}\right) - \frac{\partial L}{\partial q^i}\right) = 0, \quad k = 1, \dots, n-p, \tag{2.2}$$

where $a_k^i \frac{\partial}{\partial q^i}$ for $k = 1, \dots, n-p$ span the annihilator of P, denoted $\operatorname{ann} P$.

It is useful to think in terms of the associated submanifold $\mathrm{E} \subset J^2(\mathbb{R}, Q)$ of the second-order jet space (see [15]), which geometrically describes such a second-order system of equations. E has codimension $n-p$ and in local coordinates $(t, q, \dot{q}, \ddot{q})$ is

cut out by the common zeros of the functions

$$a_k^i \left(\frac{\partial^2 L}{\partial \dot{q}^i \partial \dot{q}^j} \ddot{q}^j + \frac{\partial^2 L}{\partial \dot{q}^i \partial q^j} \dot{q}^j - \frac{\partial L}{\partial q^i} \right), \quad k = 1, \dots, n-p.$$

Let $q \in Q$ be a point and let $y : U \subset Q \to \mathbb{R}^p$ be a submersion locally defined around q. Let $y = (y^1, \dots, y^p)$. We say $y^1, \dots, y^p$ are *differentially independent* around q if $y^1, \dots, y^p$ do not have to satisfy an ODE along solutions local to q. More precisely, when restricted to E , $dy^1, \dots, dy^p, d\dot{y}^1, \dots, d\dot{y}^p, d\ddot{y}^1, \dots, d\ddot{y}^p$ are linearly independent for generic points on $\pi_2^{-1}(V) \cap \mathrm{E}$ where $V \subset U$ is an open neighborhood of q and $\pi_2 : J^2(\mathbb{R}, Q) \to Q$ is the standard projection. If $dy^1, \dots, dy^p, d\dot{y}^1, \dots, d\dot{y}^p, d\ddot{y}^1, \dots, d\ddot{y}^p$ are linearly dependent when restricted to E , for points on $\pi_2^{-1}(V) \cap \mathrm{E}$ where $V \subset U$ is an open neighborhood of q, then $y^1, \dots, y^p$ are *differentially dependent* around q.

Suppose $y^1, \dots, y^p$ are differentially independent around q. If there are functions f^i and a neighborhood W of q such that along a generic solution $c : \mathbb{R} \to W \subset Q$,

$$(z^i \circ c)(t) = f^i \left((y \circ c)(t), \dots, \frac{d^r}{dt^r}(y \circ c)(t) \right), \quad i = 1, \dots, n-p, \tag{2.3}$$

where $z^1, \dots, z^{n-p}$ are any complementary coordinates to $y^1, \dots, y^p$, then $y^1, \dots, y^p$ are said to be *configuration flat outputs* around q and the system is *configuration flat* around q. In other words, given $y^1(t), \dots, y^p(t)$ we can determine a (locally) unique trajectory for the Lagrangian system (2.2).

We present the following lemma which will be of use later.

LEMMA 2.1. *Let $q \in Q$, U be an open neighborhood of q, and $y : U \to \mathbb{R}^p$ be a configuration flat output. Then generically the set of solutions $c : \mathbb{R} \to U$ that project down to the same curve $y \circ c$ are all isolated.*

Proof. By definition of flatness, along generic solutions, given $y(t)$ the complementary coordinates $z(t)$ are locally uniquely determined by equations (2.3). ∎

3. Mechanical systems with n degrees of freedom and $n-1$ controls. Consider the mechanical system whose Lagrangian is given by

$$L(v) = \frac{1}{2} g(v, v) - V \circ \tau_Q(v), \tag{3.1}$$

where g is the Riemannian metric (assumed to be nondegenerate) corresponding to kinetic energy and V is the potential energy function on Q and $\tau_Q : TQ \to Q$ is the tangent bundle projection. Suppose the number of controls $p = n-1$, in other words $\dim P = n-1$. In this section we shall present a method for determining if this system is configuration flat. If the system is configuration flat our approach provides us with a constructive method for finding all possible (configuration) flat outputs.

Before proceeding further we present some concepts from Riemannian geometry. Given a metric g we have a notion of differentiation of objects on the manifold such as functions, vector fields, differential forms, and tensors along a given vector field Z. This is the covariant derivative ∇ given by the Levi–Cività connection (see [1]). ∇_Z denotes covariant derivative along a vector field Z and is related to parallel (with respect to metric) transport of objects along the integral curves of Z. The covariant derivative of a function f along Z denoted $\nabla_Z f$ is just the familiar directional derivative $Z(f)$ or the Lie derivative. But the covariant derivative of a vector field X along Z denoted $\nabla_Z X$ is not the same as the Lie derivative $[Z, X]$. Some properties of ∇

are

$$\nabla_Z(X_1 + X_2) = \nabla_Z X_1 + \nabla_Z X_2, \tag{3.2}$$

$$\nabla_Z(fX) = \nabla_Z X + Z(f)X, \tag{3.3}$$

$$\nabla_{fZ} X = f\nabla_Z X, \tag{3.4}$$

$$\nabla_Z X - \nabla_X Z = [Z, X], \tag{3.5}$$

where X, X_1, X_2, Z are arbitrary vector fields and f is an arbitrary function on the manifold. In a coordinate system $(q^1, \ldots, q^n)$ on manifold Q the covariant derivatives are calculated with the aid of Christoffel symbols Γ^i_{jk}, where $i, j, k = 1, \ldots, n$ and Christoffel symbols are defined by

$$\nabla_{\frac{\partial}{\partial q^j}} \frac{\partial}{\partial q^k} = \Gamma^i_{jk} \frac{\partial}{\partial q^i}. \tag{3.6}$$

From the properties (3.5) of ∇ it follows that $\Gamma^i_{jk} = \Gamma^i_{kj}$. The symbols Γ^i_{jk} can be computed from metric g by the formula

$$\Gamma^m_{jk} = \frac{1}{2}\left(\frac{\partial g_{ik}}{\partial q^j} + \frac{\partial g_{ij}}{\partial q^k} - \frac{\partial g_{jk}}{\partial q^i}\right) g^{im}, \quad j, m = 1, \ldots, n, \tag{3.7}$$

where $g^{ik} g_{kj} = \delta^i_j$ (g^{ik} are components of the inverse of matrix g_{ik}). Then the covariant derivative of vector field $X = X^k \frac{\partial}{\partial q^k}$ along $Z = Z^j \frac{\partial}{\partial q^j}$ is given by

$$\nabla_Z X = Z^j X^k \Gamma^i_{jk} \frac{\partial}{\partial q^i} + Z^j \frac{\partial X^k}{\partial q^j} \frac{\partial}{\partial q^k}. \tag{3.8}$$

For the mechanical system under consideration let us define an associated distribution D by

$$D = \operatorname{span}\{\xi, \nabla_Z \xi \ : \ Z \in \mathfrak{X}(Q)\}, \tag{3.9}$$

where ξ is any vector field such that $\operatorname{ann} P = \operatorname{span}\{\xi\}$ and $\mathfrak{X}(Q)$ is the set of all smooth vector fields on Q.

It is easy to check that D doesn't depend on the choice of $\xi \in \operatorname{ann} P$. By the linearity of covariant derivative it follows that

$$D = \operatorname{span}\{\xi, \nabla_{\frac{\partial}{\partial q^i}} \xi \ : \ i = 1, \ldots, n\}, \tag{3.10}$$

where $(q^1, \ldots, q^n)$ are any set of coordinates. Hence D is easily calculated using equations (3.7), (3.8), and (3.10). The following proposition characterizes configuration flat outputs $y^1, \ldots, y^p$ by conditions on $\ker Ty$, which in coordinates is the null space of the Jacobian of the map y.

PROPOSITION 3.1. *Let q be a point on Q and U be an open neighborhood of q, and suppose $y : U \subset Q \to \mathbb{R}^p$ is a submersion. If $y^1, \ldots, y^p$ are configuration flat outputs, then*

$$g(\ker Ty, D) = 0. \tag{3.11}$$

Conversely if $g(\ker Ty, D) = 0$ and if a certain regularity condition holds at q, then $y^1, \ldots, y^p$ are configuration flat outputs around q.

The regularity condition is that the ratios of functions in the following set should not all be the same at q:

$$\{\nabla_\eta(g(\xi,Z)) : g(\xi,Z), \nabla_\eta(g(\nabla_{Z_1}Z_2,\xi)) : g(\nabla_{Z_1}Z_2,\xi), \nabla_\eta(\xi(V)) : \xi(V)\}, \tag{3.12}$$

where Z, Z_1, Z_2 are arbitrary vector fields around q that are y-related to some vector field on $\mathbb{R}^p$ and ξ, η are fixed nonvanishing vector fields such that $\operatorname{ann} P = \mathit{span}\{\xi\}$ *and* $\ker Ty = \mathit{span}\{\eta\}$.

REMARK 3.2. *Proposition* 3.1 *states the conditions for configuration flatness in intrinsic geometric terms. In coordinates the algorithm for deciding if the system is configuration flat is as follows. Calculate D using equation* (3.10). *If $D = TQ$, then system is not configuration flat, since for any y, one can find some vector field $Z \in D = TQ$, such that $g(\ker Ty, Z) \neq 0$. Suppose* $\dim D \leq n-1$. *Then choose a one-dimensional distribution, say spanned by a vector field η, that is orthogonal to D. Since a one-dimensional distribution is integrable locally, one can find independent functions $y^1, \dots, y^p$ ($p = n-1$) around q that "cut out" the leaves of the corresponding foliation. These will be flat outputs provided the regularity conditions are met.*

The regularity conditions can be checked in coordinates as follows. Choose a function z that completes $y^1, \dots, y^p$ to a coordinate system. Then $y^1, \dots, y^p$ will be flat outputs if the following ratios of functions are not all identically equal in a local neighborhood:

$$\begin{aligned}
&\frac{\partial}{\partial z}\left(g\left(\xi, \frac{\partial}{\partial y^j}\right)\right) : g\left(\xi, \frac{\partial}{\partial y^j}\right), && j = 1, \dots, p,\\
&\frac{\partial}{\partial z}\left(g\left(\nabla_{\frac{\partial}{\partial y^k}}\frac{\partial}{\partial y^j}, \xi\right)\right) : g\left(\nabla_{\frac{\partial}{\partial y^k}}\frac{\partial}{\partial y^j}, \xi\right), && j,k = 1, \dots, p,\\
&\frac{\partial}{\partial z}(\xi(V)) : \xi(V).
\end{aligned} \tag{3.13}$$

If these are all identically equal that means $y^1, \dots, y^p$ are differentially dependent and another one-dimensional distribution must be tried.

REMARK 3.3. *It is readily seen that configuration flatness is determined primarily by the kinetic energy metric g since the role of potential function V only enters via the regularity conditions. This explains why in many known examples* (see [11]) *the presence or absence of gravity does not alter the configuration flat outputs but only the solution curves where singularities occur. However, we present an example in the next section where the potential function plays a crucial role via the regularity conditions.*

Proof of Proposition 3.1. Given a submersion $y : Q \to \mathbb{R}^p$, one can choose a local coordinate chart on Q such that y is the canonical submersion of $\mathbb{R}^n$ onto $\mathbb{R}^p$. Let the corresponding coordinates on Q be $(q^1, \dots, q^n)$. Then, $y^j(q) = q^j$ for $j = 1, \dots, p = n-1$. Let $\xi = a^i \frac{\partial}{\partial q^i}$ span ann P. Then all solutions of the system satisfy the single ODE

$$a^i\left(\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}^i}\right) - \frac{\partial L}{\partial q^i}\right) = 0. \tag{3.14}$$

Suppose in these coordinates g is given by g_{ij}. Then we can rewrite equation (3.14) as

$$a^i\left(g_{ij}\ddot{q}^j + \frac{\partial g_{ik}}{\partial q^j}\dot{q}^j\dot{q}^k - \frac{1}{2}\frac{\partial g_{jk}}{\partial q^i}\dot{q}^j\dot{q}^k + \frac{\partial V}{\partial q^i}\right) = 0. \tag{3.15}$$

Using the formula (3.7) for the Christoffel symbols and using $q^j = y^j$ for $j = 1, \dots, p$ to separate the terms involving $\dot{q}^n$ and $\ddot{q}^n$, we rewrite equation (3.15) as

$$a^i \left(g_{ij}\ddot{y}^j + \Gamma^m_{jk} g_{mi} \dot{y}^j \dot{y}^k + \frac{\partial V}{\partial q^i} + g_{in}\ddot{q}^n + \Gamma^m_{nn} g_{mi} (\dot{q}^n)^2 + \Gamma^m_{jn} g_{mi} \dot{y}^j \dot{q}^n \right) = 0, \tag{3.16}$$

where the range of summation of various indices is clear.

Necessity. Suppose that y are flat outputs. Then it follows that the coefficient of $\ddot{q}^n$ in the above ODE must to be zero. Otherwise we can rewrite the equation as

$$\frac{d\dot{q}^n}{dt} = f(y, \dot{y}, \ddot{y}, q^n, \dot{q}^n)$$

for some smooth function f, and by the existence theorem of solutions to ODEs, given any curve $y(t)$ we get a 2-parameter family of solutions $q(t)$ (parametrized by initial conditions $q^n(t_0), \dot{q}^n(t_0)$) that project to $y(t)$, and they are not isolated from each other and hence by Lemma 2.1 y cannot be flat, contradicting our assumption. So $a^i g_{in} = 0$ and this leaves us with an ODE of the form

$$A(y)(\dot{q}^n)^2 + B(y, \dot{y})\dot{q}^n + C(y, \dot{y}, \ddot{y}, q^n) = 0.$$

A similar reasoning tells us that the term $\dot{q}^n$ should be absent; in other words, $A(y) = 0$ and $B(y, \dot{y}) = 0$. Here A and B are given by

$$A = a^i \Gamma^m_{nn} g_{mi}, \quad B = a^i \Gamma^m_{jn}\, g_{mi} \dot{y}^j.$$

Observe that B is linear in terms $\dot{y}$ with coefficients that are functions only of (y, q^n). Hence the condition $B = 0$ can be written as $n - 1$ equations that set the coefficients of $\dot{y}^j$ to be zero. The equation $A = 0$ has the same form as these, and we get the following n equations:

$$a^i \Gamma^m_{jn} g_{im} = 0, \quad j = 1, \dots, n.$$

So, all together, flatness of y implies the following equations:

$$\begin{aligned} a^i g_{in} &= 0, \\ a^i \Gamma^m_{jn} g_{im} &= 0, \quad j = 1, \dots, n. \end{aligned} \tag{3.17}$$

If $\ker Ty = \operatorname{span}\{\eta\}$, then in our choice of coordinates $\eta = \lambda \frac{\partial}{\partial q^n}$ where λ is some nonvanishing function on Q. Hence, $g(\xi, \eta) = a^i g_{in} = 0$ by the first condition, where $\xi = a^i \frac{\partial}{\partial q^i}$ spans $\operatorname{ann} P$. Also since

$$\nabla_{\frac{\partial}{\partial q^j}} \eta = \lambda \Gamma^m_{jn} \frac{\partial}{\partial q^m} + \frac{\partial \lambda}{\partial q^j} \frac{\partial}{\partial q^n},$$

it follows that

$$g(\nabla_{\frac{\partial}{\partial q^j}} \eta, \xi) = \lambda a^i \Gamma^m_{jn} g_{im} + \frac{\partial \lambda}{\partial q^j} a^i g_{in} = 0.$$

But, by the derivation property,

$$\nabla_Z (g(\xi, \eta)) = (\nabla_Z g)(\xi, \eta) + g(\nabla_Z \xi, \eta) + g(\xi, \nabla_Z \eta)$$

and since $\nabla_Z g = 0$ for any $Z \in \mathfrak{X}(Q)$ (by the property of Levi–Cività connection) and since $g(\eta, \xi) = 0$ it follows that

$$g(\nabla_{\frac{\partial}{\partial q^j}} \xi, \eta) = 0, \quad j = 1, \dots, n.$$

By linearity of ∇ it follows that

$$g(\nabla_Z \xi, \eta) = 0 \quad \forall Z \in \mathfrak{X}(Q).$$

Hence, $\ker Ty$ is orthogonal to D.

Sufficiency. Conversely, if $\ker Ty$ is orthogonal to D, previous reasoning shows that in the same coordinate system equations (3.17) hold. As seen before these imply that the solution curves of the system are given by the ODE

$$E(q^n, y, \dot{y}, \ddot{y}) = 0,$$

where

$$E = a^i g_{ij} \ddot{y}^j + a^i g_{im} \Gamma^m_{jk} \dot{y}^j \dot{y}^k + a^i \frac{\partial V}{\partial q^i}.$$

This is not sufficient for flatness of $y^1, \dots, y^p$ since it is possible that $y^1, \dots, y^p$ are differentially dependent and this happens when E does not depend on q^n. More precisely $y^1, \dots, y^p$ are differentially dependent around q when there exists a neighborhood V of q such that $\frac{\partial E}{\partial q^n}$ is identically zero on $(\pi_2^{-1}(V) \cap \{E = 0\}) \subset J^2(\mathbb{R}, Q)$ where $\pi_2 : J^2(\mathbb{R}, Q) \to Q$ is the standard projection. The functions E and $\frac{\partial E}{\partial q^n}$ are both affine in $\ddot{y}$ and quadratic in $\dot{y}$ with the coefficients functions only of (y, q^n), and E depends on $\ddot{y}$ nontrivially since metric g is nondegenerate. Hence either $\frac{\partial E}{\partial q^n}$ is identically zero on $\pi_2^{-1}(q) \cap \{E = 0\}$ or it is non zero for generic points on $\pi_2^{-1}(q) \cap \{E = 0\}$. Furthermore, $\frac{\partial E}{\partial q^n}$ is identically zero on $\pi_2^{-1}(q) \cap \{E = 0\}$ if and only if it is a multiple of E as a polynomial in $\dot{y}$ and $\ddot{y}$ for points on $\pi_2^{-1}(q)$. Hence the regularity condition we impose is that $\frac{\partial E}{\partial q^n}$ is a not a multiple of E as a polynomial in $\ddot{y}$ and $\dot{y}$ for points on $\pi_2^{-1}(q)$. Then it would follow from continuity and the implicit function theorem that for generic points on $\pi_2^{-1}(V) \cap \{E = 0\}$ where V is some neighborhood of q, q^n can be locally solved for in terms of $y, \dot{y}, \ddot{y}$, implying flatness around q.

The rest of the proof is concerned with showing that this condition translates to the regularity condition stated in the proposition. It is sufficient to show that $\frac{\partial E}{\partial q^n}$ is a multiple of E as a polynomial in $\dot{y}, \ddot{y}$ with the ratio being a smooth function on Q in a neighborhood of q if and only if the set of ratios of functions (3.12) are all identically equal in a neighborhood of q.

Let η span $\ker Ty$. Then $\eta = \lambda \frac{\partial}{\partial q^n}$ for some nonvanishing function λ. Also let $\xi = a^i \frac{\partial}{\partial q^i}$ span $\operatorname{ann} P$. Suppose $\frac{\partial E}{\partial q^n} = fE$ for some function f defined in a neighborhood of q on Q. Considering coefficients of $\ddot{y}^j$ terms we get

$$\frac{\partial}{\partial q^n}(a^i g_{ij}) = f a^i g_{ij}, \quad j = 1, \dots, p. \tag{3.18}$$

Also observe that any vector field Z on Q is y-related if and only if it has the form $Z^j(y) \frac{\partial}{\partial y^j} + Z^n(y, q^n) \frac{\partial}{\partial q^n}$. Hence

$$\nabla_\eta(g(\xi, Z)) = \lambda \frac{\partial}{\partial q^n}(Z^j a^i g_{ij})$$

$$= \lambda Z^j \frac{\partial}{\partial q^n}(a^i g_{ij}) = \lambda f Z^j a^i g_{ij},$$

where we have used $a^i g_{in} = 0$ and equation (3.18). Hence equation (3.18) is equivalent to

$$\nabla_\eta(g(\xi, Z)) = \lambda f g(\xi, Z), \tag{3.19}$$

where Z is any arbitrary y-related vector field.

Considering coefficients of $\dot{y}^j \dot{y}^k$ we get

$$\frac{\partial}{\partial q^n}(a^i g_{im} \Gamma^m_{jk}) = f a^i g_{im} \Gamma^m_{jk}, \quad j, k = 1, \ldots, p. \tag{3.20}$$

Assuming equation (3.18), this is equivalent to

$$\nabla_\eta(g(\nabla_{Z_1} Z_2, \xi)) = f\lambda g(\nabla_{Z_1} Z_2, \xi), \tag{3.21}$$

where Z_1, Z_2 are arbitrary y-related vector fields. This is because substituting $Z_l = Z_l^j(y)\frac{\partial}{\partial y^j} + Z_l^n(y, q^n)\frac{\partial}{\partial q^n}$ for $l = 1, 2$ we get

$$g(\nabla_{Z_1} Z_2, \xi) = Z_1^j Z_2^k g\left(\Gamma^m_{jk}\frac{\partial}{\partial y^m}, \xi\right) + Z_1^j \frac{\partial Z_2^k}{\partial y^j} g\left(\frac{\partial}{\partial y^k}, \xi\right),$$

where we have used $a^i g_{in} = 0$, $a^i \Gamma^m kng_{im} = 0$ (since $\ker Ty$ is orthogonal to D) and $\frac{\partial Z_2^k}{\partial q^n} = 0$ for $k = 1, \ldots, p$. Hence

$$\begin{aligned}
&\nabla_\eta(g(\nabla_{Z_1} Z_2, \xi)) \\
&= \lambda Z_1^j Z_2^k \frac{\partial}{\partial q^n}(a^i g_{im} \Gamma^m_{jk}) + \lambda Z_1^j \frac{\partial Z_2^k}{\partial y^j}\frac{\partial}{\partial q^n}(a^i g_{ik}) \\
&= \lambda f Z_1^j Z_2^k a^i g_{im} \Gamma^m_{jk} + \lambda f Z_1^j \frac{\partial Z_2^k}{\partial y^j} a^i g_{ik},
\end{aligned}$$

where we have used equations (3.18) and (3.20). This simplifies to

$$\nabla_\eta(g(\nabla_{Z_1} Z_2, \xi)) = \lambda f g(\nabla_{Z_1} Z_2, \xi). \tag{3.22}$$

Finally considering the coefficients of the terms independent of $\dot{y}$ and $\ddot{y}$ we get

$$\frac{\partial}{\partial q^n}\left(a^i \frac{\partial V}{\partial q^i}\right) = f a^i \frac{\partial V}{\partial q^i}.$$

Clearly this is equivalent to

$$\nabla_\eta(\xi(V)) = \lambda f \xi(V), \tag{3.23}$$

completing the proof. □

4. Systems with n degrees of freedom, $n-1$ controls, and symmetry. In this section we shall consider systems of the type considered in the previous section that also exhibit symmetries. We shall suppose that a Lie group G acts on our configuration space Q with action Φ_h corresponding to $h \in G$ and that

$$\Phi_h^* g = g, \qquad \Phi_h^* P = P \quad \forall h \in G. \tag{4.1}$$

In other words the kinetic energy of the system as well as the range of control forces both are invariant under the group action. However, we do not assume that V is

invariant under the group action. Many mechanical systems fall under this category. Rigid body systems moving in Euclidean space actuated by body fixed forces are typical examples where the group is $G = SE(3)$, even though the equations of motion often do not have $SE(3)$ as a symmetry group since potential forces due to gravity break the symmetry. But since V plays a very limited role in configuration flatness we may expect that when the system is configuration flat that it would be possible to find flat outputs that reflect this symmetry. We believe this to be true and shall prove it for the case $\dim D = n-1$. The general case $\dim D < n-1$ has not yet been resolved completely (see Remark 4.4).

LEMMA 4.1. *Consider a system satisfying* (4.1). *Let* D *be defined as in* (3.9). *Then* $\Phi_{h_*} D = D$.

Proof. Let ξ span $\operatorname{ann} P$. Clearly $\Phi_{h_*}(\operatorname{ann} P) = \operatorname{ann} P$. Hence $\Phi_{h_*}\xi = \lambda_h \xi \in D$ where λ_h is some smooth function. Since Φ_h is an isometry by (4.1), it follows that $\Phi_{h_*}(\nabla_Z \xi) = \nabla_{\Phi_{h_*} Z}(\Phi_{h_*}\xi)$ by properties of ∇ (see, for example, [5, p. 161]). Hence

$$\begin{aligned} \Phi_{h_*}\nabla_Z \xi &= \nabla_{\Phi_{h_*} Z}(\lambda_h \xi) \\ &= \lambda_h \nabla_{\Phi_{h_*} Z}\xi + (\nabla_{\Phi_{h_*} Z}\lambda_h)\xi \in D. \end{aligned} \tag{4.2}$$

So we have $\Phi_{h_*} D \subset D$. Since Φ_h is a diffeomorphism, the result follows by dimension count. ∎

Let $y : Q \to \mathbb{R}^p$ be a map defined locally around $q \in Q$. We shall say $y^1, \ldots, y^p$ are *G-equivariant* if

$$\Phi_{h_*} \ker Ty = \ker Ty.$$

This means level sets of y are mapped to level sets by the group action.

PROPOSITION 4.2. *Consider a system satisfying* (4.1). *Suppose* $\dim D = n-1$ *and that the system is configuration flat. Then the flat outputs are G-equivariant.*

Proof. The proof follows from the fact that $\ker Ty$ is the orthogonal complement to D and Lemma 4.1. ∎

REMARK 4.3. *The case* $\dim D = n-1$ *is not as restrictive as it may seem. Typically* $\dim D = n$, *implying that the system is not configuration flat. When the system is configuration flat* ($\dim D \leq n-1$), *most likely* $\dim D = n-1$. *In fact, many examples of systems that are configuration flat fall into this category including the first example in next section as well as the "ducted fan with stand" in* [20] *and the "planar coupled rigid bodies" example in* [13].

REMARK 4.4. *In the case when* $\dim D < n-1$, *given the system is flat with flat outputs* $y : Q \to \mathbb{R}^p$ *around* $q \in Q$, *it is possible to construct outputs* $\tilde{y} : Q \to \mathbb{R}^p$ *around* q *that are G-equivariant and satisfy* $g(\ker T\tilde{y}, D) = 0$. *But it hasn't been resolved whether it is possible to construct* $\tilde{y}$ *in such a way that it also satisfies the regularity conditions* (3.12). *The authors are currently trying to resolve this technical issue but suspect that at least in typical cases this construction should work. The second example in the next section falls into the case* $\dim D = n-2$, *and we see that it possesses G-equivariant flat outputs.*

5. Examples. In this section we shall consider some examples to illustrate the theory developed in the previous section.

5.1. Underwater vehicle. We shall study a simple model of an underwater vehicle (see Figure 5.1) that is controlled by a force applied through a fixed point P on the body whose magnitude and direction can be independently controlled.

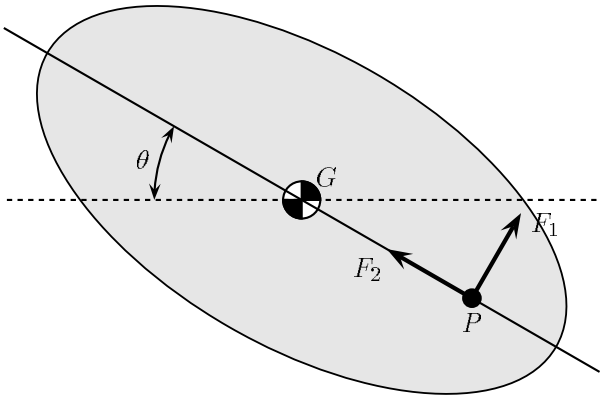


FIG. 5.1. *Underwater vehicle in* $\mathbb{R}^2$.

Only the motion in the vertical plane is considered and hence our configuration space is $SE(2) = \mathbb{R}^2 \times S^1$. This is reasonable when the vehicle has symmetries about three orthogonal planes. In addition if we assume that the center of buoyancy is coincident with the center of mass, the kinetic energy is given by

$$\frac{1}{2}(m+\delta m)(\dot{x}_1\cos\theta - \dot{x}_2\sin\theta)^2 + \frac{1}{2}(m-\delta m)(\dot{x}_1\sin\theta + \dot{x}_2\cos\theta)^2 + \frac{1}{2}I(\dot{\theta})^2, \tag{5.1}$$

where (x_1, x_2) are horizontal and vertical coordinates of the center of mass G, θ is the orientation (measured clockwise) of line PG with respect to horizontal axis, $m = M + (m_1 + m_2)/2$ and $\delta m = (m_1 - m_2)/2$, where M is the mass of the vehicle and m_1 and m_2 are added mass terms that take into account inertia of the fluid, and I is the effective moment of inertia taking into account the fluid. This model assumes an incompressible, irrotational flow and neglects viscosity effects. It is assumed that the motion of the fluid is entirely due to that of the solid. The body and the fluid together are considered to form a dynamical system, and the kinetic energy is the combined energy of body and fluid. See [7] and [6] for details. The analysis in [7] assumes a neutrally buoyant model, but we need not make this assumption since this only alters the form of the potential function but does not affect the kinetic energy. In fact for the first part of the analysis we shall not assume any specific form for potential V. If the vehicle is in air (strictly speaking vacuum) $m_1 = m_2 = 0$, so $m = M$ and $\delta m = 0$ and the kinetic energy takes the familiar form

$$\frac{1}{2}(m(\dot{x}_1)^2 + m(\dot{x}_2)^2 + I(\dot{\theta})^2),$$

where I is the usual moment of inertia and the model is the same as that of VTOL (see [10]).

The metric g in coordinates x_1, x_2, θ is given by the matrix

$$\begin{bmatrix} m+\delta m\cos 2\theta & -\delta m\sin 2\theta & 0 \\ -\delta m\sin 2\theta & m-\delta m\cos 2\theta & 0 \\ 0 & 0 & I \end{bmatrix}.$$

The control forces lie in the codistribution

$$\begin{aligned} P &= \operatorname{span}\{d(x_1 + R\cos\theta), d(x_2 - R\sin\theta)\} \\ &= \operatorname{span}\{dx_1 - R\sin\theta d\theta, dx_2 - R\cos\theta\ d\theta\} \end{aligned}$$

and $\xi = \frac{\partial}{\partial\theta} + R\sin\theta\frac{\partial}{\partial x_1} + R\cos\theta\frac{\partial}{\partial x_2}$ spans $\operatorname{ann} P$, where R is the length of PG.

The Christoffel symbols Γ^i_{jk} can be computed from g using equation (3.7). Then using formula (3.8) we see that

$$\begin{aligned}
\nabla_{\frac{\partial}{\partial x_1}}\xi &= -\frac{m\delta m}{m^2-(\delta m)^2}\sin 2\theta\frac{\partial}{\partial x_1} - \frac{\delta m}{m^2-(\delta m)^2}(\delta m + m\cos 2\theta)\frac{\partial}{\partial x_2} \\
&\quad + \frac{R\delta m\cos\theta}{I}\frac{\partial}{\partial\theta}, \\
\nabla_{\frac{\partial}{\partial x_2}}\xi &= -\frac{\delta m}{m^2-(\delta m)^2}(-\delta m + m\cos 2\theta)\frac{\partial}{\partial x_1} + \frac{m\delta m}{m^2-(\delta m)^2}\sin 2\theta\frac{\partial}{\partial x_2} \\
&\quad - \frac{R\delta m\sin\theta}{I}\frac{\partial}{\partial\theta}, \\
\nabla_{\frac{\partial}{\partial\theta}}\xi &= \frac{mR\cos\theta}{m+\delta m}\frac{\partial}{\partial x_1} - \frac{mR\sin\theta}{m+\delta m}\frac{\partial}{\partial x_2}. \qquad (5.2)
\end{aligned}$$

It can be seen by computation that the above vector fields together with ξ span the full tangent space for generic points and generic parameter values $m, \delta m, I, R$. Since by equation (3.10)

$$D = \operatorname{span}\{\nabla_{\frac{\partial}{\partial x_1}}\xi, \nabla_{\frac{\partial}{\partial x_2}}\xi, \nabla_{\frac{\partial}{\partial\theta}}\xi, \xi\},$$

it follows that $D = TQ$ for generic points on Q and for generic parameter values and hence the system is not configuration flat for generic parameter values regardless of the potential energy function.

However, for the case $\delta m = 0$ we see that

$$D = \operatorname{span}\left\{R\cos\theta\frac{\partial}{\partial x_1} - R\sin\theta\frac{\partial}{\partial x_2}, R\sin\theta\frac{\partial}{\partial x_1} + R\cos\theta\frac{\partial}{\partial x_2} + \frac{\partial}{\partial\theta}\right\}.$$

Hence $\dim D = 2$ and $\eta = \frac{\partial}{\partial\theta} - \frac{I}{mR}\sin\theta\frac{\partial}{\partial x_1} - \frac{I}{mR}\cos\theta\frac{\partial}{\partial x_2}$ spans the orthogonal complement to D. Since D has codimension 1, up to a diffeomorphism there is at most 1 set of flat outputs. One set of functions that "cut out" the foliation due to η is

$$y_1 = x_1 - \frac{I}{mR}\cos\theta, \quad y_2 = x_2 + \frac{I}{mR}\sin\theta.$$

To ensure that y_1, y_2 are indeed flat outputs we must check the regularity conditions (3.13). Let us choose $z = \theta$ as a complementary coordinate to y_1, y_2. Then,

$$\begin{aligned}
\frac{\partial}{\partial y_1} &= \frac{\partial}{\partial x_1}, \quad \frac{\partial}{\partial y_2} = \frac{\partial}{\partial x_2}, \\
\frac{\partial}{\partial z} &= -\frac{I}{mR}\sin\theta\frac{\partial}{\partial x_1} - \frac{I}{mR}\sin\theta\frac{\partial}{\partial x_1} + \frac{\partial}{\partial\theta}.
\end{aligned}$$

Hence

$$\begin{aligned}
\frac{\partial}{\partial z}\left(g\left(\xi, \frac{\partial}{\partial y_1}\right)\right) : g\left(\xi, \frac{\partial}{\partial y_1}\right) &= -\sin z : \cos z, \\
\frac{\partial}{\partial z}\left(g\left(\xi, \frac{\partial}{\partial y_2}\right)\right) : g\left(\xi, \frac{\partial}{\partial y_2}\right) &= \cos z : \sin z. \qquad (5.3)
\end{aligned}$$

So at any point $q = (y_1, y_2, z)$ these two ratios are unequal. This ensures that y_1, y_2 are indeed flat outputs everywhere.

When the vehicle is in air (strictly speaking vacuum) $\delta m = 0$, and in this case it is already known to be flat (see [10, 11]). We have just shown that up to a diffeomorphism these are the only configuration flat outputs. Also we have covered the case of an underwater vehicle of spherical shape (since then $m_1 = m_2$), and this result is independent of any assumptions we make on the potential function V.

Now let us suppose that the system is moving under gravity in air and the potential energy is given by $V = mgx_2$, where $g \approx 9.8$ m/s^2 is the acceleration due to gravity. Then the solutions of the system in coordinates y_1, y_2, z satisfy the ODE

$$\ddot{y}_1 \sin z + \ddot{y}_2 \cos z + g \cos z = 0.$$

So along generic solution curves we get

$$z(t) = \tan^{-1} \frac{\ddot{y}_2 + g}{\ddot{y}_1}$$

or

$$z(t) = \tan^{-1} \frac{\ddot{y}_2 + g}{\ddot{y}_1} + \pi.$$

The exception being the singularity at $\ddot{y}_1 = 0, \ddot{y}_2 + g = 0$. Note that this singularity is not a point on Q but corresponds to a submanifold in the jet space $J^2(\mathbb{R}, Q)$, the space with coordinates $(t, q, \dot{q}, \ddot{q})$, and such singularities are very common in practical examples. We still want to regard such systems as flat, and this is the reason why our definition of flatness refers to generic curves as opposed to all curves. Also note that although potential V does not affect the flat outputs of the system it influences where the singularities occur.

We also see that the general system (no assumptions on δm) possesses an $SE(2)$ symmetry when the potential function is ignored. If we consider translating and rotating our spatial frame of reference the expression for kinetic energy as well as the expression for P are invariant. We may state this more precisely as follows. Consider the following action of $SE(2)$ on $Q = SE(2)$. Given $h = (\alpha_1, \alpha_2, \phi) \in SE(2)$ the action Φ_h corresponds to first rotating the spatial frame counterclockwise by ϕ about its origin and then with respect to this frame translating the frame without rotation by $(-\alpha_1, -\alpha_2)$. Hence if $q = (x_1, x_2, \theta) \in Q$ then

$$\Phi_h(q) = (x_1 \cos\phi + x_2 \sin\phi + \alpha_1, -x_1 \sin\phi + x_2 \sin\phi + \alpha_2, \theta + \phi).$$

The corresponding tangent map $T\Phi_h$ is given by

$$\begin{aligned} \frac{\partial}{\partial x_1} &\to \cos\phi \frac{\partial}{\partial x_1} + \sin\phi \frac{\partial}{\partial x_2}, \\ \frac{\partial}{\partial x_2} &\to -\sin\phi \frac{\partial}{\partial x_1} + \cos\phi \frac{\partial}{\partial x_2}, \\ \frac{\partial}{\partial \theta} &\to \frac{\partial}{\partial \theta}. \end{aligned} \tag{5.4}$$

It is easy to verify this preserves g. Recalling that $\xi = \frac{\partial}{\partial\theta} + R\sin\theta\frac{\partial}{\partial x_1} + R\cos\theta\frac{\partial}{\partial x_2}$ spans $\operatorname{ann} P$, we see that $\Phi_{h_*}\xi = \xi$, implying $\Phi_h^* P = P$. In particular these statements are true for the $\delta m = 0$ case as well. Hence by Proposition 4.2 the flat outputs are G-equivariant. This is indeed true since $\eta = \frac{\partial}{\partial\theta} - \frac{I}{mR}\sin\theta\frac{\partial}{\partial x_1} - \frac{I}{mR}\cos\theta\frac{\partial}{\partial x_2}$ spans $\ker Ty$ and $\Phi_{h_*}\eta = \eta$.

5.2. Particle in the force field. This example does not necessarily correspond to an engineering example but illustrates the regularity conditions. We consider a particle of unit mass moving in three-dimensional Euclidean space in the presence of a potential field $V = x_2x_3$. Hence the kinetic energy metric is given by the 3×3 identity matrix in orthogonal coordinates x_1, x_2, x_3. Suppose that we control independently the forces along x_1 and x_3 directions. Hence $P = \operatorname{span}\{dx_1, dx_3\}$ and $\xi = \frac{\partial}{\partial x_2}$ spans $\operatorname{ann} P$. We see that Christoffel symbols are all zero by (3.7) (which is a feature of Euclidean space), and using (3.8) and (3.10) we obtain $D = \operatorname{span}\{\frac{\partial}{\partial x_2}\}$; hence the orthogonal complement to D is $\operatorname{span}\{\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_3}\}$, which is two dimensional. Hence we have infinitely many "candidates" for flat outputs that are not equivalent via a diffeomorphism. But these "candidates" may not satisfy the regularity conditions (3.13). Following the method outlined in Remark 3.2 we pick some η, say $\eta = \frac{\partial}{\partial x_3}$, which is orthogonal to D. Then $y_1 = x_1, y_2 = x_2$ are possible choices of corresponding "candidates" for flat outputs (since they cut out the one-dimensional foliation by η). We may choose $z = x_3$ to complete the coordinate system, and then we see that the ratio of functions $\frac{\partial}{\partial z}(\xi(V)) : \xi(V)$ in the set (3.13) is $1 : x_3$, whereas the ratio of $\frac{\partial}{\partial z}(g(\xi, \frac{\partial}{\partial y^2})) : g(\xi, \frac{\partial}{\partial y^2})$ is $0 : 1$. Hence x_1, x_2 are configuration flat outputs (globally). But alternatively another choice could have been $\eta = \frac{\partial}{\partial x_1}$ with corresponding candidates $y_1 = x_2, y_2 = x_3$. Choosing $z = x_1$ we see that all the ratios in (3.13) are zero and hence equal. Hence x_2, x_3 are not flat outputs as they are differentially dependent. This example is simple enough that the above conclusions can be reached by inspecting the equations of motion for the system

$$\ddot{x}_1 - \frac{\partial V}{\partial x_1} = F_1, \tag{5.5}$$

$$\ddot{x}_2 - \frac{\partial V}{\partial x_2} = 0, \tag{5.6}$$

$$\ddot{x}_3 - \frac{\partial V}{\partial x_3} = F_3, \tag{5.7}$$

where F_1, F_3 are the forces along x_1, x_3 directions. Equation (5.6) alone characterizes all solution trajectories of system and substituting $V = x_2x_3$ we obtain

$$\ddot{x}_2 - x_3 = 0. \tag{5.8}$$

It is clear from the equation that x_2, x_3 are differentially dependent and hence are not flat outputs. However, it is also clear from the equations that x_1, x_2 are flat outputs since along solution curves

$$x_3(t) = \frac{d^2x_2(t)}{dt^2}$$

and x_1, x_2 do not satisfy an ODE.

Also note that the system is globally controllable since it is globally flat. However if $V = 0$ then the system is not configuration flat and not even locally accessible.

It is easy to see that translations by the group $\mathbb{R}^3$ leave g and P invariant. But Proposition 4.2 does not apply since $\dim D = n - 2$. However, as mentioned in Remark 4.4 we see that G-equivariant flat outputs exist. In fact $y = (x_1, x_2)$ are G-equivariant, although not all (configuration) flat outputs are G-equivariant, since $\tilde{y} = (f(x_1, x_3), x_2)$, where f is an arbitrary smooth function with $\frac{\partial f}{\partial x_1} \neq 0$, are not G-equivariant for a typical f but are configuration flat outputs.

6. Conclusions and future work. We have presented a method for determining configuration flatness of Lagrangian control systems with n degrees of freedom and $n-1$ controls. Our method is constructive and provides a way for finding configuration flat outputs if they exist. We assumed a Lagrangian of the form "kinetic energy minus potential." We also assumed that the range of control forces depends only on configuration. These assumptions are not unreasonable since a wide class of systems falls into this category. However $n-1$ controls is a special case and is the simplest case next to fully actuated (n controls) systems which are always flat. In that sense we regard this as a first step toward a general theory of configuration flatness of Lagrangian systems. The authors are currently working on generalizing this result to an arbitrary number of controls.

Acknowledgment. It is a pleasure to thank Jerrold Marsden for valuable comments and Naomi Leonard for important corrections on the underwater vehicle example. The authors would also like to thank the reviewers for the useful comments and corrections.

REFERENCES

[1] R. Abraham and J. Marsden, *Foundations of Mechanics*, 2nd ed., Addison–Wesley, Reading, MA, 1985.
[2] B. Charlet, J. Lévine, and R. Marino, *On dynamic feedback linearization*, Systems Control Lett., 13 (1989), pp. 143–151.
[3] M. Fliess, J. Lévine, P. Martin, and P. Rouchon, *Flatness and defect of nonlinear systems: Introductory theory and examples*, Internat. J. Control, 61 (1995), pp. 1327–1361.
[4] M. Fliess, J. Levine, P. Martin, and P. Rouchon, *Linéarisation par bouclage dynamique et transformations de Lie-Bäcklund*, C. R. Acad. Sci. Paris Sèr. I Math., 317 (1993), pp. 981–986.
[5] S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry*, Vol. 1, Interscience Publishers, New York, 1963.
[6] H. Lamb, *Hydrodynamics*, Dover Publications, New York, 1945.
[7] N. Leonard, *Compensation for actuator failures: Dynamics and control of underactuated underwater vehicles*, in Proceedings of the 9th International Symposium on Unmanned Untethered Submersible Technology, Durham, NH, 1995; Autonomous Undersea Systems Institute, Lee, NH.
[8] A. D. Lewis and R. M. Murray, *Configuration controllability of simple mechanical control systems*, SIAM J. Control Optim., 35 (1997), pp. 766–790.
[9] P. Martin, *Contribution à l'étude des systèmes differentiellement plats*, Ph.D. thesis, École des Mines de Paris, 1992.
[10] P. Martin, S. Devasia, and B. Paden, *A different look at output tracking: Control of a vtol aircraft*, Automatica, 32 (1996), pp. 101–107.
[11] R. Murray, M. Rathinam, and W. Sluis, *Differential flatness of mechanical control systems*, in Proceedings of the 1995 ASME International Mechanical Engineering Congress and Exposition, San Francisco, 1995.
[12] J.-B. Pomet, *A differential geometric setting for dynamic equivalence and dynamic linearization*, in Geometry in Nonlinear Control and Differential Inclusions, B. Jakubczyk, W. Respondek, and T. Rzeżuchowski, eds., Warsaw, Banach Center Publications, 1995, pp. 319–339.
[13] M. Rathinam and W. Sluis, *A test for differential flatness by reduction to single input systems*, in Proceedings of IFAC 96, Vol. E, 1996, pp. 257–262; Technical Memorandum CIT-CDS 95-018, California Institute of Technology, Pasadena, CA 91125, 1995.
[14] P. Rouchon, M. Fliess, J. Levine, and P. Martin, *Flatness, motion planning and trailer systems*, in Proceedings of the IEEE Conference on Decision and Control, San Antonio, 1993.
[15] D. Saunders, *The Geometry of Jet Bundles*, Cambridge University Press, London, 1989.

[16] W. Shadwick, *Absolute equivalence and dynamic feedback linearization*, Systems Control Lett., 15 (1990), pp. 35–39.
[17] W. Sluis, *Absolute Equivalence and Its Applications to Control Theory*, Ph.D. thesis, University of Waterloo, Waterloo, Canada, 1993.
[18] W. M. Sluis and D. Tilbury, *A Bound on the Number of Integrators Needed to Linearize a Control System*, Systems Control Lett., 29 (1996), pp. 43–50.
[19] D. Tilbury, R. M. Murray, and S. S. Sastry, *Trajectory generation for the N-trailer problem using Goursat normal form*, IEEE Trans. Automat. Control, 40 (1995), pp. 802–819.
[20] M. van Nieuwstadt and R. Murray, *Approximate trajectory generation for differentially flat systems with zero dynamics*, in Proceedings of the 34th IEEE Conference on Control and Decision, New Orleans, LA, December 1995, pp. 4224–4230.
[21] M. van Nieuwstadt, M. Rathinam, and R. Murray, *Differential flatness and absolute equivalence*, in Proceedings IEEE Conference on Control and Decision, Lake Buena Vista, FL, 1994, pp. 326–332.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 1, pp. 180–192, January 1998

010

PARTIAL DISTURBANCE REJECTION WITH INTERNAL STABILITY AND H_∞ NORM BOUND*

VASFI ELDEM†, HITAY ÖZBAY‡, HASAN SELBUZ†, AND KADRI ÖZÇALDIRAN†§

Abstract. Complete disturbance rejection problems are equivalent to zeroing (cancelling) all the Markov parameters of the closed loop system between the disturbance and the controlled output. When this is not possible, one might consider partial disturbance rejection which can be defined as zeroing the first, say k, Markov parameters. In this article, our objective is to present general solvability conditions for the partial disturbance rejection problem by dynamic output feedback under the constraint of internal stability. With this solution we also obtain a suitable parametrization for the set of all solutions of the problem which is then used to obtain an H_∞ norm bound on the closed loop system. In the first part of the paper, a natural framework for the partial disturbance rejection problem is introduced. This framework consists of the ring of stable and proper rational functions and its quotient rings. Thus, the solvability conditions and the set of all solutions to the problem are easily obtained. The parametrization of the set of all solutions provides an opportunity to pursue further design goals. Along this line, H_∞ minimization has been incorporated into the problem. The upper and lower bounds on the H_∞ norm of the closed loop transfer function is obtained and compared with direct H_∞ disturbance attenuation. The results are illustrated with a simple example.

Key words. disturbance rejection, Markov parameters, internal stability, H_∞ norm

AMS subject classifications. 93A30, 93B50, 93D25

PII. S0363012995287659

1. Introduction.

In this work, linear, time-invariant systems described as

$$\begin{bmatrix} y \\ z \end{bmatrix} = \begin{bmatrix} Z_1 & Z_2 \\ Z_3 & Z_4 \end{bmatrix} \begin{bmatrix} u \\ d \end{bmatrix} \tag{1.1}$$

are considered. Here, y, z, u, and d take values from p-, q-, m-, and r-dimensional linear spaces. Z_1, Z_2, Z_3, and Z_4 are proper transfer matrices of appropriate dimensions. The literature contains a rich variety of control problems related to the system given above. The essential motivation in these problems is to design the closed loop transfer function T_{zd} between the disturbance d and the controlled output z. The early attempts along this line were devoted to cancelling (zeroing) the effect of the disturbance on the controlled output ($T_{zd} = 0$). This problem is usually referred to as the disturbance rejection (or disturbance decoupling) problem and is abbreviated to DRP. The solvability conditions for DRP can be expressed as matching of the zeros of certain subsystems. The reader may refer to Willems and Commault (1981), Özgüler and Eldem (1985), and Eldem and Özgüler (1988) for the details of the "zero-matching" conditions. Although DRP has been very useful from a purely algebraic point of view, it did not have much to offer for a practical design, because

*Received by the editors June 13, 1995; accepted for publication (in revised form) October 31, 1996.

http://www.siam.org/journals/sicon/36-1/28765.html

†Department of Mathematics, Research Institute for Basic Sciences, Marmara Research Center, Tübitak, Gebze, Kocaeli 41470, Turkey (vasfi@yunus.mam.tubitak.gov.tr).

‡Department of Electrical Engineering, Ohio State University, 205 Dreese Laboratory, 2015 Neil Ave., Columbus, OH 43210-1272 (ozbay@ee.eng.ohio-state.edu).

§Department of Electrical and Electronics Engineering, Boğaziçi University, 80815 Bebek, Istanbul, Turkey.

the rigid algebraic solvability conditions for DRP are hardly met in practical cases. This is probably one of the basic reasons why an alternative design procedure which minimizes the H_∞ norm of T_{zd} has been very popular during the last decade. The H_∞ design techniques offer a breakaway from the rigid algebraic structure of DRP and at the same time guarantee a certain disturbance attenuation level at all frequencies.

In this work, we consider yet another alternative design to DRP, which has not attracted much attention in the literature, namely, partial disturbance rejection (PDR). PDR can be defined as zeroing the first, say k, Markov parameters of T_{zd}. More specifically, in PDR problems (PDRPs) a feedback control which makes the least infinite zero order of T_{zd} greater then k is being sought. This problem was initially introduced by Emre and Silverman (1980) as a dynamic cover problem. Later it has been considered by Malabre and Martinez Garcia (1993) and Martinez, Malabre, and Kucera (1995). In the latter, dynamic state feedback together with disturbance feedthrough was used to achieve PDR with internal stability. Thus, the controller has "full information" structure which is equivalent to the so-called "one-sided model matching" problem. Here, we consider the case with dynamic measurement feedback, which is equivalent to "two-sided model matching." It turns out that the solvability conditions of these problems are quite similar and can also be expressed as zero-matching conditions. The first and an incomplete version of the results in this paper is given in Eldem (1994).

An interesting feature of PDR is that it can be interpreted as optimization at infinity. In fact, increasing the infinite zero order of T_{zd} results in increased attenuation levels at high frequencies. In H_∞ design, on the other hand, all frequencies are taken into account. This implies that PDR and H_∞ can be used together. More specifically, after obtaining a certain attenuation level at high frequencies via PDR, the remaining flexibility can be used to minimize the response at low frequencies. This has been first pointed out and illustrated by an example in Martinez, Malabre, and Kucera (1995). Here, after characterizing the set of all solutions to PDR, we also emphasize the incorporation of H_∞ point of view to PDR and show that an upper bound on $\| T_{zd} \|_\infty$ can be obtained in terms of the problem data. The results are illustrated by simple examples.

It is true that attenuation at high frequencies (or high frequency rolloff as usually called) can also be achieved by employing suitable weights on the disturbance in standard H_∞ optimization problems. Here, we exploit the algebraic structure of the system to achieve the same goal. The assessment of the superiority of one method over the other naturally calls for further study, and perhaps it depends on the particular system being considered. If, for instance, the algebraic structure of the system allows a partial disturbance rejecting design, then our method might at least give more intuition and flexibility to pursue further design goals.

2. Notation and problem formulation. As usual $\mathbf{R}$ denotes the field of real numbers. The ring of polynomials in the indeterminant s, with coefficients from $\mathbf{R}$, is denoted as $\mathbf{R}[s]$. The field of fractions of $\mathbf{R}[s]$, i.e., the field of rational functions, is represented by $\mathbf{R}(s)$. We will use $\mathbf{R}^{p\times m}(s)$ to denote $p \times m$ matrices with entries from $\mathbf{R}(s)$. The Laurent series expansion of $N(s)$ in $\mathbf{R}^{p\times m}(s)$ is given as

$$N(s) = \sum_{i=-k}^{\infty} N_i s^{-i}, \tag{2.1}$$

where N_i's are $p\times m$ matrices over $\mathbf{R}$. $N(s)$ is called proper if $N_{-k} = N_{-(k-1)} = \cdots = N_{-1} = 0$. It is called strictly proper, if it is proper and $N_0 = 0$. If N_0 is square and

nonsingular, it is called biproper. Strictly polynomial, strictly proper, polynomial, and proper parts of N are denoted as N^+, N^-, N_+, and N_-, respectively. More specifically,

$$N^+ := \sum_{i=-k}^{-1} N_i s^{-i}, \quad N_+ := \sum_{i=-k}^{0} N_i s^{-i}, \tag{2.2}$$

$$N^- := \sum_{i=1}^{\infty} N_i s^{-i}, \quad N_- := \sum_{i=0}^{\infty} N_i s^{-i}. \tag{2.3}$$

$\mathbf{R}_p(s)$ and $\mathbf{R}_{ps}(s)$ will be used to denote the ring of proper rational functions and the ring of proper and stable rational functions, respectively. If we let $\partial(.)$ denote the usual degree function for the polynomial ring, then $\mathbf{R}_p(s)$ and $\mathbf{R}_{ps}(s)$ become Euclidean domains with the following degree functions:

$$\partial_p(p/q) := \partial(q) - \partial(p), \tag{2.4}$$

$$\partial_{ps}(p/q) := \partial(q) - \partial(p) + \text{number of unstable zeros of } p. \tag{2.5}$$

$\partial(.)_p$ is sometimes called "the relative degree." Recall that $\mathbf{R}_p^{m\times m}(s)$ and $\mathbf{R}_{ps}^{m\times m}(s)$ are rings with the usual definition of matrix addition and multiplication. An element in a ring is called a unit if it has an inverse in the ring. The units in $\mathbf{R}_p^{m\times m}(s)$ are called biproper matrices. A similar terminology will be used for the units in $\mathbf{R}_{ps}^{m\times m}(s)$, and they will be called *biproper* and *bistable*. Let us now go back to the PDRP. The objective in this problem, given the system (1.1), is to find a proper Z_c such that under the control law

$$u = -Z_c y \tag{2.6}$$

the closed loop transfer matrix between z and d, T_{zd} has first k Markov parameters equal to zero. Note that T_{zd} can be written as

$$T_{zd} = Z_4 - Z_3 Z_c (I + Z_1 Z_c)^{-1} Z_2. \tag{2.7}$$

Thus, if we assume that Z_1 is strictly proper, $(1 + Z_1 Z_c)^{-1}$ exists and it is proper. In view of this, the partial disturbance rejection problem via measurement feedback (PDRPM) can be defined as follows in Definition 2.1.

DEFINITION 2.1 (PDRPM). *Given proper Z_4, Z_3, and Z_2, find a proper X such that*

$$Z_4 - Z_3 X Z_2 = \frac{1}{s^{k+1}} P \tag{2.8}$$

for some proper P.

For the case with internal stability, a similar definition can be used. Before doing that, we shall assume without loss of generality, that Z_4, Z_3, and Z_2 are both proper and stable. (If not, Youla parametrization of internally stabilizing compensators could be used to get a similar equation as above with Z_4, Z_3, Z_2 stable and proper and with X as the free parameter matrix over $\mathbf{R}_{ps}(s)$. The reader is referred to Francis (1987) for the details.) In view of this observation, the partial disturbance rejection problem by measurement feedback with internal stability (PDRPMS) can be defined as follows in Definition 2.2.

DEFINITION 2.2 (PDRPMS). *Given proper, stable matrices Z_4, Z_3, and Z_2, find a proper, stable Q such that*

$$Z_4 - Z_3 Q Z_2 = \frac{1}{\pi(s)} P \tag{2.9}$$

for some stable, proper P and a stable polynomial $\pi(s)$ of degree $k+1$.

Observe that the solvability condition of PDRPMS is independent of the polynomial $\pi(s)$. More precisely, if Q is a solution for some polynomial $\pi_1(s)$ and a proper P_1, i.e., $Z_4 - Z_3 Q Z_2 = P_1(s)/\pi_1(s)$, then it is also a solution for an arbitrary stable polynomial $\pi_2(s)$ of degree $k+1$, i.e., $Z_4 - Z_3 Q Z_2 = P_2(s)/\pi_2, (P_2 := \pi_2 P_1/\pi_1)$. Therefore, we can fix $\pi(s)$ and define PDRPMS accordingly in Definition 2.3.

DEFINITION 2.3 (PDRPMS). *Given stable, proper matrices Z_4, Z_3, and Z_2 and a stable polynomial $\pi(s)$ of degree $k+1$, find a stable, proper Q such that*

$$Z_4 - Z_3 Q Z_2 = \frac{1}{\pi} P \tag{2.10}$$

for some proper and stable P.

Also observe that the above definition, where the polynomial is fixed as π, gives rise to a very natural algebraic setup for PDRPs. This consists of the ideal $(1/\pi)R_{ps}(s)$ and the quotient ring $\mathbf{R}_{ps}(s)/(1/\pi)R_{ps}(s)$ which is extensively investigated in the next section.

3. Preliminary results. In this section, a simple version of the problem is introduced in order to give an insight for the algebraic setup behind the partial model matching problems. More specifically, the problem to be considered can be defined in Definition 3.1.

DEFINITION 3.1 (Problem P1). *Given m and n in $\mathbf{R}_{ps}(s)$ and a stable polynomial π of degree $k+1$, find q in $\mathbf{R}_{ps}(s)$ such that*

$$n - mq = \frac{1}{\pi} p \tag{3.1}$$

for some p in $\mathbf{R}_{ps}(s)$.

Consider the ideal $(1/\pi)R_{ps}(s)$. Using this ideal, let us define an equivalence relation on $\mathbf{R}_{ps}(s)$ as follows: *n_1 and n_2 in $\mathbf{R}_{ps}(s)$ are said to be equivalent if $n_1 - n_2$ is in $(1/\pi)R_{ps}(s)$*. This relation gives the quotient ring $\mathbf{R}_{ps}(s)/(1/\pi)R_{ps}(s)$. We will denote this quotient ring by $\mathbf{R}_{ps}(\pi)$.

LEMMA 3.2. *$\mathbf{R}_{ps}(\pi)$ is a $(k+1)$-dimensional $\mathbf{R}$-linear space. A basis for $\mathbf{R}_{ps}(\pi)$ can be given as*

$$\frac{s^i}{\pi}, \quad i = 1, 2, ..., k+1. \tag{3.2}$$

Proof. The reader may refer to, for instance, Özgüler (1994). □

For a given n_1 in $\mathbf{R}_{ps}(s)$, n_1 mod $(1/\pi)\mathbf{R}_{ps}(s)$ can be calculated easily as follows:

$$n_1(\pi) := n_1 \bmod \frac{1}{\pi}\mathbf{R}_{ps}(s) = \frac{1}{\pi}(\pi n_1)^+. \tag{3.3}$$

It is clear from the above equation that $\{s^i/\pi, \ \ i = 1, ..., k+1\}$ is a basis for $\mathbf{R}_{ps}(\pi)$. Furthermore, for any n_1 in $\mathbf{R}_{ps}(s)$ we have

$$n_1 = \frac{1}{\pi}(\pi n_1)^+ \ + \ \frac{1}{\pi}(\pi n_1)_-. \tag{3.4}$$

LEMMA 3.3. *P*1 *is solvable for given* n, m, *and a stable polynomial* $\pi(s)$ *of degree* $k+1$ *iff it is solvable for* $n(\pi)$ *and* $m(\pi)$.

Proof. The proof directly follows from the definition of $n(\pi)$ and $m(\pi)$ and equation (3.4). ∎

For a given stable and proper m, let us now consider the set of equivalence classes generated by mq as q ranges over $\mathbf{R}_{ps}(s)$. It is clear that this set is equivalent to the quotient ring $mR_{ps}(s)/(1/\pi)R_{ps}(s)$ which we denote as $\mathbf{R}_{ps}(m\pi)$.

LEMMA 3.4. $\mathbf{R}_{ps}(m\pi)$ *is an R-linear subspace of* $\mathbf{R}_{ps}(\pi)$ *of dimension* $k+1-\partial_p(m)$.

Proof. Let q be in $\mathbf{R}_{ps}(s)$. Then $mq \bmod 1/\pi R_{ps}(s)$ can be calculated as

$$mq \bmod \left(\frac{1}{\pi}R_{ps}\right) = \frac{1}{\pi}(\pi mq)^+. \tag{3.5}$$

Note that since q and m are proper $\partial(\pi mq)^+ \le k+1-\partial_p(m)$ and the equality is achieved if q is biproper. Consequently, $\mathbf{R}_{ps}(m\pi)$ has dimension $k+1-\partial_p(m)$. Now we have to show that for any given polynomial δ with $\partial(\delta) \le k+1-\partial_p(m)$, there exists a q in $\mathbf{R}_{ps}$ such that

$$mq \bmod \left(\frac{1}{\pi}R_{ps}\right) = \frac{\delta}{\pi}. \tag{3.6}$$

To this end, let $m = m_1/m_2$ where m_1 *and* m_2 are in $\mathbf{R}[s]$. Also let $\hat{m}_1$ be a Hurwitz polynomial such that $\partial(\hat{m}_1) = \partial(m_1)$. Now we can choose q as $q := m_2\beta/\pi\hat{m}_1$ where $\beta := ((\hat{m}_1/m_1)\delta)^+$. Note that q is stable and furthermore, since $\partial(\beta) = \partial(\delta) \le k+1-\partial(m_2)+\partial(m_1)$, it follows that it is also proper. Then

$$(\pi mq)^+ = \left(\pi\frac{m_1}{m_2}\frac{m_2\beta}{\pi\hat{m}_1}\right)^+ = \left(\frac{m_1\beta}{\hat{m}_1}\right)^+ = \delta \tag{3.7}$$

$$\Rightarrow mq \bmod \frac{1}{\pi}R_{ps} = \frac{\delta}{\pi}. \tag{3.8}$$

Hence the proof is complete. ∎

LEMMA 3.5. *P*1 *is solvable iff* $n(\pi)$ *is in* $\mathbf{R}_{ps}(m\pi)$.

Proof. The proof directly follows from the definition of $n(\pi)$, $\mathbf{R}_{ps}(m\pi)$, and Lemma 3.3. ∎

Remark 3.6. *Note that if* $\partial_p(n) \ge k+1$, *then P*1 *is trivially solved by* $q=0$. *Therefore, assume that* $\partial_p(n) < k+1$. *Let* $n(\pi) = \psi/\pi$ *and* $m(\pi) = \phi/\pi$ *where* ψ *and* ϕ *are in* $\mathbf{R}[s]$. *The above lemma implies that the inequality* $\partial(\psi) \le k+1-\partial_p(m)$ $(=\partial(\phi))$ *is a necessary and sufficient condition for the solvability of P*1. *Thus, an equivalent condition for the solvability of P*1 *is* $\partial_p(n(\pi)) \ge \partial_p(m(\pi))$, *which, in turn, is equivalent to* $\partial_p(n) \ge \partial_p(m)$. *This condition can be interpreted as a matching condition for the infinite zero orders. Consequently, we have the following lemma.*

LEMMA 3.7. *P*1 *is solvable iff* $\partial_p(n) \ge \min\{\partial_p(m), k+1\}$.

4. The solution of PDRPMS. In this section, the solvability conditions for PDRPMS is presented. Since Z_4, Z_3, and Z_2 are matrices with entries from $\mathbf{R}_{ps}(s)$, there exist biproper and bistable matrices B_1, B_2, C_1, and C_2 such that

$$B_1Z_3B_2 := \begin{bmatrix} \Lambda & 0 \\ 0 & 0 \end{bmatrix}, \qquad C_1Z_2C_2 := \begin{bmatrix} \Gamma & 0 \\ 0 & 0 \end{bmatrix}, \tag{4.1}$$

where Λ and Γ are diagonal matrices defined as

$$\Lambda := \text{diag}\{\lambda_i\}, \qquad \Gamma := \text{diag}\{\gamma_i\}. \tag{4.2}$$

Clearly, λ_i and γ_i are in $\mathbf{R}_{ps}(s)$. Also partition $B_1 Z_4 C_2$ compatibly as

$$B_1 Z_4 C_2 := \begin{bmatrix} Z_{41} & Z_{42} \\ Z_{43} & Z_{44} \end{bmatrix}.$$

Let π be a maximum-degree stable polynomial such that

$$\pi \begin{bmatrix} 0 & Z_{42} \\ Z_{43} & Z_{44} \end{bmatrix} \tag{4.3}$$

is proper and denote the degree of π as $\partial(\pi) = k^* + 1$. Then Theorem 4.1 holds.

THEOREM 4.1. *PDRPMS is solvable for some integer k iff*

1. $k \le k^*$,
2. $\min\{\partial_p(\lambda_i) + \partial_p(\gamma_j), k+1\} \le \partial_p(Z_{41})_{ij}$,

where $(.)_{ij}$ *denotes the* ij*th entry.*

Proof. Suppose that the problem is solvable. Then, there exist stable and proper Q and P such that $Z_4 - Z_3 Q Z_2 = (1/\pi)P$. Using B_1, B_2, C_1, and C_2 defined by equation (4.1) and defining

$$B_2^{-1} Q C_1^{-1} := \begin{bmatrix} Q_1 & Q_2 \\ Q_3 & Q_4 \end{bmatrix}, \qquad B_1 P C_2 := \begin{bmatrix} P_1 & P_2 \\ P_3 & P_4 \end{bmatrix}, \tag{4.4}$$

we obtain the following equality:

$$\begin{bmatrix} Z_{41} & Z_{42} \\ Z_{43} & Z_{44} \end{bmatrix} - \begin{bmatrix} \Lambda & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} Q_1 & Q_2 \\ Q_3 & Q_4 \end{bmatrix} \begin{bmatrix} \Gamma & 0 \\ 0 & 0 \end{bmatrix} = \frac{1}{\pi} \begin{bmatrix} P_1 & P_2 \\ P_3 & P_4 \end{bmatrix}. \tag{4.5}$$

Since

$$\begin{bmatrix} 0 & Z_{42} \\ Z_{43} & Z_{44} \end{bmatrix} = \frac{1}{\pi} \begin{bmatrix} 0 & P_2 \\ P_3 & P_4 \end{bmatrix}, \tag{4.6}$$

it is clear that $k \le k^*$. Furthermore, since

$$(Z_{41})_{ij} - \lambda_i \gamma_j (Q_1)_{ij} = \frac{1}{\pi}(P_1)_{ij} \tag{4.7}$$

Lemma 3.7 implies Theorem 4.1(2).

For sufficiency, note that if $\min\{\partial_p(\lambda_i) + \partial_p(\gamma_j), k+1\} = k+1$, then $(Q_1)_{ij}$ can be chosen as zero. Otherwise, Lemma 3.3 can be used to construct $(Q_1)_{ij}$. Finally, Q can be obtained from Q_1 via equation (4.1). ∎

In order to characterize the set of all solutions of PDRPMS, a different quotient ring is employed. To this end, let $\sigma(s)$ be a stable polynomial of degree equal to $\partial(\phi)(= k+1-\partial_p(m))$ and consider the quotient ring $\mathbf{R}_{ps}(\sigma)$. This quotient ring is also an R-linear space of dimension $k+1-\partial_p(m)$ with a basis $s^i/\sigma, i = 1, 2, ..., k+1-\partial_p(m)$.

LEMMA 4.2. *Let n and m be stable proper rationals and* π *be a stable polynomial of degree* $k+1$. *Suppose that* $\partial_p(n) \ge \partial_p(m)$ *(P*1 *is solvable for all k). Then, there exists a unique solution* q^* *of P*1 *in* $\mathbf{R}_{ps}(\sigma)$ *and furthermore every solution q can be represented as*

$$q = q^* + \frac{1}{\sigma} R_{ps}(s). \tag{4.8}$$

Proof. Note that any solution q can be uniquely decomposed as in (3.4)

$$q = q^* + \frac{1}{\sigma}\hat{q}, \tag{4.9}$$

where q^* is in $\mathbf{R}_{ps}(\sigma)$ and $\hat{q}$ is stable and proper. This implies that

$$\frac{1}{\pi}p = n - mq = n - mq^* - m\frac{1}{\sigma}\hat{q} \;\;\Rightarrow\;\; n - mq^* = \frac{1}{\pi}\left(p + \frac{m\pi}{\sigma}\hat{q}\right). \tag{4.10}$$

Since $m\pi/\sigma$ is biproper, it follows that q^* is a solution in $\mathbf{R}_{ps}(\sigma)$. q^* can be expressed as α^*/σ for some polynomial α^* of degree $\leq k + 1 - \partial_p(m)$. Note that

$$(\pi n)^+ = \left(\pi m\frac{\alpha^*}{\sigma}\right)^+ \;\;\Rightarrow\;\; \alpha^* = \left(\frac{\sigma n}{m}\right)^+; \tag{4.11}$$

i.e., α^* is unique. Consequently, the set of all solutions can be obtained as $q = q^* + (1/\sigma)\hat{q}$ as $\hat{q}$ ranges over $\mathbf{R}_{ps}$. ∎

Remark 4.3. *Note that by the above result the order of the compensator increases with k, the lower bound for the relative degree of the closed loop transfer function. This can be seen by observing that the unique part q^* of the solution q is expressed as $q^* = \alpha^*/\sigma$, where $\partial(\sigma) = k + 1 - \partial_p(m)$. This increase basically comes from the nature of the problem rather than the method being employed. In order to see this, let n and m be as in the above lemma with m biproper and bistable; i.e., PDRPMS is solvable for all k. Then it follows that $(1/\pi)p/m = n/m - q$. This implies that the first k Markov parameters of n/m are cancelled with the first k Markov parameters of q. This is only possible if q has order greater than k. The increase in the order of the compensator could be taken as a disadvantage of the approach presented here. However, it will be shown in Corollary* 5.11 *that if π is chosen to have real and negative zeros which tend to $-\infty$, then one can get arbitrarily close to the best achievable $\mathcal{H}_\infty$ performance. This fact holds true regardless of the degree $(k+1)$ of π (provided of course that the degree is high enough to yield a solution). Thus, one might not need to increase k too much to get a better performance.*

Using the above characterization and Theorem 4.1 we can also determine the set of all solutions in the general case. Note that by equation (4.1) we have

$$(Z_{41})_{ij} - \lambda_i\gamma_j(Q_1)_{ij} = \frac{1}{\pi}(P_1)_{ij}, \tag{4.12}$$

where $(Q_1)_{ij}$ can be expressed as

$$(Q_1)_{ij} = (Q_1^*)_{ij} + \frac{w_j}{\tau_i}(\hat{Q}_1)_{ij}, \tag{4.13}$$

where τ_i and w_j are stable polynomials with degrees $\partial(\tau_i) = k + 1 - \partial_p(\lambda_i)$, $\partial(w_j) = \partial_p(\gamma_j)$. Let us now define the solution Q as follows:

$$Q = B_2\begin{bmatrix} Q_1^* + \Omega_1\hat{Q}_1\Omega_2 & Q_2 \\ Q_3 & Q_4 \end{bmatrix}C_1, \tag{4.14}$$

where $\Omega_1 := \text{diag}\,\{1/\tau_i\}$, $\Omega_2 := \text{diag}\,\{w_j\}$, and B_2 and C_1 are stable, biproper matrices defined in equation (4.1). Then, we have

$$Z_4 - Z_3QZ_2 = Z_4 - B_1^{-1}\begin{bmatrix} \Lambda & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} Q_1^* + \Omega_1\hat{Q}_1\Omega_2 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} \Gamma & 0 \\ 0 & 0 \end{bmatrix}C_2^{-1} \tag{4.15}$$

$$= \frac{1}{\pi}P_1^* - B_1^{-1}\begin{bmatrix} \Lambda\Omega_1\hat{Q}_1\Omega_2\Gamma & 0 \\ 0 & 0 \end{bmatrix}C_2^{-1}, \tag{4.16}$$

where

$$\frac{1}{\pi}P_1^* := Z_4 - B_1^{-1}\begin{bmatrix} \Lambda Q_1^*\Gamma & 0 \\ 0 & 0 \end{bmatrix} C_2^{-1}. \tag{4.17}$$

Let

$$\tilde{B}_1 := B_1^{-1}\begin{bmatrix} \Lambda\Omega_1 \\ 0 \end{bmatrix} \text{ and } \tilde{C}_2 = [\Omega_2\Gamma \;\; 0]C_2^{-1}. \tag{4.18}$$

Note that $\tilde{C}_2$ is left biproper and $\pi\tilde{B}_1$ is right biproper. Furthermore, the set of all closed loop transfers T_{zd} of PDRPMS can be generated by

$$T_{zd} = \frac{1}{\pi}(P_1^* - \pi\tilde{B}_1\hat{Q}_1\tilde{C}_2) \tag{4.19}$$

as $\hat{Q}_1$ ranges over $\mathbf{R}_{ps}$.

5. A comparison of PDRPMS with H_∞ disturbance attenuation. A very natural question which can be raised about the parametrization of all possible closed loop transfers T_{zd} given in the previous section is the infimum value of $\| T_{zd} \|_\infty$ that can be obtained via PDRPMS. In the following discussion, this infimum will be denoted by γ^*_{pdrs}. That is to say,

$$\gamma^*_{pdrs} := \inf_{\hat{Q}_1 \in R_{ps}} \left\| \frac{1}{\pi}(P_1^* - \pi\tilde{B}_1\hat{Q}_1\tilde{C}_2) \right\|_\infty . \tag{5.1}$$

We also define γ^*_π by

$$\gamma^*_\pi := \inf_{\hat{Q}_1 \in R_{ps}} \| P_1^* - \pi\tilde{B}_1\hat{Q}_1\tilde{C}_2 \|_\infty . \tag{5.2}$$

Clearly,

$$\gamma^*_{pdrs} \le \left\| \frac{1}{\pi} \right\|_\infty \gamma^*_\pi. \tag{5.3}$$

As PDRPMS is nothing but a restricted model matching problem, it follows that

$$\gamma^* \le \gamma^*_{pdrs}, \tag{5.4}$$

where γ^* denotes the $\mathcal{H}_\infty$ optimal model mismatch, i.e.,

$$\gamma^* := \inf_{Q \in R_{ps}} \| Z_4 - Z_3 Q Z_2 \|_\infty . \tag{5.5}$$

Thus, it follows from equations (5.3) and (5.4) that

$$\gamma^* \le \gamma^*_{pdrs} \le \left\| \frac{1}{\pi} \right\|_\infty \gamma^*_\pi. \tag{5.6}$$

In this section, we shall establish that, under a number of assumptions, γ^*_{pdrs} tends to γ^* as the roots of π tend to infinity. The assumptions that we adopt to this end are as follows.

A1. Z_3 has full row rank.

A2. Z_2 has full column rank.

A3. Z_2 and Z_3 do not have any finite imaginary axis zeros.

Remark 5.1. *Note that the assumptions above are adopted only for the comparison of PDRPMS with H_∞ disturbance attenuation and they have no relation to the solvability conditions of the PDRPMS presented in the previous sections. The first two of these assumptions are adopted to keep the comparison simple, because with these two assumptions we would be dealing with a* one-block *problem which is the simplest case in H_∞ disturbance attenuation. Without these assumptions, we would be considering* two-block or four-block *problems, which would make the comparison quite lengthy and complicated. We believe, however, that such a comparison is possible and would probably yield results similar to those we obtain for the* one-block *case.*

The third assumption above, on the other hand, is less restrictive than its counterpart in H_∞ disturbance attenuation. Assumption A3 *allows for zeros at infinity whereas in H_∞ disturbance attenuation zeros at infinity are not usually allowed as they give rise to improper compensators. This is then taken care of by employing high gain feedback. Therefore, PDRPMS can be viewed also as an alternative way (to high gain feedback) of dealing with zeros at infinity.*

Now, we let $Z_3 := Z_{3,i}Z_{3,0}$ and $Z_2 := Z_{2,co}Z_{2,ci}$ be inner-outer and co-inner-co-outer factorizations of Z_3 and Z_2, respectively. Under the assumptions introduced above, $Z_{3,i}^{-1}$ and $Z_{2,ci}^{-1}$ exist and are from $\mathcal{L}_\infty$.

We first recall the following result from Francis (1987).

LEMMA 5.2. $\gamma^* = \| Z_{3,i}^{-1} Z_4 Z_{2,ci}^{-1} \|_H$ *where* $\| (\cdot) \|_H$ *denotes the Hankel norm of* $(\cdot)$.

Similarly, we have Lemma 5.3.

LEMMA 5.3. $\gamma_\pi^* = \| \pi Z_{3,i}^{-1} Z_4 Z_{2,ci}^{-1} \|_H$.

Proof. Recall that $Z_3 = B_1^{-1}[\Lambda \ : \ 0]B_2^{-1}$. Thus,

$$Z_{3,i}^{-1} Z_3 = Z_{3,i}^{-1} B_1^{-1}[\Lambda \ : \ 0]B_2^{-1} \tag{5.7}$$

is outer. Since B_2 is bistable and biproper and Ω_1 is stable, it follows that

$$Z_{3,i}^{-1} B_1^{-1}[\Lambda\Omega_1] \tag{5.8}$$

is also outer. Also recall that $\Omega_1 = \mathrm{diag}\{1/\tau_i\}$, where τ_i's are Hurwitz polynomials of degree $k + 1 - \partial(\lambda_i)$. This implies that

$$\pi Z_{3,i}^{-1} \tilde{B}_1 := Z_{3,i}^{-1} B_1^{-1}[\Lambda\Omega_1] \tag{5.9}$$

is bistable and biproper. Using a similar argument, it is easy to see that $\tilde{C}_2 Z_{2,ci}^{-1}$ is also bistable and biproper. Then, it follows that γ_π^* is the solution of a Nehari problem. More specifically,

$$\gamma_\pi^* := \inf_{\hat{Q}_1 \in R_{ps}} \| P_1^* - \pi \tilde{B}_1 \hat{Q}_1 \tilde{C}_2 \|_\infty \tag{5.10}$$

$$= \inf_{\hat{Q}_1 \in R_{ps}} \| Z_{3,i}^{-1} P_1^* Z_{2,ci}^{-1} - \pi Z_{3,i}^{-1} \tilde{B}_1 \hat{Q}_1 \tilde{C}_2 Z_{2,ci}^{-1} \|_\infty \tag{5.11}$$

$$= \| Z_{3,i}^{-1} P_1^* Z_{2,ci}^{-1} \|_H = \| \pi Z_{3,i}^{-1} Z_4 Z_{2,ci}^{-1} \|_H \ . \tag{5.12}$$

Furthermore, because $\pi Z_{3,i}^{-1} \tilde{B}_1$ and $\tilde{C}_2 Z_{2,ci}^{-1}$ are bistable and biproper, we can solve for $\hat{Q}_1$. ∎

Remark 5.4. *In case Z_i, $i = 1, 2, 3, 4$, are SISO transfer functions with distinct unstable zeros, the Nevanlinna–Pick interpolation theory can be used to show that* (*see, e.g., Foias, Özbay, and Tannenbaum* (1996)) *the Hankel norm above can be computed as* $[\lambda_{max}(S^*V^*SV)]^{1/2}$ *where* $S_{ij} := 1/(a_i + \bar{a_j})$ *and* $V = \mathrm{diag}\{\alpha_i\}$ *where* α_i*'s and* a_i*'s come from the partial fraction expansion*

$$Z_{3,i}^{-1} Z_4 Z_{2,ci}^{-1} = \sum_{k=1}^{l} \frac{\alpha_k}{(a_k - s)}. \tag{5.13}$$

However, to bypass the difficulties associated with the interpolation problem for multivariable systems, in the sequel we shall adopt a slightly different approach to the problem.

Before computing γ_π^* explicitly, we recall the computation of γ^*. Let (C, A, B) be a minimal realization of $(Z_{3,i}^{-1} Z_4 Z_{2,ci}^{-1})_u$, the strictly proper and antistable part of $Z_{3,i}^{-1} Z_4 Z_{2,ci}^{-1}$. Let L_c and L_o denote the controllability and observability Gramians which are given as the unique solutions of

$$AL_c + L_c A^T = BB^T, \tag{5.14}$$

$$A^T L_o + L_o A = C^T C. \tag{5.15}$$

Then, we have (see Francis (1987))

$$\gamma^* = \| Z_{3,i}^{-1} Z_4 Z_{2,ci}^{-1} \|_H = [\lambda_{max}(L_c L_o)]^{\frac{1}{2}}. \tag{5.16}$$

To use this fact in computing γ_π^*, we first note the following fact.

LEMMA 5.5. *Let (C, A, B) be a minimal realization of $(Z_{3,i}^{-1} Z_4 Z_{2,ci}^{-1})_u$ with controllability and observability Gramians L_c and L_o.*

1. *$(C, A, \pi(A)B)$ is a minimal realization of $(\pi Z_{3,i}^{-1} Z_4 Z_{2,ci}^{-1})_u$.*
2. *Controllability and observability Gramians of $(C, A, \pi(A)B)$ are given by $\pi(A)L_c\pi(A^T)$ and L_o, respectively.*

Proof. Using the Laurent series expansion we have

$$(Z_{3,i}^{-1} Z_4 Z_{2,ci}^{-1})_u = \sum_{i=1}^{\infty} CA^{i-1} B s^{-i}, \tag{5.17}$$

$$(\pi Z_{3,i}^{-1} Z_4 Z_{2,ci}^{-1})_u = [\pi (Z_{3,i}^{-1} Z_4 Z_{2,ci}^{-1})_u]^- = \left[\pi(s) \sum_{i=1}^{\infty} CA^{i-1} B s^{-i} \right]^- \tag{5.18}$$

$$= \sum_{i=1}^{\infty} CA^{i-1} \pi(A) B s^{-i}. \tag{5.19}$$

Note that minimality of (C, A, B) implies minimality of $(C, A, \pi(A)B)$ iff the set of roots of π does not intersect the spectrum of A. In this case, A is antistable and π is Hurwitz. Thus, $(C, A, \pi(A)B)$ is minimal.

To prove Lemma 5.5(2), simply note that

$$A[\pi(A)]L_c\pi(A^T) + \pi(A)L_c\pi(A^T)A^T = \pi(A)BB^T\pi(A^T), \tag{5.20}$$

$$A^T L_o + L_o A = C^T C. \tag{5.21}$$

This completes the proof. □

Thus, we conclude from equation (5.16) and the lemma above that the following corollary holds.

COROLLARY 5.6. $\gamma_\pi^* = \| \pi Z_{3,i}^{-1} Z_4 Z_{2,ci}^{-1} \|_H = [\lambda_{max}(\pi(A)L_c\pi(A^T)L_o)]^{1/2}$.

COROLLARY 5.7. *Let (C, A, B) be a balanced realization of $(Z_{3,i}^{-1} Z_4 Z_{2,ci}^{-1})_u$ so that $L_c = L_o = D$, where D is a diagonal matrix. Then*

1. $\gamma^* = \lambda_{max}(D)$.
2. $\gamma_\pi^* = [\lambda_{max}(\pi(A)D\pi(A^T)D)]^{1/2}$.

In the sequel it will be assumed, without loss of generality, that (C, A, B) is a balanced realization. This assumption together with the corollary above immediately yield an upper bound for γ_π^* as shown in the following lemma.

LEMMA 5.8. $\lambda_{max}(\pi(A)D\pi(A^T)D) \le \lambda_{max}(D^2)\lambda_{max}(\pi(A)\pi(A^T))$.

Proof. The proof depends on two simple observations:

1. For square nonsingular matrices X and Y, the eigenvalues of XY and YX are the same.
2. If P is positive semidefinite then $\lambda_{max}(Q^TPQ) \le \lambda_{max}(P)\lambda_{max}(Q^TQ)$ (see Marcus and Minc (1992)).

Then, using the properties above successively, we get

$$\lambda_{max}(\pi(A)D\pi(A^T)D) = \lambda_{max}(D^{\frac{1}{2}}\pi(A)D\pi(A)D^{\frac{1}{2}}) \tag{5.22}$$

$$\le \lambda_{max}(D)\lambda_{max}(D^{\frac{1}{2}}\pi(A)\pi(A^T)D^{\frac{1}{2}}) \tag{5.23}$$

$$= \lambda_{max}(D)\lambda_{max}(\pi(A^T)D\pi(A)) \tag{5.24}$$

$$\le \lambda_{max}(D^2)\lambda_{max}(\pi(A^T)\pi(A)). \tag{5.25}$$

This completes the proof. □

Thus, it follows from Corollary 5.6, Lemma 5.5, and the fact $[\lambda_{max}(D^2)]^{1/2} = \lambda_{max}(D)$ that Corollary 5.9 holds.

COROLLARY 5.9. $\gamma_\pi^* \le \gamma^*[\lambda_{max}(\pi(A)\pi(A^T))]^{\frac{1}{2}}$.

Let us assume without loss of generality that $\pi(s)$ is a monic polynomial with real distinct roots. We can now present the main result of this section.

THEOREM 5.10. *Let δ_i denote the roots of π which are assumed to be distinct. Then*

$$\gamma^* \le \gamma_{pdrs}^* \le \gamma^* \left[\lambda_{max} \left\{ \prod_{i=1}^{k+1} \left(\frac{A^T}{\delta_i} + I \right)' \prod_{j=1}^{k+1} \left(\frac{A}{\delta_i} + I \right) \right\} \right]^{\frac{1}{2}}.$$

Proof. Note that $\| 1/\pi \|_\infty = (\prod_{i=1}^{k+1} \delta_i)^{-1}$. Then

$$\gamma_{pdrs}^* \le \left(\prod_{i=1}^{k+1} \delta_i \right)^{-1} \gamma^*[\lambda_{max}(\pi(A)\pi(A^T))]^{\frac{1}{2}} \tag{5.26}$$

$$= \left(\prod_{i=1}^{k+1} \delta_i \right)^{-1} \gamma^* \left[\lambda_{max} \left[\left(\prod_{i=1}^{k+1} (A + \delta_i I) \prod_{i=1}^{k+1} (A^T + \delta_i I) \right) \right] \right]^{\frac{1}{2}} \tag{5.27}$$

$$= \left(\prod_{i=1}^{k+1}\delta_i\right)^{-1}\gamma^*\left[\lambda_{max}\left[\left(\prod_{i=1}^{k+1}\delta_i\right)^2\prod_{i=1}^{k+1}\left(\frac{A}{\delta_i}+I\right)\prod_{i=1}^{k+1}\left(\frac{A^T}{\delta_i}+I\right)\right]\right]^{\frac{1}{2}} \tag{5.28}$$

$$= \left(\prod_{i=1}^{k+1}\delta_i\right)^{-1}\gamma^*\left[\left(\prod_{i=1}^{k+1}\delta_i\right)^2\lambda_{max}\left[\prod_{i=1}^{k+1}\left(\frac{A}{\delta_i}+I\right)\prod_{i=1}^{k+1}\left(\frac{A^T}{\delta_i}+I\right)\right]\right]^{\frac{1}{2}}. \tag{5.29}$$

Thus, we conclude that

$$\gamma^* \le \gamma^*_{pdrs} \le \gamma^*\left[\lambda_{max}\left[\prod_{i=1}^{k+1}\left(\frac{A}{\delta_i}+I\right)\prod_{i=1}^{k+1}\left(\frac{A^T}{\delta_i}+I\right)\right]\right]^{\frac{1}{2}} \tag{5.30}$$

and this completes the proof. □

Noting that the matrix on the right-hand side of the theorem above tends to identity as $\delta_i \to \infty$, we conclude that Corollary 5.11 holds.

COROLLARY 5.11. *As* $\delta_i \to \infty$, $i = 1, \ldots, k+1$, $\gamma^*_{pdrs} \to \gamma^*$.

The same fact is also emphasized in Martinez, Malabre, and Kucera (1995), without proof. We now illustrate this by the following example.

Example. Let $N := Z_4 = (s-2)/(s+1)^2$ and $M := Z_2Z_3 = (s-1)/(s+1)^2$. Fix $k = 1, \pi = (s+1)^2$. Then $Q_1^* = 1$, which yields $P_1^* = -1$. Now consider the problem defined by (5.40) with $\sigma = s + a$:

$$\gamma^*_\pi = \min\left\|P_1^* - \frac{\pi M}{\sigma}\hat{Q}_1\right\|_\infty = \min\left\|1 + \frac{s-1}{s+a}\hat{Q}_1\right\|_\infty = 1. \tag{5.31}$$

If we choose $\hat{Q}_1 = -1$, then we have the following parametric expression:

$$\left\|P_1^* - \frac{\pi M}{\sigma}\hat{Q}_1\right\|_\infty = \left\|\frac{a+1}{s+a}\right\|_\infty \;\Rightarrow\; \| T_{zd} \|_\infty \le \left\|\frac{1}{(s+1)^2}\right\|_\infty \left\|\frac{a+1}{s+a}\right\|_\infty. \tag{5.32}$$

Thus, the limit, as a goes to infinity, of the H_∞ norm of T_{zd} becomes 1 (note that the optimal value for the H_∞ norm of T_{zd} is 1/4). In this way, the parametrization of σ as $s + a$ yields a suboptimal solution for the problem defined by (5.39), where at high frequencies attenuation is provided by the term $1/(s+1)^2$ and for the lower frequencies the parameter a could be used to minimize the response.

Now let us parametrize π as $\pi = (s+a)^2$. Also let $\sigma = s + a$ and choose $\hat{Q}_1 = d(s+c)/(s+a)$, where c and d are free parameters. Note that with the above choice, $P_1^* = ((s+a)/(s+1))^2$ which is independent of σ. Then

$$P_1^* - \frac{\pi M}{\sigma}\hat{Q}_1 = \frac{(s+a)^2}{(s+1)^2} - \frac{d(s+c)(s-1)}{(s+1)^2}. \tag{5.33}$$

Now if we choose d and c as $d := (a^2+2a-3)/4$ and $c := (3a^2-2a-1)/(a^2+2a-3)$, we obtain

$$P_1^* - \frac{\pi M}{\sigma}\hat{Q}_1 = \frac{(a+1)^2}{4} \;\Rightarrow\; \| T_{zd} \|_\infty \le \left\|\frac{(a+1)^2}{4(s+a)^2}\right\|_\infty \tag{5.34}$$

$$\Rightarrow \lim_{a\to\infty} \| T_{zd} \|_\infty \le \lim_{a\to\infty}\left\|\frac{(a+1)^2}{4(s+a)^2}\right\|_\infty = \frac{1}{4}. \tag{5.35}$$

Therefore, optimal attenuation is obtained in the limit. Using this second parametrization, a trade-off value for the parameter a can be chosen. If a is small, then a good attenuation at high frequencies is obtained, but there is sacrifice at the lower frequencies. If a is large, then the low frequency attenuation level is close to the optimal attenuation, but the response at high frequencies increases in magnitude.

6. Conclusions. In this work, partial disturbance rejection problems are investigated in a natural framework which consists of the subrings of $\mathbf{R}_{ps}(s)$ of a given relative degree and their quotient rings. First, the solvability conditions and the characterization of the set of all solutions are obtained for the scalar case. It is shown that these solvability conditions immediately imply the solvability condition in the case when dynamic measurement feedback is used. Furthermore, it is demonstrated by an example that partial disturbance rejection can be used together with H_∞ design to get stronger attenuation at high frequencies while minimizing the response (as much as possible) at lower frequencies. At this point, an upper bound for the infinity norm of the closed loop transfer function is also calculated.

REFERENCES

V. Eldem (1994), *On partial disturbance rejection by measurement feedback with internal stability*, in Proceedings 33rd IEEE CDC Conference, Orlando, FL, pp. 853–854.

V. Eldem and A. B. Özgüler (1988), *Disturbance decoupling problems by measurement feedback: A characterization of all solutions and fixed modes*, SIAM J. Control, 26, pp. 168–185.

E. Emre and L. H. Silverman (1980), *Partial model matching of linear systems*, IEEE Trans. Automat. Control, AC-25, pp. 280–281.

C. H. Foias, H. Özbay, and A. Tannenbaum (1996), *Robust Control of Infinite Dimensional Systems: Frequency Domain Methods*, Lecture Notes in Control and Inform. Sci., Springer-Verlag, New York.

B. A. Francis (1987), *A Course in H_∞ Control Theory*, Lecture Notes in Control and Inform. Sci., 88, Springer-Verlag, New York.

M. Malabre and J. C. Martinez Garcia (1993), *The partial model matching or the partial disturbance rejection problem: Geometric and structural solutions*, in Proceedings of the MTNS'93, Regensburg, Germany, pp. 341–342.

M. Marcus and H. Minc (1992), *A Survey of Matrix Theory and Matrix Inequalities*, Dover Publications, New York.

J. C. Martinez Garcia, M. Malabre, and V. Kucera (1995), *The partial model matching with stability*, Systems Control Lett., 24, pp. 61–74.

A. B. Özgüler (1994), *Linear Multi-Channel Control: A System Matrix Approach*, Prentice–Hall, Englewood Cliffs, NJ.

A. B. Özgüler and V. Eldem (1985), *Disturbance decoupling problems via dynamic output feedback*, IEEE Trans. Automat. Control, AC-30, pp. 756–764.

J. C. Willems and C. Commault (1981), *Disturbance decoupling by measurement feedback with internal stability or pole placement*, SIAM J. Control Optim., 19, pp. 490–504.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 1, pp. 193–212, January 1998

011

BIFURCATION AND NORMAL FORM OF NONLINEAR CONTROL SYSTEMS, PART I*

WEI KANG†

Abstract. The bifurcations of control systems with a single input are studied. Based on the normal forms of control systems, the equilibrium sets are classified. A set of quadratic invariants for control systems is found. Sufficient conditions for a system to be linearly controllable or stabilizable near a bifurcation point are given in terms of the quadratic invariants.

Key words. nonlinear systems, bifurcations, normal forms, invariants, linearly controllable, stabilizable

AMS subject classifications. 93C10, 93C15

PII. S0363012995290288

1. Introduction.

Bifurcation theory studies the changes in qualitative structure of the flow of a dynamic system as parameters are varied. Local bifurcation theory focuses on the stability of the bifurcating solution [6], [15]. In this paper, some bifurcation problems for control systems are addressed. Given a control system with parameters and control inputs, the location of the equilibrium points depends on the values of the parameters and control inputs. The set of equilibrium points is not necessarily a smooth manifold in the state and parameter space. Furthermore, the fundamental properties such as stabilizability and controllability change as the equilibrium point is varied. Understanding the change of these properties is important in feedback design. For instance, a bifurcation occurs in the system of axial flow compressor (see [13]). On one branch of the equilibria, the system is linearly controllable. On the other branch, the system is not stabilizable on one side of the bifurcation point. In fact, feedback is found to achieve the desired stability pattern on the controllable branch [12], [11]. The information about controllability and stabilizability along a set of equilibrium points is also helpful when gain scheduling methods are applied to a nonlinear system. In this paper, local bifurcations of equilibrium sets are classified based on the normal forms of control systems. The controllability and stabilizability at points in the equilibrium set depend on the values of the quadratic invariants. Recent research shows that the bifurcation in the Moore and Greitzer model of the axial flow compressor is equivalent to the two branch bifurcation given in Theorems 3.2 and 4.2.

The behavior of any Hopf bifurcation can be reduced to a few different cases. This is possible because a nonlinear dynamic system can be transformed into a simplified normal form based on Poincaré's theory. Since the bifurcation phenomenon is invariant under change of coordinates, one can study the bifurcation of dynamic systems by focusing on their normal forms. This idea simplifies the problem. An affine control system consists of two vector fields (the drift vector and the control vector). To study the bifurcations of control systems, it is necessary to find normal forms in which both vectors are simplified. For linear systems, a normal form is the controller form. In

*Received by the editors August 14, 1995; accepted for publication (in revised form) October 31, 1996. Research supported in part by AFOSR-95-1-0169.
http://www.siam.org/journals/sicon/36-1/29028.html
†Department of Mathematics, Naval Postgraduate School, Monterey, CA 93943 (wkang@math.nps.navy.mil).

this paper, a set of nonlinear control system normal forms is found. All the results on equilibrium sets, controllability, and stabilizability are proved based on these normal forms and their invariants. The normal forms in this paper generalized the work in [8] to parameter-dependent systems which are not linearly controllable. Since the control system normal forms have the Brunovsky form in their linearization and the triangle structure in the quadratic parts, the study of their controllability and stabilizability is simple.

In general, a control system can have more than one equilibrium point even without parameters. This is because of the existence of control inputs. In the presence of parameters, the bifurcation of a control system has at least two-dimensional freedom. This paper is organized in the following way. In Part I, we focus on control systems without parameters. For single input systems, this is a one-dimensional bifurcation problem. In Part II [16], the problem is addressed for control systems with one parameter, which is a two-dimensional bifurcation problem. In both parts, we only consider systems with a single input, although the formulation of the problems are given for general multi-input systems.

In section 1 of Part I, the problem is formulated from bifurcation viewpoint and then an intuitive example is given. In section 2, the quadratic normal forms and invariants of control systems are introduced; these play a key role in the proofs of the main theorems. The problems formulated in section 1.1 are addressed in sections 3–5 for nonlinear systems without parameters. Sufficient conditions in terms of quadratic invariants for controllability of linearization and local stabilizability are proved. In [16], the problems formulated in this section will be addressed for systems with a single parameter.

Some interesting problems related to control system in the presence of bifurcation are addressed in [1], [2], and [5].

1.1. Problem formulation. Classic bifurcation theory studies the changes in qualitative properties such as stability of a dynamic system about bifurcating constant solutions as parameters are varied. More specifically, a system with parameters is defined by

$$\dot{x} = f_\mu(x) \tag{1.1}$$

where $x \in \mathbb{R}^n$ is the state variable and μ is the parameter. For different values of μ, the behavior of the dynamic flows can be qualitatively different. For instance, the equilibrium point x_0 defined by $f_\mu(x_0) = 0$ depends on the value of μ. Furthermore, the stability of the system around x_0 can be different if the value of μ is changed.

Consider a control system

$$\dot{x} = f(x, \mu) + g(x, \mu)u \tag{1.2}$$

where $x \in \mathbb{R}^n$ is the state variable, $u \in \mathbb{R}^m$ is the control input, and μ is the parameter. The performance of the system depends on the values of μ and u. For instance, the equilibrium point x_0 of the system defined by

$$f(x_0, \mu) + g(x_0, \mu)u = 0$$

changes if the values of μ and u are changed. Furthermore, more than one branch of equilibrium points can occur. The controllability of the system at these equilibrium points also changes. In this paper, we classify the bifurcations of equilibrium sets. The change of properties such as controllability and stabilizability is also studied.

Given a system (1.2), assume that

$$f(0,0)=0,$$

and we only consider local bifurcation around $(x,u,\mu)=(0,0,0)$. Assume the rank of the matrix $g(0,0)$ is m. In a local neighborhood of $(x,u,\mu)=(0,0,0)$, if

$$f(x_0,\mu_0)+g(x_0,\mu_0)u_0=0, \tag{1.3}$$

then u_0 is the unique value of u for which the vector field in (1.2) vanishes at (x_0,u,μ_0).

DEFINITION 1.1. *The set*

$$E=\{(x,\mu)|\exists u_0\in\mathbb{R}\ such\ that f(x,\mu)+g(x,\mu)u_0=0\} \tag{1.4}$$

is called the equilibrium set of (1.2).

In this definition of the equilibrium set, the parameter μ is treated as a variable satisfying $\dot{\mu}=0$. A point in E is called an *equilibrium point* of system (1.2). It is known that feedback of the form $u=u(x)$ can change the closed-loop system equilibria. The set E consists of all the possible closed-loop equilibria under state feedbacks. Understanding the topology of E is fundamental in the study of control of stationary bifurcations by state feedback. A special case of equilibrium set is given by systems without parameters. Consider a system

$$\dot{x}=f(x)+g(x)u \tag{1.5}$$

which is independent of the parameter μ. The equilibrium point $x=0$, $u=0$ is not unique even if the matrix

$$\frac{\partial f}{\partial x}(0)$$

has full rank. This is caused by the presence of input variable u.

DEFINITION 1.2. *The equilibrium set of* (1.5) *is defined by*

$$E=\{x|\exists u_0\in\mathbb{R}\ such\ that f(x)+g(x)u_0=0\}.$$

Dynamic bifurcation theory is always connected with the problem of stability, in particular, the stability of the bifurcating solution. For control systems, it makes more sense to study the controllability and stabilizability of control systems around the equilibrium points in E. The general concept of controllability of nonlinear systems is not addressed in this paper. We focus on the property of controllability of the linearization. Given a point $(x_0,\mu_0)\in E$. Suppose $u=u_0$ is the unique value of u satisfying (1.3). The *linearization* of (1.2) at (x_0,μ_0) is defined to be the pair $(A_{x_0\mu_0},B_{x_0\mu_0})$ in which

$$A_{x_0\mu_0}=\left.\frac{\partial f(x,\mu)+g(x,\mu)u_0}{\partial x}\right|_{\substack{x=x_0\\ \mu=\mu_0}},\ B_{x_0\mu_0}=g(x_0,\mu_0).$$

DEFINITION 1.3. *Given a point* (x_0,μ_0) *in* E*, the control system* (1.2) *is called linearly controllable at* (x_0,μ_0) *if its linearization* $(A_{x_0\mu_0},B_{x_0\mu_0})$ *defines a controllable linear system.*

In this paper, the term "controllability" is used for controllability of the linearization. In the following, the problems addressed in this paper are formulated from the bifurcation viewpoint.

Question 1. Find a classification of equilibrium sets. This is similar to the problem of finding all the "bifurcation diagrams" in the bifurcation theory of ODEs.

Question 2. Is the system linearly controllable at the equilibrium points in E?

Question 3. Is the system locally stabilizable by state feedback at the equilibrium points in E?

Remark. In this paper, Questions 1–3 are addressed only for systems which are not linearly controllable at the origin. In fact, if a system is linearly controllable at $(x_0, \mu_0) = (0, 0)$, then the system is always linearly controllable at all points in E near $(x_0, \mu_0) = (0, 0)$. So, the answers to Question 2 and 3 are trivial. For linearly controllable systems, the solution of Question 1 is simple. For instance, if a system has a single input and a single parameter, from the Brunovsky form of its linearization and the implicit function theorem it can be proved that the equilibrium set E is (locally) a smooth manifold of dimension two. Any small values of x_1 and μ uniquely determine a point in E. Therefore, all the equilibrium sets of such systems are diffeomorphic to each other.

Questions 2 and 3 are closely related in the sense that a linearly controllable system must be locally stabilizable by state feedback. As mentioned above, a control system usually has more than one equilibrium point. Questions 1–3 are applicable to a control system even if the system is independent of any parameter. In Part I, we address these problems for systems without parameters. In Part II, systems with a single parameter are considered.

1.2. An example of control system with bifurcation. In the following, an example is given for which the equilibrium set is found. The answer to Question 2 is given. The stabilizability at the origin is also proved. This example is a bifurcation discussed in section 3. In fact, this system is in normal form.

Example. Consider a two-dimensional control system without parameter

$$\begin{aligned} \dot{z} &= zx + z^2, \\ \dot{x} &= u. \end{aligned} \tag{1.6}$$

The origin $(z, x) = (0, 0)$ is an equilibrium point of the system; however, it is certainly not the only one. The system is not linearly controllable at the origin since the linearization (A, B) is

$$A = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

However, at the equilibrium points nearby, the system can be linearly controllable. First of all, where are the equilibrium points? They are determined by $zx + z^2 = 0$ and they have the following parametrization

$$E = \{(z, x) | x = \nu \text{ and either } z = -\nu \text{ or } z = 0\}.$$

The graph of E is shown in Figure 1.1. From Figure 1.1, it is obvious that a bifurcation occurs at the origin. The equilibrium set E has two branches, which intersect at $(z, x) = (0, 0)$. In the following, the notation E_- and E_+ are used to represent the subsets of E for $z = -\nu$ and $z = 0$, respectively.

To answer the second question, it is necessary to find the linearization and its controllability matrix at any point in E. In fact, the controllability matrix is

$$\begin{bmatrix} 0 & z \\ 1 & 0 \end{bmatrix}.$$

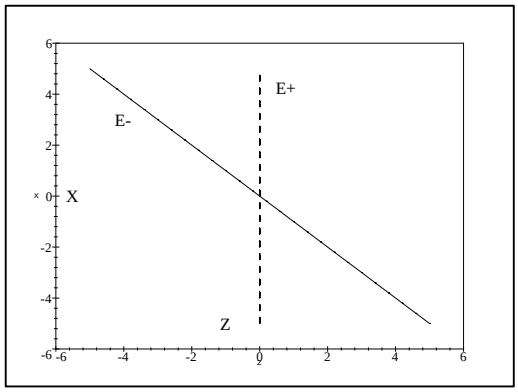


FIG. 1.1. *The equilibrium set in zx_1-plane. On the solid line, the system is linearly controllable. On the dotted line, the system is not linearly controllable.*

Therefore, the system is linearly controllable at points in the branch $E_- \backslash \{(0,0)\}$, and the system is not linearly controllable at points in E_+.

At points in E_-, the system can be stabilized locally by state feedback because it is linearly controllable. However, at a point in E_+, the stabilizability depends on the nonlinear part of the system. For instance, the system is stabilizable at the origin. One stabilizing feedback is

$$u = -x - z - z^2.$$

To show that the closed-loop system is asymptotically stable, one can check that the reduced dynamic system on the center manifold is

$$\dot{z} = -z^3 + O(z)^4.$$

This implies that the closed-loop system is asymptotically stable at $(z, x) = (0, 0)$.

2. Normal forms and invariants. The normal forms of control systems are introduced in this section. They are a tool in the proofs of the main theorems on bifurcation of control systems. Normal forms for linearly controllable systems have been introduced in [8] and [10]. In [9], normal forms for control systems which are not linearly controllable were introduced. The techniques and results in these papers generalized Poincaré's normal form of ODE to control systems. In this section, we also introduce a set of quadratic invariants. The advantage of introducing invariants is that the normal form of a given system can be found without finding the change of coordinates and feedback. Furthermore, these invariants provide information about the parametrization of equilibrium set E and the properties such as stabilizability or controllability of the system. In fact, most conditions in the main theorems of this paper are given in terms of the quadratic invariants.

2.1. Assumptions. In Part I of this paper we consider only control systems without a parameter. A control system is defined by

$$\dot{\xi} = f(\xi) + g(\xi)v, \tag{2.1}$$

where the variable $\xi \in \mathbb{R}^n$ is the state of the system. Assume that $f(0) = 0$ and $g(0) \neq 0$. Also assume that $v \in \mathbb{R}$; i.e., the system has a single input. All the vectors

and functions in this paper are assumed to be C^k for some sufficiently large k. As pointed out in the remark in section 1.1, we assume that the linearization of system (2.1) at $x = 0$, $u = 0$,

$$(A, B) = \left(\frac{\partial f}{\partial \xi}(0), g(0)\right)$$

is not controllable. Furthermore, the controllability index of (A, B) is assumed to be $n - 1$. Equivalently,

Assumption. We have that

$$\operatorname{rank}(\begin{bmatrix} B & AB & A^2B & \cdots & A^{n-1}B \end{bmatrix}) = n - 1. \tag{2.2}$$

The origin is in the equilibrium set. Near the origin, there is a unique value u_0 satisfying $f(\xi) + g(\xi)u_0 = 0$ for any $\xi \in E$ because $g(0) \neq 0$. So, given a point in E, the linearization of the system at this point is unique. The transformations used in this section are change of coordinates and feedback in the form

$$\begin{aligned} x &= \phi(\xi), \\ u &= \alpha(\xi) + \beta(\xi)v, \end{aligned} \tag{2.3}$$

in which $\phi(\xi)$ is a diffeomorphism near the origin $\xi = 0$ and $\beta(0) \neq 0$. Before we introduce the normal forms and invariants, it is necessary to make sure that Questions 1–3 are well proposed under the transformations of form (2.3). In fact, it is well known that changes of coordinates and state feedbacks do not change the controllability (of the linearization) and the local stabilizability of a control system [14]. So if one system is transformed into another by (2.3), the answers to Question 2 and 3 for these two systems are the same. If (2.3) is considered as a map from (ξ, v) to (x, u), it is a local diffeomorphism. Therefore, in a local neighborhood of $\xi = 0$, $\xi_0 \in E$ if and only if $x_0 = \phi(\xi_0)$ is an equilibrium point for the resulting system. Here, the equilibrium set is invariant under the transformations (2.3). So, if a class of nonlinear control systems can be simplified into a normal form, the bifurcation problems for this class of systems are equivalent to the same problems for a system in the normal form.

2.2. Normal forms. Given a system (2.1) satisfying assumption (2.2), it is well known (see, for instance, [7], [9]) that the system can be transformed into the following form by a linear change of coordinates and feedback:

$$\begin{aligned} \dot{z} &= \lambda z + f_1^{[2]}(z, x) + g_1^{[1]}(z, x)u + O(x, z, u)^3, \\ \dot{x} &= A_2 x + B_2 u + f_2^{[2]}(z, x) + g_2^{[1]}(z, x)u + O(x, z, u)^3, \end{aligned} \tag{2.4}$$

where $z \in \mathbb{R}$ and $x \in \mathbb{R}^{n-1}$. The pair (A_2, B_2) is in the following (Brunovsky) form:

$$A_2 = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}_{(n-1)\times(n-1)}, \qquad B_2 = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}_{(n-1)\times 1}. \tag{2.5}$$

The superscripts of $f_i^{[2]}$ and $g_i^{[1]}$, $i = 1$ or 2, denote that $f_i^{[2]}$ and $g_i^{[1]}$ are homogeneous polynomials of second and first degree, respectively. Similar superscripts will also be

applied to other vector fields and functions (e.g., $\alpha^{[2]}$ or $\beta^{[1]}$). The notation $O(z,x,u)^3$ represents nonlinear terms of third and higher degrees. In (2.4) the linearization is already in its normal form; the next step of finding normal form is to simplify the quadratic part of (2.4) while leaving its linear part invariant. Following the idea in [8] and [9], we use the quadratic transformation (or quadratic change of coordinates and feedback)

$$\begin{aligned} \begin{bmatrix} z \\ x \end{bmatrix} &= \begin{bmatrix} \bar{z} \\ \bar{x} \end{bmatrix} + \phi^{[2]}(\bar{z},\bar{x}), \\ u &= \bar{u} + \alpha^{[2]}(\bar{z},\bar{x}) + \beta^{[1]}(\bar{z},\bar{x})\bar{u} \end{aligned} \tag{2.6}$$

to simplify the quadratic part. In [9] it is proved that (2.4) can be transformed into the following normal form. The coefficients in its normal form are uniquely determined by the quadratic part of (2.4). For the reason of simplicity, we still use z, x, and u as state and control variables.

If $\lambda \neq 0$,

$$\begin{aligned} \dot{z} &= \lambda z + \sum_{i=1}^{n-1} \gamma_{x_i x_i} x_i^2 + \gamma_{z x_1} z x_1 + O(x,z,u)^3, \\ \dot{x} &= A_2 x + B_2 u + \tilde{f}_2^{[2]}(x) + O(x,z,u)^3; \end{aligned} \tag{2.7}$$

if $\lambda = 0$,

$$\begin{aligned} \dot{z} &= \sum_{i=1}^{n-1} \gamma_{x_i x_i} x_i^2 + \gamma_{z x_1} z x_1 + \gamma_{zz} z^2 + O(x,z,u)^3, \\ \dot{x} &= A_2 x + B_2 u + \tilde{f}_2^{[2]}(x) + O(x,z,u)^3. \end{aligned} \tag{2.8}$$

In (2.7) and (2.8), the vector $\tilde{f}_2^{[2]}(x)$ is in the extended quadratic controller form introduced in [8]

$$\begin{aligned} &\tilde{f}_2^{[2]}(x) = [\tilde{f}_{21}(x), \tilde{f}_{22}(x), \ldots, \tilde{f}_{2\,n-1}(x)]^T, \quad \tilde{f}_{2i}(x) = \sum_{j=i+2}^{n-1} a_{i\,n-j} x_j^2 \ \text{ for } \ 1 \le j \le n-3, \\ &\tilde{f}_{2\,n-2}(x) = \tilde{f}_{2\,n-1}(x) = 0. \end{aligned} \tag{2.9}$$

The symbols $\gamma_{x_i x_i}$, $\gamma_{z x_1}$ and γ_{zz} denote the constant coefficients of the quadratic terms. These normal forms will be used as a tool in the proofs of the theorems in sections 3, 4, and 5.

Given a system, formulas for finding the coefficients in its normal form are given in Definition 2.1. In fact, they are a complete set of invariants. A change of coordinates transforming a system to its normal form can be found by solving a set of linear algebraic equations, which are called homological equations [9].

2.3. Invariants. In the following, we introduce quadratic invariants. They are the "intrinsic parameters" of a system which completely determine the equivalent class of quadratic parts of a system under quadratic transformations of the form (2.6). Conditions in many theorems of this paper will be given in terms of these invariants.

Although the invariants are defined at an equilibrium point (we define them at $(z,x)=(0,0)$), they carry important information of controllability and stabilizability of a system at all the equilibrium points near the origin.

Denote by C_x, C_z, and X_z the following n-dimensional row or column vectors:

$$\begin{aligned} C_x &= [0 \quad 1 \quad 0 \quad \cdots \quad 0], \qquad C_z = [1 \quad 0 \quad 0 \quad \cdots \quad 0], \\ X_z &= [1 \quad 0 \quad 0 \quad \cdots \quad 0]^T. \end{aligned} \tag{2.10}$$

Given two vector fields $X(x)$ and $Y(x)$ defined in $\mathbb{R}^n$, the operator ad_X is defined by $ad_X(Y)=[X,Y]$, where the right-hand side is the Lie bracket of two vector fields which is defined by

$$\frac{\partial Y}{\partial x}X - \frac{\partial X}{\partial x}Y.$$

The Lie operator L_X is defined by

$$L_X(\phi(x)) = \frac{\partial \phi}{\partial x}X$$

for C^1 functions defined in $\mathbb{R}^n$. In Definition 2.1, we use the notation $f(z,x)+g(z,x)u$ to represent the right side of a system (2.4). The notation A represents the matrix in the linearization of $f(z,x)$, i.e.,

$$A = \left.\frac{\partial f}{\partial (z,x)}\right|_{\substack{z=0\\x=0}} = \begin{bmatrix} \lambda & 0 \\ 0 & A_2 \end{bmatrix}.$$

DEFINITION 2.1. *Given a system* (2.4)*, the* quadratic invariants *are defined by*

$$\begin{aligned} a_{tr} &= \frac{1}{2}C_x A^{t-1}[ad_f^r(g), ad_f^{r-1}(g)]\Big|_{z=0,x=0}, \qquad \begin{matrix} 1 \le r \le n-3, \\ 1 \le t \le n-r-2, \end{matrix} \\ \gamma_{x_{n-r}x_{n-r}} &= \frac{1}{2}C_z[ad_f^r(g), ad_f^{r-1}(g)]\Big|_{z=0,x=0}, \quad 1 \le r \le n-1, \\ \gamma_{zx_1} &= (-1)^{n-1}C_z[X_z, ad_f^{n-1}(g)]\Big|_{z=0,x=0}, \\ \gamma_{zz} &= \frac{1}{2}C_z ad_{X_z}^2(f)\Big|_{z=0,x=0}. \end{aligned} \tag{2.11}$$

THEOREM 2.2. *Given a control system satisfying assumption* (2.2)*, assume that its linearization is in the form of* (2.4)*.*

(i) *The quadratic transformation* (2.6) *does not change the values of the quadratic invariants.*

(ii) *The quadratic invariants* (2.11) *are equal to the coefficients of the quadratic terms of the normal form* (2.7), (2.8)*.*

(iii) *Given two systems in the form of* (2.4) *with the same linearization (i.e., they have the same* λ*), the quadratic part of one system can be transformed into that of another system by a suitable transformation* (2.6) *if and only if they have the same quadratic invariants.*

Proof. (i) The proof of (i) has two parts. In the first part, we only consider changes of coordinates without feedback, i.e., $u = \bar{u}$ in the quadratic transformation (2.6). In the second part, we prove that the invariants cannot be changed by any quadratic feedback. Suppose that system (2.4) is transformed into the following system by a quadratic change of coordinates:

$$\begin{aligned} \dot{\bar{z}} &= \lambda\bar{z} + \bar{f}_1^{[2]}(\bar{z},\bar{x}) + \bar{g}_1^{[1]}(\bar{z},\bar{x})\bar{u} + O(\bar{x},\bar{z},\bar{u})^3, \\ \dot{\bar{x}} &= A_2\bar{x} + B_2\bar{u} + \bar{f}_2^{[2]}(\bar{z},\bar{x}) + \bar{g}_2^{[1]}(\bar{z},\bar{x})\bar{u} + O(\bar{x},\bar{z},\bar{u})^3. \end{aligned} \tag{2.12}$$

Denote the invariants of (2.4) and (2.12) by a_{tr}, $\gamma_{x_ix_i}$, γ_{zx_1}, γ_{zz} and $\bar{a}_{tr}$, $\bar{\gamma}_{x_ix_i}$, $\bar{\gamma}_{zx_1}$, $\bar{\gamma}_{zz}$, respectively. Notice that if we treat X_z, $f(z,x)$ and $g(z,x)$ as vector fields in $\mathbb{R}^n$, then f and $\bar{f}$ represent the same vector field. Similarly, g and $\bar{g}$ represent the same vector field. Since Lie bracket and Lie operators are independent of the choice of coordinate systems, sometimes we use f and g to represent these two vector fields without mentioning the coordinate system (z,x or $\bar{z},\bar{x}$). The vectors X_z and $\overline{X}_z$ are defined based on coordinate systems (see (2.10)). The invariants can be expressed in the following way using Lie bracket and Lie operators:

$$\begin{aligned} a_{tr} &= \frac{1}{2}\, L_{[ad_f^r(g),ad_f^{r-1}(g)]}L_f^{t-1}(x_1)\Big|_{\substack{z=0\\x=0}}, \\ \gamma_{x_{n-r}x_{n-r}} &= \frac{1}{2}\, L_{[ad_f^r(g),ad_f^{r-1}(g)]}(z)\Big|_{\substack{z=0\\x=0}}, \\ \gamma_{zx_1} &= (-1)^{n-1}L_{[X_z,ad_f^{n-1}(g)]}(z). \end{aligned} \tag{2.13}$$

Under the new coordinates, we have

$$x_1 = \bar{x}_1 + O(\bar{z},\bar{x})^2, \quad z = \bar{z} + O(\bar{z},\bar{x})^2, \quad X_z = \overline{X}_z + O(\bar{z},\bar{x}). \tag{2.14}$$

From (2.13) and (2.14),

$$a_{tr} = \frac{1}{2}\, L_{[ad_f^r(g),ad_f^{r-1}(g)]}L_f^{t-1}(\bar{x}_1)\Big|_{\substack{z=0\\x=0}} + \frac{1}{2}\, L_{[ad_f^r(g),ad_f^{r-1}(g)]}L_f^{t-1}(O(\bar{z},\bar{x})^2)\Big|_{\substack{z=0\\x=0}}.$$

In this relation, the second term on the right side is zero. The first term on the right side is $\bar{a}_{tr}$. This proves that $a_{tr} = \bar{a}_{tr}$. Similarly, we can prove that $\gamma_{x_ix_i} = \bar{\gamma}_{x_ix_i}$.

Now, let's consider γ_{zx_1}. By (2.13) and (2.14), we have

$$\begin{aligned} \gamma_{zx_1} &= L_{[\overline{X}_z,ad_f^{n-1}(g)]}(\bar{z})\Big|_{\substack{\bar{z}=0\\\bar{x}=0}} + L_{[\overline{X}_z,ad_f^{n-1}(g)]}(O(\bar{z},\bar{x})^2)\Big|_{\substack{\bar{z}=0\\\bar{x}=0}} \\ &+ L_{[O(\bar{z},\bar{x}),ad_f^{n-1}(g)]}(\bar{z}+O(\bar{z},\bar{x})^2)\Big|_{\substack{\bar{z}=0\\\bar{x}=0}}. \end{aligned} \tag{2.15}$$

From (2.13), we know that

$$L_{[\overline{X}_z,ad_f^{n-1}(g)]}(\bar{z})\Big|_{\substack{\bar{z}=0\\\bar{x}=0}} = \bar{\gamma}_{zx_1}. \tag{2.16}$$

It is easy to check that

$$L_{[\overline{X}_z,ad_f^{n-1}(g)]}(O(\bar{z},\bar{x})^2)\Big|_{\substack{\bar{z}=0\\\bar{x}=0}} = 0 \tag{2.17}$$

and

$$ad_f^r(g) = (-1)^r \begin{bmatrix} \lambda^r & 0 \\ 0 & A_2^r \end{bmatrix} \begin{bmatrix} 0 \\ B_2 \end{bmatrix} + O(\bar{z}, \bar{x}).$$

Therefore,

$$ad_f^{n-1}(g) = O(\bar{z}, \bar{x}).$$

So

$$L_{[O(\bar{z},\bar{x}),ad_f^{n-1}(g)]}(\bar{z} + O(\bar{z}, \bar{x})^2)\Big|_{\substack{\bar{z}=0 \\ \bar{x}=0}} = 0. \tag{2.18}$$

Equations (2.16), (2.17), and (2.18) imply $\gamma_{zx_1} = \bar{\gamma}_{zx_1}$. If $\lambda = 0$, there is another invariant γ_{zz}. By the separation principle in [9, Lemma 4.2], it is enough to show that γ_{zz} is invariant under the change of coordinates $z = z + \phi^{[2]}(z)$. However, this is a well-known result in dynamic systems. In fact, z^2 is a resonant term in the dynamic system of z (the definition can be found in [3]). In Poincaré's theory of normal forms for dynamic systems, the coefficient of a quadratic resonant term cannot be changed by quadratic change of coordinates. This shows that γ_{zz} is invariant under a quadratic change of coordinates.

If feedback is applied to system (2.4), then the new vector fields in the resulting system are

$$\bar{f}(z, x) = f(z, x) + \alpha^{[2]}(z, x)g(z, x), \quad \bar{g}(z, x) = g(z, x) + \beta^{[1]}(z, x)g(z, x).$$

It is obvious that γ_{zz} will not be changed by the feedback. By mathematical induction, it can be proved that

$$ad_{\bar{f}}^r(\bar{g}) = ad_f^r(g) + \sum_{i=0}^{r} h_{ri}(z, x) ad_f^i(g) + O(z, x)^2, \quad h_{ri}(z, x) = O(z, x),$$

$$\left[ad_{\bar{f}}^r(\bar{g}), ad_{\bar{f}}^{r-1}(\bar{g})\right] = [ad_f^r(g), ad_f^{r-1}(g)] + \sum_{i=1}^{r} q_{ri}(z, x) ad_f^i(g) + O(z, x),$$

$$C_x A^{t-1} ad_f^i(g)|_{z=0,x=0} = 0 \quad \text{if } t + i \le n - 2,$$

$$C_z ad_f^i(g)|_{z=0,x=0} = 0 \qquad \text{if } 0 \le i \le n - 1.$$

Substituting these relations into the definition of invariants (2.11), it shows that $\bar{a}^{tr}$, $\bar{\gamma}_{x_i x_i}$, $\bar{\gamma}_{zx_1}$ are equal to a_{tr}, $\gamma_{x_i x_i}$, γ_{zx_1}.

(ii) The proof of the second part is based on calculation. By mathematical induction, it can be proved that, if $1 \le r < n - 2$,

$$ad_f^r(g) = (-1)^r \left(\left\{ \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \right\} = n - r \quad + \left\{ \begin{bmatrix} 2\gamma_{x_{n-r}x_{n-r}} x_{n-r} \\ 2a_{1\,r} x_{n-r} \\ \vdots \\ 2a_{n-r-2\,r} x_{n-r} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \right\} = n - r - 1 \right)$$

$$+ h_r(x_{n-r+1}, x_{n-r+2}, \ldots, x_{n-1}) + O(x, z)^2. \tag{2.19}$$

Therefore,

$$[ad_f^r(g), ad_f^{r-1}(g)] = \begin{bmatrix} 2\gamma_{x_{n-r}x_{n-r}} & 2a_{1\,r} & \cdots & 2a_{n-r-2\,r} & 0 & \cdots & 0 \end{bmatrix}^T + O(z,x) \tag{2.20}$$

for $r < n-2$. Equations (2.20) and (2.11) imply

$$a_{tr} = \frac{1}{2}\, C_x A^{t-1} [ad_f^r(g), ad_f^{r-1}(g)]\Big|_{z=0,x=0},$$
$$\gamma_{x_{n-r}x_{n-r}} = \frac{1}{2}\, C_z [ad_f^r(g), ad_f^{r-1}(g)]\Big|_{z=0,x=0}$$

for $1 \le r \le n-3$ and $1 \le t \le n-r-2$. Similarly, we can show that

$$ad_f^{n-2}(g) = (-1)^{n-2} \left(\begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \begin{bmatrix} 2\gamma_{x_2x_2}x_2 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \right) + h_{n-2}(x_3, x_4, \ldots, x_{n-1}) + O(x,z)^2,$$

$$ad_f^{n-1}(g) = (-1)^{n-1} \begin{bmatrix} 2\gamma_{x_1x_1}x_1 + \gamma_{zx_1}z \\ 0 \\ \vdots \\ 0 \end{bmatrix} + h_{n-1}(x_2, x_3, \ldots, x_{n-1}) + O(x,z)^2.$$

So, it is easy to check that

$$\gamma_{x_2x_2} = \frac{1}{2} C_z [ad_f^{n-2}(g), ad_f^{n-3}(g)]\Big|_{\substack{z=0\\x=0}}, \quad \gamma_{x_1x_1} = \frac{1}{2} C_z [ad_f^{n-1}(g), ad_f^{n-2}(g)]\Big|_{\substack{z=0\\x=0}},$$

$$\gamma_{zx_1} = (-1)^{n-1} C_z [X_z, ad_f^{n-1}(g)]\Big|_{\substack{z=0\\x=0}}.$$

If $\lambda = 0$, then γ_{zz} is the coefficient of z^2 because of Definition 2.1.

(iii) From the result in [9], we know that a system (2.4) can be transformed into a normal form (2.7) or (2.8). Therefore, two systems can be transformed from one to the other if and only if they have the same normal form. From (ii), the coefficients in normal form and the invariants have one-to-one correspondence. So, the two systems have the same normal form if and only if they have the same invariants. This concludes the proof of the theorem. □

Theorem 2.2 implies that the coefficients of the normal form can be computed without finding the transformation. Furthermore, it will be shown in the next three sections that the properties such as controllability and stabilizability are closely related to these invariants.

3. Classification of equilibrium sets. In this section, Question 1 is addressed. Different systems have different equilibrium sets. However, the equilibrium sets of systems with the same normal form are diffeomorphic to one other. Based on the normal forms in section 2, the equilibrium sets of systems satisfying (2.2) are classified to three different classes. For each class of equilibrium sets, a parametrization of the equilibrium set is also found, which is a linear approximation of E.

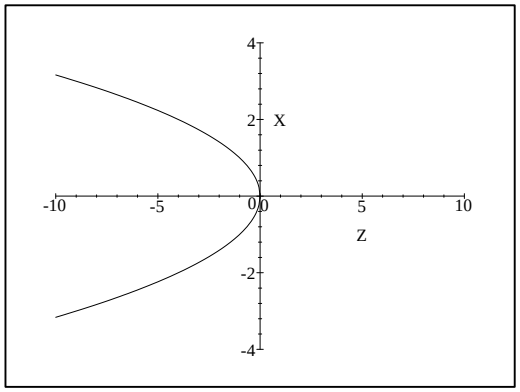


FIG. 3.1. *The equilibrium set in zx_1-plane.*

Given a nonlinear control system, if its linearization has controllability index $n-1$, then it can be transformed into a system in the form of (2.4). So, we only consider system (2.4) in sections 3, 4, and 5.

THEOREM 3.1. *Given a system of the form* (2.4)*, if $\lambda \neq 0$, then there exists an open neighborhood U of $(z, x) = (0, 0)$ such that the points in $E \cap U$ satisfy*

$$\begin{aligned} &x_1 = \nu, \\ &z = O(\nu)^2, \\ &x_i = O(\nu)^2 \ \text{ for } \ i = 2, 3, \ldots, n-1. \end{aligned} \tag{3.1}$$

Remark. This theorem shows that, in a neighborhood of the origin, there exists a unique equilibrium point of the system for a given value of x_1. The set E is a smooth curve tangent to the x_1-axis at the origin. Therefore, the equilibrium set does not show an obvious bifurcation. However, in Theorem 4.1 it is proved that the controllability of the system changes near the origin for different equilibrium points. A typical graph of the equilibrium set for the systems with $\lambda \neq 0$ is shown in Figure 3.1 for the following system in normal form:

$$\begin{aligned} &\dot{z} = z + x_1^2 + x_2^2, \\ &\dot{x}_1 = x_2, \\ &\dot{x}_2 = u. \end{aligned} \tag{3.2}$$

The equilibrium set E is $x_1 = \nu$, $x_2 = 0$, $z = -\nu^2$.

Proof of Theorem 3.1. For a point (z, x) to be an equilibrium point of system (2.4), it must satisfy

$$\begin{aligned} &\lambda z + f_1^{[2]} + g_1^{[1]} u + O(z, x, u)^3 = 0, \\ &A_2 x + B_2 u + f_2^{[2]} + g_2^{[1]} u + O(z, x, u)^3 = 0. \end{aligned} \tag{3.3}$$

Denote the left side of the equations by $G(z, x, u)$. This is a system of equations and

$$\left.\frac{\partial G}{\partial(z, x_2, ..., x_{n-1}, u)}\right|_{z=0, x=0, u=0} = \begin{bmatrix} \lambda & 0 \\ 0 & I \end{bmatrix}$$

where I is the $(n-1) \times (n-1)$ identity matrix. By the implicit function theorem, in a local neighborhood of the origin there exists a unique set of functions

$z(x_1), x_2(x_1), \ldots, x_{n-1}(x_1), u(x_1)$ satisfying equations (3.3). Since $\partial G/\partial x_1$ is the zero matrix at the origin, these functions do not contain linear terms in x_1. This proves the theorem. □

The topology of the equilibrium sets for systems with $\lambda = 0$ depends on the quadratic part of its normal form. The quadratic function of z, x_1 of the uncontrollable dynamics in the normal form has an associated symmetric matrix, which is

$$Q = \begin{bmatrix} \gamma_{zz} & \frac{1}{2}\gamma_{zx_1} \\ \frac{1}{2}\gamma_{zx_1} & \gamma_{x_1x_1} \end{bmatrix}.$$

Denote d_1 and d_2 the eigenvalues of this matrix. Then there is an orthonormal matrix T, the column vectors of which are unit eigenvectors of Q, such that

$$Q = T \begin{bmatrix} d_1 & 0 \\ 0 & d_2 \end{bmatrix} T^T. \tag{3.4}$$

THEOREM 3.2. *Given a system* (2.4) *with* $\lambda = 0$*, the following hold.*
(i) *If*

$$\det(Q) > 0, \tag{3.5}$$

then there is no equilibrium point other than $(z, x) = (0, 0)$ *near the origin.*
(ii) *If*

$$\det(Q) < 0, \tag{3.6}$$

then the equilibrium set has the following parametrization

$$x_i = O(\nu)^2, \qquad for \;\; i = 2, \ldots, n-1,$$

$$\begin{bmatrix} z \\ x \end{bmatrix} = T \begin{bmatrix} 1 \\ \pm\sqrt{-\dfrac{d_1}{d_2}} \end{bmatrix} \nu + O(\nu)^2 \tag{3.7}$$

in an open neighborhood of the origin.

Remark. The relation (3.7) implies that if (3.6) holds, the equilibrium set has two branches. At the origin, the two branches have tangent vectors

$$T \begin{bmatrix} 1 & \pm\sqrt{-\dfrac{d_1}{d_2}} \end{bmatrix}^T. \tag{3.8}$$

A typical example of such equilibrium set is given by the system (1.6) in section 1.2 which satisfies (3.6). The system is in normal form. The bifurcation diagram is shown in Figure 1.1.

Proof of Theorem 3.2. A system (2.4) can be simplified to (2.8). A transformation (2.6) does not change the linear part of the functions in (3.7). Therefore, it is sufficient to prove the theorem for the normal form (2.8). In this case, the equilibrium point (z, x) satisfies

$$x_2 + O(z, x, u)^2 = 0, \ldots, x_{n-1} + O(z, x, u)^2 = 0, \quad u + O(z, x, u)^2 = 0.$$

TABLE 3.1
The classification of equilibrium sets.

Condition	Equilibrium set	Example
$\lambda \neq 0$	Smooth 1-d manifold tangent to x_1-axis	Figure 3.1
$\lambda = 0$, $\det(Q) > 0$	Single point	
$\lambda = 0$, $\det(Q) < 0$	Two smooth curves tangent to vectors (3.8) at origin	Figure 1.1

This implies that $x_i = O(x_1, z)^2$, for $i = 2, \ldots, n-1$, and $u = O(x_1, z)^2$. Substituting this into the relation

$$\sum_{i=2}^{n-1} \gamma_{x_i x_i} x_i^2 + \gamma_{z x_1} z x_1 + \gamma_{zz} z^2 + O(z, x, u)^3 = 0$$

yields

$$\gamma_{x_1 x_1} x_1^2 + \gamma_{z x_1} z x_1 + \gamma_{zz} z^2 + O(x_1, z)^3 = 0. \tag{3.9}$$

If the left side of (3.9) is denoted by $F(z, x_1)$, then

$$F(z, x_1) = [z, x_1] Q \begin{bmatrix} z \\ x_1 \end{bmatrix} + O(z, x_1)^3.$$

The condition (3.5) is equivalent to the fact that the matrix Q in (3.4) is sign definite. Therefore, except for the point $(z, x_1) = (0, 0)$, the value of $F(z, x_1)$ is not zero near the origin. This implies that $(z, x) = (0, 0)$ is an isolated equilibrium point. The first part of the theorem is proved.

If (3.6) holds, then the matrix (3.4) is not sign definite and it has full rank. By the change of coordinates

$$\begin{bmatrix} w_1 & w_2 \end{bmatrix} = \begin{bmatrix} z & x_1 \end{bmatrix} T, \tag{3.10}$$

the equation $F(z, x) = 0$ becomes $d_1 w_1^2 + d_2 w_2^2 + O(w_1, w_2)^3 = 0$. Therefore,

$$w_2 = \pm \sqrt{-\frac{d_1}{d_2}} w_1 + O(w_1)^2.$$

Substituting this into (3.10) and denoting ν as variable w_1 yield equation (3.7). This completes the proof of the second part. □

We summarize the classification of equilibrium sets in Table 3.1. The uncontrollable eigenvalue λ (a linear invariant) and the quadratic invariants determine the class of the equilibrium set.

4. Controllability. In this section and the next section, we study problems related to Question 2 and 3. Suppose we choose an equilibrium point in E. If it is not the origin, then the controllability of its linearization depends on the quadratic part of the system. Sufficient conditions for a system to be linearly controllable at equilibrium points are given in this section.

THEOREM 4.1. *Given a system in the form of* (2.4) *with* $\lambda \neq 0$, *if* $\gamma_{x_1 x_1} \neq 0$, *then there is a neighborhood* U *of* $(z, x) = (0, 0)$ *such that the system is linearly controllable at all the equilibrium points in* U *except the origin.*

Remark. Given a system with $\lambda \neq 0$, if $\lambda > 0$, then the system cannot be stabilized at the origin by C^1 state feedback. However, an interesting corollary of Theorem 4.1 is that, if $\gamma_{x_1x_1} \neq 0$, there is a neighborhood U of the origin such that the system is locally stabilizable at all equilibrium points in U except $(z, x) = (0, 0)$.

Proof. Given a system (2.4). It can be transformed into its normal form (2.7). Since a change of coordinates and feedback does not change the controllability of the linearization, it is enough to prove the theorem for normal forms. Denote the linearization of the normal form at an equilibrium (3.1) by (A_ν, B_ν). Using (3.1) it is easy to check that, at an equilibrium point in E, we have

$$A_\nu = \begin{bmatrix} \lambda & 0 \\ 0 & A_2 \end{bmatrix} + \begin{bmatrix} \gamma_{zx_1}\nu & 2\gamma_{x_1x_1}\nu & 0 \\ 0 & 0 & 0 \end{bmatrix}_{n\times n} + O(\nu)^2,$$

$$B_\nu = \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 \end{bmatrix}^T + O(\nu)^2.$$

Therefore,

$$\begin{aligned} A_\nu^k B_\nu &= [\overbrace{0 \quad 0 \quad \cdots \quad 1}^{n-k} \quad 0 \quad \cdots \quad 0]^T + O(\nu)^2, \quad 0 \le k \le n-2, \\ A_\nu^{n-1} B_\nu &= \begin{bmatrix} 2\gamma_{x_1x_1}\nu & 0 & \vdots & 0 \end{bmatrix}^T + O(\nu)^2. \end{aligned} \tag{4.1}$$

Equations in (4.1) imply that the controllability matrix $[B_\nu, A_\nu B_\nu, \ldots, A_\nu^{n-1} B_\nu]$ has full rank for small nonzero values of ν if $\gamma_{x_1x_1} \neq 0$. From the assumption in the theorem, the system is linearly controllable near the origin. □

System (3.2) satisfies the condition of Theorem 4.1. In fact, $\gamma_{x_1x_1} = 1$. Therefore, the system is linearly controllable at all points in E near $z = 0$, $x = 0$ except the origin. The condition in Theorem 4.1 is sufficient but not necessary. If $\gamma_{x_1x_1} = 0$, the controllability of the system depends on both the quadratic and higher degree terms.

If $\lambda = 0$, then the equilibrium set may have two branches. The following theorem studies the controllability of such systems.

THEOREM 4.2. *Consider a system in the form of* (2.4) *with* $\lambda = 0$. *Assume that inequality* (3.6) *holds. If* $\gamma_{x_1x_1} \neq 0$, *then the system is linearly controllable at all the equilibrium points in* $E \setminus \{(0, 0)\}$ *near the origin.*

Proof. It is simpler to consider a system in normal form (2.8). The linearization of (2.8) at an equilibrium point $(z, x) \in E$ is

$$\begin{aligned} A_\nu &= \begin{bmatrix} 0 & 0 \\ 0 & A_2 \end{bmatrix} + \begin{bmatrix} \gamma_{zx_1}x_1 + 2\gamma_{zz}z & 2\gamma_{x_1x_1}x_1 + \gamma_{zx_1}z & 0 \\ 0 & 0 & 0 \end{bmatrix}_{n\times n} + O(\nu)^2, \\ B_\nu &= \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 \end{bmatrix}^T + O(\nu)^2. \end{aligned} \tag{4.2}$$

By calculation, it is easy to check that

$$\begin{aligned} A_\nu^r B_\nu &= [\overbrace{0 \quad \cdots \quad 0 \quad 1}^{n-r} \quad 0 \quad \cdots \quad 0]^T + O(\nu)^2, \quad 1 \le r \le n-2, \\ A_\nu^{n-1} B_\nu &= \begin{bmatrix} 2\gamma_{x_1x_1}x_1 + \gamma_{zx_1}z & 0 & \cdots & 0 \end{bmatrix}^T + O(\nu)^2. \end{aligned}$$

From (ii) of Theorem 3.2, the controllability matrix at the equilibrium point is in the

following form:

$$[B_\nu, A_\nu B_\nu, \ldots, A_\nu^{n-1} B_\nu] = \begin{bmatrix} 0 & 0 & \cdots & 0 & p\nu \\ 0 & 0 & \cdots & 1 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 1 & 0 & \cdots & 0 & 0 \end{bmatrix} + O(\nu)^2, \tag{4.3}$$

where

$$p = \begin{bmatrix} \gamma_{zx_1} & 2\gamma_{x_1x_1} \end{bmatrix} T \begin{bmatrix} 1 \\ \pm\sqrt{-\dfrac{d_1}{d_2}} \end{bmatrix}.$$

The result in the theorem follows the following claim.

Claim. If $\gamma_{x_1x_1} \neq 0$, then $p \neq 0$.

Proof of the claim. Assume that $p = 0$. Then

$$\begin{bmatrix} 0 & 1 \end{bmatrix} QT \begin{bmatrix} 1 & \pm\sqrt{-\dfrac{d_1}{d_2}} \end{bmatrix}^T = 0.$$

By (3.4), this equation is equivalent to

$$\begin{bmatrix} 0 & 1 \end{bmatrix} T \begin{bmatrix} d_1 & 0 \\ 0 & d_2 \end{bmatrix} \begin{bmatrix} 1 & \pm\sqrt{-\dfrac{d_1}{d_2}} \end{bmatrix}^T = 0.$$

Without loss of generality we assume $d_1 < 0$; then

$$\begin{aligned} \begin{bmatrix} 0 & 1 \end{bmatrix} T \begin{bmatrix} d_1 & 0 \\ 0 & d_2 \end{bmatrix} &= s \begin{bmatrix} -\sqrt{-d_1} & \pm\sqrt{d_2} \end{bmatrix}, \\ \begin{bmatrix} 0 & 1 \end{bmatrix} T &= s \begin{bmatrix} \frac{1}{\sqrt{-d_1}} & \frac{\pm 1}{\sqrt{d_2}} \end{bmatrix} \end{aligned} \tag{4.4}$$

for some $s \in \mathbb{R}$. From (3.4) and (4.4),

$$\gamma_{x_1x_1} = \begin{bmatrix} 0 & 1 \end{bmatrix} T \begin{bmatrix} d_1 & 0 \\ 0 & d_2 \end{bmatrix} T^T \begin{bmatrix} 0 \\ 1 \end{bmatrix} = s^2 \begin{bmatrix} -\sqrt{-d_1} & \pm\sqrt{d_2} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{-d_1}} \\ \frac{\pm 1}{\sqrt{d_2}} \end{bmatrix} = 0.$$

It is a contradiction. Therefore, $p \neq 0$. The claim is proved. □

Remark. In the case of $\lambda = 0$ and $\det(Q) < 0$, if $\gamma_{x_1x_1} = 0$, one branch of the equilibrium set E is tangent to $z = 0$. The controllability of the linearization at points in this branch depends on the cubic and higher degree terms of the system. If the system is not linearly controllable on this branch, it is proved in Part II that the stabilizability of the system changes as the equilibrium point passing through the origin along this curve. On the other branch of E, the system is linearly controllable if this branch is not tangent to $x_1 = 0$ (i.e., $\gamma_{zz} \neq 0$). This result can be proved by finding the controllability matrix of the normal form. If $\gamma_{zz} = 0$, then the controllability of the linearization at points in this branch depends on cubic and higher degree terms.

Example. Consider system (1.6) in the example of section 1.2. The invariants of the system are $\gamma_{x_1x_1} = 0$, $\gamma_{zx_1} = 1$, and $\gamma_{zz} = 1$. The matrix Q is

$$Q = \begin{bmatrix} 1 & \dfrac{1}{2} \\ \dfrac{1}{2} & 0 \end{bmatrix}. \tag{4.5}$$

Since E_- is not tangent to $x = 0$ (see Figure 1.1), the system is linearly controllable in E_- except the origin. The branch E_+ is tangent to $z = 0$; the controllability in E_+ depends on higher degree terms. The system (1.6) is not linearly controllable at points in E_+ because there is no cubic term in the nonlinear system.

5. Stabilizability of control systems. In this section, we prove a sufficient condition in terms of quadratic invariants for stabilizability of control systems. Consider a system of form (2.4). The system is stabilizable when $\lambda < 0$, and it is not stabilizable by C^1 state feedback if $\lambda > 0$. So we consider only the case in which λ is zero. If the system is nonlinear, its center manifold can have different shapes under different feedback. In the following we prove that if the quadratic invariants satisfy certain conditions, then there exists feedback so that the reduced dynamics on the center manifold are asymptotically stable. The center manifold theory can be found in [4]. Therefore, the feedback renders the closed-loop system locally asymptotically stable. The following theorem is a partial answer to Question 3 in the sense that the sufficient condition for stabilizability of a control system at a special equilibrium point—the origin—is given.

THEOREM 5.1. *Suppose $\lambda = 0$ in system (2.4). Suppose (3.6) holds. If $\gamma_{zx_1} \neq 0$, then there exists C^1 state feedback which locally asymptotically stabilizes the system at the origin.*

Proof. Since the system can be transformed into its normal form (2.8), we only prove the theorem for systems in normal form. Use the feedback

$$u(z, x) = F_1 x_1 + F_2 x_2 + \cdots + F_{n-1} x_{n-1} + \alpha z + \beta z^2, \tag{5.1}$$

where $F = (F_1, F_2, \ldots, F_{n-1})$ stabilizes the controllable part; i.e.,

$$A_2 + B_2 F \tag{5.2}$$

is a Hurwitz matrix. Since F_1 is the constant term in the characteristic polynomial of the matrix (5.2), $F_1 \neq 0$. The center manifold of the closed-loop system is

$$x = \pi(z) = \begin{bmatrix} \pi_1(z) & \cdots & \pi_{n-1}(z) \end{bmatrix}^T \tag{5.3}$$

such that

$$\begin{bmatrix} \pi_2 & \pi_3 & \cdots & \pi_{n-1} & F\pi \end{bmatrix}^T + \begin{bmatrix} 0 & 0 & \cdots & 0 & \alpha z + \beta z^2 \end{bmatrix}^T + \tilde{f}^{[2]}(\pi_3, \ldots, \pi_{n-1})$$

$$= \frac{d\pi}{dz} \left(\sum_{i=1}^{n-1} \gamma_{x_i x_i} \pi_i^2 + \gamma_{z x_1} z \pi_1 + \gamma_{zz} z^2 \right) + O(z)^3. \tag{5.4}$$

Solving the equation for the linear part of π we get

$$\pi_1 = -\frac{\alpha}{F_1} z + \pi_1^{[2]}(z) + O(z)^3,$$

$$\pi_i(z) = O(z)^2 \qquad \text{for } i = 2, \ldots, n-1.$$

Substituting this result into (5.4), we get the equation for the quadratic part of $\pi(z)$:

$$\begin{bmatrix} \pi_2 \\ \pi_3 \\ \vdots \\ F_1 \pi_1^{[2]} + \sum_{i=2}^{n-1} F_i \pi_i \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \beta z^2 \end{bmatrix} = \left(\frac{\gamma_{x_1 x_1} \alpha^2}{F_1^2} - \frac{\gamma_{z x_1} \alpha}{F_1} + \gamma_{zz} \right) z^2 \begin{bmatrix} -\frac{\alpha}{F_1} \\ 0 \\ \vdots \\ 0 \end{bmatrix} + O(z)^3.$$

Therefore, $\pi_i(z) = O(z)^3$ for $i = 3, \ldots, n-1$. Furthermore,

$$\begin{aligned}\pi_1(z) &= -\frac{\alpha}{F_1}z + \left(-\frac{\beta}{F_1} + \frac{\alpha F_2}{F_1^2}\left(\frac{\gamma_{x_1x_1}\alpha^2}{F_1^2} - \frac{\gamma_{zx_1}\alpha}{F_1} + \gamma_{zz}\right)\right)z^2 + O(z)^3,\\ \pi_2 &= -\frac{\alpha}{F_1}\left(\frac{\gamma_{x_1x_1}\alpha^2}{F_1^2} - \frac{\gamma_{zx_1}\alpha}{F_1} + \gamma_{zz}\right)z^2 + O(z)^3.\end{aligned} \tag{5.5}$$

We can choose α so that

$$\frac{\gamma_{x_1x_1}\alpha^2}{F_1^2} - \frac{\gamma_{zx_1}\alpha}{F_1} + \gamma_{zz} = 0. \tag{5.6}$$

In fact, if $\gamma_{zz} = 0$, we take $\alpha = 0$, if $\gamma_{zz} \neq 0$; the value of α is determined by

$$F_1 = \frac{\gamma_{zx_1} \pm \sqrt{\gamma_{zx_1}^2 - 4\gamma_{x_1x_1}\gamma_{zz}}}{2\gamma_{zz}}\alpha. \tag{5.7}$$

The sign "+" or "−" will be determined later. The condition $\det(Q) < 0$ implies $\gamma_{zx_1}^2 - 4\gamma_{x_1x_1}\gamma zz > 0$. Therefore, F_1 must be a real number. Substituting (5.6) into (5.5), we get

$$\begin{aligned}\pi_1(z) &= -\frac{\alpha}{F_1}z - \frac{\beta}{F_1}z^2 + O(z)^3,\\ \pi_i(z) &= O(z)^3 \qquad\qquad \text{for } i = 2, \ldots, n-1.\end{aligned}$$

Substituting this into the equation of z in (2.8), we get the reduced dynamic system of the closed-loop system on the center manifold

$$\dot{z} = -\beta\frac{\gamma_{zx_1}F_1 - 2\alpha\gamma_{x_1x_1}}{F_1^2}z^3 + ez^3 + O(z)^4 \tag{5.8}$$

where e is a constant depending on F_1, α, and the coefficients of z^3 in (2.8). However, it is independent of β. The value of β is determined by the following method. If $\gamma_{zz} = 0$, then $\alpha = 0$. The number $\gamma_{zx_1}F_1 - 2\alpha\gamma_{x_1x_1}$ is not zero because $\gamma_{zx_1} \neq 0$ is an assumption. By choosing β, we can make the coefficient of z^3 be negative. The dynamics are asymptotically stable. If $\gamma_{zz} \neq 0$, from (5.7)

$$\gamma_{zx_1}F_1 - 2\alpha\gamma_{x_1x_1} = \frac{\gamma_{zx_1}(\gamma_{zx_1} \pm \sqrt{\gamma_{zx_1}^2 - 4\gamma_{x_1x_1}\gamma_{zz}}) - 4\gamma_{x_1x_1}\gamma_{zz}}{2\gamma_{zz}}\alpha.$$

From (5.7) and the fact $F_1 \neq 0$, we know that $\alpha \neq 0$. Since $\gamma_{zx_1}^2 - 4\gamma_{x_1x_1}\gamma_{zz} > 0$, we can always choose "+" or "-" in this expression so that this number does not equal zero. Therefore, by a suitable choice of β, we can make the coefficient of z^3 less than zero. So, it is proved that there is feedback such that the reduced dynamic system on the center manifold is asymptotically stable. By the center manifold theorem, the closed-loop system is asymptotically stable. □

Remark. The proof, in fact, shows a method of designing stabilizing feedback for systems in normal form. It is given by

$$u(z, x) = Fx + \alpha z + \beta z^2,$$

where F stabilizes (A_2, B_2). The number α satisfies

$$\alpha = 0 \qquad \text{if } \gamma_{zz} = 0,$$

$$\alpha = \frac{2\gamma_{zz}F_1}{\gamma_{zx_1} \pm \sqrt{\gamma_{zx_1}^2 - 4\gamma_{x_1x_1}\gamma_{zz}}} \qquad \text{if } \gamma_{zz} \neq 0,$$

where the sign is chosen such that

$$\gamma_{zx_1} \pm \sqrt{\gamma_{zx_1}^2 - 4\gamma_{x_1x_1}\gamma_{zz}} \neq 0,$$

$$\gamma_{zx_1}(\gamma_{zx_1} \pm \sqrt{\gamma_{zx_1}^2 - 4\gamma_{x_1x_1}\gamma_{zz}}) - 4\gamma_{x_1x_1}\gamma_{zz} \neq 0.$$

The number β satisfies

$$\beta(\gamma_{zx_1}F_1 - 2\alpha\gamma_{x_1x_1}) > 0,$$

and the absolute value of β is sufficiently large.

6. Conclusion. Problems formulated from the bifurcation viewpoint concerning equilibrium sets, controllability, and stabilizability of control systems are introduced. Normal forms and invariants of control systems are employed in the analysis. The topology of the equilibrium set and the properties such as controllability and stabilizability of a control system point are proved to be closely related to the invariants. The local bifurcations of equilibrium sets are classified, and the set is linearly approximated by a parametrization. Typical diagrams of bifurcation equilibrium sets are shown by examples of systems in normal forms. Sufficient conditions given by invariants for controllability and stabilizability at the points in equilibrium sets are found.

In Part II, the same problems will be addressed for control systems with a single parameter. The equilibrium set is two dimensional. More complex bifurcations occur in this case.

Acknowledgment. The author would like to thank Professor Arthur J. Krener for his creative suggestions and comments.

REFERENCES

[1] E. Abed and J.-H. Fu, *Local feedback stabilization and bifurcation control*, I. *Hopf bifurcation*, Systems Control Lett., 7 (1986), pp. 11–17.
[2] E. Abed and J.-H. Fu, *Local feedback stabilization and bifurcation control,* II. *Stationary bifurcation.*, Systems Control Lett., 8 (1987), pp. 467–473.
[3] V. I. Arnold, *Geometrical Methods in the Theory of Ordinary Differential Equations*, 2nd ed., Springer-Verlag, Berlin, New York, 1988.
[4] J. Carr, *Application of Center Manifold Theory*, Springer-Verlag, New York, 1981.
[5] F. Colonius and W. Kliemann, *Controllability and stabilization of one-dimensional systems near bifurcation points*, Systems Control Lett., 24 (1995), pp. 87–95.
[6] J. Hale and H. Kocak, *Dynamics and Bifurcations*, Springer-Verlag, Berlin, New York, 1991.
[7] T. Kailath, *Linear Systems*, Prentice–Hall, Inc., Englewood Cliffs, NJ, 1980.
[8] W. Kang and A. J. Krener, *Extended quadratic controller normal form and dynamic feedback linearization of nonlinear systems*, SIAM J. Control Optim., 30 (1992), pp. 1319–1337.
[9] W. Kang, *Quadratic normal forms of nonlinear control systems with uncontrollable linearization*, in Proc. IEEE Conf. on Decision and Control, New Orleans, 1995, pp. 608–612.
[10] W. Kang, *Extended controller form and invariants of nonlinear control systems with a single input*, J. Math. Systems Estim. Control, 4 (1994), pp. 253–256.

[11] M. Krstic, J. M. Protz, J. D. Paduano, and P. V. Kokotovic, *Backstepping designs for jet engine stall and surge control*, Proc. IEEE Conf. Decision and Control, New Orleans, 1995, pp. 3049–3055.
[12] D.-C. Liaw and E. H. Abed, *Stability analysis and control of rotating stall*, in Proc. IFAC Nonlinear Control Systems Design Symposium, Bordeaux, France, June 1992.
[13] F. E. McCaughan, *Bifurcation analysis of axial flow compressor stability*, SIAM J. Appl. Math., 50 (1990), pp. 1232–1253.
[14] H. Nijmeijer and A. J. van der Schaft, *Nonlinear Dynamical Control Systems*, Springer-Verlag, New York, 1990.
[15] S. Wiggins, *Introduction to Applied Nonlinear Dynamical Systems and Chaos*, Springer-Verlag, 1990.
[16] W. Kang, *Bifurcation and normal form of nonlinear control systems, Part* II, SIAM J. Control Optim., 36 (1998), pp. 213–232.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 1, pp. 213–232, January 1998

012

BIFURCATION AND NORMAL FORM OF NONLINEAR CONTROL SYSTEMS, PART II*

WEI KANG†

Abstract. The normal forms and invariants of control systems with a parameter are found. Bifurcations of equilibrium sets are classified. The changes of properties such as controllability of the linearization or stabilizability near a bifurcation point of a control system are studied.

Key words. nonlinear systems, bifurcations, normal forms, invariants, linearly controllable, stabilizable

AMS subject classifications. 93C10, 93C15

PII. S036301299529029X

1. Introduction. In this paper, we continue the study of bifurcation problems formulated in Part I [10]. We focus on systems with a parameter. Comparing with the results in Part I, systems with a parameter have three types of normal forms instead of two. Another difference is that, in the presence of a parameter, the equilibrium sets are not curves. In fact, they are two-dimensional surfaces in the state-parameter space. In general, there always exists a curve on the equilibrium set such that the system is not linearly controllable at any point on it.

The following nonlinear system with parameter μ is considered:

$$\dot{\xi} = f(\xi, \mu) + g(\xi, \mu)v. \tag{1.1}$$

The variable $\xi \in \mathbb{R}^n$ is the state, $v \in \mathbb{R}$ is the input variable, and the parameter is $\mu \in \mathbb{R}$. The vector fields $f(\xi, \mu)$ and $g(\xi, \mu)$ are assumed to be C^k for some sufficiently large k. Our attention is focused on local bifurcation near the origin $(\xi, \mu) = (0, 0)$. Assume

$$f(0, 0) = 0, \quad g(0, 0) \neq 0.$$

Following Part I, the equilibrium set E is defined to be

$$E = \{(x, \mu) | \exists v_0 \text{ such that } f(x, \mu) + g(x, \mu)v_0 = 0\}. \tag{1.2}$$

The linearization of the system at the origin is (A, B)

$$A = \frac{\partial f}{\partial x}(0, 0), \quad B = g(0, 0).$$

If the system is linearly controllable at $\xi = 0$, $\mu = 0$, then the system is linearly controllable for all equilibrium points in E near $(x, \mu) = (0, 0)$. Therefore, Questions 1–3 formulated in Part I are interesting only if (1.1) is not linearly controllable at the origin. Similar to Part I, we always assume

Assumption.

$$\operatorname{rank}(\begin{bmatrix} B & AB & A^2B & \cdots & A^{n-1}B \end{bmatrix}) = n - 1. \tag{1.3}$$

*Received by the editors August 14, 1995; accepted for publication (in revised form) October 31, 1996. Research supported in part by AFOSR-95-1-0169.

http://www.siam.org/journals/sicon/36-1/29029.html

†Department of Mathematics, Naval Postgraduate School, Monterey, CA 93943 (wkang@math.nps.navy.mil).

To classify the equilibrium sets and their bifurcations, it is necessary to introduce normal forms and transformations. For systems with parameters, a transformation consists of a change of coordinates and feedback. Both can be parameter related. More specifically, a transformation is given by

$$\begin{aligned} x &= \phi(\xi,\mu),\\ u &= \alpha(\xi,\mu)+\beta(\xi,\mu)v \end{aligned} \tag{1.4}$$

in which $\frac{\partial\phi}{\partial x}(0,0)$ is nonsingular and $\beta(0,0)\neq 0$. If a transformation is applied to (1.1), denote the equilibrium set of the resulting system by $\bar{E}$. Then, in a neighborhood of $(\xi,\mu)=(0,0)$, a point (ξ,μ) is in E if and only if $(x,\mu)=(\phi(\xi,\mu),\mu)$ is in $\bar{E}$. Therefore, in the sense of diffeomorphism, the transformation (1.4) does not change the equilibrium set E (locally). Furthermore, the change of coordinates and feedback (1.4) does not change the properties of our interest such as controllability (of the linearization) or stabilizability [8].

Following the idea used in Part I, we simplify nonlinear control systems by the transformations of form (1.4). This will be done in section 2. In sections 3–5 the parametrization of equilibrium sets and the controllability at points in E are discussed for systems with different normal forms. In section 6 the problem of stabilizability is addressed for systems with a certain type of normal form. In this paper, the discussion is focused on systems with a control input. For systems without control, the classical bifurcation theory and some interesting applications can be found in [1], [2], [4], and [9].

2. Normal forms and quadratic invariants. The first step of finding normal forms is to simplify the linear part. Since the controllability index of (A,B) is $n-1$, there exists a linear change of coordinates and feedback independent of μ transforming the system (1.1) into

$$\begin{aligned} \dot{z} &= \lambda z+\gamma\mu+O(z,x,\mu,u)^2,\\ \dot{x} &= A_2x+\Gamma\mu+B_2u+O(z,x,\mu,u)^2, \end{aligned} \tag{2.1}$$

where $\Gamma=\begin{bmatrix}\gamma_1 & \gamma_2 & \cdots & \gamma_{n-1}\end{bmatrix}^T$, $x\in\mathbb{R}^{n-1}$, and $z\in\mathbb{R}$. The pair (A_2,B_2) is in Brunovsky form:

$$A_2=\begin{bmatrix} 0 & 1 & 0 & \cdots & 0\\ 0 & 0 & 1 & \cdots & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & 0 & \cdots & 1\\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}_{(n-1)\times(n-1)},\quad B_2=\begin{bmatrix} 0\\ 0\\ \vdots\\ 0\\ 1 \end{bmatrix}_{(n-1)\times 1}.$$

To further simplify the linearization of the system, consider another change of coordinates

$$\begin{aligned} \bar{x}_1 &= x_1,\\ \bar{x}_i &= x_i+\gamma_{i-1}\mu, \quad i=2,\dots,n,\\ \bar{u} &= u+\gamma_n\mu, \end{aligned} \tag{2.2}$$

which transforms system (2.1) into (2.3). For the reason of simplicity, we still use (z,x) and u to represent the state variables and control input for the new system:

$$\begin{aligned} \dot{z} &= \lambda z+\gamma\mu+O(z,x,\mu,u)^2,\\ \dot{x} &= A_2x+B_2u+O(z,x,\mu,u)^2. \end{aligned} \tag{2.3}$$

If $\lambda \neq 0$, the equation for z can be simplified by $\bar{z} = z + \frac{\gamma}{\lambda}\mu$. In the resulting system, the equation for $\bar{z}$ is

$$\dot{\bar{z}} = \lambda \bar{z} + O(\bar{z}, x, \mu, u)^2.$$

If $\lambda = 0$ and $\gamma \neq 0$, the change of coordinate $\bar{z} = \frac{1}{\gamma} z$ transforms the equation for z into

$$\dot{\bar{z}} = \mu + O(\bar{z}, x, \mu, u)^2.$$

The normal forms for the linear part of (1.1) are summarized in the following lemma.

LEMMA 2.1. *Given a system* (1.1) *satisfying assumption* (1.3)*, there exists a linear change of coordinates and feedback which transforms* (1.1) *into one of the following forms:*

$\lambda \neq 0$

$$\begin{aligned} \dot{z} &= \lambda z + f_1^{[2]}(z, x, \mu) + g_1^{[1]}(z, x, \mu)u + O(z, x, \mu, u)^3, \\ \dot{x} &= A_2 x + B_2 u + f_2^{[2]}(z, x, \mu) + g_2^{[1]}(z, x, \mu)u + O(z, x, \mu, u)^3; \end{aligned} \tag{2.4}$$

$\lambda = 0$*, Jordan form*

$$\begin{aligned} \dot{z} &= \mu + f_1^{[2]}(z, x, \mu) + g_1^{[1]}(z, x, \mu)u + O(z, x, \mu, u)^3, \\ \dot{x} &= A_2 x + B_2 u + f_2^{[2]}(z, x, \mu) + g_2^{[1]}(z, x, \mu)u + O(z, x, \mu, u)^3; \end{aligned} \tag{2.5}$$

$\lambda = 0$*, diagonal form*

$$\begin{aligned} \dot{z} &= f_1^{[2]}(z, x, \mu) + g_1^{[1]}(z, x, \mu)u + O(z, x, \mu, u)^3, \\ \dot{x} &= A_2 x + B_2 u + f_2^{[2]}(z, x, \mu) + g_2^{[1]}(z, x, \mu)u + O(z, x, \mu, u)^3. \end{aligned} \tag{2.6}$$

In the normal form, λ is an uncontrollable eigenvalue of the linearization at $(z, x, \mu) = (0, 0, 0)$. If $\lambda = 0$ and if μ is considered as a state variable with $\dot{\mu} = 0$, the matrix of uncontrollable dynamic system is a zero matrix or it can be simplified into a Jordan block. These two cases are shown in (2.5) and (2.6).

The following quadratic transformations are employed to simplify the quadratic part of a system into its normal form while leaving the linear part invariant:

$$\begin{aligned} \begin{bmatrix} \bar{z} & \bar{x} \end{bmatrix}^T &= \begin{bmatrix} z & x \end{bmatrix}^T + \phi^{[2]}(z, x, \mu), \\ \bar{u} &= u + \alpha^{[2]}(z, x, \mu) + \beta^{[1]}(z, x, \mu)u. \end{aligned} \tag{2.7}$$

The normal forms are given in the following theorem. The notation $\tilde{f}^{[2]}(x)$ in the theorem represents the extended controller form (see [10, equation (2.9)], [5], [7]).

THEOREM 2.2. *Consider a control system satisfying assumption* (1.3)*. Suppose that its linearization is in the form given by* (2.4)*,* (2.5)*, or* (2.6)*. Then there exists a quadratic change of coordinates and feedback* (2.7) *which transforms the system into one of the following normal forms.*

(i) *For* (2.4)*, the normal form is*

$$\begin{aligned}\dot{z} &= \lambda z + \sum_{i=1}^{n-1} \gamma_{x_i x_i} x_i^2 + \gamma_{z x_1} z x_1 + \gamma_{x_1 \mu} x_1 \mu + \gamma_{z\mu} z\mu + O(z, x, \mu, u)^3,\\ \dot{x} &= A_2 x + B_2 u + \tilde{f}^{[2]}(x) + O(z, x, \mu, u)^3.\end{aligned} \tag{2.8}$$

(ii) *For* (2.5) *the normal form is*

$$\begin{aligned}\dot{z} &= \mu + \sum_{i=1}^{n-1} \gamma_{x_i x_i} x_i^2 + \gamma_{z x_1} z x_1 + \gamma_{x_1 \mu} x_1 \mu + \gamma_{zz} z^2 + O(z, x, \mu, u)^3,\\ \dot{x} &= A_2 x + B_2 u + \tilde{f}^{[2]}(x) + O(z, x, \mu, u)^3.\end{aligned} \tag{2.9}$$

(iii) *For* (2.6)*, the normal form is*

$$\begin{aligned}\dot{z} &= \sum_{i=1}^{n-1} \gamma_{x_i x_i} x_i^2 + \gamma_{z x_1} z x_1 + \gamma_{x_1 \mu} x_1 \mu + \gamma_{z\mu} z\mu + \gamma_{zz} z^2 + \gamma_{\mu\mu}\mu^2 + O(z, x, \mu, u)^3,\\ \dot{x} &= A_2 x + B_2 u + \tilde{f}^{[2]}(x) + O(z, x, \mu, u)^3.\end{aligned} \tag{2.10}$$

Proof. The theorem shows three quadratic normal forms for systems with different linearizations. We prove them separately.

(i) Suppose the linearization has the same form as (2.4). We consider the following extended system in which μ is treated as a state variable:

$$\begin{aligned}\dot{z} &= \lambda z + f_1^{[2]}(z, x, \mu) + g_1^{[1]}(z, x, \mu)u + O(z, x, \mu, u)^3,\\ \dot{x} &= A_2 x + B_2 u + f_2^{[2]}(z, x, \mu) + g_2^{[1]}(z, x, \mu)u + O(z, x, \mu, u)^3,\\ \dot{\mu} &= 0.\end{aligned}$$

By the result in [6], there exists a quadratic transformation

$$\begin{aligned} z &= \bar{z} + \phi_1^{[2]}(\bar{z}, \bar{x}, \bar{\mu}), \quad x = \bar{x} + \phi_2^{[2]}(\bar{z}, \bar{x}, \bar{\mu}),\\ \mu &= \bar{\mu} + \phi_3^{[2]}(\bar{z}, \bar{x}, \bar{\mu}), \quad u = \bar{u} + \alpha^{[2]}(\bar{z}, \bar{x}, \bar{\mu}) + \beta(\bar{z}, \bar{x}, \bar{\mu})\bar{u}\end{aligned}$$

so that, under the new coordinates, the dynamics of $\bar{z}$ and $\bar{x}$ are in their quadratic normal forms given in [6], which are

$$\begin{aligned}\dot{\bar{z}} &= \lambda \bar{z} + \sum_{i=1}^{n-1} \gamma_{x_i x_i} \bar{x}_i^2 + \gamma_{z x_1} \bar{z}\bar{x}_1 + \gamma_{x_1 \mu} \bar{x}_1 \bar{\mu} + \gamma_{z\mu} \bar{z}\bar{\mu} + O(\bar{z}, \bar{x}, \bar{\mu}, \bar{u})^2,\\ \dot{\bar{x}} &= A\bar{x} + B\bar{u} + \tilde{f}^{[2]}(\bar{x}) + O(\bar{z}, \bar{x}, \bar{\mu}, \bar{u})^3.\end{aligned} \tag{2.11}$$

Since transformation (2.7) does not change the last variable μ, we substitute the relation

$$\mu = \bar{\mu} + \phi_3^{[2]}(\bar{z}, \bar{x}, \bar{\mu})$$

back into (2.11). It is obvious that this will not change the linear and quadratic parts in the dynamics of $\bar{z}$ and $\bar{x}$. Notice that the system (2.11) is in the same form as (2.8).

(ii) If a system has the same linearization as (2.5), the results in [6] do not provide a complete normal form for the system since the linearization of the uncontrollable part (including the dynamics of z and μ) is not diagonal. To simplify the proof by using results in [6], let's assume that $\mu = 0$. Then, system (2.5) is

$$\begin{aligned}\dot{z} &= f_1^{[2]}(z,x,0) + g_1^{[1]}(z,x,0)u + O(z,x,u)^3,\\ \dot{x} &= A_2x + B_2u + f_2^{[2]}(z,x,0) + g_2^{[1]}(z,x,0)u + O(z,x,u)^3.\end{aligned}$$

By a suitable transformation

$$\begin{aligned}&\bar{z} = z + \phi_1^{[2]}(z,x), \quad \bar{x} = x + \phi_2^{[2]}(z,x),\\ &\bar{u} = u + \alpha^{[2]}(z,x) + \beta^{[2]}(z,x)u,\end{aligned} \tag{2.12}$$

the system can be transformed into the following normal form given in [6]. To simplify the notation, we still use (z,x) instead of $(\bar{z},\bar{x})$ as the state variables:

$$\begin{aligned}\dot{z} &= \sum_{i=1}^{n-1}\gamma_{x_ix_i}x_i^2 + \gamma_{zx_1}zx_1 + \gamma_{zz}z^2 + O(z,x,u)^3,\\ \dot{x} &= A_2x + B_2u + \tilde{f}^{[2]}(x) + O(z,x,u)^3.\end{aligned}$$

So, transformation (2.12) transforms (2.5) into

$$\begin{aligned}\dot{z} &= \mu + \sum_{i=1}^{n-1}\gamma_{x_ix_i}x_i^2 + \gamma_{zx_1}zx_1 + \gamma_{zz}z^2 + \gamma_{z\mu}z\mu\\ &\quad + \sum_{i=1}^{n-1}b_{x_i\mu}x_i\mu + \gamma_{\mu\mu}\mu^2 + b_{\mu u}\mu u + O(z,x,\mu,u)^3,\\ \dot{x} &= A_2x + B_2u + \tilde{f}^{[2]}(x) + d_{z\mu}z\mu + \sum_{i=1}^{n-1}d_{x_i\mu}x_i\mu + d_{\mu\mu}\mu^2 + d_{\mu u}\mu u + O(z,x,\mu,u)^3,\end{aligned} \tag{2.13}$$

where $d_{z\mu}$, $d_{x_i\mu}$, $d_{\mu\mu}$, and $d_{\mu u}$ are constant vectors of dimension $n-1$. By a transformation $\bar{z} = z - b_{\mu u}\mu x_{n-1}$, $\bar{x} = x - d_{\mu u}\mu x_{n-1}$ the quadratic part μu in (2.13) can be cancelled. So, we focus on a system of the following form:

$$\begin{aligned}\dot{z} &= \mu + \sum_{i=1}^{n-1}\gamma_{x_ix_i}x_i^2 + \gamma_{zx_1}zx_1 + \gamma_{zz}z^2 + \gamma_{z\mu}z\mu + \sum_{i=1}^{n-1}b_{x_i\mu}x_i\mu + \gamma_{\mu\mu}\mu^2 + O(z,x,\mu,u)^3,\\ \dot{x} &= A_2x + B_2u + \tilde{f}^{[2]}(x) + d_{z\mu}z\mu + \sum_{i=1}^{n-1}d_{x_i\mu}x_i\mu + d_{\mu\mu}\mu^2 + O(z,x,\mu,u)^3.\end{aligned} \tag{2.14}$$

Let's consider a transformation

$$\bar{z} = z, \quad \bar{x} = x + \phi^{[2]}(z,x,\mu), \quad \bar{u} = u + \alpha^{[2]}(z,x,\mu) + \beta^{[1]}(z,x,\mu)u, \tag{2.15}$$

where $\phi^{[2]}(z,x,0) = 0$ and $\alpha^{[2]}(z,x,0) = 0$. To leave B_2 invariant, we also assume

$$\frac{\partial\phi_i^{[2]}}{\partial x_{n-1}} = 0 \tag{2.16}$$

for all $1 \le i \le n-2$, where $\phi_i^{[2]}$ is the ith component of $\phi^{[2]}$. By the separation principle in [6], this transformation will not change the quadratic part of the uncontrollable

dynamic system (the equation of z); it will not change the normal form $\tilde{f}^{[2]}$. Only the quadratic part related to μ in the controllable dynamic system is affected by this transformation. Denote by $f_\mu(z,x,\mu)$ the quadratic terms with μ in the dynamics of x in (2.14)

$$f_\mu(z,x,\mu) = d_{z\mu} z\mu + \sum_{i=1}^{n-1} d_{x_i\mu} x_i \mu + d_{\mu\mu}\mu^2. \tag{2.17}$$

Similarly, if the transformation (2.15) is applied to (2.14), the quadratic terms with μ in the resulting dynamics of $\bar{x}$ is denoted by $\bar{f}_\mu(z,x,\mu)$. The relation between f_μ and $\bar{f}_\mu$ is defined by the following homological equation from [6]:

$$\bar{f}_\mu + \Pi(\phi^{[2]}, \alpha^{[2]}) = f_\mu, \tag{2.18}$$

where Π is the linear operator

$$\Pi(\phi^{[2]}, \alpha^{[2]}) = \frac{\partial \phi^{[2]}}{\partial x} A_2 x + \frac{\partial \phi^{[2]}}{\partial z}\mu - A_2\phi^{[2]} - B_2\alpha^{[2]}.$$

Define a linear space W to be the space consisting of quadratic vectors f_μ in the form of (2.17). Define V to be a linear space consisting of the elements $(\phi^{[2]}, \alpha^{[2]})$ satisfying (2.16). Then Π is a linear map from V to W. The kernel of Π is

$$\ker(\Pi) = \left\{ (\phi^{[2]}, \alpha^{[2]}) \;\middle|\; \begin{array}{l} \phi_1^{[2]} = a_1 z\mu + a_2 x_1 \mu + a_3 \mu^2 \\ \\ \phi_i^{[2]} = \dfrac{\partial \phi_{i-1}^{[2]}}{\partial x} A_2 x + \dfrac{\partial \phi_{i-1}^{[2]}}{\partial z}\mu \;\text{ for }\; i = 1,\ldots,n-1 \\ \\ \alpha^{[2]} = \dfrac{\partial \phi_{n-1}^{[2]}}{\partial x} A_2 x + \dfrac{\partial \phi_{n-1}^{[2]}}{\partial z}\mu \end{array} \right\}.$$

Therefore, the dimension of $\ker(\Pi)$ is 3. The dimension of the image space under Π is $\dim(\Pi(V)) = \dim(V) - \dim(\ker(\Pi)) = n^2 - 1 = \dim(W)$. This implies that the map Π is onto. So, there exists a transformation given by $(\phi^{[2]}, \alpha^{[2]})$ in V which solves the homological equation (2.18) for $\bar{f}_\mu = 0$. Therefore, the vector field f_μ can be cancelled. By the homological equations of $g^{[1]}$ in [6] and condition (2.16), a suitable choice of $\beta(z,x,\mu)$ will avoid the quadratic terms involving u. Therefore, the system can be simplified into the following form without f_μ:

$$\begin{aligned} \dot{z} &= \mu + \sum_{i=1}^{n-1} \gamma_{x_i x_i} x_i^2 + \gamma_{z x_1} z x_1 + \gamma_{zz} z^2 + \gamma_{z\mu} z\mu + \sum_{i=1}^{n-1} b_{x_i\mu} x_i \mu + \gamma_{\mu\mu}\mu^2 + O(z,x,\mu,u)^3, \\ \dot{x} &= A_2 x + B_2 u + \tilde{f}^{[2]}(x) + O(z,x,\mu,u)^3. \end{aligned} \tag{2.19}$$

In the system, all quadratic parts are in normal forms except the terms with μ in the equation of z. This part can be simplified by

$$\bar{z} = z + \sum_{i=1}^{n-2} c_{x_i\mu} x_i \mu + c_{z\mu} z\mu + c_{\mu\mu}\mu^2 + c_{zz} z^2. \tag{2.20}$$

By the separation principle in [6], this transformation does not change the terms in the controllable part. The equation of z is transformed into

$$\begin{aligned}\dot{\tilde{z}} &= \mu + \sum_{i=1}^{n-1} \gamma_{x_i x_i} x_i^2 + \gamma_{z x_1} z x_1 + \gamma_{zz} z^2 + \gamma_{z\mu} z\mu + \sum_{i=1}^{n-1} b_{x_i\mu} x_i \mu + \gamma_{\mu\mu}\mu^2 \\ &+ \sum_{i=2}^{n-1} c_{x_{i-1}\mu} x_i \mu + c_{z\mu}\mu^2 + 2c_{zz} z\mu + O(z,x,\mu)^3.\end{aligned} \tag{2.21}$$

By choosing the transformation such that $b_{x_i\mu} = -c_{x_{i-1}\mu}$ for $2 \le i \le n-1$, $\gamma_{\mu\mu} = -c_{z\mu}$, and $\gamma_{z\mu} = -2c_{zz}$, the quadratic part of system (2.19) can be simplified into the normal form (2.9).

(iii) The argument similar to the proof of (i) can be applied to (iii). Given a system (2.6). If μ is considered as a state variable, the extended system is in the following form:

$$\begin{aligned}\dot{z} &= f_1^{[2]}(z,x,\mu) + g_1^{[1]}(z,x,\mu)u + O(z,x,\mu,u)^3, \\ \dot{x} &= A_2 x + B_2 u + f_2^{[2]}(z,x,\mu) + g_2^{[1]}(z,x,\mu)u + O(z,x,\mu,u)^3, \\ \dot{\mu} &= 0.\end{aligned}$$

From [6], this system can be simplified into a normal form

$$\begin{aligned}\dot{\bar{z}} &= \sum_{i=1}^{n-1} \gamma_{x_i x_i} \bar{x}_i^2 + \gamma_{z x_1} \bar{z}\bar{x}_1 + \gamma_{x_1\mu}\bar{x}_1\bar{\mu} + \gamma_{z\mu}\bar{z}\bar{\mu} + \gamma_{zz}\bar{z}^2 + \gamma_{\mu\mu}\bar{\mu}^2 + O(\bar{z},\bar{x},\bar{\mu},\bar{u})^3, \\ \dot{\bar{x}} &= A\bar{x} + B\bar{u} + \tilde{f}^{[2]}(\bar{x}) + O(\bar{z},\bar{x},\bar{\mu},\bar{u})^3.\end{aligned} \tag{2.22}$$

The quadratic transformation is

$$\begin{aligned}\bar{z} &= z + \phi_1^{[2]}(z,x,\mu), \quad \bar{x} = x + \phi_2^{[2]}(z,x,\mu), \\ \bar{\mu} &= \mu + \phi_3^{[2]}(z,x,\mu), \quad \bar{u} = u + \alpha^{[2]}(z,x,\mu) + \beta(z,x,\mu)u.\end{aligned}$$

Since the transformation for our purpose does not change μ, we substitute $\bar{\mu} = \mu + \phi_3^{[2]}(z,x,\mu)$ back into (2.22). It is easy to see that this will not change the linear and quadratic parts of the dynamics of z and x. Notice that the system (2.22) is in the same form as (2.10). □

In Part I, it is shown that the normal form of a system is completely determined by the invariants. The computation of invariants is straightforward. This implies that, given a control system, the normal form can be found without finding the change of coordinates. A similar result holds for systems with parameters, which is proved in the rest of the section. For the reason of simplicity, we assume that the linearization of the system is in the form of (2.4)–(2.6). The parameter μ is treated as a state variable such that $\dot{\mu} = 0$. In the following, the *extended system* (including the original system and $\dot{\mu} = 0$) is denoted by

$$\dot{x}_e = f_e(x_e) + g_e(x_e)u. \tag{2.23}$$

If a system is in one of the forms given by (2.4)–(2.6), then the extended system (2.23) has state variables

$$x_e = \begin{bmatrix} z & x & \mu \end{bmatrix}^T.$$

Denote by C_z, C_x, X_z, and X_μ the following row and column vectors in $\mathbb{R}^{n+1}$:

$$\begin{aligned} C_z &= \begin{bmatrix} 1 & 0 & 0 & \cdots 0 \end{bmatrix}, \quad C_x = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \end{bmatrix}, \\ X_z &= \begin{bmatrix} 1 & 0 & 0 & \cdots 0 \end{bmatrix}^T, \quad X_\mu = \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 \end{bmatrix}^T. \end{aligned} \tag{2.24}$$

The linearization of the extended system at $z = 0, x = 0, \mu = 0$ is denoted by (A_e, B_e),

$$A_e = \left.\frac{\partial f_e}{\partial x_e}\right|_{z=0,x=0,\mu=0}, \quad B_e = g_e(0).$$

DEFINITION 2.3. *Given a control system satisfying* (1.3)*. Suppose that its linearization is in the form of* (2.4)–(2.6)*. The quadratic invariants are defined to be*

$$\begin{aligned} a_{tr} &= \frac{1}{2}\, C_x A_e^{t-1}[ad_{f_e}^r(g_e), ad_{f_e}^{r-1}(g_e)]\Big|_{z=0,x=0,\mu=0}, \qquad \begin{matrix} 1 \le r \le n-3, \\ 1 \le t \le n-r-2, \end{matrix} \\ \gamma_{x_{n-r}x_{n-r}} &= \frac{1}{2}\, C_z[ad_{f_e}^r(g_e), ad_{f_e}^{r-1}(g_e)]\Big|_{z=0,x=0,\mu=0}, \quad 1 \le r \le n-1, \\ \gamma_{zx_1} &= (-1)^{n-1}\, C_z[X_z, ad_{f_e}^{n-1}(g_e)]\Big|_{z=0,x=0,\mu=0}, \\ \gamma_{x_1\mu} &= (-1)^{n-1}\, C_z[X_\mu, ad_{f_e}^{n-1}(g_e)]\Big|_{z=0,x=0,\mu=0}, \end{aligned} \tag{2.25}$$

and for (2.8)

$$\gamma_{z\mu} = C_z ad_{X_z} ad_{X_\mu}(f)\big|_{z=0,x=0,\mu=0}, \tag{2.26}$$

for (2.9)

$$\gamma_{zz} = \frac{1}{2} C_z ad_{X_z}^2(f)\Big|_{z=0,x=0,\mu=0}, \tag{2.27}$$

for (2.10)

$$\begin{aligned} \gamma_{z\mu} &= C_z ad_{X_z} ad_{X_\mu}(f)\big|_{z=0,x=0,\mu=0}, \\ \gamma_{zz} &= \frac{1}{2} C_z ad_{X_z}^2(f)\Big|_{z=0,x=0,\mu=0}, \\ \gamma_{\mu\mu} &= \frac{1}{2} C_z ad_{X_\mu}^2(f)\Big|_{z=0,x=0,\mu=0}. \end{aligned} \tag{2.28}$$

THEOREM 2.4. *Given a system satisfying* (1.3)*, suppose its linearization is one of* (2.4)–(2.6)*.*

(i) *Quadratic transformations defined by* (2.7) *do not change the values of the quadratic invariants.*

(ii) *The quadratic invariants of normal form* (2.8)–(2.10) *are the corresponding coefficients of the quadratic terms.*

Proof. (i) The quadratic invariants a_{tr}, $\gamma_{x_ix_i}$, γ_{zx_1}, and $\gamma_{x_1\mu}$ are defined in the same way as the invariants for systems without parameters because the parameters

in the system are treated as state variables. Following the same argument used in the proof of Theorem 2.1 in Part I, one can prove that a_{tr}, $\gamma_{x_i x_i}$, $\gamma_{z x_1}$, and $\gamma_{x_1 \mu}$ are invariant under the quadratic change of coordinates and feedback (2.7). For systems in the form of (2.4), $\gamma_{z\mu}$ is the coefficient of $z\mu$ in the uncontrollable dynamics, which is a resonant term (see, for instance, [3]). It cannot be changed by change of coordinates of the form

$$\bar{z} = z + \phi^{[2]}(z, \mu).$$

By the separation principle in [6], the coefficient of resonant term does not change under any quadratic transformation of the form (2.7). Similarly, one can prove that, for system (2.6), γ_{zz}, $\gamma_{z\mu}$, and $\gamma_{\mu\mu}$ are invariant under the quadratic transformations because they are the coefficients of resonant terms. If system (2.5) is under consideration, (2.21) shows that γ_{zz} is invariant under (2.20). By the separation principle in [6], transformations other than (2.20) do not change the coefficient γ_{zz}.

(ii) By the definition of invariants, it is obvious that $\gamma_{z\mu}, \gamma_{zz}$, and $\gamma_{\mu\mu}$ are the coefficients of $z\mu$, z^2, and μ^2, respectively. For the other invariants, we will prove the result for system (2.9). The other two cases are similar. Keep in mind that the state variables are in the order of z, x, μ for the extended system. The Lie brackets of f_e and g_e are

$$\begin{aligned} ad^r_{f_e}(g_e) = (-1)^r[\overbrace{0 \quad \cdots \quad 0 \quad 1}^{n-r} \quad 0 \quad \cdots \quad 0] \\ +(-1)^r[\overbrace{2\gamma_{x_{n-r}x_{n-r}}x_{n-r} \quad 2a_{1\,r}x_{n-r} \quad \cdots \quad 2a_{n-r-2\,r}x_{n-r}}^{n-r-1} \quad 0 \quad \cdots \quad 0] \\ +h_r(x_{n-r+1}, x_{n-r+2}, \ldots, x_{n-1}) + O(z, x, \mu)^2 \end{aligned}$$

for $1 \le r < n-2$. Furthermore,

$$\begin{aligned} ad^{n-2}_{f_e}(g_e) = (-1)^{n-2} \left(\begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \end{bmatrix}^T + \begin{bmatrix} 2\gamma_{x_2x_2}x_2 & 0 & 0 & \cdots & 0 \end{bmatrix}^T \right) \\ +h_{n-2}(x_3, x_4, \ldots, x_{n-1}) + O(z, x, \mu)^2 \end{aligned}$$

and

$$\begin{aligned} ad^{n-1}_{f_e}(g_e) = (-1)^{n-1} \begin{bmatrix} 2\gamma_{x_1x_1}x_1 + \gamma_{zx_1}z + \gamma_{x_1\mu}\mu \\ 0 \\ \vdots \\ 0 \end{bmatrix} \\ +h_{n-1}(x_2, x_3, \ldots, x_{n-1}) + O(z, x, \mu)^2. \end{aligned}$$

Substituting these relations into the right side of (2.25), they are equal to the corresponding coefficients. □

Given a system satisfying assumption (1.3), the linear part of the system can be transformed into a system in one of the forms given by (2.4), (2.5), or (2.6). They have different bifurcation patterns which are addressed in the following sections.

3. Systems with $\lambda \neq 0$. In this section, we assume that the uncontrollable mode λ is not zero. By a linear change of coordinates, such systems can be transformed into (2.4). The main theorem in this section gives a parametrization of the equilibrium

set and the answer to Question 2 in Part I, namely, the controllability of the system at points in E.

THEOREM 3.1. *Consider a system of form* (2.4).

(i) *The equilibrium set E satisfies*

$$\begin{aligned} &x_1 = \nu, \\ &z = O(\nu,\mu)^2, \\ &x_i = O(\nu,\mu)^2, \quad 2 \le i \le n-1. \end{aligned} \tag{3.1}$$

(ii) *There exists a function $c(x_1,\mu)$ in the following form:*

$$c(x_1,\mu) = 2\gamma_{x_1x_1}x_1 + \gamma_{x_1\mu}\mu + O(x_1,\mu)^2 \tag{3.2}$$

such that the system is linearly controllable at $(z,x,\mu) \in E$ if and only if $c(x_1,\mu) \neq 0$.

Remark. The theorem implies that the equilibrium set is a two-dimensional manifold. At the origin, the manifold is tangent to the $x_1\mu$-space. For any fixed μ_0, the set of equilibrium points with $\mu = \mu_0$ is a smooth curve in the state space. Therefore, the equilibrium set does not have bifurcation. However, part (ii) of Theorem 3.1 shows that the controllability of the system changes as the equilibrium points are varied. In fact, if $\gamma_{x_1x_1}^2 + \gamma_{x_1\mu}^2 \neq 0$, the system is linearly controllable at all the equilibrium points in E except a one-dimensional submanifold. This submanifold is tangent to the subspace

$$2\gamma_{x_1x_1}x_1 + \gamma_{x_1\mu}\mu = 0$$

at the origin. If both $\gamma_{x_1x_1}$ and $\gamma_{x_1\mu}$ are zero, the controllability of the system depends on the higher degree terms.

Remark. If $\gamma_{x_1x_1} \neq 0$, then the submanifold $c(x_1,\mu) = 0$ is transversal to the set $E_0 = \{(x,0) \in E\}$. Therefore, the system is always linearly controllable at any equilibrium point in E_0 except the origin. This is actually the result shown in Theorem 4.1 in Part I.

Example. To show a typical example, we consider the following system in normal form:

$$\begin{aligned} \dot z &= -z + 5x_1^2 + x_2^2 + zx_1 + z\mu - 10x_1\mu, \\ \dot x_1 &= x_2, \\ \dot x_2 &= u. \end{aligned}$$

The equilibrium set is

$$x_2 = 0, \quad z = -\frac{5x_1^2 - 10x_1\mu}{-1 + x_1 + \mu}.$$

This manifold is tangent to $x_1\mu$-plane at the origin. In $zx_1\mu$-space, the graph of E is a saddle as shown in Figure 3.1. The curve on the surface E is the subset satisfying the condition $c(x_1,\mu) = 0$, where

$$c(x_1,\mu) = 10x_1 - \frac{5x_1^2 - 10x_1\mu}{-1 + x_1 + \mu} - 10\mu.$$

The system is not linearly controllable at the points on the curve.

Proof of Theorem 3.1. (i) Given any quadratic change of coordinates and feedback (2.7), under the new coordinates, the equations in (3.1) are equivalent to

$$\bar x_1 = \bar\nu, \quad \bar z = O(\bar\nu,\mu)^2, \quad x_i = O(\bar\nu,\mu)^2,$$

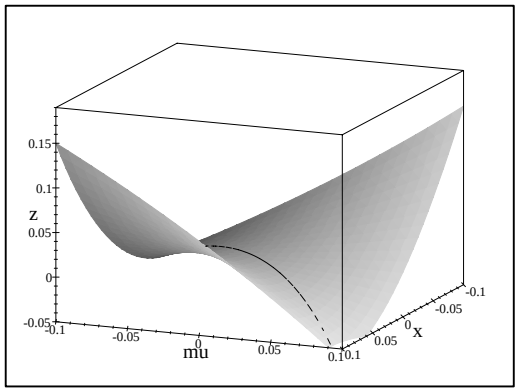


FIG. 3.1. *The equilibrium set in $zx_1\mu$-space.*

which have the same form as (3.1). Therefore, property (i) of Theorem 3.1 is invariant under quadratic transformations. To prove (i), it is enough to show the result for systems in the normal form. Consider system (2.8); it is obvious that a point in the equilibrium set satisfies

$$\begin{aligned} &x_i + O(z,x,\mu)^2 = 0, \quad 2 \le i \le n-1, \\ &\lambda z + O(z,x,\mu)^2 = 0. \end{aligned}$$

The solution of these equations for x_i, $i = 2, \ldots, n-1$ and z is in the form of (3.1).

(ii) Once again, we only prove the result for normal forms since the controllability and the linear part of $c(x_1,\mu)$ are invariant under quadratic transformations. Modulo higher degree terms, the matrix A and B in the linearization of (2.8) at any point $(z,x,\mu) \in E$ satisfies

$$A = \begin{bmatrix} \lambda & 0 \\ 0 & A_2 \end{bmatrix}$$

$$+ \begin{bmatrix} \gamma_{zx_1}x_1 + \gamma_{z\mu}\mu & 2\gamma_{x_1x_1}x_1 + \gamma_{zx_1}z + \gamma_{x_1\mu}\mu & 2\gamma_{x_2x_2}x_2 & 2\gamma_{x_3x_3}x_3 & \cdots & 2\gamma_{x_{n-1}x_{n-1}}x_{n-1} \\ 0 & 0 & 0 & 2a_{1\,n-3}x_3 & \cdots & 2a_{1\,1}x_{n-1} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 2a_{n-3\,1}x_{n-1} \\ 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 \end{bmatrix},$$

$$B = \begin{bmatrix} 0 & 0 & \cdots & 1 \end{bmatrix}^T + O(z,x_1,\mu)^2.$$

From relation (3.1), the matrix A satisfies

$$A = \begin{bmatrix} \lambda & 0 \\ 0 & A_2 \end{bmatrix} + \begin{bmatrix} \gamma_{zx_1}x_1 + \gamma_{z\mu}\mu & 2\gamma_{x_1x_1}x_1 + \gamma_{x_1\mu}\mu & 0 \\ 0 & 0 & 0 \end{bmatrix}_{n\times n} + O(x_1,\mu)^2 \ .$$

Therefore, the controllability matrix at an equilibrium point is

$$\begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix} = \begin{bmatrix} 0 & 2\gamma_{x_1x_1}x_1 + \gamma_{x_1\mu}\mu \\ J & 0 \end{bmatrix} + O(x_1,\mu)^2,$$

where

$$J = \begin{bmatrix} 0 & 0 & \cdots & 1 \\ \vdots & \vdots & & \vdots \\ 0 & 1 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \end{bmatrix}. \tag{3.3}$$

By suitable row and column operations, all the quadratic and higher degree terms can be cancelled except the element at the up right corner; the resulting matrix is

$$R = \begin{bmatrix} 0 & c(x_1, \mu) \\ J & 0 \end{bmatrix},$$

and $c(x_1, \mu)$ satisfies

$$c(x_1, \mu) = 2\gamma_{x_1 x_1} x_1 + \gamma_{x_1 \mu}\mu + O(x_1, \mu)^2.$$

It is obvious that the matrix R has full rank (i.e., the system is linearly controllable) if and only if $c(x_1, \mu) \neq 0$. This proves the second part of the theorem.

4. Systems with $\boldsymbol{\lambda = 0}$ and Jordan form in the uncontrollable dynamics. In this section, we consider system (2.5). In this case, the values of x_1 and z are used as the parameters of the equilibrium set E. Furthermore, since the transformation (2.7) does not change μ, all the quadratic terms in the parametric equation of μ can be found. The bifurcation addressed in the present section has a natural relation with the saddle node bifurcation of dynamic systems. Consider system (2.9) without cubic terms. Then $x = 0$ defines the zero dynamics of the system for output $y = x_1$. From the normal form (2.9), it is easy to show that its zero dynamics have a saddle node bifurcation at the origin.

THEOREM 4.1. *Given a system* (2.5)*, we have the following conditions.*

(i) *Its equilibrium set satisfies*

$$\begin{aligned} &x_1 = \nu_1, \\ &z = \nu_2, \\ &\mu = -\gamma_{x_1 x_1}\nu_1^2 - \gamma_{z x_1}\nu_1\nu_2 - \gamma_{zz}\nu_2^2 + O(\nu_1, \nu_2)^3, \\ &x_i = O(\nu_1, \nu_2)^2, \qquad\qquad 2 \leq i \leq n-1. \end{aligned} \tag{4.1}$$

(ii) *There exists a function $c(z, x_1)$ in the form*

$$c(z, x_1) = \gamma_{z x_1} z + 2\gamma_{x_1 x_1} x_1 + O(z, x_1)^2 \tag{4.2}$$

such that the system is linearly controllable at a point $(z, x, \mu) \in E$ if and only if $c(z, x_1) \neq 0$.

Remark. The projection of E to $zx_1\mu$-space is approximately a quadratic surface. It is a paraboloid or a saddle. In fact, if the matrix

$$Q_1 = \begin{bmatrix} \gamma_{x_1 x_1} & \dfrac{\gamma_{z x_1}}{2} \\ \dfrac{\gamma_{z x_1}}{2} & \gamma_{zz} \end{bmatrix} \tag{4.3}$$

is sign definite ($\det(Q_1) > 0$), E is approximately a paraboloid. If (4.3) is not sign definite but it has full rank ($\det(Q_1) < 0$), E is approximately a saddle. In any case, E has bifurcation near $\mu = 0$. More specifically, we define

$$E_{\mu_0} = \{(x, \mu) \in E | \mu = \mu_0\}. \tag{4.4}$$

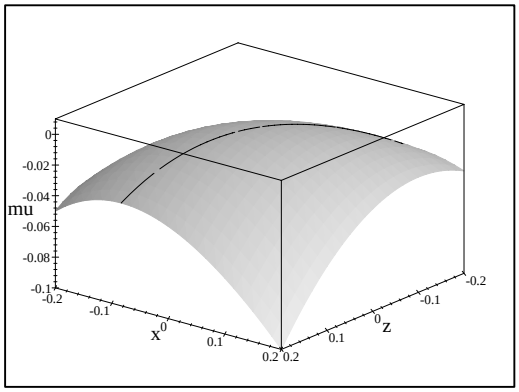


FIG. 4.1. *The equilibrium set in $zx_1\mu$-space.*

Then the topology of E_μ changes as μ passing through zero. If E is approximately a paraboloid, E_μ is empty for the values of μ on one side of zero, and it is a closed curve if μ is on the other side. If E is approximately a saddle, then E_μ is approximately two lines which meet at the origin for $\mu = 0$. It is a connected set. However, E_μ is approximately a hyperbola for $\mu \neq 0$ which is not a connected set. In the following, two examples are given to show the bifurcation diagrams of systems in normal form with $\det(Q_1) > 0$ and $\det(Q_1) < 0$.

Example. Consider the system

$$\begin{aligned} z &= \mu + x_1^2 + x_2^2 + zx_1 + x_1\mu + z^2, \\ \dot{x}_1 &= x_2, \\ \dot{x}_2 &= u. \end{aligned} \tag{4.5}$$

It is easy to check that $\det(Q_1) = 3/4$. Therefore, the graph of E is a paraboloid. In fact, the equilibrium set is

$$x_2 = 0, \quad \mu = -\frac{x_1^2 + zx_1 + z^2}{1 + x_1}.$$

The function $c(z, x_1)$ is

$$c(z, x_1) = 2x_1 + z - \frac{x_1^2 + zx_1 + z^2}{1 + x_1}.$$

The graph of E and the curve $c(z, x_1) = 0$ on E is shown in Figure 4.1. The system is linearly controllable at all points in E except the equilibria on the curve given by $c(z, x_1) = 0$.

From Figure 4.1, the bifurcation of E is obvious. If $\mu > 0$, the set E_μ is empty. If $\mu < 0$, the set E_μ is a closed curve around the origin.

Example. Consider the system

$$\begin{aligned} z &= \mu - x_1^2 + x_2^2 + zx_1 + x_1\mu + z^2, \\ \dot{x}_1 &= x_2, \\ \dot{x}_2 &= u. \end{aligned} \tag{4.6}$$

It is easy to check that $\det(Q_1) = -5/4$. Therefore, the graph of E is a saddle. In fact, the equilibrium set is

$$x_2 = 0, \quad \mu = -\frac{-x_1^2 + zx_1 + z^2}{1 + x_1}.$$

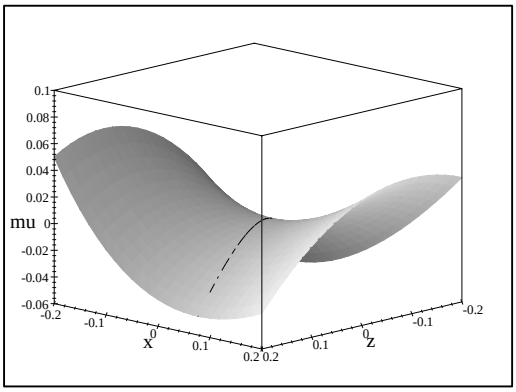


FIG. 4.2. *The equilibrium set in $zx_1\mu$-space.*

The function $c(z, x_1)$ is

$$c(z, x_1) = -2x_1 + z - \frac{-x_1^2 + zx_1 + z^2}{1 + x_1}.$$

The graph of E and the curve $c(z, x_1) = 0$ on E are shown in Figure 4.2. The system is linearly controllable at all points in E except the equilibria on the curve $c(z, x_1) = 0$.

The bifurcation of E is different from the previous example. In this case, E_μ consists of two lines through the origin when $\mu = 0$. If $\mu \neq 0$, the set E_μ is a hyperbola.

Proof of Theorem 4.1. (i) Modulo the higher degree terms in (4.1) $(O(\nu_1, \nu_2)^3$ in the equation of μ and $O(\nu_1, \mu_2)^2$ in the equations of x_i, $i = 2, \ldots, n-1)$, the functions are invariant under transformation (2.7). Therefore, we consider only a system in normal form, which is (2.9). Any point in the equilibrium set satisfies

$$u = O(x_1, \mu)^2, \quad x_i = O(x_1, \mu)^2, \quad 2 \leq i \leq n-1. \tag{4.7}$$

Substituting this relation into the equation

$$\mu + \sum_{i=1}^{n-1} \gamma_{x_i x_i} x_i^2 + \gamma_{z x_1} z x_1 + \gamma_{x_1 \mu} x_1 \mu + \gamma_{zz} z^2 + O(z, x, \mu, u)^3 = 0$$

we get

$$\mu + \gamma_{x_1 x_1} x_1 + \gamma_{z x_1} z x_1 + \gamma_{x_1 \mu} x_1 \mu + \gamma_{zz} z^2 + O(z, x_1, \mu)^3 = 0.$$

Define $x_1 = \nu_1$ and $z = \nu_2$. The equation has a unique solution for μ near $\mu = 0$, and its solution is in the form of (4.1).

(ii) Based on (4.1), the linearization of system (2.9) at a point in E is

$$A = \begin{bmatrix} \gamma_{z x_1} x_1 + 2\gamma_{zz} z & 2\gamma_{x_1 x_1} x_1 + \gamma_{z x_1} z & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & 0 & \cdots & 0 \end{bmatrix} + O(z, x_1)^2, \tag{4.8}$$

$$B = \begin{bmatrix} 0 & 0 & \cdots & 1 \end{bmatrix}^T + O(z, x_1)^2.$$

The controllability matrix is

$$R = \begin{bmatrix} 0 & \gamma_{zx_1} z + 2\gamma_{x_1x_1} x_1 \\ J & 0 \end{bmatrix} + O(z, x_1)^2, \tag{4.9}$$

where J is a matrix defined in (3.3). By elementary matrix operations, the rank of the matrix equals the rank of a matrix in the following form:

$$\begin{bmatrix} 0 & c(z, x_1) \\ J & 0 \end{bmatrix},$$

where the function $c(z, x_1)$ is in the form of (4.2). Therefore, the system is linearly controllable if and only if $c(z, x_1) \neq 0$. □

5. Systems with $\lambda = 0$ and diagonal form in uncontrollable dynamics. Comparing with the previous section, the systems considered in this section have a fundamental difference. The linear part in the uncontrollable dynamic system is zero. To find a parametrization for the set E, the matrix of the quadratic part in the uncontrollable dynamics is very important. This matrix for normal form (2.10) is

$$Q = \begin{bmatrix} \gamma_{zz} & \frac{1}{2}\gamma_{zx_1} & \frac{1}{2}\gamma_{z\mu} \\ \frac{1}{2}\gamma_{zx_1} & \gamma_{x_1x_1} & \frac{1}{2}\gamma_{x_1\mu} \\ \frac{1}{2}\gamma_{z\mu} & \frac{1}{2}\gamma_{x_1\mu} & \gamma_{\mu\mu} \end{bmatrix}. \tag{5.1}$$

The eigenvalues of Q are denoted by d_1, d_2, d_3. Suppose T_1, T_2, T_3 are three column unit eigenvectors of Q associated with the eigenvalues d_1, d_2, d_3, respectively. In the parametrization of E, new variables w_1, w_2, w_3 are used. Their relation with the state variables and parameters are

$$\begin{bmatrix} w_1 & w_2 & w_3 \end{bmatrix}^T = \begin{bmatrix} T_1 & T_2 & T_3 \end{bmatrix}^T \begin{bmatrix} z \\ x_1 \\ \mu \end{bmatrix}. \tag{5.2}$$

THEOREM 5.1. *Given a system satisfying* (1.3)*, suppose its linearization is in the form of* (2.6)*.*

(i) *If Q is positive definite or negative definite, then $(z, x, \nu) = (0, 0, 0)$ is an isolated equilibrium point.*

(ii) *If Q is not sign definite and if it has full rank, then the equilibrium set satisfies*

$$\begin{aligned} w_3 &= \pm\sqrt{-\frac{d_1 w_1^2 + d_2 w_2^2}{d_3}} + O(w_1, w_2)^2, \\ x_i &= O(w_1, w_2)^2, \qquad\qquad 2 \leq i \leq n-1 \end{aligned} \tag{5.3}$$

where d_1 and d_2 represent the two eigenvalues with the same sign and w_i, $i = 1, 2, 3$, is defined by (5.2)*.*

(iii) *If the conditions in* (ii) *are satisfied, then there exists a function $c(z, x_1, \mu)$ in the following form:*

$$c(z, x_1, \mu) = \gamma_{zx_1} z + 2\gamma_{x_1x_1} x_1 + \gamma_{x_1\mu}\mu + O(z, x_1, \mu)^2 \tag{5.4}$$

such that the system is linearly controllable at a point $(z, x, \mu) \in E$ if and only if $c(z, x_1, \mu) \neq 0$.

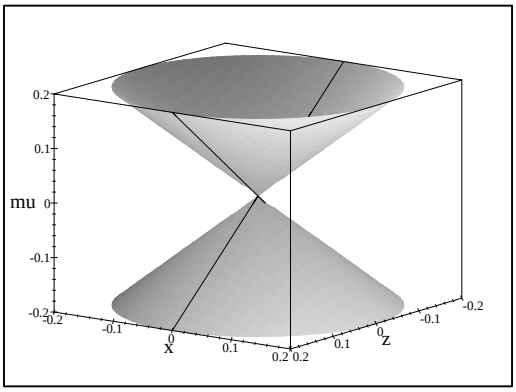


FIG. 5.1. *The equilibrium set in $zx_1\mu$-space.*

Remark. In case (ii), the graph of the equilibrium set in $zx\mu$-plane is approximated by a cone

$$d_1 w_1^2 + d_2 w_2^2 + d_3 w_3^2 = 0.$$

The center line of the cone is parallel to the eigenvector of d_3. For any fixed value of μ, the set of equilibrium points E_μ (defined in (4.4)) is approximately a conic curve. The shape of E_μ depends on the orientation of the cone in $zx_1\mu$-space. Furthermore, for small values of μ, the topology of E_μ at $\mu = 0$ is always different from that of E_μ at $\mu \neq 0$.

The surface $c(z, x_1, \mu) = 0$ is tangent to the plane $\gamma_{zx_1} z + 2\gamma_{x_1x_1} x_1 + \gamma_{x_1\mu}\mu = 0$ in $zx_1\mu$-space. In general, the intersection of such plane with the cone has a single point or it consists of two different lines. In the following, an example is shown in which the center line of the cone is μ-axis. The subset of E satisfying $c(z, x_1, \mu) = 0$ consists of two lines through the origin.

If a system in normal form (2.10) has no cubic and higher degree terms, then $x = 0$ defines its zero dynamics for the output $y = x_1$. Suppose that the eigenvectors of d_3 are not normal to $z\mu$-space, then the intersection between E and zx_1-plane consists of two lines when the bifurcation is a cone. This implies that the zero dynamics have a transcritical bifurcation.

Example. Consider the system

$$\begin{aligned} \dot{z} &= x_1^2 + x_2^2 + z^2 - \mu^2, \\ \dot{x}_1 &= x_2, \\ \dot{x}_2 &= u. \end{aligned}$$

The equilibrium set is

$$x_2 = 0, \quad x_1^2 + z^2 - \mu^2 = 0.$$

The function $c(z, x_1, \mu)$ is $2x_1$. In $zx_1\mu$-space, it is a cone shown in Figure 5.1. The two lines on the cone are the set $c(z, x_1, \mu) = 0$. The system is linearly controllable at all points in E except those on the two lines.

Proof of Theorem 5.1. (i) Consider the system (2.10). The points in equilibrium set satisfy equation (4.7). Substituting (4.7) into the equation

$$\sum_{i=1}^{n-1} \gamma_{x_i x_i} x_i^2 + \gamma_{zx_1} z x_1 + \gamma_{x_1\mu} x_1 \mu + \gamma_{z\mu} z\mu + \gamma_{zz} z^2 + \gamma_{\mu\mu}\mu^2 + O(z, x_1, \mu)^3 = 0$$

we get

$$(5.5)\quad \gamma_{x_1x_1}x_1^2+\gamma_{zx_1}zx_1+\gamma_{x_1\mu}x_1\mu+\gamma_{z\mu}z\mu+\gamma_{zz}z^2+\gamma_{\mu\mu}\mu^2+O(z,x_1,\mu)^3=0.$$

The matrix of the quadratic function in (5.5) is given by (5.1). Therefore, if the matrix Q is positive definite or negative definite, so is the function on the left side of (5.5). It has no nontrivial solution near $(z,x_1,\mu)=(0,0,0)$.

(ii) If the matrix Q in (5.1) is not sign definite, and if it has full rank, then the matrix has three nonzero eigenvalues d_1, d_2, and d_3. Furthermore, we can assume that $d_1d_1>0$. Then d_3 has different sign from d_1 and d_2. By a change of coordinates (5.2), equation (5.5) becomes

$$d_1w_1^2+d_2w_2^2+d_3w_3^2+O(w_1,w_2,w_3)^3=0,$$

solving the equation for w_3. The solution is in the form of (5.3).

(iii) Given an equilibrium point (z,x,μ) in E, keep in mind that x_i, $i=2,\dots,n-1$, has no linear terms of z,x_1,μ. Therefore, the linearization of the system at the point with respect to z,x is

$$A=\begin{bmatrix} \gamma_{zx_1}x_1+\gamma_{z\mu}\mu+2\gamma_{zz}z & 2\gamma_{x_1x_1}x_1+\gamma_{zx_1}z+\gamma_{x_1\mu}\mu & 0 & 0 & \cdots & 0\\ 0 & 0 & 1 & 0 & \cdots & 0\\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & 0 & 0 & \cdots & 1\\ 0 & 0 & 0 & 0 & \cdots & 0 \end{bmatrix}+O(z,x_1,\mu)^2,$$

$$B=\begin{bmatrix} 0 & 0 & \cdots & 1 \end{bmatrix}^T+O(z,x_1,\mu)^2.$$

The controllability matrix is

$$\begin{bmatrix} 0 & 2\gamma_{x_1x_1}x_1+\gamma_{zx_1}z+\gamma_{x_1\mu}\mu\\ J & 0 \end{bmatrix}+O(z,x_1,\mu)^2$$

where J is defined by (3.3). Using elementary row and column operations, it can be proved that the matrix has the same rank as the following matrix R:

$$R=\begin{bmatrix} 0 & 2\gamma_{x_1x_1}x_1+\gamma_{zx_1}z+\gamma_{x_1\mu}\mu+O(z,x_1,\mu)^2\\ J & 0 \end{bmatrix}.$$

Define $c(z,x_1,\mu)$ to be the nonconstant entry

$$c(z,x_1,\mu)=2\gamma_{x_1x_1}x_1+\gamma_{zx_1}z+\gamma_{x_1\mu}\mu+O(z,x_1,\mu)^2.$$

In a neighborhood of the origin, the matrix R has full rank or, equivalently, the system is linearly controllable at the equilibrium point if and only if $c(z,x_1,\mu)\neq 0$. □

One conclusion from the results in sections 3–5 is that the topology and bifurcation of E has five different cases. This is summarized in Table 5.1.

6. Stabilizability. In this section, we prove a theorem on stabilizability of control systems around an uncontrollable equilibrium point. Given a system with controllability index $n-1$ at the origin, if the uncontrollable mode is positive (i.e., the case of $\lambda>0$), then the uncontrollable mode will be positive at all uncontrollable

TABLE 5.1
The classification of equilibrium sets.

Condition	Equilibrium set	Example
$\lambda \neq 0$	smooth 2-d manifold tangent to $x_1\mu$-plane	Figure 3.1
$\lambda = 0$, Jordan form $\det(Q_1) > 0$	paraboloid	Figure 4.1
$\lambda = 0$, Jordan form $\det(Q_1) < 0$	saddle	Figure 4.2
$\lambda = 0$, diagonal form Q is sign definite	single point	
$\lambda = 0$, diagonal form Q is indefinite, $\det(Q) \neq 0$	cone	Figure 5.1

equilibrium points near the origin. On the other hand, if $\lambda < 0$, the system is always stabilizable at all the equilibrium points near the origin. Therefore, the interesting case is $\lambda = 0$. In this section, we focus on the case in which the uncontrollable dynamics have Jordan form, i.e., the case of paraboloid or saddle bifurcations.

If one or both of the quadratic invariants $\gamma_{x_1x_1}$ and γ_{zx_1} are not zero, there is a curve $c(z, x_1) = 0$ in the equilibrium set E such that the system is not linearly controllable at all the points on the curve. In the following, we focus on the problem of feedback stabilization at the uncontrollable equilibrium, which is the set $E \cap \{c(z, x_1) = 0\}$. This set is denoted by $E_u = \{(z, x, \mu) \in E | c(z, x_1) = 0\}$.

THEOREM 6.1. *Consider a system* (2.5). *Suppose* $\det(Q_1) \neq 0$. *If* $\gamma_{x_1x_1}^2 + \gamma_{zx_1}^2 \neq 0$, *then the origin divides the curve* E_u *into two pieces. The system is stabilizable by* C^1 *state feedback on one piece and it is not stabilizable by any* C^1 *state feedback on the other piece. More specifically, the system is stabilizable around a point* (z, x, μ) *in* E_u *if* $\gamma_{zx_1}x_1 + 2\gamma_{zz}z < 0$, *and the system is not stabilizable at a point in* E_u *if* $\gamma_{zx_1}x_1 + 2\gamma_{zz}z > 0$.

Remark. The theorem shows that, in general, the property of stabilizability switches from stabilizable to unstabilizable or vice versa as equilibrium points in E_u pass through the origin. For instance, the system given in (4.5) satisfies $\gamma_{x_1x_1} = 1$, $\gamma_{zx_1} = 1$, and $\gamma_{zz} = 1$. The conditions in Theorem 6.1 are fulfilled. The stabilizable $(x_1 + 2z < 0)$ and unstabilizable $(x_1 + 2z > 0)$ equilibrium points in E_u are shown in Figure 6.1 by solid and dotted curves, respectively.

Proof of Theorem 6.1. It is sufficient to prove the theorem for the normal forms. Consider a system (2.9). Given an equilibrium point (z, x, μ) in E_u, the linearization of the system at the equilibrium point is given by (4.8). Since at least one of $\gamma_{x_1x_1}$ and γ_{zx_1} is not zero, we assume $\gamma_{x_1x_1} \neq 0$ (the proof for the case $\gamma_{zx_1} \neq 0$ is similar). From (4.2), the point (z, x, μ) in E_u satisfies

$$x_1 = -\frac{\gamma_{zx_1}}{2\gamma_{x_1x_1}} z + O(z)^2. \tag{6.1}$$

Substituting this relation into (4.8), we get the linearization of the system at the given equilibrium point:

$$A = \begin{bmatrix} -\dfrac{\gamma_{zx_1}^2}{2\gamma_{x_1x_1}} z + 2\gamma_{zz}z & 0 \\ 0 & A_2 \end{bmatrix} + O(z)^2, \quad B = \begin{bmatrix} 0 & 0 & \cdots & 1 \end{bmatrix}^T + O(z)^2. \tag{6.2}$$

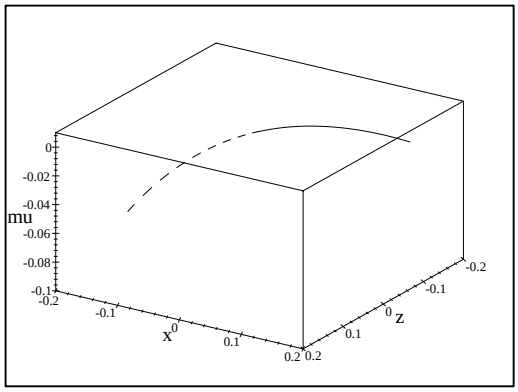


FIG. 6.1. *The set E_u. Dotted line: unstabilizable equilibrium points. Solid line: stabilizable equilibrium points.*

The controllability matrix is in the following form

$$R(z) = \begin{bmatrix} 0 & 0 \\ J & 0 \end{bmatrix} + O(z)^2.$$

In E_u, the rank of the matrix $R(z)$ is $n-1$. There is a vector valued function $C(z) = \begin{bmatrix} c_1(z) & \cdots & c_n(z) \end{bmatrix}$ such that $C(z)R(z) = 0$, $C(0) = [1, 0, \ldots, 0]$. It is easy to check that the function $C(z)$ has the form

$$C(z) = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix} + O(z). \tag{6.3}$$

By the linear control theory, the vector $C(z)$ is normal to the controllability subspace of the linear system (A, B). The uncontrollable mode of the linearization at a point in E_u is a number $\bar{\lambda}$ satisfying

$$C(z)A = \bar{\lambda}C(z). \tag{6.4}$$

From (6.2) and (6.3), relation (6.4) is equivalent to

$$-\frac{\gamma_{zx_1}^2}{2\gamma_{x_1x_1}}z + 2\gamma_{zz}z + O(z)^2 = \bar{\lambda}(1 + O(z)).$$

Therefore, the uncontrollable mode of the linearization at equilibrium point (z, x, μ) in E_u satisfies

$$\bar{\lambda} = \frac{4\gamma_{zz}\gamma_{x_1x_1} - \gamma_{zx_1}^2}{2\gamma_{x_1x_1}}z + O(z)^2. \tag{6.5}$$

Except for $z = 0$, this number is nonzero near the origin because $\det(Q_1) \neq 0$. The sign of the function changes as z passes through zero. From (6.1), on the curve E_u we have

$$\frac{4\gamma_{zz}\gamma_{x_1x_1} - \gamma_{zx_1}^2}{2\gamma_{x_1x_1}}z = \gamma_{zx_1}x_1 + 2\gamma_{zz}z + O(z).$$

This implies that, near the origin, the uncontrollable mode λ of the system at an equilibrium point in E_u has the same sign as $\gamma_{zx_1}x_1 + 2\gamma_{zz}z$. A nonlinear system

is stabilizable by C^1 feedback if the uncontrollable mode is less than zero, and it is not stabilizable by any C^1 state feedback if the uncontrollable mode is greater than zero. Therefore, the system is stabilizable if $\gamma_{zx_1}x_1 + 2\gamma_{zz}z < 0$ and the system is not stabilizable if $\gamma_{zx_1}x_1 + 2\gamma_{zz}z > 0$. Relation (6.5) implies that the property of stabilizability by C^1 feedback changes as z moving across zero. □

7. Conclusion. For control systems satisfying assumption (1.3), their equilibrium sets are classified under changes of coordinates and feedback. There are five different classes. They are summarized in Table 5.1. The equilibrium sets in different classes have either a different topology or a different bifurcation.

Another topic addressed in the paper is the relationship between quadratic invariants and controllability or stabilizability. In general, the uncontrollable equilibrium points form a one-dimensional curve. The tangent line of the curve is uniquely determined by quadratic invariants. Stabilizability at uncontrollable equilibrium points is discussed in section 6. If E is a paraboloid or a saddle, then it is proved that the stabilizability changes as an uncontrollable equilibrium point passes through the origin.

REFERENCES

[1] E. Abed and J.-H. Fu, *Local feedback stabilization and bifurcation control,* I. *Hopf bifurcation*, Systems Control Lett., 7 (1986), pp. 11–17.
[2] E. Abed and J.-H. Fu, *Local feedback stabilization and bifurcation control,* II. *Stationary bifurcation*, Systems Control Lett., 8 (1987), pp. 467–473.
[3] V. I. Arnold, *Geometrical Methods in the Theory of Ordinary Differential Equations*, 2nd ed., Springer-Verlag, Berlin, New York, 1988.
[4] J. Hale and H. Kocak, *Dynamics and Bifurcations*, Springer-Verlag, Berlin, New York, 1991.
[5] W. Kang and A. J. Krener, *Extended quadratic controller normal form and dynamic feedback linearization of nonlinear systems*, SIAM J. Control Optim., 30 (1992), pp. 1319–1337.
[6] W. Kang, *Quadratic normal forms of nonlinear control systems with uncontrollable linearization*, in Proc. IEEE Conf. on Decision and Control, New Orleans, 1995, pp. 608–612.
[7] W. Kang, *Extended controller form and invariants of nonlinear control systems with a single input*, J. Math. Systems Estim. Control, 4 (1994), pp. 253–256.
[8] H. Nijmeijer and A. J. van der Schaft, *Nonlinear Dynamical Control Systems*, Springer-Verlag, Berlin, New York, 1990.
[9] S. Wiggins, *Introduction to Applied Nonlinear Dynamical Systems and Chaos*, Springer-Verlag, Berlin, New York, 1990.
[10] W. Kang, *Bifurcation and normal form of nonlinear control systems, Part* I, SIAM J. Control Optim., 36 (1998), pp. 193–212.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 1, pp. 233–252, January 1998

013

OPTIMAL RESIDENCE TIME CONTROL OF HAMILTONIAN SYSTEMS PERTURBED BY WHITE NOISE*

JAMES P. DUNYAK† AND MARK I. FREIDLIN‡

Abstract. Optimal control of perturbed Hamiltonian systems in $\Re^2$ is studied. Systems are considered with a control term scaling with the size of a small perturbing noise. The dynamics are shown to converge in a certain sense to a diffusion on a graph. Using the approach developed in [M. I. Freidlin and A. D. Wentzell, *Mem. Amer. Math. Soc.*, 109 (1994), pp. 1–82] and [M. I. Freidlin and A. D. Wentzell, *Ann. Probab.*, 21 (1993), pp. 2215–2245] for random perturbations of Hamiltonian systems, a convergence theorem is discussed. An optimal control theorem is then developed to maximize the expected exit time from a domain. This control is asymptotically robust for small noise. Several examples are provided.

Key words. control, Hamiltonian, stochastic

AMS subject classification. 49B60

PII. S0363012995291658

1. Introduction.

Consider a Hamiltonian system in the plane

$$\begin{aligned} \dot{\tilde{X}}_t &= \overline{\nabla} H(\tilde{X}_t), \\ \tilde{X}_0 &= (\tilde{X}_0^1, \tilde{X}_0^2) = x \ \in \Re^2, \\ \overline{\nabla} H(x) &= \left[\frac{\partial H}{\partial x_2}(x), -\frac{\partial H}{\partial x_1}(x) \right]. \end{aligned} \tag{1}$$

Let $H(x) \to \infty$ as $|x| \to \infty$. The trajectories of this system describe a collection of oscillations which may be grouped in families based on the specific extrema of H they circle. For example, let $H(x)$ be shown in Figure 1(a) and the corresponding state-space trajectories be shown in Figure 1(b). The trajectories or oscillations in Figure 1(b) can be grouped into three families. The ∞-shaped curve Γ_4 is the separatrix, which goes through the Hamiltonian function's saddle point. Two families are inside the ∞-shaped curve Γ_4: the family circling O_1 (which includes the curve Γ_1) and the family circling O_3 (which includes the curve Γ_3). The third family of oscillations encircles the entire ∞-shaped curve Γ_4, including the points O_1, O_2, and O_3. The trajectory Γ_2 is in this family.

Of course, if there is no noise, the system will preserve the oscillations determined by the initial integral $H(x)$. Our interest is in the situation with a small perturbing noise and a similarly scaled control. Let the white noise $\sqrt{\epsilon}\dot{\tilde{W}}_t$ perturb the system, $0 < \epsilon \ll 1$,

$$\begin{aligned} \dot{\tilde{X}}_t^\epsilon &= \overline{\nabla} H(\tilde{X}_t^\epsilon) + \epsilon^{1/2} \dot{\tilde{W}}_t, \\ \tilde{X}_0^\epsilon &= x \ \in \Re^2. \end{aligned}$$

*Received by the editors September 11, 1995; accepted for publication (in revised form) October 31, 1996. This research was supported in part by ARO grant DAAL03-92-G-0219.
http://www.siam.org/journals/sicon/36-1/29165.html

†Department of Mathematics, Texas Tech University, Lubbock, TX 79409 (dunyak@math.ttu.edu). The research of this author was supported in part by the Texas Advanced Research Program.

‡Department of Mathematics, University of Maryland, College Park, MD 20742 (mif@athena.umd.edu).

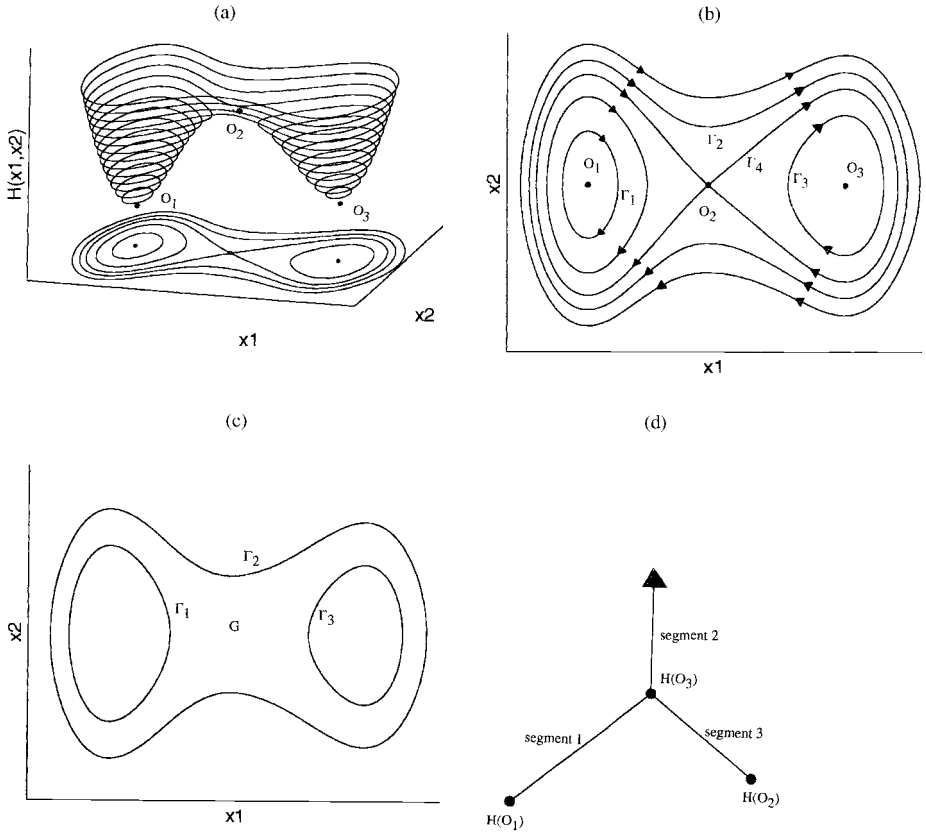


FIG. 1.

Then the perturbed trajectory will, with probability 1, sooner or later leave any bounded domain. Suppose we would like to preserve the oscillations of not very large and not very small amplitudes as long as possible. For example, a domain G is shown in Figure 1(c) bounded by the unperturbed system trajectories Γ_1, Γ_2, and Γ_3, and our goal is to keep the trajectory of the perturbed system inside G as long as possible. Suppose that to achieve this goal we can control the vector field in the dynamical system, and the size of the control is also scaled by ϵ:

$$\begin{aligned} \dot{\tilde{X}}_t^{\epsilon,c} &= \overline{\nabla} H(\tilde{X}_t^{\epsilon,c}) + \epsilon\, c(\tilde{X}_t^{\epsilon,c}) + \epsilon^{1/2}\dot{\tilde{W}}_t, \\ \tilde{X}_0^{\epsilon,c} &= x \ \in \Re^2. \end{aligned} \tag{2}$$

It can be shown that the exit time from the domain G is of order $\frac{1}{\epsilon}$. To make it of order 1 as $\epsilon \downarrow 0$ we rescale time $\frac{t}{\epsilon} \to t$. Then the equation for $X_t^{\epsilon,c} = \tilde{X}_{t/\epsilon}^{\epsilon,c}$ has the form

$$\begin{aligned} \dot{X}_t^{\epsilon,c} &= \frac{1}{\epsilon}\overline{\nabla} H(X_t^{\epsilon,c}) + c(X_t^{\epsilon,c}) + \dot{W}_t, \\ X_0^{\epsilon,c} &= x \ \in \Re^2. \end{aligned} \tag{3}$$

We assume that the vector field $c(x)$ is twice continuously differentiable (with uniformly bounded derivatives) everywhere except possibly at points located on a

finite number of trajectories of the nonperturbed system in equation (1). Krylov [13], among others, discusses stochastic differential equations with coefficients with simple discontinuities [14]. To bound the magnitude of the control, we allow only c such that $c^t(x)M(x)c(x) \leq K^2$ for some constant $K > 0$ and some matrix function $M(x)$ which is uniformly positive definite over G. $M(x)$ may be taken as symmetric since $c^t(x)M(x)c(x) = \frac{1}{2}c^t(x)(M(x)+M^t(x))c(x)$ for any matrix M. $M(x)$ is also assumed to be twice continuously differentiable with bounded derivatives. Such control vector fields we call permissible, and we denote the class of permissible fields by Π_K. This class Π_K is much narrower than is actually required but allows more direct application of the results in [1, 2]. The concluding section addresses this issue.

Let $\tau^{\epsilon,c}$ be the exit time from the domain G; $\tau^{\epsilon,c} = \min\{t : \ X_t^{\epsilon,c} \notin G\}$. Our goal is to make $E_x\tau^{\epsilon,c}$ as big as possible by choosing the control $c \in \Pi_K$. One can, of course, consider this problem for a fixed ϵ. If $c^{*,\epsilon} = c^{*,\epsilon}(x) \in \Pi_K$ is the optimal control and $V^{*,\epsilon}(x) = E_x\tau^{\epsilon,c^*}$, then one can write (for $M(x) = I$) the Bellman equation for $V^*(x)$:

$$\frac{1}{2}\Delta V^{*,\epsilon}(x) + \frac{1}{\epsilon}\overline{\nabla}H(x)\cdot\nabla V^{*,\epsilon}(x) + K|\nabla V^{*,\epsilon}(x)| = -1, \quad x \in G, \tag{4}$$

$$V^{*,\epsilon}(x)|_{\partial G} = 0.$$

Then $c^{*,\epsilon}(x) = \nabla V^{*,\epsilon}(x)$ and $V^{*,\epsilon}(x) = E_x\tau^{\epsilon,c^*} \geq E_x\tau^{\epsilon,c}$ for all $c \in \Pi_K$. Equation (4) requires the solution of a nonlinear boundary value problem in $\Re^2$ for each value of ϵ. We show in this paper how this problem may be replaced, for small ϵ, with an ordinary differential equation that is independent of ϵ. If $0 < \epsilon \ll 1$, one can give a more explicit answer which may be more useful in application. One can calculate a control $c^*(x) \in \Pi_k$ such that for any $c \in \Pi_K$,

$$E_x\tau^{\epsilon,c^*} \geq E_x\tau^{\epsilon,c}$$

for $x \in G$ and ϵ small enough. So the control c^* may be considered optimal in an asymptotic sense; c^* will outperform any other control if the noise is small enough.

We should comment first on the relative scaling of the noise and control. Of course, our theoretical result is a limit as $\epsilon \to 0$, but in applications we expect to have a fixed noise size (which implies a fixed ϵ). If this noise is "small," then the appropriately normalized exit time should be well approximated by the theoretical limit. This measure of "smallness" is in practice easy to recognize: the perturbed dynamical system should make many oscillations (very nearly following the deterministic unperturbed trajectories) before the noise causes much change in the value of the Hamiltonian $H(\tilde{X}_t^\epsilon)$. Thus, for short times, the Hamiltonian dynamics will clearly dominate the noise perturbations.

The control is also scaled by ϵ in equation (2). Again, in applications we anticipate a fixed (but "small") value of ϵ to be in effect. Scaling both noise and control together reflects the fact that they are both small and both approximately of the same order of magnitude. Example 1 below provides an example of how our optimum control performs for a fixed (but "small") size for the noise and control.

Of course, other relative scalings of the noise and control are possible and in some cases might be more appropriate. If the control is not small, then the underlying Hamiltonian dynamics are corrupted. These asymptotic control problems involving residence time and probability control have been considered before for non-Hamiltonian systems. Freidlin and Wentzell used large deviation theory to analyze

residence time control problems for non-Hamiltonian systems in [3, 4]. Results of this type were also discussed by Fleming and Souganidis in [5] and extended by Dupuis and Kushner [6] to minimizing escape probabilities in a special class of degenerate noise. Minimizing expected escape time or probability using large deviations for the linear, constant coefficient equation has been explored in detail by Meerkov, Runolfsson, and Kim in [7, 8, 9, 10, 11]. Kappos [12] also considered similar problems. Nonasymptotic stochastic control problems in $\Re^N$, allowing progressively measurable controls and more general performance measures, have been discussed in detail by Krylov [13] and in many other references. In general, difficult computations are involved in solving two-dimensional optimal stabilization problems using any of these methods, while the methods of this paper (applicable only in the Hamiltonian case) are quite implementable. The notable exception is in case of linear dynamics which is so effectively analyzed by Meerkov, Runolfsson, and Kim.

For Hamiltonian systems the exit from the domain is not a large deviation, and the large deviation approach is too rough for such problems. A graph Γ can be constructed that is homeomorphic to the set of connected components of the level sets of the Hamiltonian. The vertices of the graph correspond to the critical points of the Hamiltonian: exterior vertices correspond to extrema, interior vertices to the saddle points and their associated ∞-shaped trajectories. Interior points of the graph segments correspond to the periodic trajectories of the system. Our example in Figure 1 illustrates the map. Associated with the three trajectory families is the graph shown in Figure 1(d). The ∞-shaped separatrix Γ_4 is mapped to the joint in the center of the graph. The lower-left-hand segment corresponds to the family of trajectories encircling O_1, with the equilibrium O_1 mapping to the lower segment endpoint. Hence the curve Γ_1 maps to a point on segment 1. Similarly, the lower right segment in Figure 1(d) corresponds to oscillations circling O_3, with O_3 mapping to the lower segment endpoint. The curve Γ_3 maps to a point on segment 3. The oscillations circling the ∞- shaped curve Γ_4 map to the upper segment in the graph, so that Γ_2 maps to segment 2. Such a graph for description of the perturbed Hamiltonian system was introduced by Freidlin and Wentzell [1, 2], and we refer to this paper for a rigorous definition. Each point $x \in \Re^2$ with $H(x) = H$ belongs to a connected component $C_i(H)$ of the level set $\{x : \ H(x) = H\}$. Such a set can consist of one or several connected components. Let $Y(x) : \ x \to (i(x), H(x))$ be the mapping from $\Re^2$ into Γ. Then, for $x \in C_i(H)$, $Y(x)$ is the point of the graph Γ corresponding to $C_i(H)$.

As is shown in [1, 2], the stochastic process $Y(X_t^{\epsilon,0}) = Y_t^{\epsilon,0}$ in the case $c(x) = 0$ converges weakly to a diffusion process on Γ. Inside each edge the process is defined by averaging the diffusion in the plane along the periodic unperturbed Hamiltonian trajectories. At the vertices some gluing conditions should be added (see below). Let $\Gamma_G = Y(G)$ as shown, for example, in Figures 1(c) and 1(d). By a modification of the proof given in [1] we show here that if the drift $c(x)$ is added, as in equation (3), $Y_t^{\epsilon,c} = Y(X_t^{\epsilon,c})$ also converges weakly to a Markov process Y_t^c on the graph Γ. Then we can check that

$$\lim_{\epsilon \downarrow 0} E_x \tau^{\epsilon,c} = E_{Y(x)} \tau^c,$$

where τ^c is the exit time for Y_t^c from the set Γ_G. The function $E_{Y(x)}\tau^c = v(x)$ is the unique solution of a simple ordinary differential equation on Γ satisfying some natural boundary conditions and gluing conditions at the vertices. Using this fact, one can calculate explicitly the control c^* maximizing the $E_z\tau^c$, $z \in \Gamma_G$ in $c \in \Pi_K$. It turns out that in general $|c^*| = K$ everywhere, and c^* changes sign at a finite number of

points on the graph Γ. The number of these change points is less than the number of edges. A simple procedure to calculate these change points can be given in many cases. Then one can prove that the control

$$c^*(x) = K \ \ \text{sign}(v'(i(x), H(x))) \ \ M^{-1}(x)\nabla H(x)/(\nabla H^t(x) M^{-1}(x)\nabla H(x))^{1/2}$$

is asymptotically the best. Note, as expected, that this may be viewed as a "bang-bang" control law because the control is always saturated with $c^t Mc = K^2$. Here $v(i(x), H(x))$ is the solution to a differential equation on the graph discussed below. For each $c \in \Pi_K$ there is an ϵ_0 dependent on c with $E_x \tau^{\epsilon, c^*} \geq E_x \tau^{\epsilon, c}$ for any $\epsilon \leq \epsilon_0$.

2. The convergence theorem. Formulation of the control problem requires an appropriate convergence theorem. The theorem here is a modification of the convergence theorem in [1], allowing for the discontinuous drift term.

We use the following notation (see [1]):

I_i is the interior of a segment of the graph.

O_k is a vertex or joint on the graph.

D_i is the set of all points $x \in \Re^2$ such that $Y(x)$ belongs to the interior of segment I_i.

$C_k = \{x : \ Y(x) = O_k\}$.

$C_{ki} = C_k \bigcap \partial D_i$.

$C(H) = \{x : \ H(x) = H\}$ for H in the range of $H(x)$.

$C_i(H) = \{x \in D_i : \ H(x) = H\}$.

$D_i(H_1, H_2) = \{x \in D_i : \ H_1 < H(x) < H_2\}$.

$D_k(\pm\delta)$ is the connected component of $\{x : \ H(O_k) - \delta < x < H(O_k) + \delta\}$ containing C_k.

$C_{ki}(\delta) = \{x \in D_i : \ H(x) = H(O_k) \pm \delta\}$.

The time-scaled equation (3) is addressed in the theorem. Note that, for small ϵ, the trajectory of $X_t^{\epsilon, c}$ will circle many times before $H(X_t^{\epsilon, c})$ will change significantly. This leads to an averaging effect, which defines the limiting process inside the edges. The limiting process is a diffusion process on the graph Γ corresponding to the Hamiltonian.

This limiting diffusion will be described first. See [1, 2] for a technical description of processes on graphs. Associated with each point on each segment I_i on Γ there is a connected curve $C_i(H) \subset \Re^2$. This curve is a portion of the level set $\{x : \ H(x) = H\}$. With each segment I_i of the graph, associate a differential operator

$$L_i^c f(i, H) = \frac{1}{2} A_i^c(H) f''(i, H) + B_i^c(H) f'(i, H) \tag{5}$$

with

$$A_i^c(H) = \frac{\oint_{C_i(H)} \frac{|\nabla H(x)|^2}{|\overline{\nabla} H(x)|} dl}{\oint_{C_i(H)} \frac{1}{|\overline{\nabla} H(x)|} dl} \tag{6}$$

and

$$B_i^c(H) = \frac{\oint_{C_i(H)} \frac{\frac{1}{2}\Delta H(x) + \nabla H(x) \cdot c(x)}{|\overline{\nabla} H(x)|} dl}{\oint_{C_i(H)} \frac{1}{|\overline{\nabla} H(x)|} dl}. \tag{7}$$

Here $f'(i, \cdot)$ indicates the derivative with respect to the local coordinate H on the ith segment.

Associate with each graph joint a gluing condition

$$\sum_{i:I_i\sim O_k} (\pm\beta_{ki}) f'(i, H(x_k)) = 0. \tag{8}$$

The sign $\pm$ of the coefficients is given by $\pm\beta_{ki} \geq 0$ if $H \geq H(x_k)$ on I_i and $\pm\beta_{ki} \leq 0$ if $H \leq H(x_k)$ on I_i. These coefficients are computed as

$$\beta_{ki} = \int_{C_{ki}} |\nabla H(x)| dl$$

for curve C_{ki} corresponding to the portion of C_k in the inverse image of graph segment I_i. For example, in Figure 1 with the joint associated with the saddle point O_2 called joint 2, curve C_{21} is the left half of the ∞-shaped curve through O_2, curve C_{22} is the entire ∞-shaped curve, and curve C_{23} is the right half of the ∞-shaped curve.

The outer vertices must also be considered. If the integral

$$\int \exp\left[-\int \frac{2B_i^c(H)}{A_i^c(H)} dH\right] dH \tag{9}$$

diverges at the end $H(x_k)$, then the vertex is inaccessible. Noting that

$$|\nabla H(x)|^2 / |\overline{\nabla} H(x)| \;=\; |\nabla H(x)|$$

and the bound on the magnitude of $c(x)$, for $H \in I_i$,

$$\int \frac{2 \frac{\oint_{C_i(H)} \frac{\nabla H(x)\cdot c(x)}{|\overline{\nabla} H(x)|} dl}{\oint_{C_i(H)} \frac{1}{|\overline{\nabla} H(x)|} dl}}{A_i^c(H)} dH$$

is bounded. The article [1] establishes that the outer vertices for the uncontrolled system are inaccessible. As a result, the integral in (9) diverges, and the addition of control does not alter the accessibility of the outer vertices.

THEOREM 1. *Let the Hamiltonian $H(x)$, $x \in \Re^2$, be four times continuously differentiable; $H(x) \geq A_1|x|^2$, $|\nabla H(x)| \geq A_2|x|$, $\Delta H(x) \geq A_3$ for sufficiently large $|x|$, where A_1, A_2, A_3 are positive constants. Let $H(x)$ have a finite number of critical points x_1, x_2, ..., x_N at which the matrix of second derivatives $(\frac{\partial^2 H(x)}{\partial x^i \partial x^j})$ is nondegenerate. Also assume that each level curve, $C_k = \{x : Y(x) = Y(x_k)\}$, contains at most one critical point.*

Let $c(x)$, $x \in \Re^2$, meet two conditions: c is twice continuously differentiable (with uniformly bounded derivatives) except at points in a finite number of level sets of $H(x)$, and $|c(x)| \leq K$ for some positive constant K.

Let $(X_t^{\epsilon,c}, P_x^\epsilon)$ be the diffusion process on $\Re^2$ connected with problem (3)*, and define $Y_t^{\epsilon,c} = Y(X_t^{\epsilon,c})$ as the corresponding process on graph Γ. Let Y_t^c be the process on the graph corresponding to the differential operator on the graph Γ in problems* (5) *through* (8)*.*

Then the process $Y_t^{\epsilon,c}$ converges weakly to the Markov process Y_t^c.

The proof of Theorem 1 is essentially the same as the proof of Theorem 2.2 in [1], so the reader is directed to the reference for the proof. Note that, for a trajectory in the interior of a segment, $X_t^{\epsilon,c}$ rotates many times along the trajectories before

$H(X_t^{\epsilon,t})$ changes much. Consider Ito's equation for the value of the Hamiltonian. Since $\nabla H(X_t^{\epsilon,t}) \cdot \overline{\nabla} H(X_t^{\epsilon,t}) = 0$,

$$H(X_t^{\epsilon,t}) = H(X_t^{\epsilon,0}) + \int_0^t \nabla H(X_t^{\epsilon,s}) \cdot dW_s + \int_0^t \left(\frac{1}{2} \Delta H(X_t^{\epsilon,s}) + \nabla H(X_t^{\epsilon,t}) \cdot c(H(X_t^{\epsilon,s})) \right) ds. \tag{10}$$

The resulting averaging, in light of equation (10), will lead to the differential equation for the interior of the segment. This is proved rigorously for $c = 0$ in [1]; the proofs require little modification to handle the additional control term.

The joints provide more complications. The addition of the control does not change the gluing conditions. This may also be shown following the proof in [1] but noting that the proof uses a series of five lemmas, labeled 3.1 through 3.5. The last four require some modification in the method of proof. The new proofs follow by repeated application of the Girsanov–Cameron–Martin theorem relating the measure induced by X_t^ϵ and the uncontrolled trajectory Z_t^ϵ with

$$\dot{Z}_t^\epsilon = \frac{1}{\epsilon} \overline{\nabla} H(Z_t^\epsilon) + \dot{W}_t, \qquad Z_0^\epsilon = x, \tag{11}$$

and

$$\frac{d\mu_{Z_x^\epsilon}}{d\mu_{X_x^{\epsilon,c}}}(X_\cdot) = \exp\left\{ - \int_0^1 c(Z_t^\epsilon) \cdot dW_t - \frac{1}{2} \int_0^1 |c(Z_t^\epsilon)|^2 dt \right\}. \tag{12}$$

Results for Z_t^ϵ in equation (11) from [1] are then applied in the controlled case via equation (12).

COROLLARY 1. *Let the Hamiltonian and control $c(x)$ meet the conditions in Theorem* 1.

Let G_Γ be a closed and connected set on the graph Γ corresponding to the Hamiltonian $H(x)$, and let $G = Y^{-1}(G_\Gamma)$ be the corresponding inverse image in $\Re^2$. Assume G is a bounded set.

Then, for $x \in G$ with

$$\tau^{\epsilon,c} = \inf\{t : X_t^{\epsilon,c} \in \partial G\},$$

and

$$\tau^c = \inf\{t : Y_t^c \in \partial G_\Gamma\},$$

$$\lim_{\epsilon \to 0} E_x\{\tau^{\epsilon,c}\} = E_{Y(x)}\{\tau^c\}.$$

Proof. From Ito's formula,

$$H(X_t^{\epsilon,c}) = H(X_0^{\epsilon,c}) + \int_0^t \nabla H(X_s^{\epsilon,c}) \cdot dW_s + \int_0^t \left(\frac{1}{2} \Delta H(X_s^{\epsilon,c}) + \nabla H(X_s^{\epsilon,c}) \cdot c(X_s^{\epsilon,c}) \right) ds.$$

For a region U such that the closure of $Y(U)$ contains no vertices, standard results bound the exit time from U for all $\epsilon \in (0, 1]$ and $x \in U$ (see Theorem 4 of Chapter

2, [13]). Lemma 3.4 in [1] gives an upper bound for exit from a domain belonging to a neighborhood of a vertex. These are used to establish $E_x\{\tau^{\epsilon,c}\} < A < \infty$ for any $\epsilon \in (0,1]$, $x \in G$.

Now, from Theorem 1, for $T > 0$,

$$\lim_{\epsilon\to 0} E_x\left\{\tau^{\epsilon,c} \bigwedge T\right\} = E_x\left\{\tau^{c} \bigwedge T\right\}.$$

Also, using the Markov property through conditioning on $X_T^{\epsilon,c}$ and Chebyshev's inequality,

$$\begin{aligned} 0 \le E_x\{\tau^{\epsilon,c}\} - E_x\{\tau^{\epsilon,c} \bigwedge T\} &= E_x\{(\tau^{\epsilon,c} - \tau^{\epsilon,c} \bigwedge T)\, I_{\tau^{\epsilon,c}>T}\} \\ &\le A\, P_x\{\tau^{\epsilon,c} > T\} \\ &\le A^2/T. \end{aligned}$$

We see that

$$\lim_{T\to\infty} E_x\left\{\tau^{\epsilon,c} \bigwedge T\right\} = E_x\{\tau^{\epsilon,c}\}$$

uniformly in $\epsilon \in (0,1]$. Then immediately

$$\lim_{T\to\infty}\lim_{\epsilon\to 0} E_x\left\{\tau^{\epsilon,c} \bigwedge T\right\} = \lim_{\epsilon\to 0}\lim_{T\to\infty} E_x\left\{\tau^{\epsilon,c} \bigwedge T\right\}.$$

The corollary follows from Theorem 1 and the dominated convergence theorem. □

3. Solving the control problem. Theorem 1 provides the needed tool for solution of the control problem.

THEOREM 2. *Let G_Γ be a closed and connected set on the graph Γ corresponding to the Hamiltonian $H(x)$, and let $G = Y^{-1}(G_\Gamma)$ be the corresponding inverse image in $\Re^2$. Assume G is a bounded set.*

Let $H(x)$, $x \in \Re^2$, be four times continuously differentiable in G. Let $H(x)$ have a finite number of critical points $x_1, x_2, \ldots, x_N$ in G at which the matrix of the second derivatives is nondegenerate. Also assume that each level curve $C_k = \{x : Y(x) = Y(x_k)\}$ contains at most one critical point.

Let $v(i,y)$ be a solution, continuous and twice differentiable in the interior of the graph segments, of

$$\begin{aligned} &\frac{1}{2}A_i v''(i,y) + B_i v'(i,y) + KE_i|v'(i,y)| + 1 = 0, \qquad (i,y) \in G_\Gamma, \\ &\textstyle\sum_{i:I_i\sim O_k}(\pm\beta_{ki})v'(i,H(y_k)) \quad = 0 \quad \textit{for all interior vertices}, \\ &v(i,y) = 0 \quad \textit{on the boundary } \partial G_\Gamma \end{aligned} \tag{13}$$

with

$$A_i(H) = \frac{\oint_{C_i(H)} |\nabla H(x)| dl}{\oint_{C_i(H)} \frac{1}{|\overline{\nabla} H(x)|} dl}, \qquad B_i(H) = \frac{\oint_{C_i(H)} \frac{\frac{1}{2}\Delta H(x)}{|\overline{\nabla} H(x)|} dl}{\oint_{C_i(H)} \frac{1}{|\overline{\nabla} H(x)|} dl},$$

$$E_i(H) = \frac{\oint_{C_i(H)} \frac{(\nabla H^t(x) M^{-1}(x) \nabla H(x))^{1/2}}{|\overline{\nabla} H(x)|} dl}{\oint_{C_i(H)} \frac{1}{|\overline{\nabla} H(x)|} dl},$$

and

$$\beta_{ki} = \int_{C_{ki}} |\nabla H(x)| dl.$$

The sign $\pm$ *of the coefficients meet* $\pm\beta_{ki} \geq 0$ *if* $H \geq H(x_k)$ *on* I_i *and* $\pm\beta_{ki} \leq 0$ *if* $H \leq H(x_k)$ *on* I_i.

Define the optimal control $c^*(x)$ *by*

$$c^*(x) = K \ \ sign(v'(Y(x))) \ \ \frac{M^{-1}(x)\nabla H(x)}{(\nabla H^t(x) M^{-1}(x) \nabla H(x))^{1/2}} \tag{14}$$

and the set of permissible controls Π_K *as the set of functions* c *meeting two conditions:* c *is twice continuously differentiable (with uniformly bounded derivatives) in* G *except at points in a finite number of level sets of* $H(x)$*; and* $c^t(x)M(x)c(x) \leq K^2$ *for constant* $K > 0$*, with* $M(x)$ *twice continuously differentiable (with bounded derivatives),* $M(x)$ *uniformly positive definite on* $x \in G$*, and* $M(x)$ *assumed (without loss of generality) to be symmetric. Then, with*

$$\tau_x^{\epsilon,c} = \inf\{t : X_t^{\epsilon,c} \in \partial G\},$$

for each $c(x) \in \Pi_K$ *there exists an* ϵ_0 *(dependent on* c*) such that*

$$E\{\tau_x^{\epsilon,c}\} \leq E\{\tau_x^{\epsilon,c^*}\}$$

for all $\epsilon \leq \epsilon_0$*. The optimal control* c^* *is independent of initial conditions* x *in the interior of* G.

Comment. Note that all permissible control vector fields which are equal almost everywhere (in the Lebesgue sense) are considered equivalent.

Proof. The theorem follows easily from Theorem 1 and a representation of the solution to the diffusion equation on the graph as the expectation of a functional of the process on the graph. Consider the diffusion process on the graph governed by the operators (5) inside the edges and the gluing conditions (8) on the vertices. Let $\tau_y^{\epsilon,c}$ be the hitting time of the set G_Γ. Note that

$$u^c(i,y) = E\{\tau_y^c\} \tag{15}$$

is the solution of

$$\begin{gathered} L_i^c u^c(i,y) + 1 = 0, \qquad (i,y) \in G_\Gamma, \\ \textstyle\sum_{i:I_i \sim O_k} (\pm\beta_{ki}) u^{c\prime}(i, H(y_k)) = 0 \quad \text{for all interior vertices,} \\ u^c(i,y) = 0 \quad \text{on the boundary } \partial G_\Gamma \end{gathered} \tag{16}$$

(see [2]). Consider $\text{sign}(v'(i,x))$ as a fixed function and apply equations (13) and (16) to see that $u^{c^*}(i,x) = v(i,x)$. Let, for $c \in \Pi_K$,

$$e^c(i,y) = u^{c^*}(i,y) - u^c(i,y),$$

which is the solution to

$$L_i^c e^c(i,y)$$

$$+\left[K\frac{\oint_{C_i(y)}\frac{(\nabla H^t(x)M^{-1}(x)\nabla H(x))^{1/2}}{|\overline{\nabla}H(x)|}dl}{\oint_{C_i(y)}\frac{1}{|\overline{\nabla}H(x)|}dl}|v'(i,y)| - \frac{\oint_{C_i(y)}\frac{\nabla H(x)\cdot c(x)}{|\overline{\nabla}H(x)|}dl}{\oint_{C_i(y)}\frac{1}{|\overline{\nabla}H(x)|}dl}v'(i,y)\right]$$

$$= 0 \quad \text{for } (i,y)\in G_\Gamma,$$

$$\textstyle\sum_{i:I_i\sim O_k}(\pm\beta_{ki})e^c{}'(i,H(y_k)) = 0 \quad \text{for all interior vertices,}$$

$$e^c(i,y) = 0 \quad \text{on the boundary } \partial G_\Gamma.$$

Hence there is a functional integral representation for $e^c(i,y)$ with, for appropriate stochastic process (i_t, Z_t) on the graph,

$$e^c(i,y) = E_{i,y}\left\{\int_0^\tau\left[K\frac{\oint_{C_i(Z_t)}\frac{(\nabla H^t(x)M^{-1}(x)\nabla H(x))^{1/2}}{|\overline{\nabla}H(x)|}dl}{\oint_{C_i(Z_t)}\frac{1}{|\overline{\nabla}H(x)|}dl}|v'(i_t,Z_t)|\right.\right.$$

$$\left.\left. - \frac{\oint_{C_i(Z_t)}\frac{\nabla H(x)\cdot c(x)}{|\overline{\nabla}H(x)|}dl}{\oint_{C_i(Z_t)}\frac{1}{|\overline{\nabla}H(x)|}dl}v'(i_t,Z_t)\right]dt\right\}. \tag{17}$$

The requirement that the control vector satisfies $c^t(x)M(x)c(x)\le K^2$ establishes that $e^c(i,y)>0$ when $c\neq c^*$ (i.e., $c\neq c^*$ on a set of Lebesgue measure greater than zero). To see this, consider the problem of maximizing, for fixed x and $v'(i_t,Z_t)$,

$$\sup_{c(x):\, c^t(x)M(x)c(x)\le K^2} v'\nabla H(x)\cdot c(x). \tag{18}$$

Clearly, due to the linear nature of the optimization, the maximum will occur on the boundary. Using Lagrange multiplier λ, we maximize

$$v'\nabla H(x)\cdot c(x) + \lambda(c^t(x)M(x)c(x) - K^2).$$

Taking the gradient with respect to c and solving for c,

$$c(x) = -\left(\frac{2v'}{\lambda}M^{-1}(x)\nabla H(x)\right).$$

Differentiating with respect to λ yields the constraint and

$$\lambda = \pm\frac{2|v'|(\nabla H^t(x)M^{-1}(x)\nabla H(x))^{1/2}}{K}.$$

The sign is resolved by considering equation (18). Note that only the properties of $H(x)$ inside the set G are required to determine the process behavior until the exit from the set G. Corollary 1 then completes the proof. □

The following corollary follows immediately from Theorem 2 and shows that the optimal control c^* is asymptotically robust. Let τ^{c^*} be the exit time for the limit process $Y_t^{c^*}$ on the graph.

COROLLARY 2.

(1) *For any* $\epsilon_0>0$,

$$\inf_{\epsilon\in(0,\epsilon_0]} E_x[\tau^{\epsilon,c^*}] \le \sup_{c\in\Pi_K}\inf_{\epsilon\in(0,\epsilon_0]} E_x[\tau^{\epsilon,c}] \le E_x[\tau^{c^*}] \le \sup_{\epsilon\in(0,\epsilon_0]} E_x[\tau^{\epsilon,c^*}].$$

(2)

$$\lim_{\epsilon_0\downarrow 0}\sup_{c\in\Pi_K}\inf_{\epsilon\in(0,\epsilon_0]} E_x[\tau^{\epsilon,c}] = \lim_{\epsilon_0\downarrow 0} E_x[\tau^{\epsilon,c^*}] = E_x[\tau^{c^*}].$$

4. Comments on applying the control law. Several extensions of the results in Theorems 1 and 2 are available. An important generalization is the consideration of general conservation laws $H(x)$. Since the skew gradient $\overline{\nabla}H$ is orthogonal to the gradient ∇H, the unperturbed dynamics can always be written in terms of a scalar function $\beta(x)$ as

$$\dot{\tilde{X}}_t = \beta(\tilde{X}_t)\overline{\nabla}H(\tilde{X}_t),$$
$$\tilde{X}_0 = (\tilde{X}_0^1, \tilde{X}_0^2) = x \ \in \Re^2.$$

This equation is discussed in more detail in Borodin and Freidlin [15]. A controlled version of the perturbed dynamic system is then written as

$$\dot{\tilde{X}}_t^\epsilon = \beta(\tilde{X}_t^\epsilon)\overline{\nabla}H(\tilde{X}_t^\epsilon) + \epsilon c(\tilde{X}_t^\epsilon) + \epsilon^{1/2}\dot{\tilde{W}}_t,$$
$$\tilde{X}_0^\epsilon = x \ \in \Re^2.$$

As long as $\beta(x)$ is not equal to zero anywhere in G, then the results above need only small modification. The diffusions will again converge to diffusions on a graph, with the graph having the same structure as discussed above. A convergence result similar to Theorem 1 will hold for $Y_t^{\epsilon,c} = Y(X_t^{\epsilon,c})$, with $Y_t^{\epsilon,c} \to Y_t^c$. However, the operator coefficients will be determined differently. We will have a theorem of the following form, which is proved in the same way as Theorem 2.

THEOREM 3. *Let G_Γ be a closed and connected set on the graph Γ corresponding to the Hamiltonian $H(x)$, and let $G = Y^{-1}(G_\Gamma)$ be the corresponding inverse image in $\Re^2$. Assume G is a bounded set.*

Let $H(x)$, $x \in \Re^2$, be four times continuously differentiable in G. Let $H(x)$ have a finite number of critical points x_1, x_2, ..., x_N in G at which the matrix of second derivatives is nondegenerate. Also assume that each level curve $C_k = \{x : Y(x) = Y(x_k)\}$ contains at most one critical point. Finally, assume $\beta(x)$ is four times continuously differentiable and $\beta(x) \neq 0$ anywhere in G.

Let $v(i,y)$ be a solution, continuous and twice differentiable in the interior of the graph segments, of

$$\tfrac{1}{2}A_i v''(i,y) + B_i v'(i,y) + KE_i|v'(i,y)| + 1 = 0, \qquad (i,y) \in G_\Gamma,$$
$$\textstyle\sum_{i:I_i \sim O_k} (\pm\beta_{ki}) v'(i, H(y_k)) \quad = 0 \quad \textit{for all interior vertices},$$
$$v(i,y) = 0 \quad \textit{on the boundary } \partial G_\Gamma,$$

with

$$A_i(H) = \frac{\oint_{C_i(H)} \frac{|\nabla H(x)|^2}{|\beta(x)\overline{\nabla}H(x)|}dl}{\oint_{C_i(H)} \frac{1}{|\beta(x)\overline{\nabla}H(x)|}dl}, \qquad B_i(H) = \frac{\oint_{C_i(H)} \frac{\frac{1}{2}\Delta H(x)}{|\beta(x)\overline{\nabla}H(x)|}dl}{\oint_{C_i(H)} \frac{1}{|\beta(x)\overline{\nabla}H(x)|}dl},$$

$$E_i(H) = \frac{\oint_{C_i(H)} \frac{(\nabla H^t(x) M^{-1}(x) \nabla H(x))^{1/2}}{|\beta(x)\overline{\nabla}H(x)|}dl}{\oint_{C_i(H)} \frac{1}{|\beta(x)\overline{\nabla}H(x)|}dl},$$

and

$$\beta_{ki} = \int_{C_{ki}} |\beta(x)\nabla H(x)|dl.$$

The sign $\pm$ *of the coefficients meet* $\pm\beta_{ki} \geq 0$ *if* $H \geq H(x_k)$ *on* I_i *and* $\pm\beta_{ki} \leq 0$ *if* $H \leq H(x_k)$ *on* I_i.

Define the optimal control $c^*(x)$ *by*

$$c^*(x) = K \ \ sign(v'(Y(x))) \ \ \frac{M^{-1}(x)\nabla H(x)}{(\nabla H^t(x) M^{-1}(x) \nabla H(x))^{1/2}}$$

and the set of permissible controls Π_K *as the set of functions* c *meeting two conditions:* c *is twice continuously differentiable (with uniformly bounded derivatives) in* G *except at points in a finite number of level sets of* $H(x)$*; and* $c^t(x)M(x)c(x) \leq K^2$ *for constant* $K > 0$*, with* $M(x)$ *twice continuously differentiable (with bounded derivatives),* $M(x)$ *uniformly positive definite on* $x \in G$*, and* $M(x)$ *assumed (without loss of generality) to be symmetric. Then, with*

$$\tau_x^{\epsilon,c} = \inf\{t : X_t^{\epsilon,c} \in \partial G\},$$

for each $c(x) \in \Pi_K$ *there exists an* ϵ_0 *(dependent on* c*) such that*

$$E\{\tau_x^{\epsilon,c}\} \leq E\{\tau_x^{\epsilon,c^*}\}$$

for all $\epsilon \leq \epsilon_0$*. The optimal control* c^* *is independent of initial conditions* x *in the interior of* G.

Note that Corollary 2 will also apply, so the control will again be asymptotically robust. The formulation of this problem is more complex if $\beta(x) = 0$ somewhere in G. In this case, the structure of the resulting graph is changed.

Finally, we consider a problem of our most general form. Consider a more general noise model dependent on a nonsingular matrix σ. For scalar function β, let

$$\dot{\hat{X}}_t^\epsilon = \beta(\hat{X}_t^\epsilon)\overline{\nabla} H(\hat{X}_t^\epsilon) + \epsilon c(\hat{X}_t^\epsilon) + \epsilon^{1/2}\sigma\dot{\hat{W}}_t,$$
$$\hat{X}_0^\epsilon = x \ \in \Re^2.$$

By doing a simple transformation $\tilde{X}_t = \sigma^{-1}\hat{X}_t$, the equations become

$$\dot{\tilde{X}}_t^\epsilon = \beta(\sigma\tilde{X}_t^\epsilon)\sigma^{-1}\overline{\nabla} H(\sigma\tilde{X}_t^\epsilon) + \epsilon\sigma^{-1}c(\sigma\tilde{X}_t^\epsilon) + \epsilon^{1/2}\dot{\tilde{W}}_t,$$
$$\tilde{X}_0^\epsilon = x \ \in \Re^2.$$

Now the conservation law is $H(\sigma x)$. Defining $H_2(x) = H(\sigma x)$, the term $\beta(\sigma x)\sigma^{-1}\overline{\nabla} H(\sigma x)$ can be written as $\beta_2(x)\overline{\nabla} H_2(x)$ for some scalar function $\beta_2(x)$. In the same manner, define $c_2(x) = \sigma^{-1}c(\sigma x)$. The original domain $\hat{G}$ is transformed to $G_2 = \{x : \ \sigma x \in G\}$, and the original control limit matrix $\hat{M}$ is transformed to $M_2(x) = \sigma^t M(\sigma^{-1}x)\sigma$. Our optimization problem then becomes

$$\dot{\tilde{X}}_t^\epsilon = \beta_2(\tilde{X}_t^\epsilon)\overline{\nabla} H_2(\tilde{X}_t^\epsilon) + \epsilon c_2(\tilde{X}_t^\epsilon) + \epsilon^{1/2}\dot{\tilde{W}}_t,$$
$$\tilde{X}_0^\epsilon = x \ \in \Re^2$$

with region G_2 and control magnitude limit $c_2(x)^t M_2(x) c_2(x) \leq K^2$. Note in particular that if $\beta(x) \neq 0$ in G, then $\beta_2(x) \neq 0$ in G_2 so that Theorem 3 and then Corollary 2 can be applied.

Also note the special structure of the set G: the boundary ∂G is composed of level sets of the Hamiltonian H. This special structure is completely natural because

of the underlying Hamiltonian dynamics. If, for some $x \in G$, the unperturbed and uncontrolled solution $\tilde{X}_t$ with initial condition x leaves G, then no control $\epsilon c(\cdot)$ is able to prevent exit from G when ϵ is small. Thus for a more general domain $\tilde{G}$, we should decompose $\tilde{G} = G_1 \bigcup G_2$ with $G_1 = \{x : \ \tilde{X}_t \in G \ \forall t \geq 0 \text{ when } \tilde{X}_0 = x\}$ and $G_2 = \tilde{G} - G_1$. Then Theorem 2 is applied in G_1 and initial points in G_2 are not asymptotically stabilizable.

5. Examples. Several examples are given to illustrate application of the theorem. The first provides a simple numerical example to illustrate how the theory might be applied. The other examples discuss solving for the optimal control problem in a more general setting.

Example 1. Consider control of a simple linear oscillator perturbed by a small noise and controlled by a small control:

$$\begin{pmatrix} \dot{\tilde{X}}^c_{1\,t} \\ \dot{\tilde{X}}^c_{2\,t} \end{pmatrix} = \begin{pmatrix} \tilde{X}^c_{2\,t} \\ -\tilde{X}^c_{1\,t} \end{pmatrix} + 0.0001\ c(\tilde{X}^c_t) + 0.0001^{1/2}\dot{\tilde{W}}_t, \tag{19}$$
$$\tilde{X}^c_0 = x \ \in \Re^2.$$

Note that

$$\begin{pmatrix} \tilde{X}^c_{2\,t} \\ -\tilde{X}^c_{1\,t} \end{pmatrix} = \overline{\nabla} H(\tilde{X}^c_t)$$

for Hamiltonian function

$$H(x) = \frac{1}{2}(x_1^2 + x_2^2).$$

Here, clearly, is the situation we have discussed above with $\epsilon = 0.0001$. Choice of the constant K (in $c^t(x)M(x)c(x) \leq K^2$) will be used to constrain the precise size of the control with respect to the noise and Hamiltonian dynamics. In order to fit into the mathematical format above, we rescale time by $\frac{t}{0.0001} \to t$, with the resulting dynamics

$$\begin{pmatrix} \dot{X}^c_{1\,t} \\ \dot{X}^c_{2\,t} \end{pmatrix} = \frac{1}{0.0001}\begin{pmatrix} X^c_{2\,t} \\ -X^c_{1\,t} \end{pmatrix} + c(X^c_t) + \dot{W}_t, \tag{20}$$
$$X^c_0 = x \ \in \Re^2.$$

Our goal is to stabilize X^c_t so that $1 \leq H(X^c_t) \leq 2$ for as long as possible. We can see the geometry of the situation in Figure 2. The Hamiltonian and its level curves are shown in Figure 2(a) and 2(b). Our region G is shown in Figure 2(c), and G has boundaries $H(x) = 1$ and $H(x) = 2$. Figure 2(d) shows the segment of the graph corresponding to G. Note that the segment endpoints correspond to the trajectories $H(x) = 1$ and $H(x) = 2$.

We should first informally verify that application of our theory to this problem is reasonable: we should establish that X^0_t (note the control $c = 0$) generally makes a number of revolutions before the Hamiltonian $H(X^0_t)$ changes a great deal. Figure 3 shows a sample path for X^0_t and $H(X^0_t)$. Only a few sample paths need be examined to convince us that use of the theory is reasonable.

Now we can easily find the optimal control for various control size limits K ($c^t(x)M(x)c(x) \leq K^2$ with $M(x) = I$). For this example, we used $K = 2$ and

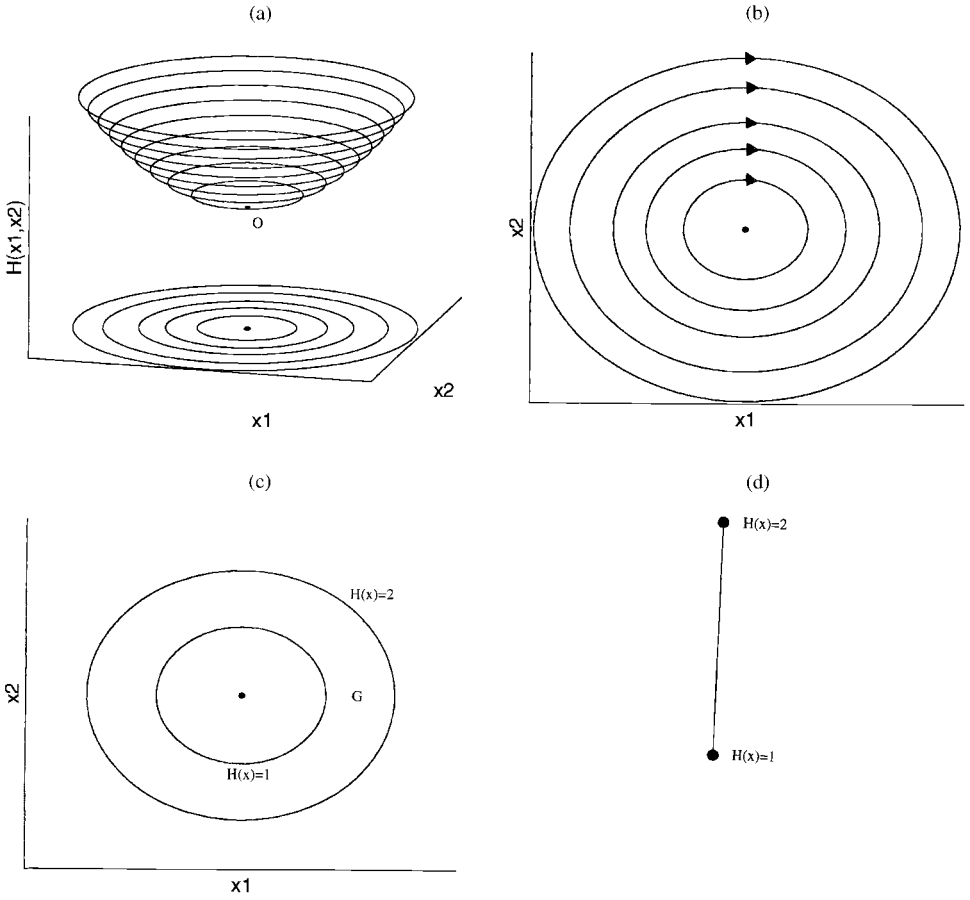


FIG. 2.

$K = 4$. The solution of the optimal control requires, from equation (13), the solution of

$$\begin{aligned} hv''(h) + v'(h) + \sqrt{2h}K|v'(h)| + 1 = 0, \quad h \in (1,2), \\ v(1) = v(2) = 0. \end{aligned} \tag{21}$$

Note that there is only one graph segment, so that the segment notation has been dropped. Then the control c is given according to equation (14). This problem was easily solved in MATLAB and required only a few minutes of coding and execution. The solution v also provides (for small ϵ and a specific K) the approximation

$$v(H(X_0^{\epsilon,c^*})) \approx E_x(\tau^{\epsilon,c^*})$$

as follows from equations (15) and (16). Note that solving equation (21) with $K = 0$ yields an estimate of the uncontrolled exit time. Figure 4 shows the theoretical limit (as $\epsilon \to 0$) of the mean exit times for no control, optimal control with $K = 2$, and optimal control with $K = 4$. These are, respectively, the solid, dashed, and dotted lines. The regions where v increases or decreases determine whether the control is along or opposed to the gradient ∇H, as seen in equation (14).

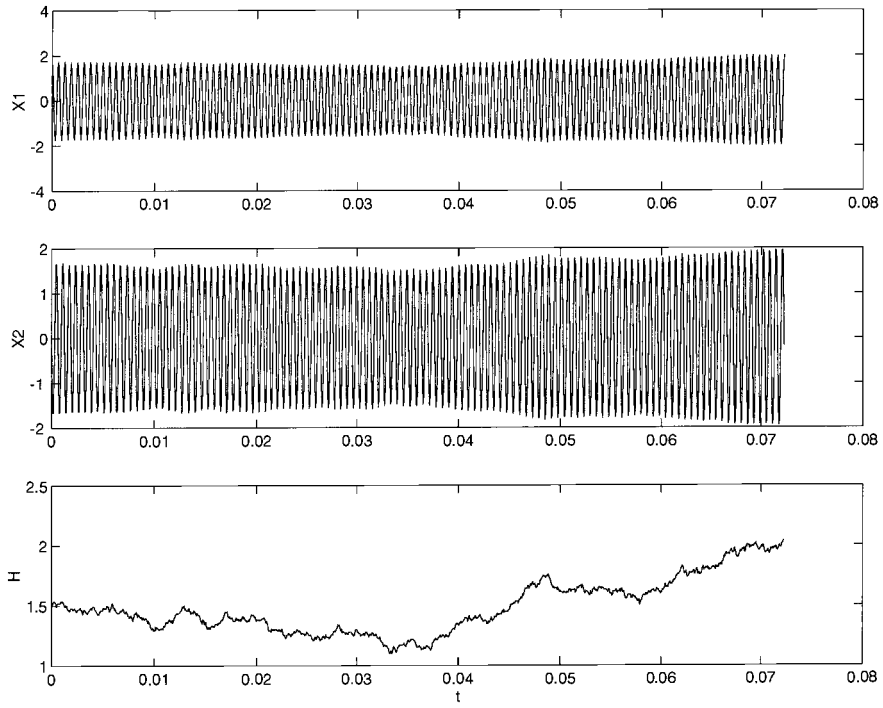


FIG. 3.

To test whether $\epsilon = 0.0001$ is indeed "small," Monte Carlo numerical simulations of equation (20) were also run in MATLAB. These computations turned out to be delicate: since the random trajectories generally circled hundreds or thousands of times before leaving $G = \{x : \ 1 \leq \frac{1}{2}(x_1^2 + x_2^2) \leq 2\}$, numerical errors in the integration scheme caused significant drift in $H(X_t^{c^*})$. This was counteracted by noticing that, since the Hamiltonian drifted slowly, equation (20) could be approximated as a linear equation over short time intervals. A state transition matrix was calculated and applied over each of these short time intervals. Using this viewpoint, 250 sample trajectories were calculated for each of the three controls and initial conditions corresponding to initial Hamiltonian values of 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, and 1.9. The mean exit time over the 250 samples is plotted in Figure 4 as "o" for no control, "x" for optimal control with $K = 2$, and "+" for optimal control with $K = 4$.

Although there is some sampling error and probably still some error from the numerical integration scheme, the agreement in Figure 4 is good. The theoretical limit (as $\epsilon \to 0$) does predict the actual exit times for $\epsilon = 0.0001$, leading us to believe that the asymptotically optimal controls are good choices for controlling the exit time in equation (19).

Example 2. The second example is the simplest and most obvious. Consider the simple Hamiltonian function described in Figures 1(a) and 1(b), along with the corresponding graph in Figure 1(d). Let G be a set whose boundary is a level set of $H(x)$ and assume G contains all three extrema, O_1, O_2, and O_3. Let the matrix function $M(x) = I$ in the control constraint $c^t(x)M(x)c(x) \leq K^2$.

Theorem 2 may be easily applied to maximize the expected exit time from G. First note that the differential equation in (13) may be viewed as a first-order equation

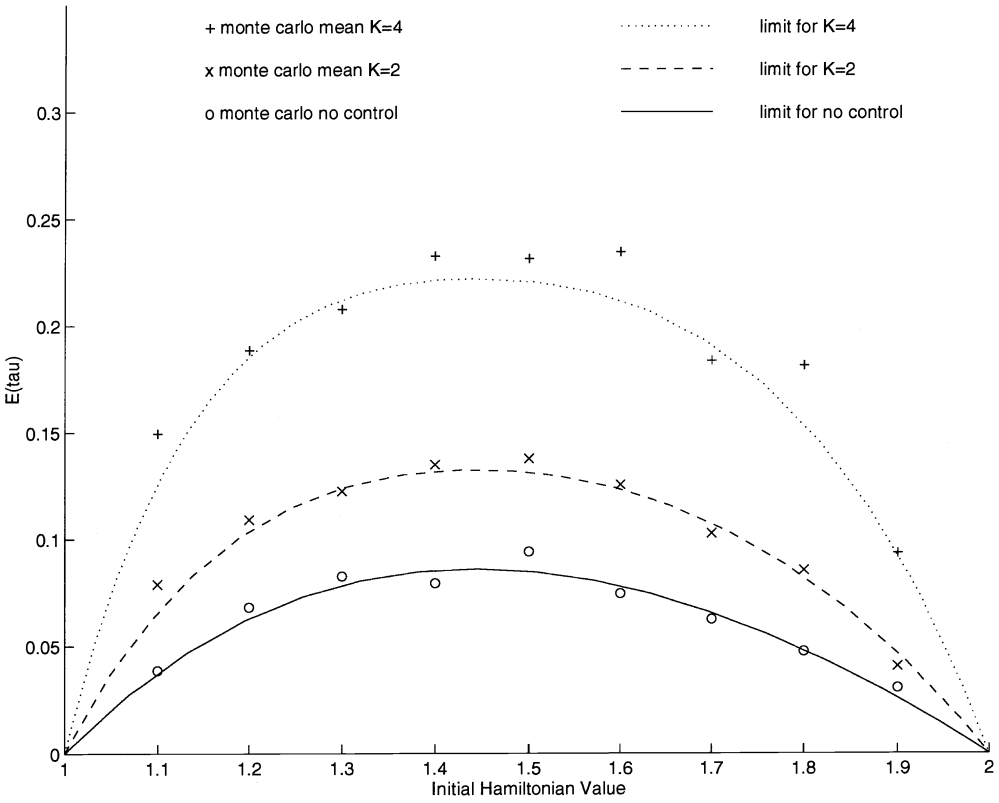


FIG. 4.

involving v' and its derivative v'', with no direct dependence on v. As long as v' is negative, the solution in the interior of a segment may be written as

$$
\begin{aligned}
v'(i,y) &= \exp\left\{\int_{y_0}^{y} -\frac{2}{A_i(\hat{y})}\left(B_i(\hat{y}) - KE_i(\hat{y})\right)d\hat{y}\right\} v'(i,y_0) \\
&+ \int_{y_0}^{y} -\frac{2}{A_i(\tilde{y})} \exp\left\{\int_{\tilde{y}}^{y} -\frac{2}{A_i(\hat{y})}\left(B_i(\hat{y}) - KE_i(\hat{y})\right)d\hat{y}\right\} d\tilde{y}.
\end{aligned} \tag{22}
$$

Consider the lower left segment of the graph in Figure 1(d). The inaccessibility condition in equation (9) establishes that $v'(1,y) < 0$ as y approaches the lower segment endpoint. Then, since $A_i(y) \geq 0$, $v'(1,y) < 0$ on the left segment. Similarly, $v'(3,y) < 0$ on the lower right segment. The boundary condition on the graph center node then requires that $v'(2,y) < 0$ for y at the center joint. Again applying (22) establishes $v'(3,y) < 0$ on the upper segment of the graph. Thus

$$c^*(x) = -K\ \nabla H(x)/|\nabla H(x)|$$

for Example 1. This result is expected; the control pushes away from the boundary of G along the gradient of $H(x)$.

Example 3. Let $H(x)$ have the general structure illustrated in Figures 1(a), 1(b), and 1(d). Let G be a set with the structure illustrated in Figure 1(c). The goal is to keep the dynamical system inside G; the system oscillations should not be too

far from the ∞-shaped curve passing through the saddle point O_2. Let the matrix function $M(x) = I$ in the control constraint $c^t(x)M(x)c(x) \le K^2$.

This problem is more complex, but a straightforward method may be constructed for its solution. For convenience, let $(1, y_1)$, $(2, y_2)$, and $(3, y_3)$ be the boundary of the image of G on the lower left graph segment, the upper graph segment, and the lower right graph segment, respectively. Note that in Theorem 2 the boundary conditions require $v(i, y_i) = 0$ for the three boundary points. Also use the notation y_n to indicate the value of y at the central node. Our problem requires six boundary and gluing conditions provided by

$$\begin{cases} v(1, y_1) = 0, \\ v(2, y_2) = 0, \\ v(3, y_3) = 0, \\ v(1, y_n) = v(2, y_n), \\ v(2, y_n) = v(3, y_n), \\ \beta_{12} v'(2, y_n) = \beta_{11} v'(1, y_n) + \beta_{13} v'(3, y_n) \end{cases}$$

with $\beta_{11} > 0$, $\beta_{12} > 0$, and $\beta_{13} > 0$.

Solution representations such as (22) show that the sign of v' can change sign at most once in the interior of an interval. Also, for Example 2, $v'(1, y_1) > 0$, $v'(2, y_2) < 0$, and $v'(3, y_3) > 0$ follow from the fact that $v(i, y) \ge 0$.

It is possible that $v'(1, y_n) = v'(2, y_n) = v'(3, y_n) = 0$. This can be tested immediately by using solution representations such as (22) to solve the resulting two-point boundary value problems using $v(i, x_i) = 0$ and $v'(i, y_n) = 0$. Each boundary value problem only requires solution of a linear algebraic equation. Continuity then requires $v(1, y_n) = v(2, y_n) = v(3, y_n)$. If this condition is met, then the optimal control is

$$c^*(x) = \begin{cases} K \frac{\nabla H(x)}{|\nabla H(x)|} & \text{for } x \in Y^{-1}(I_1), \\ -K \frac{\nabla H(x)}{|\nabla H(x)|} & \text{for } x \in Y^{-1}(I_2), \\ K \frac{\nabla H(x)}{|\nabla H(x)|} & \text{for } x \in Y^{-1}(I_3). \end{cases}$$

For the remainder of this example we consider the case when $v'(1, y_n) = v'(2, y_n) = v'(3, y_n) = 0$ does not hold. The node gluing condition then shows that $v'(i, y)$ must change sign on one or two of the graph segments. Changing sign on no segments or changing sign on all three segments leads to a violation of the joint gluing condition. The fact that at least one of the three segments does not change sign, and therefore is described by a linear differential equation, allows explicit solution without solving a nonlinear boundary value problem.

Now a method of solution may be constructed, based on solving for a single parameter via integration of one-dimensional differential equations. First solve the linear equation

$$L_i^{\hat{c}} u^{\hat{c}}(i, y) + 1 = 0, \qquad (i, y) \in G_\Gamma,$$

$$\sum_{i: I_i \sim O_k} (\pm \beta_{ki}) u^{\hat{c}\prime}(i, H(y_k)) = 0 \quad \text{for all interior vertices,}$$

$$u^{\hat{c}}(i, y) = 0 \quad \text{on the boundary } \partial G_\Gamma,$$

with

$$\hat{c}(x) = \begin{cases} K\frac{\nabla H(x)}{|\nabla H(x)|} & \text{for } x \in Y^{-1}(I_1), \\ -K\frac{\nabla H(x)}{|\nabla H(x)|} & \text{for } x \in Y^{-1}(I_2), \\ K\frac{\nabla H(x)}{|\nabla H(x)|} & \text{for } x \in Y^{-1}(I_3). \end{cases}$$

This corresponds to a control pushing toward the graph joint. The problem is easily solved by viewing $u^{\hat{c}'}(1, y_1)$, $u^{\hat{c}'}(2, y_2)$, and $u^{\hat{c}'}(3, y_3)$ as unknowns. Then three linear algebraic equations can be written for these unknowns. First define the state transition matrix for each segment:

$$\Phi_i^{\hat{c}'}(y, y_i) = \begin{bmatrix} -\frac{2}{A_i^{\hat{c}}(y)} B_i^{\hat{c}}(y) & 0 \\ 1 & 0 \end{bmatrix} \Phi(y, y_i), \qquad y \in I_i,$$

$$\Phi_i^{\hat{c}}(y_i, y_i) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Then

$$\begin{bmatrix} u^{\hat{c}'}(i, y) \\ u^{\hat{c}}(i, y) \end{bmatrix} = \Phi_i^{\hat{c}}(y, y_i) \begin{bmatrix} u^{\hat{c}'}(i, y_i) \\ 0 \end{bmatrix} - \int_{y_i}^{y} \Phi_i^{\hat{c}}(y, \hat{y}) \begin{bmatrix} \frac{2}{A_i^{\hat{c}}(\hat{y})} \\ 0 \end{bmatrix} d\hat{y}, \qquad y \in I_i. \tag{23}$$

Evaluating these three solutions at the joint y_n and applying the gluing and continuity conditions establishes three linear conditions in the three unknowns $u^{\hat{c}'}(1, y_1)$, $u^{\hat{c}'}(2, y_2)$, and $u^{\hat{c}'}(3, y_3)$. In the case that $u^{\hat{c}'}(1, y_n) = u^{\hat{c}'}(2, y_n) = u^{\hat{c}'}(3, y_n) = 0$, $\hat{c}$ is the optimal control. Due to the gluing conditions, at least one of the $u^{\hat{c}'}(i, y)$ does not change sign. For convenience, assume $u^{\hat{c}'}(3, y) > 0$ on I_3. Then, for the optimal control c^*, $u^{\hat{c}}(3, y) \leq u^{c^*}(3, y)$. This, together with

$$\begin{aligned} u^{\hat{c}'}(3, y) &= \exp\left\{ \int_{y_3}^{y} -\frac{2}{A_i^{\hat{c}}(\hat{y})} B_i^{\hat{c}}(\hat{y}) d\hat{y} \right\} u^{\hat{c}'}(3, y_3) \\ &\quad + \int_{y_3}^{y} -\frac{2}{A_i^{\hat{c}}(\tilde{y})} \exp\left\{ \int_{\tilde{y}}^{y} -\frac{2}{A_i^{\hat{c}}(\hat{y})} B_i^{\hat{c}}(\hat{y}) d\hat{y} \right\} d\tilde{y}, \end{aligned} \tag{24}$$

indicates that $u^{\hat{c}'}(3, y_3) \leq u^{c^*\prime}(3, y_3)$. We see then that $u^{c^*\prime}(3, y) = v'(3, y) > 0$ on segment I_3. This provides the key to solving the nonlinear control problem in equation (13).

The solution can now be described as solving a single equation for a single unknown α. For notational purposes, represent this nonlinear algebraic equation by

$$F(\alpha) = 0,$$

where evaluation of $F(\cdot)$ is best described by a procedure rather than an explicit formula. To calculate $F(\alpha)$, perform the following steps:

(1) Set $v'(1, y_1) = \alpha$.

(2) Use (13) and integrate the nonlinear equation along I_1 to the node. This results in boundary conditions at the node of $v(1, y_n)(\alpha) = v(2, y_n)(\alpha) = v(3, y_n)(\alpha)$ and $v'(1, y_n)(\alpha)$. The dependence on α is made explicit in the notation.

(3) Now solve the linear two-point boundary value problem along I_3. Because $v'(3,y) > 0$, $c^*(x) = \hat{c}(x)$ for $x \in Y^{-1}I_3$. Using the values $v(3,y_n)(\alpha)$ and $v(3,y_3) = 0$, the solution representation in equation (23) may be used to immediately calculate the resulting $v'(3,y_n)(\alpha)$.

(4) Continuity and gluing conditions in equation (13) immediately yield values for $v(2,y_n)(\alpha)$ and $v'(2,y_n)(\alpha)$. Use equation (13) on segment I_2 and integrate to finally find an endpoint value $v(2,y_2)(\alpha)$. The function $F(\alpha)$ may then be expressed by

$$F(\alpha) = v(2,y_2)(\alpha).$$

After a solution to $F(\alpha) = 0$ is found, the functions $v(i,y)$ are then calculated by setting $v'(1,y_1) = \alpha$ and integrating along the graph segments again. See the above four steps. The direction of optimal control is then along the gradient of $H(x)$, with direction determined by $sign(v'(i(x),H(x)))$ and magnitude equal to K.

This solution has an interesting interpretation. The unperturbed Hamiltonian oscillations at which

$$sign(v'(i(x),H(x)))$$

changes are in a sense the "safest" paths, and the control pushes the system trajectory toward these "most stable" solutions. If only one sign change of $v'(i,x)$ occurs on the graph in G_Γ, then all points in G are attracted to the resulting single most stable trajectory. If two sign changes of $v'(i,x)$ occur on the graph in G_Γ, then points in $x \in G$ are attracted to the "safe" path which is least risky to reach.

6. Conclusions. The results of this paper are easy to apply and have significant intuitive appeal: the optimal control is in the direction of the gradient of the Hamiltonian, with sign determined by solution of a simple boundary value problem on the corresponding graph. However, three important directions of expansion of these results are possible.

The admissible control class Π_K is not the broadest possible choice. A better choice would be all functionals $c_t = c(X_s^{\epsilon,c}, 0 \le s \le t)$ with $c^t(x)M(x)c(x) \le K^2$ and $c(\cdot)$ measurable with respect to the system trajectory up to time t. This appears possible, but it is complicated by the fact that the convergence occurs on the graph instead of in $\Re^2$. Many analytical tools have not yet been developed for processes on graphs. It is worth noting that for each fixed ϵ the optimal control given by equation (4) is in Π_K.

The noise process in our problem is two-dimensional white noise $\dot{W}_t$. These results can possibly be extended for a more general noise term of the form $\sigma(X_t^{\epsilon,c})\dot{W}_t$ (with spatial dependence in the noise field) for positive definite $\sigma(\cdot)$. Degenerate noise is much more problematic.

Finally, controls for perturbed Hamiltonian systems in $\Re^n$, with $n > 2$, are of course of great interest. Systems with more than one integral of motion could be considered; in this case, the random motion would occur not on a graph but on some higher-dimensional object. The geometry would be complicated, and the nonperturbed system trajectories would not in general provide the required averaging conditions. More can be said about a special case in $\Re^n$ in which averaging is accomplished by a second noise process. Consider, for example, a system such as

$$\dot{\tilde{X}}_t^{\epsilon,c} = f(\tilde{X}_t^{\epsilon,c}) + \zeta_t(\tilde{X}_t^{\epsilon,c}) + \epsilon\, c(\tilde{X}_t^{\epsilon,c}) + \epsilon^{1/2}\dot{\tilde{W}}_t,$$
$$\tilde{X}_0^{\epsilon,c} = x \ \in \Re^2.$$

Here ζ_t is a random process. As written here, ζ_t is smooth enough to allow stochastic integration, but a white noise process could also be used. We are interested in systems with an integral of motion $H(\cdot)$ for the uncontrolled and unperturbed system

$$\begin{aligned} \dot{\tilde{X}}_t &= f(\tilde{X}_t) + \zeta_t(\tilde{X}_t), \\ \tilde{X_0^{\epsilon,c}} &= x \ \in \Re^2, \end{aligned} \tag{25}$$

so

$$\begin{aligned} \nabla H(x) \cdot f(x) &= 0, \\ \nabla H(x) \cdot \zeta_t(x) &= 0. \end{aligned}$$

Conditions can be given for the noise process ζ_t so that the system in equation (25) is ergodic in the level sets of $H(\cdot)$. Then averaging will occur, and the approach used in this paper could be applied.

REFERENCES

[1] M. I. Freidlin and A. D. Wentzell, *Random perturbations of Hamiltonian systems*, Mem. Amer. Math. Soc., 109 (1994), pp. 1–82.

[2] M. I. Freidlin and A. D. Wentzell, *Diffusion processes on graphs and the averaging principle*, Ann. Probab., 21 (1993), pp. 2215–2245.

[3] A. D. Wentzell and M. I. Freidlin, *Some problems concerning stability under small random perturbations*, Theory Probab. Appl., 17 (1972), pp. 269–283.

[4] M. I. Freidlin and A. D. Wentzell, *Random Perturbations of Dynamical Systems*, Springer-Verlag, Berlin, New York, 1984.

[5] W. H. Fleming and P. E. Souganidis, *PDE-viscosity solution approach to some problems of large deviations*, Ann. Scoula. Norm. Sup. Pisa Cl. Sci. (4), 13 (1986), pp. 171–192.

[6] P. Dupuis and H. Kushner, *Minimizing escape probabilities: A large deviations approach*, SIAM J. Control Optim., 27 (1989), pp. 432–445.

[7] S. M. Meerkov and T. Runolfsson, *Residence time control*, IEEE Trans. Automat. Control, 33 (1988), pp. 323–332.

[8] S. M. Meerkov and T. Runolfsson, *Theory of residence time control by output feedback*, in Proc. of the 28th IEEE Conference on Decision and Control, Tampa, Florida, 1989, pp. 1175–1179.

[9] S. M. Meerkov and T. Runolfsson, *Output residence time control*, IEEE Trans. Automat. Control, 34 (1989), pp. 1171–1176.

[10] T. Runolfsson, *Residence time control of systems subject to measurement noise*, J. Math. Anal. Appl., 145 (1990), pp. 289–308.

[11] S. Kim, S. M. Meerkov, and T. Runolfsson, *Residence probability control*, Comput. Math. Appl., 19 (1990), pp. 121–125.

[12] E. Kappos, *Optimal control problems arising in large deviation theory*, in Modern Optimal Control, E.P. Roxin, ed., Dekker, New York, 1989, pp. 203–215.

[13] N. V. Krylov, *Controlled Diffusion Processes*, Springer-Verlag, Berlin, New York, 1980.

[14] I. I. Gihman and A. V. Skorohod, *The Theory of Stochastic Processes* III, Springer-Verlag, Berlin, New York, 1979.

[15] A. N. Borodin and I. M. Freidlin, *Fast oscillating random perturbations of dynamical systems with conservation laws*, Ann. Inst. H. Poincaré, Probab. Stat., 31 (1995), pp. 485–525.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 1, pp. 253–272, January 1998

014

OUTPUT DEAD BEAT CONTROL FOR A CLASS OF PLANAR POLYNOMIAL SYSTEMS*

D. NEŠIĆ†, I. M. Y. MAREELS‡, G. BASTIN§, AND R. MAHONY†

Abstract. Output dead beat control for a class of nonlinear discrete time systems, which are described by a single input-output (I-O) polynomial difference equation, is considered. The class of systems considered is restricted to systems with a two-dimensional state space description. It is assumed that the highest degree with which the present input appears in the equation is odd. Necessary and sufficient conditions for the existence of output dead beat control and for the stability of the zero output constrained dynamics are presented. We also design a minimum time output dead beat control algorithm (feedback controller) which yields stable zero dynamics, whenever this is feasible. A number of interesting phenomena are discussed and illustrated with examples.

Key words. polynomial systems, dead beat, controllability

AMS subject classifications. 93B05, 93B27, 93C55, 93D99

PII. S0363012995286381

1. Introduction.

Linear dead beat control has received a great deal of attention in the past 30 years [16]. The discoveries in the area of linear dead beat control resulted in a better understanding of linear systems theory and a number of very successful applications. The fact that very often the dynamics of a plant cannot successfully be modeled using linear time invariant equations, provide motivation for considering nonlinear dead beat control. Dead beat control or controllability for special classes of nonlinear systems has been addressed by many authors [1, 7, 8, 9, 10, 12, 20, 21, 22]. Nevertheless, a wealth of open questions remain to be explored.

Polynomial I-O systems of the form $y_{k+1} = f(y_k, \ldots, y_{k-s}, u_{k-t}, \ldots, u_k)$ are often used [13, 14, 5] for system identification in black-box mode (see also [23, 24]). y, u, and k are, respectively, the output, input, and time index. The function f is a polynomial in all its arguments. This is an obvious generalization of linear ARMA models. Although a number of applications of I-O polynomial systems have been reported, e.g., [5, 14], their control properties are not well understood.

In this paper we consider a class of I-O polynomial systems of the following form:

$$y_{k+1} = f(y_k, u_{k-1}, u_k). \tag{1.1}$$

We assume throughout the paper that the highest exponent of the argument u_k in the polynomial f is an odd integer. An application of this class of systems can be found in [5] where a subsystem of a radiator and fan is identified in this form.

*Received by the editors May 22, 1995; accepted for publication (in revised form) November 9, 1996. This research was supported by the Cooperative Research Centre for Adaptive and Robust Systems, Australia.

http://www.siam.org/journals/sicon/36-1/28638.html

†Department of Systems Engineering, RSISE, Australian National University, Canberra, ACT 0200, Australia (dragan.nesic@anu.edu.au, robert.mahony@anu.edu.au).

‡Engineering Department, FEIT, Australian National University, Canberra, ACT 0200, Australia (i.mareels@ee.mu.oz.au).

§CESAME, Bâtiment Euler, Avenue G. Lemaitre 4, 1348 Louvain-La-Neuve, Belgium (bastin@auto.ucl.ac.be). This research was performed while the author was visiting the Australian National University.

The control question that we are interested in is minimum time output dead beat regulation. In particular, we want to design a control law of the form

$$u_k = c(y_k, u_{k-1}) \tag{1.2}$$

such that $y_k = 0\ \forall k \geq t$ for some integer t and such that the constrained dynamics[1] defined by

$$\begin{aligned} 0 &= f(0, u_{k-1}, u_k), \\ u_k &= c(0, u_{k-1}) \end{aligned} \tag{1.3}$$

are stable in a sense to be specified later. The paper deals with two questions: output dead beat controllability and stability of constrained dynamics for (1.1).

Some pioneering work on controllability for a class of discrete time bilinear systems can be found in [11]. Papers [7, 8, 12] provide complete conditions for controllability for the same class of systems. Invariance of the control independent set was investigated in [11]. We show that a new notion of strongly invariant sets, first introduced in [20, 21], is crucial for output dead beat controllability of (1.1). We take a similar approach as in [20, 21], where dead beat controllability of scalar polynomial systems is considered. The output controllability result of this paper can be viewed as a generalization of some results on odd systems in [20, 21]. In the conference version [22] of this paper, we provided the output controllability test for (1.1). However, the design of a feasible dead beat controller and stability of constrained dynamics are analyzed in the sequel.

Output dead beat control of recursive nonlinear systems was investigated in [1, 3]. Existence of constrained dynamics together with a number of interesting phenomena were studied. The considered class of systems was, however, large and results are consequently weak. The notion of criterion of choice is introduced in the context of stability of constrained dynamics in [1, 3]. This notion is also important in our discussions. Stability of one-dimensional explicit constrained dynamics $u_k = f(u_{k-1})$ was investigated [2, 3]. Our paper extends these results to the case of implicitly defined polynomial dynamics (1.3) and we present necessary and sufficient conditions of the existence of a criterion of choice that leads to stable constrained dynamics. We point out that the stability of an interval that we consider was investigated in [3] and in [15], where this property is referred to as "permanence." In [15], global stability properties of a number of nonlinear explicit systems of the form $y_{k+1} = F(y_k, \ldots, y_{k-s})$ were investigated (see also [19] for continuous time results). We, however, consider the implicit difference equation in (1.3). We emphasize that the notion of constrained dynamics considered here differs from the concept of zero dynamics introduced in [17, 18]. Moreover, the notion of zero dynamics [17, 18] appears not to be sufficiently general to be applicable to the stabilizing dead beat control problem considered here.

This paper provides an explicit test for verifying the existence of an output dead beat control law which yields stable constrained dynamics for the system of the form (1.1). Furthermore, a constructive design method is provided to find any such feedback law. A purpose of this paper is to show the difficulties that one may face when tackling the output dead beat control problem for the simple class of I-O polynomial systems (1.1) and to present a number of interesting phenomena.

[1]The definition of stable zero output constrained dynamics that we analyze is more general than the usual definition of zero dynamics found in literature [17]. To make the distinction more obvious we refer to our definition as constrained dynamics and to the definition in [17] as zero dynamics.

The paper is organized as follows. In section 2 we present some notation, and in section 3 we define the problem and the class of systems that we consider. The question of the existence of dead beat control is addressed in section 4. Sections 5 and 6 are, dedicated to, respectively, the stability of constrained dynamics and a method to check the existence of a dead beat control law which yields stable constrained dynamics. The modified dead beat control law which zeros the output in minimum time and also yields stable constrained dynamics is then presented in section 7. In section 8, we present several examples which illustrate our methods. The summary and conclusion are given in the last section.

2. Mathematical preliminaries. We use the standard definitions of rings and fields [6]. We work over the field of real numbers which is denoted as $\Re$. $\Re^n$ is a set of all n-tuples of elements of $\Re$, where n is a nonnegative integer. The ring of polynomials in n variables over the real field $\Re$ is denoted as $\Re[x_1, x_2, \ldots, x_n]$. Let $f_1, f_2, \ldots, f_s$ be polynomials in $\Re[x_1, x_2, \ldots, x_n]$. Then we define

$$V(f_1, f_2, \ldots, f_s) = \{(a_1, a_2, \ldots, a_n) \in \Re^n : f_i(a_1, a_2, \ldots, a_n) = 0 \text{ for all } 1 \le i \le s\}.$$

We call $V(f_1, f_2, \ldots, f_s)$ the real algebraic set or real variety defined by $f_1, f_2, \ldots, f_s$. Since the defining polynomials of a real variety are often clear from the context, it is often denoted simply as V.

DEFINITION 2.1. *A real variety $V \subset \Re^n$ is* irreducible *if whenever V is written in the form $V = V_1 \cup V_2$, where V_1 and V_2 are real varieties, then either $V_1 = V$ or $V_2 = V$* [6, p. 196].

THEOREM 2.2 (see [6, p. 202]). *Let $V \subset \Re^n$ be a real variety. Then V can be written as a finite union of irreducible varieties $V = V_1 \cup V_2 \cup \cdots \cup V_m$ where each V_i is an irreducible variety.*

Let $f, g \in \Re[x_1, x_2, \ldots, x_n]$. $f|g$ means that g divides f; that is, there exists a polynomial $h \in \Re[x_1, x_2, \ldots, x_n]$ such that $f = hg$. $f \equiv g|h$ means that f is divisible by h modulo g; that is, given polynomials h and g, $deg(g) < deg(f)$ there exists a polynomial $h_1 \in \Re[x_1, x_2, \ldots, x_n]$ such that $f = h_1 h + g$. Also, $f \not| g$ and $f \equiv g \not| h$ denotes, respectively, that f is not divisible by g and f is not divisible by h modulo g.

We say that a variety $V \subset \Re^2$ has special form if

$$V = \left\{ (y, v) \in \Re^2 : y - \sum_{i=0}^{n-1} b_i v^i = 0,\ b_i \in \Re,\ i = 0, 1, \ldots, n-1 \right\}.$$

Varieties of special form are irreducible because they can be parametrized by polynomials [6, p. 197].

3. Definition of the system. We consider systems described by the following recursive I-O polynomial equation:

$$y_{k+1} = f(y_k, u_{k-1}, u_k), \tag{3.1}$$

where y_k is the output of the system at time k and u_k is the input to the system at time k. The function f is a polynomial, $f \in \Re[y, v, u]$. We assume that the highest exponent of u in $f(y, v, u)$ is an odd integer. A system (3.1) with this property is referred to as an *odd system*.

It is always possible to rewrite (3.1) in the following form:

$$y_{k+1} = g_n(y_k, u_{k-1})u_k^n + g_{n-1}(y_k, u_{k-1})u_k^{n-1} + \cdots + g_0(y_k, u_{k-1}), \tag{3.2}$$

where $g_n \not\equiv 0$ and n is an odd positive integer.

ASSUMPTION 1. *Constrained dynamics are defined as*

$$\forall v \in \Re, \quad \exists u \in \Re \text{ such that } f(0, v, u) = 0. \tag{3.3}$$

A sequence of controls is denoted as

$$\mathcal{U} = \{u_0, u_1, \ldots\}.$$

The truncation to a sequence of length $p+1$ is denoted as $\mathcal{U}_p = \{u_0, u_1, \ldots, u_p\}$. The composition of the function f in equation (3.1) under the action of a control sequence $\mathcal{U}_p$ which starts from $(y_0, u_{-1}) \in \Re^2$ is denoted as

$$f^{\mathcal{U}_p}(y_0, u_{-1}) = \underbrace{f(f(\ldots f(f}_{p \text{ times}}(y_0, u_{-1}, u_0), u_0, u_1), \ldots, u_{p-1}, u_p).$$

Obviously $y_{p+1} = f^{\mathcal{U}_p}(y_0, u_{-1})$ is the output at time $p+1$, given the starting point (y_0, u_{-1}) and the input $\mathcal{U}_p$.

We can introduce the state variables $x_1(k) = y_k$ and $x_2(k) = u_{k-1}$ and write accordingly the model in state space format. In the sequel, we refer to $(y_0, u_{-1}) \in \Re^2$ as an initial state although we work with the input output equation (3.1).

We are interested in output dead beat control in Definition 3.1.

DEFINITION 3.1. *The system* (3.1) *is output dead beat controllable if for every* $(y_0, u_{-1}) \in \Re^2$ *there is a sequence* $\mathcal{U} = \{u_0, u_1, \ldots\}$ *such that the output of the system* (3.1) *is driven to zero in finite time, that is,* $y_k = 0, \forall k \geq t$, *where* t *is a nonnegative integer.*

DEFINITION 3.2. *A feedback controller, given by* $u_k = c(y_k, u_{k-1})$, *is an output dead beat controller if there exists a positive integer* P *such that* $\forall (y_0, u_{-1}) \in \Re^2$ *and* $k \geq P$ *we have* $y_k = 0$, *where* $y_{k+1} = f(y_k, u_{k-1}, c(y_k, u_{k-1}))$.

Because of Assumption 1, we can split the dead beat control problem into two parts. Indeed, the control sequence $\mathcal{U}$ in Definition 3.1 may be split into two parts. $\mathcal{U}_t$ is the part of the sequence $\mathcal{U}$ that transfers the output to the origin and $\{u_{t+1}, \ldots\}$ the part which keeps the output at the origin. Section 4 is concerned with the existence of the sequence $\mathcal{U}_t$, which naturally leads to the construction of an (feedback) output dead beat controller. In section 5 we consider the properties of the obtained control laws, which settles the usefulness of the approach.

4. Output dead beat controllability of recursive polynomial systems. In this section, we consider when it is possible to transfer the output of the system (3.1) to the origin in finite time, starting from an arbitrary initial state $(y_0, u_{-1}) \in \Re^2$. The following definition is used in the sequel.

DEFINITION 4.1. *The one-step reachable set from an initial state* $(y_0, u_{-1}) \in \Re^2$ *is defined as*

$$V_r(y_0, u_{-1}) = \{(y, u) \in \Re^2 : y - f(y_0, u_{-1}, u) = 0\}.$$

We also define the projection of the one-step reachable set onto the first coordinate axis as

$$\Pi V_r(y_0, u_{-1}) = \{y \in \Re : \exists v \in \Re : (y, v) \in V_r(y_0, u_{-1})\}$$

and call it the set of one-step reachable outputs.

Observe that the one-step reachable set is a real variety and it has special form for any initial state in $\Re^2$. Moreover, since the systems is odd, the only states from

which it may not be possible to zero the output in one step belong to the real variety V_C defined by

$$V_C = \{(y, v) \in \Re^2 : g_n(y, v) = 0\}. \tag{4.1}$$

Notice that dim $V_C < 2$.

DEFINITION 4.2. *The variety V_C given by* (4.1) *is called the* critical variety.

DEFINITION 4.3. *The number of varieties of special form that are contained in V_C is denoted by N.*

Let V and W be varieties. We introduce the notation

$$V \xrightarrow{f} W \tag{4.2}$$

to denote that $V_r(y_0, u_{-1}) = W \ \forall (y_0, u_{-1}) \in V$. It should be emphasized that equation (4.2) means that the one-step reachable set from any initial state in V is *equal to W*.

DEFINITION 4.4. *A set $V_{I_j} \subseteq V_C$ is invariant if*

$$\forall (y, v) \in V_{I_j},\ V_r(y, v) \subseteq V_{I_j}. \tag{4.3}$$

The union of all invariant sets $V_I = \cup_j V_{I_j}$ is called the maximal invariant set.

DEFINITION 4.5. *A subset W_{I_j} of the variety V_C is strongly invariant if it is invariant and $\forall (y_0, u_{-1}) \in W_{I_j}$ there exists an integer $t \geq 0, t = t(y_0, u_{-1})$ and a sequence of controls $\mathcal{U}_t = \{u_0, u_1, \ldots, u_t\}$ which yields $(y_{t+1}, u_t) = (y_0, u_{-1})$, where $y_{t+1} = f^{\mathcal{U}_t}(y_0, u_{-1})$. The union of all strongly invariant sets $W_I = \cup_j W_{I_j}$ is called the maximal strongly invariant set.*

DEFINITION 4.6. *The number of varieties of special form that are contained in the maximal strongly invariant set W_I of V_C is denoted by L.*

The propositions below indicate some important properties of the maximal invariant and strongly invariant sets.

PROPOSITION 4.7. *The maximal strongly invariant set can be decomposed into finitely many strongly invariant subsets W_{I_j}, each of which can be decomposed into finitely many varieties of special form W_i:*

$$W_I = \underbrace{W_1 \cup \cdots \cup W_{L_1}}_{W_{I_1}} \cup \underbrace{W_{L_1+1} \cup \cdots \cup W_{L_1+L_2}}_{W_{I_2}} \cup \cdots \cup \underbrace{W_{L_1+\cdots+L_{p-1}+1} \cup \cdots \cup W_{L_1+\cdots+L_p}}_{W_{I_p}},$$

where $L_1 + L_2 + \cdots + L_p = L$. Therefore, the maximal strongly invariant set is itself a variety.

Sketch of the proof. We prove this proposition in four steps. Since $V_r(y_0, u_{-1})$ is of special form for any $(y_0, u_{-1}) \in V_C$, at least one variety of special form W_1 belongs to the maximal strongly invariant subset W_I. Then we can show that in order to have invariance one-step reachable sets from any initial state in W_1 must coincide; that is, $V_r(y_1, v_1) = V_r(y_2, v_2)\ \forall (y_1, v_1), (y_2, v_2) \in W_1$. Therefore, we show that one can write $V_r(y, v) = W_2\ \forall (y, v) \in W_1$, where W_2 is a variety of special form which is a subset of V_C. After this, we show that the union $W_1 \cup W_2 \cup \cdots \cup W_L$ is a subset of W_I. Finally, it is proved that the union $W_1 \cup W_2 \cup \cdots \cup W_L$ is equal to W_I, and the partition into smaller strongly invariant sets follows easily. For a more detailed proof see [22]. □

Proofs of the propositions below hinge on the proof of Proposition 4.7 (see [22]).

PROPOSITION 4.8. *Any invariant subset V_{I_j} of the critical variety V_C contains a strongly invariant subset W_{I_j}.*

PROPOSITION 4.9. *Any initial state that belongs to an invariant subset V_{I_j} of the critical variety V_C is transferred to a strongly invariant subset W_{I_j} (which is a subset of V_{I_j}) in finite time.*

PROPOSITION 4.10. *Any $(y_0, u_{-1}) \in V_C - V_I$ can be mapped to $\Re^2 - V_C$ in at most $N - L + 1$ time steps (see Definitions 4.3 and 4.6).*

COMMENT 1. An immediate consequence of Proposition 4.10 is that if $V_I = \emptyset$ any initial state in V_C can be mapped to $\Re^2 - V_C$ in at most $N + 1$ time steps and hence the output can be zeroed in at most $N + 2$ steps (see Definition 4.3).

PROPOSITION 4.11. *Consider the system (3.1). The critical variety V_C (4.1) contains a strongly invariant subset if and only if there exist polynomials $y - \sum_{i=0}^{n-1} b_i^p v^i$, $b_i^p \in \Re$, $p = 1, 2, \ldots, B$, $B \leq L \leq N$ such that*

$$g_n(y, v) | \left(y - \sum_{i=0}^{n-1} b_i^p v^i \right) \quad \forall p = 1, 2, \ldots, B,$$

$$g_i(y, v) \equiv b_i^{p+1} | \left(y - \sum_{i=0}^{n-1} b_i^p v^i \right) \quad \forall p = 1, 2, \ldots, B-1 \quad \forall i = 1, \ldots, n-1, \quad \textit{and}$$

$$g_i(y, v) \equiv b_i^1 | \left(y - \sum_{i=0}^{n-1} b_i^B v^i \right) \quad \forall i = 1, \ldots, n-1.$$

The above properties of invariant subsets of V_C, lead to necessary and sufficient conditions for output dead beat controllability for the class of odd systems (3.1).

THEOREM 4.12. *The odd system (3.1) is output dead beat controllable if and only if either the maximal invariant set $V_I = \emptyset$ or if $V_I \neq \emptyset$, then all irreducible components (varieties) $W_i, i = 1, 2, \ldots, L$ of the maximal strongly invariant set W_I intersect the line $y = 0$.*

Sketch of the proof. The whole state space can be partitioned as $W_I \cup (V_I - W_I) \cup (V_C - V_I) \cup (\Re^2 - V_C)$. Propositions 4.7, 4.8, 4.9, 4.10, together with the fact that any state in $V_C - V_I$ can be mapped to $\Re^2 - V_C$, give a characterization of all possible behaviors. □

COMMENT 2. It is easily verified that the conditions under which the critical variety V_C may contain invariant subsets (they are given in Proposition 4.11) are clearly not generic.

COMMENT 3. It is important to notice that Theorem 4.12 provides conditions for output controllability to the origin. If we want to check output controllability to some other point $y^* \neq 0$, then all irreducible components (varieties) W_i of the maximal strongly invariant set W_I should intersect the line $y = y^*$.

5. Stability of constrained dynamics. We examine in this section properties of the control law which keep the output of the system at zero after the output was zeroed. We extend Theorem 6.2 [3] to the class of polynomial implicitly defined systems. This theorem gives necessary and sufficient conditions for the global stability of an invariant interval for the class of explicit constrained dynamics defined by $u_k = g(u_{k-1})$ with g continuous. We consider implicitly defined polynomial systems. The equation that defines the behavior of the system is given below:

$$f(0, u_{k-1}, u_k) = 0. \tag{5.1}$$

It was noticed in [1] that the properties of the control law that keep the output at zero depend on the rule used to determine which particular solution from among the

possible alternatives u_k, satisfying (5.1), is used for any given u_{k-1}. This rule is referred to as a criterion of choice. If we have several control actions that satisfy the constraint (5.1) at our disposal, it is very important to apply "the most appropriate one."

In this section we define what we mean by stable constrained dynamics and find conditions which guarantee the existence of a "good" criterion of choice, i.e., one that leads to stable constrained dynamics. Now we give definitions for the concepts that we need in our developments.

DEFINITION 5.1. *A criterion of choice is a single valued function* $c : \Re \to \Re$ (*denoted also as* $u_k = c(u_{k-1})$) *such that*

$$f(0, v, c(v)) = 0 \quad \forall v \in \Re. \tag{5.2}$$

DEFINITION 5.2. *Consider a criterion of choice* c (*Definition* 5.1). *A bounded interval* $A \subset \Re$ *is invariant under mapping* c *if* $c(A) \subseteq A$.

DEFINITION 5.3. *Let* $A \subset \Re$ *be a bounded interval invariant under mapping* c. *Then*

1. A *is called stable if* $\forall E \subseteq \Re$, $A \subset E$, $\exists K(E) > 0$ *such that* $\forall u_{k-1} \in E$ *we have* $\sup_{u_{k-1} \in E} |c(u_{k-1})| \leq K(E) < \infty$;
2. A *is called attractive if* $\forall \Delta > 0$ $\forall u_{-1}$, $\exists T = T(\Delta, u_{-1})$ *such that* $\inf_{x \in A,} |x - u_k| < \Delta$ $\forall k > T$;
3. A *is called asymptotically stable if* 1 *and* 2 *hold.*

DEFINITION 5.4. *Implicitly defined constrained dynamics* (*equation* (5.1)) *are called stable if there exists a criterion of choice* c *such that there is a bounded interval* A *invariant under mapping* c *which is asymptotically stable.*

We emphasize that the present notion of stability is more general than allowed for in [17, 18], where only stability of equilibria is considered. Notice also that we consider a *global* stability property.

We now cite Theorem 6.2 from [3] which is used in the proof of the main result.

THEOREM 5.5 (see [3]). *Consider the map* $g : \mathbf{D} \to \mathbf{D}$, $\mathbf{D} \subset \Re$. *Let* $\mathcal{A} \stackrel{\triangle}{=} [a, b] \subset \Re$ *such that*

1. $\mathbf{D} \cap \mathcal{A}$ *is invariant under* g: $g(\mathbf{D} \cap \mathcal{A}) \subset \mathbf{D} \cap \mathcal{A}$;
2. $(\Re -]a, b[) \subset \mathbf{D}$;
3. g *is continuous on* $(\Re -]a, b[)$.

Then $\mathcal{A}$ *is globally attracting interval of the iterative map* $u(k+1) = g(u(k))$ *if and only if the following conditions hold:*

$$\begin{aligned}
&\forall x < a, && g(x) > x,\\
&\forall x > b, && g(x) < x,\\
&\forall x < a \ \text{ such that } \exists (x, z) \in G_R^{-1}, && g(x) < z,\\
&\forall x > b \ \text{ such that } \ \exists (x, z) \in G_L^{-1}, && g(x) > z.
\end{aligned} \tag{5.3}$$

The domain $\mathbf{D}$ represents the domain of definition of constrained dynamics. Other symbols used in the statement of Theorem 5.5 are given below:

$$\begin{aligned}
&G = \{(x, g(x)) : x \in \Re - [a, b]\}, \quad G_L = \{(x, g(x)) \in G : x < a\},\\
&G_R = \{(x, g(x)) \in G : x > b\}, \quad G_L^{-1} = \{(g(x), x) : (x, g(x)) \in G_L\},\\
&G_R^{-1} = \{(g(x), x) : (x, g(x)) \in G_R\}.
\end{aligned}$$

COMMENT 4. In our case the domain of definition of constrained dynamics is the whole real line; that is, $\mathbf{D} = \Re$. Therefore, condition 2 of Theorem 5.5 does not need to be verified.

Given $T \geq 0$ a real number, the following sets will be used in what follows:

$$S_1 = \{(v,u) \in \Re^2 : v < -T\}, \quad S_2 = \{(v,u) \in \Re^2 : v > T\}. \tag{5.4}$$

A very important feature of polynomial systems which is crucial for the stability of constrained dynamics is given in the theorem below.

THEOREM 5.6. *Consider the real variety V_z defined by*

$$V_z = \{(v,u) \in \Re^2 : f(0,v,u) = 0\}. \tag{5.5}$$

There exists $T \geq 0$ such that there are constant numbers n_1 and n_2 of continuous branches[2] *of variety V_z on sets S_1 and S_2 (5.4).*

Proof of Theorem 5.6. Sturm sequences can be used in order to check the exact number of distinct real roots of a univariate polynomial on any interval $[a,b]$, including $]-\infty,+\infty[$ [4]. We will regard u_{k-1} as a parameter, and for any fixed u_{k-1} we can find the number of distinct real roots u_k to (5.1). In other words, we can find the exact number of real roots to (5.1) on vertical lines $u_{k-1} = \text{const}$.

Consider the Sturm sequence of $f(0,v,u)$. It has the form

$$\begin{aligned} &A_n^0(v)u^n + \cdots + A_0^0(v), \\ &A_{n-1}^1(v)u^{n-1} + \cdots + A_0^1(v), \\ &\cdots \\ &A_0^n(v). \end{aligned} \tag{5.6}$$

The leading coefficient functions are rational functions in v. It turns out that for the number of real solutions u to (5.1) for a fixed value of the parameter v, only the leading coefficient functions are important. Actually, the signs of these functions determine the number of real roots, and since they are rational functions, we can find a set of the form $]-\infty,-D_1[\cup]D_1,+\infty[$ on which their signs do not change. The modified division algorithm which is used to determine the sequence (5.6) yields a special form of the leading coefficients in the Sturm sequence. Namely, the denominator of $A_{n-j+1}^{j+1}(v)$ has the same roots as the numerator of $A_{n-j}^j(v) \forall j > 1$. Also, $A_n^0 = g_n(0,v)$ and $A_{n-1}^1 = n\partial/\partial v[g_n(0,v)]$ are polynomials. Consequently, the set on which the $A_{n-j}^j(v)$ do not change signs can be determined considering the equations $A_{n-j}^j(v) = 0$ $\forall j = 0,\ldots,n$. We introduce the following set:

$$\mathcal{D}_1 = \{v \in \Re : A_{n-j}^j(v) = 0 \text{ for some } j = 0,\ldots,n\}. \tag{5.7}$$

Denote as D_1 the following number:

$$D_1 = \sup_{v \in \mathcal{D}_1} |v|. \tag{5.8}$$

It follows that on the set $]-\infty,-D_1[\cup]D_1,+\infty[$ all the leading coefficient functions do not change their signs. Therefore, there is a constant number of real roots u_k for every $u_{k-1} \in]-\infty,-D_1[$ and $u_{k-1} \in]D_1,+\infty[$ to (5.1). We can also say that there exists a constant number of continuous branches of V_z on sets $]-\infty,-D_1[\times\Re$ and $]D_1,+\infty[\times\Re$. This follows from the theorem on the continuity of real roots [4, p. 38]. Since $g_n(0,v) \neq 0$ ($g_n = A_n^0$) for $v \in]-\infty,-D_1[\cup]D_1,+\infty[$ and since there is a constant number of complex roots, all the conditions of the theorem are satisfied. □

[2]The term "branch of V_z" that we use corresponds to parts of irreducible varieties (curves) from which the variety V_z is composed [4, 6] that belong to sets S_1 and S_2.

COMMENT 5. Theorem 5.6 states that it is possible to find an interval $[-D_1, D_1]$ inside which all bifurcations of the variety V_z occur. Moreover, from the theorem on the continuity of roots [4, p. 38] we see that all intersections between branches of the variety V_z occur inside the same interval.

LEMMA 5.7. *A necessary condition for the existence of stable constrained dynamics is*

$$\sup_{|v|<K} \inf_{(v,u)\in V_z} |u| < +\infty \quad \forall K \in]0, +\infty[.$$

Proof of Lemma 5.7. Suppose that there exists a criterion of choice c which yields stable constrained dynamics. Suppose that there exists $v = u^*_{k-1}$ which belongs to the invariant interval such that all branches of the variety V_z have a vertical asymptote at $v = u^*_{k-1}$. In other words, the condition of Lemma 5.7 is not satisfied for any neighborhood of the origin that contains u^*_{k-1}. It is then obvious that the invariant interval must have one of the following forms: $]-\infty, +\infty[$, $[K, +\infty[$, or $]-\infty, K]$, and we have a contradiction since neither of these intervals is bounded. Suppose now that u^*_{k-1} does not belong to the invariant interval. In this case, constrained dynamics cannot be stable in the sense of Definition 5.4 because for u_{k-1} such that $u_{k-1} \to u^*_{k-1}$ we have that $|u_k| \to +\infty$, so we again obtain a contradiction. □

Now we can give definitions of maximal and minimal branches of the variety V_z.

DEFINITION 5.8. *Consider the variety V_z on sets S_1 and S_2. The maximal branch of V_z in S_2 is given by*

$$V_M^{S_2} = \{(v, u) : v \in]T, +\infty[,\ u = \max_{(v,y)\in(V_z\cap S_2),\ y<v} y\}.$$

The minimal branch of V_z in S_1 is such that

$$V_m^{S_1} = \{(v, u) : v \in]-\infty, -T[,\ u = \min_{(v,y)\in(V_z\cap S_1),\ y>v} y\}.$$

In other words, the maximal branch is the closest branch of V_z to the bisector $u = v$, which is below the bisector (on the set S_2). Notice that minimal and maximal branches are well-defined parts of irreducible varieties of V_z, following from the theorem on continuity of roots [4] and Bezout's theorem [6]. Bezout's theorem says that we can find a set $[-D_3, D_3] \times \Re$ inside which all intersections between the variety V_z and the bisector $u = v$ occur. Also notice that if there are no branches in S_2 that are below the bisector $u = v$, then by definition $V_M^{S_2} = \emptyset$.

COMMENT 6. Suppose that we can find a criterion of choice such that outside a bounded interval $[-T, T]$ all orbits are bounded, converge to the interval, and enter it in finite time from any given u_{-1}. Then it is easy to show that when Lemma 5.7 holds there exists an interval (perhaps larger than $[-T, T]$ but bounded) such that it is invariant and stable. Consequently, we will concentrate only on the existence of a bounded stable interval, and Lemma 5.7 guarantees that we can always have a criterion of choice for all $u_{-1} \in [-T, T]$ which renders the interval invariant.

Now we can state the main result.

THEOREM 5.9. *Implicitly defined constrained dynamics* (5.1) *are stable if and only if the mapping $u_k = g(u_{k-1})$ defined as*

$$u_k = \begin{cases} y & \text{such that } (u_{k-1}, y) \in V_m^{S_1} \text{ if } u_{k-1} < -T, \\ y & \text{such that } (u_{k-1}, y) \in V_M^{S_2} \text{ if } u_{k-1} > T, \\ y & \text{such that } (u_{k-1}, y) \in V_z \text{ if } u_{k-1} \in [-T, T] \text{ and} \\ & y \text{ has the smallest absolute value} \end{cases} \tag{5.9}$$

satisfies equations (5.3) *of Theorem* 5.5 *and Lemma* 5.7 *holds.*

Proof of Theorem 5.9. *Sufficiency.* Consider the criterion of choice (5.9). It is obvious that all the conditions of Theorem 5.5 are satisfied and this criterion yields stable constrained dynamics.

Necessity. We have to show only that the conditions (5.3) are necessary for stable constrained dynamics. We can find a set inside which all intersections between the variety V_z and the bisector $u = v$ occur and denote it as $[-D_3, D_3] \times \Re$. Moreover, we can find another set inside which all the intersections between V_z and $V_z^{-1} = \{(v, u) \in \Re^2 : f(0, u, v) = 0\}$ occur (modulo common components which may have infinitely many common points) and denote it as $[-D_2, D_2] \times \Re$. All the subsequent arguments are given for the sets S_1 and S_2 defined by the number $T = \max[D_1, D_2, D_3]$. Sets S_1 and S_2 (5.4) defined in this way obviously have the property that (modulo common components) there are no intersections between V_z and V_z^{-1} on the sets, there are no bifurcations of the variety V_z on the sets and, finally, minimal and maximal branches $V_m^{S_1}$ and $V_M^{S_2}$ are either parts of continuous curves or they are empty sets.

Suppose that the constrained dynamics are stable and that the first condition in (5.3) is not satisfied. Since $V_m^{S_1} = \emptyset$, all branches are below the bisector $u = v$ and as a consequence we have that $u_k \to -\infty$ as $k \to \infty$, $\forall u_{-1} \in]-\infty, -T]$. A similar situation happens when the second condition (5.3) is not satisfied and therefore the first two conditions in (5.3) are necessary to ensure stability of the constrained dynamics. In other words, a necessary condition for the stability of the implicitly defined constrained dynamics (5.1) is that $V_m^{S_1} \neq \emptyset$ and $V_M^{S_2} \neq \emptyset$.

Consider now what happens if the third condition in (5.3) is not satisfied. Since all branches of V_z in S_2 are above $V_M^{S_2}$, all their inverses will lay on the left-hand side (or below) of $(V_M^{S_2})^{-1}$. Thus, we suppose that no branch of V_z^{-1} satisfies the third condition in (5.3). Moreover, if we use pieces of branches of V_z to construct a piecewise continuous one-to-one function and use the modified Theorem 5.5 [3] we can see that no such function would satisfy the conditions of Theorem 5.5. Therefore, there does not exist a criterion of choice which yields stable constrained dynamics. The contradiction completes the proof. The last two conditions are symmetric and they are either both satisfied or not. □

6. An algebraic test for stability of constrained dynamics. We now present a method to check the conditions of Theorem 5.9. First, we provide a means of verifying the conditions of Lemma 5.7.

We write the function (5.1) as

$$f(0, v, u) = g_n(0, v)u^n + \cdots + g_0(0, v). \tag{6.1}$$

The only critical points that we have to check are the ones for which the leading coefficient $g_n(0, v)$ (6.1) vanishes [4, pp. 10, 39]. Therefore, the first step is to find all real solutions v to $g_n(0, v) = 0$. It is then necessary to check whether

$$f(0, v, u) = 0 \tag{6.2}$$

has real roots u for all critical values of v. We define the following sets:

$$\mathcal{A} = \{v : g_n(0, v) = 0\}, \tag{6.3}$$

$$\mathcal{B}(v) = \{u \in \Re : f(0, v, u) = 0\},\ v \in \mathcal{A}, \tag{6.4}$$

$$\mathcal{C} = \{(v, u) : v \in \mathcal{A},\ u \in \mathcal{B}(v)\}. \tag{6.5}$$

There must be at least one real root $u \in \mathcal{B}(v)\ \forall v \in \mathcal{A}$, otherwise Assumption 1 would not be satisfied. We can now use the implicit function theorem. For all pairs of

controls $(v,u) \in \mathcal{C}$ equation (5.1) holds. If for every $v \in \mathcal{A}$ there exists at least one $u \in \mathcal{B}(v)$ for which

$$\frac{\partial f}{\partial u} \mid_{(v,u)} \neq 0, \tag{6.6}$$

then the implicit function theorem guarantees the existence of a function $u = g(0,v)$, which is C^∞ since we deal with polynomials such that $f(0,v,g(0,v)) = 0$.

The implicit function theorem gives only sufficient conditions to check Lemma 5.7 but they are easy to check. If (6.6) does not hold, we may check whether Lemma 5.7 is satisfied. The easiest way to do this is to draw the variety V_z around every point (v,u) in $\mathcal{C}$ using Matlab (the set $\mathcal{C}$ contains finitely many points) and check whether there exists a branch of V_z which does not have a vertical asymptote at (v,u).

Before we give the classification of all possible situations we define bisectors and octants.

$$B_1 = \{(v,u) \in \Re^2 : v = u\}, \quad B_2 = \{(v,u) \in \Re^2 : -v = u\},$$

$$O_1 = \{(v,u) \in \Re^2 : v > 0,\ u > 0,\ u < v\}, \quad O_2 = \{(v,u) \in \Re^2 : v > 0,\ u > 0,\ u > v\},$$

$$O_3 = \{(v,u) \in \Re^2 : v < 0,\ u > 0,\ u > -v\}, \quad O_4 = \{(v,u) \in \Re^2 : v < 0,\ u > 0,\ u < -v\},$$

$$O_5 = \{(v,u) \in \Re^2 : v < 0,\ u < 0,\ u > v\}, \quad O_6 = \{(v,u) \in \Re^2 : v < 0,\ u < 0,\ u < v\},$$

$$O_7 = \{(v,u) \in \Re^2 : v > 0,\ u < 0,\ u < -v\}, \quad O_8 = \{(v,u) \in \Re^2 : v > 0,\ u < 0,\ u > -v\}.$$

We also use the notation A_1 and A_2 to denote, respectively, the line $v = 0$ and $u = 0$ in $\Re^2$.

A very important concept of the "inverse graph" of the variety V_z (5.5), which is given by

$$V_z^{-1} = \{(v,u) \in \Re^2 : f(0,u,v) = 0\} \tag{6.7}$$

is obtained by simply interchanging variables v and u in the defining polynomial. It is easy to check that if a point on a variety V_z is in the first octant O_1, the corresponding point on V_z^{-1} is in the second octant O_2 and vice versa. We use the following notation to summarize all possible situations:

$$O_2 \leftrightarrow O_1,\ O_3 \leftrightarrow O_8,\ O_4 \leftrightarrow O_7,\ O_5 \leftrightarrow O_6.$$

In some cases the position of branches $V_M^{S_2}$ and $V_m^{S_1}$ provide sufficient information to conclude on the stability of constrained dynamics since the conditions on the inverse graph are automatically satisfied. We summarize these trivial cases in the lemma below.

LEMMA 6.1.

1. *If one of the following conditions holds:*
   (a) $V_m^{S_1} \subset O_5$ *and* $V_M^{S_2} \subset O_1$,
   (b) $V_m^{S_1} \subset O_5$ *and* $V_M^{S_2} \subset O_8$,
   (c) $V_m^{S_1} \subset O_5$ *and* $V_M^{S_2} \subset O_7$,
   (d) $V_m^{S_1} \subset O_4$ *and* $V_M^{S_2} \subset O_1$,
   (e) $V_m^{S_1} \subset O_4$ *and* $V_M^{S_2} \subset O_8$,
   (f) $V_m^{S_1} \subset O_3$ *and* $V_M^{S_2} \subset O_1$,

then there exists a criterion of choice which yields stable constrained dynamics.

2. *If* $V_m^{S_1} \subset B_2$ *(*$V_M^{S_2} \subset B_2$*) then there exists a criterion of choice which yields stable constrained dynamics if and only if* $V_M^{S_2}$ *(*$V_m^{S_1}$*) belongs to the cone* $\{(v,u) \in \Re^2 : |v| < |u|\}$.
3. *If* $V_m^{S_1} \subset A_2$ *or* $V_M^{S_2} \subset A_2$*, the constrained dynamics are stable.*
4. *If* $V_m^{S_1} = \emptyset$ *or* $V_M^{S_2} = \emptyset$ *or* $V_m^{S_1} = \emptyset$ *and* $V_M^{S_2} = \emptyset$*, then the constrained dynamics are unstable.*
5. *If* $V_m^{S_1} \subset O_3$ *or* $V_M^{S_2} \subset O_7$ *or* $V_m^{S_1} \subset O_3$ *and* $V_M^{S_2} \subset O_7$*, then the constrained dynamics are unstable.*

It can easily be checked that the only remaining cases are

1. $V_m^{S_1} \subset O_3$ and $V_M^{S_2} \subset O_8$,
2. $V_m^{S_1} \subset O_4$ and $V_M^{S_2} \subset O_7$.

Only in these cases do we have to use "inverses" $(V_m^{S_1})^{-1}$ and $(V_M^{S_2})^{-1}$. Since we are dealing with polynomial systems, we can use the algebraic structure of these systems in order to obtain a "box" inside which all intersections between V_z and V_z^{-1} occur (modulo common components). We will use the theory of resultants to compute such a box. We denote $f_1 = f(0,v,u)$ and $f_2 = f(0,u,v)$.

Resultants procedure. First, we find the greatest common divisor of f_1 and f_2 which is denoted as $GCD(f_1,f_2) \in \Re[v,u]$. Then we compute "common components-free" polynomials:

$$f_1^{ccf} = \frac{f_1}{GCD(f_1,f_2)}, \quad f_2^{ccf} = \frac{f_2}{GCD(f_1,f_2)}. \tag{6.8}$$

Now, we can regard polynomials f_1^{ccf} and f_2^{ccf} as polynomials in v whose coefficients are polynomials in u. Now we can find the resultant of the two polynomials:

$$R(f_1^{ccf}, f_2^{ccf}) = \sum_{i=0}^{p} a_i u^i. \tag{6.9}$$

The resultant $R(f_1^{ccf}, f_2^{ccf})$ is a polynomial in u. We know that polynomials f_1^{ccf} and f_2^{ccf} have no common roots if $R(f_1^{ccf}, f_2^{ccf}) \neq 0$. We can find a number D_2 which is such that all absolute values of real roots of the resultant are less than D_2.

Second, we estimate the number D_2 using formulas for bounds on roots, e.g., $\hat{D}_2 = 1 + \sup_i |a_i|$ [4], where $a_i, i = 0,1,\ldots,p$, are coefficients of the resultant. Outside the box defined by $\{(v,u) \in \Re^2 : |v| \leq \hat{D}_2 \text{ and } |u| \leq \hat{D}_2\}$ the varieties V_z and V_z^{-1} have no intersections modulo common branches.

Third, we pick $\hat{u}$ such that $|\hat{u}| > |\hat{D}_2|$ and find sets of solutions:

$$\Sigma_1 = \{v \in \Re : f(0,v,\hat{u}) = 0\}, \quad \Sigma_2 = \{v \in \Re : f(0,\hat{u},v) = 0\}. \tag{6.10}$$

We can see that the sets Σ_1 and Σ_2 give a complete picture about the branches of varieties V_z and V_z^{-1} and therefore can be used to check whether constrained dynamics are stable for the two remaining cases. The criterion for the stability of constrained dynamics of the two last cases, which are not covered by Lemma 6.1, is given in the following lemma.

LEMMA 6.2. *If*

1. $V_m^{S_1} \subset O_3$ *and* $V_M^{S_2} \subset O_8$ *or*
2. $V_m^{S_1} \subset O_4$ *and* $V_M^{S_2} \subset O_7$,

then constrained dynamics are stable if there exist $\sigma_1 \in \Sigma_1$ *and* $\sigma_2 \in \Sigma_2$ *such that* $\sigma_1 < \sigma_2$. *In the first case sets* Σ_1 *and* Σ_2 (6.10) *are calculated using* $\hat{u} > \hat{T}$ *and in the second case* $\hat{u} < -\hat{T}$.

Proof of Lemma 6.2. It trivially follows from Theorem 5.9 and the procedure given above.

The method to check the existence of constrained dynamics consists of several steps:

1. Check the conditions of Lemma 5.7 as described before.
2. Form the Sturm sequence and find all leading coefficient functions. Using (5.8) and bounds on roots, determine the estimate $\hat{D}_1$.
3. Find the box inside which all intersections between the variety V_z and B_1, B_2, A_1, and A_2 occur. This is done in the following way. Find the following estimates:

   $$\hat{D}_3 = 1 + \max_i |n_i|,\ \hat{D}_4 = 1 + \max_i |m_i|,\ \hat{D}_5 = 1 + \max_i |k_i|,\ \hat{D}_6 = 1 + \max_i |l_i|,$$

   where $n_i,\ m_i,\ k_i,\ l_i \in \Re$ are, respectively, coefficients of polynomials $f(0, v, v)$, $f(0, v, -v)$, $f(0, 0, u)$, and $f(0, v, 0)$.
4. Find the estimate $\hat{T}$ of T using

   $$\hat{T} = \max(\hat{D}_1, \hat{D}_3, \hat{D}_4, \hat{D}_5, \hat{D}_6). \tag{6.11}$$

5. Pick any $v^* \in]-\infty, -\hat{T}[$ and compute all real roots of

   $$f(0, v^*, u) = 0. \tag{6.12}$$

   Pick any $v^{**} \in]\hat{T}, +\infty[$ and compute all real roots of

   $$f(0, v^{**}, u) = 0. \tag{6.13}$$

6. Determine to which octants do the pairs (v^*, real root to (6.12)) and (v^{**}, real root to (6.13)) belong and check whether Lemma 6.1 holds (remember that checking the position of a single point of the variety implies that the whole branch has the same position). If Lemma 6.1 is not satisfied, then proceed to the next step.
7. Compute $\hat{D}_2 = 1 + \max_i |f_i|$, where f_i are the coefficients of the resultant $R(f_1^{ccf}, f_2^{ccf})$; redefine $\hat{T} = \max(\hat{D}_1, \hat{D}_2, \hat{D}_3, \hat{D}_4, \hat{D}_5, \hat{D}_6)$; and apply the resultants procedure which is used to check conditions of Lemma 6.2.

7. Output dead beat control law with stable constrained dynamics. Propositions 4.7–4.11 can be used to design a dead beat controller (algorithm) as outlined in Figure 7.1. The obtained controller uses static feedback to compute the value of a control signal at any time instant k. The closed-loop system can be written in the form

$$\begin{aligned} y_{k+1} &= f(y_k, u_{k-1}, u_k), \\ u_k &= c(y_k, u_{k-1}). \end{aligned} \tag{7.1}$$

The control signal is obtained as a solution to a polynomial algebraic equation, and since there may be more than one solution, we need a criterion of choice to define the control law $c(y_k, u_{k-1})$. One criterion for the choice may be to apply the control signal that has the least absolute value. We may be able to shape the transient response

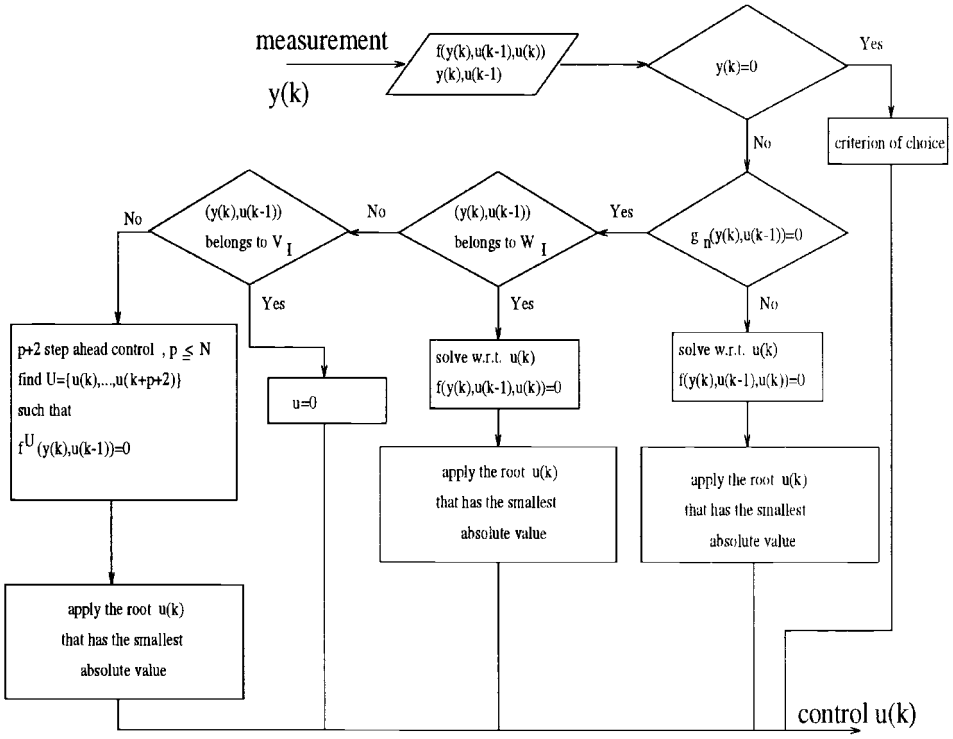


FIG. 7.1. *Output dead beat controller - algorithm.*

and keep the control signals as small as feasible, using a different criterion of choice. The question of which choice is not so critical if the output is not zero. Having zeroed the output, the criterion of choice becomes crucial for the stability of constrained dynamics and, consequently, for the stability of the closed loop system (7.1).

A criterion of choice which yields stable constrained dynamics is given by

$$u_k = \begin{cases} u \in V_m^{S_1} & \text{if } (v,u) \in S_1, \\ u \in V_M^{S_2} & \text{if } (v,u) \in S_2, \\ u \text{ s.t. it has minimum absolute value} & \text{if } v \in [-\hat{T}, \hat{T}]. \end{cases} \tag{7.2}$$

This choice does not guarantee the fastest convergence to the invariant interval, and other choices may be better in this sense than this control law. The tradeoff between the speed of convergence to the invariant interval and the shape of the transient response is a difficult problem in its own right, but very often it is possible to successfully tackle this problem on a case-by-case basis.

Notice that working with poor bounds on roots, such as the one that we have used, may yield an estimate $\hat{T}$ which is much larger than the minimal possible T, but the computations are simpler and faster to use when checking the existence of stable constrained dynamics. Computing exact roots, on the other hand, yields a smaller size of the invariant interval, which should be used when implementing the controller. Blocks in which we need to check whether $(y(k), u(k-1))$ belong to W_I or V_I are equivalent to testing whether a finite number of polynomials which define W_I and V_I are zero when evaluated at $(y(k), u(k-1))$.

8. Examples. The following example illustrates the concepts of invariant and strongly invariant subsets of the variety V_C.

EXAMPLE 1. Consider the system

$$y_{k+1} = (y_k - u_{k-1}^2 - 1)(y_k + u_{k-1}^2 + 1)[(y_k + 2)u_k^3 + u_k^2 + 1] + u_k^2 + 1. \tag{8.1}$$

Assumption 1 is satisfied. The critical variety V_C is defined by

$$V_C = \{(y, v) \in \Re^2 : (y - v^2 - 1)(y + v^2 + 1)(y + 2) = 0\}.$$

In this case we can verify that the only strongly invariant set is given by

$$W_I = \{(y, v) \in \Re^2 : (y - v^2 - 1) = 0\} \subset V_C.$$

We check the existence of strongly invariant sets via Proposition 4.11. There are three varieties of special form that are contained in V_C,

$$y - v^2 - 1, \quad y + v^2 + 1, \quad y + 2,$$

and we also have

$$g_0 = (y - v^2 - 1)(y + v^2 + 1) + 1; \quad g_1 = 0; \quad g_2 = (y - v^2 - 1)(y + v^2 + 1) + 1.$$

The only cycle of Proposition 4.11 is given by the divisions

$$g_0 \equiv 1 | (y - v^2 - 1), \quad g_1 \equiv 0 | (y - v^2 - 1), \quad g_2 \equiv 1 | (y - v^2 - 1),$$

which defines W_I. Since W_I does not intersect the line $y = 0$ according to Theorem 4.12 the system is not output dead beat controllable.

We have, therefore, $W_I \xrightarrow{f} W_I$, and t in Definition 4.5 can be chosen to be 1. From equation (8.1) it is clear that $\forall (y, v) \in V_1$, where $V_1 = \{(y, v) \in \Re^2 : (y + v^2 + 1) = 0\}$ (see Figure 8.1) we have $V_r(y, v) = W_I$. Therefore, any initial state in V_1 is transferred in one step to some point in W_I irrespective of the control that is applied. Thus, we can write

$$V_1 \xrightarrow{f} W_I \xrightarrow{f} W_I \xrightarrow{f} \dots.$$

Consider now initial states on the line $y_0 = -2$. The model of the system becomes

$$y_1 = [(-3 - u_{-1}^2)(-1 + u_{-1}^2) + 1](u_0^2 + 1).$$

Denote real solutions u_{-1} of the equations

$$[(-3 - u_{-1}^2)(-1 + u_{-1}^2) + 1] = -1,$$

$$[(-3 - u_{-1}^2)(-1 + u_{-1}^2) + 1] = 1$$

as a_i and b_i $(i = 1, 2)$, respectively. The set of one-step reachable states from $(-2, a_1)$ and $(-2, a_2)$, is V_1 and from $(-2, b_1)$ and $(-2, b_2)$ it is W_I. Notice also that $b_1 = 1$, $b_2 = -1$, and hence $(-2, b_1)$ and $(-2, b_2)$ belong to V_1. Therefore, we can write

$$(-2, a_i) \xrightarrow{f} V_1 \xrightarrow{f} W_I \xrightarrow{f} W_I \xrightarrow{f} \cdots, \ i = 1, 2.$$

The maximal invariant set V_I is

$$V_I = \{(y, v) \in \Re^2 : (y - v^2 - 1)(y + v^2 + 1) = 0\} \cup \{(-2, a_1), (-2, a_2)\}.$$

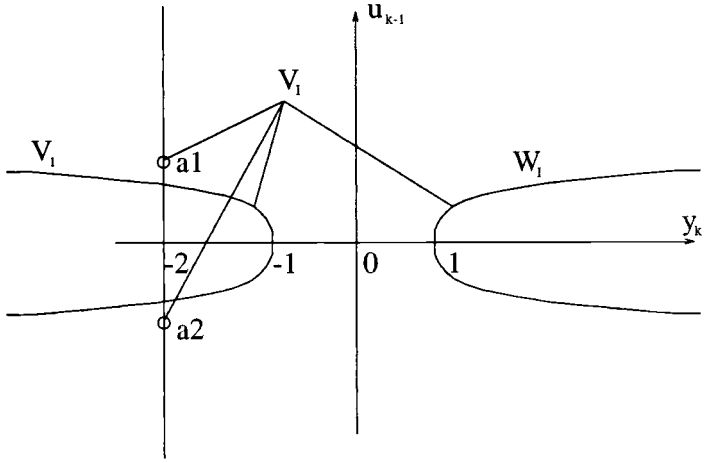


FIG. 8.1. *Invariant sets V_I and strongly invariant sets W_I (Example 1).*

Sets V_I and W_I are shown in Figure 8.1. The set $V_C - V_I$ is not invariant and there exists a control u_k which can map any initial state from it to $\Re^2 - V_C$ in one step. Observe that both V_I and W_I are real varieties, whereas $V_C - V_I$ is not. Also, initial states in V_1 are transferred to W_I in one step and the initial states $(-2, a_i),\ i = 1, 2$, are transferred to W_I in two steps.

The following example serves to illustrate why the present notion of stability of constrained dynamics is more appropriate in this context than the notion of zero dynamics introduced in [17, 18].

EXAMPLE 2. Consider the following system:

$$y_{k+1} = (u_k + 2u_{k-1} + y_k)(u_k - 0.5u_{k-1} - y_k).$$

We introduce the state variables $x_1(k) = y_k$ and $x_2(k) = u_{k-1}$ and write

$$\begin{aligned} x_1(k+1) &= (u_k + 2x_2(k) + x_1(k))(u_k - 0.5x_2(k) - x_1(k)), \\ x_2(k+1) &= u_k, \\ y(k) &= x_1(k). \end{aligned} \tag{8.2}$$

According to [17], the relative degree for system (8.3) is $d = 1$ and Assumption 1 in [17] holds. Two possible feedback laws can be used to transform the system into the form (2.6) in [17]:

$$u_k = \frac{-1.5x_2(k) + \sqrt{6.25x_2^2(k) + 10x_1(k)x_2(k) + 4x_1^2(k) + 4v(k)}}{2}, \tag{8.3}$$

$$u_k = \frac{-1.5x_2(k) - \sqrt{6.25x_2^2(k) + 10x_1(k)x_2(k) + 4x_1^2(k) + 4v(k)}}{2}, \tag{8.4}$$

where $v(k)$ is the new control input. If we use the control law (8.4), the corresponding zero dynamics are then defined as $x_2(k+1) = -2x_2(k)$ (with $x_1(k) = 0, v(k) = 0$) and are obviously not stable. If, on the other hand, we had chosen (8.3), we obtain $x_2(k+1) = 0.5x_2(k)$, which is obviously stable. In this case there are four different continuous feedback laws that transform the system into the form (2.6) in [17]. Three of them yield stable zero dynamics, and one yields unstable zero dynamics. Also, there

are infinitely many discontinuous control laws that keep the output at zero. Notice that all conditions in [17] are satisfied, and it appears that the stability of the zero dynamics depends on the choice of the feedback law. The criterion of choice that we use in the definition of stable constrained dynamics takes this phenomenon explicitly into account.

The following example illustrates the method for checking the existence of stable constrained dynamics.

EXAMPLE 3. Check the existence of stable constrained dynamics for the following system:

$$y_{k+1} = -2(1+y_k^2)u_k^5 - 2u_k^3 + 2u_k u_{k-1}(1+y_k^4) + 2u_k u_{k-1}^2 + u_{k-1}u_k^4 + u_{k-1}u_k^2 - u_{k-1}^2 - u_{k-1}^3 + y_k^3.$$

For $y_k = 0$ we have

$$-2u_k^5 - 2u_k^3 + 2u_k u_{k-1} + 2u_k u_{k-1}^2 + u_{k-1}u_k^4 + u_{k-1}u_k^2 - u_{k-1}^2 - u_{k-1}^3 = 0. \tag{8.5}$$

Therefore, the variety V_z is defined by

$$V_z = \{(v,u) \in \Re^2 : -2u^5 - 2u^3 + 2uv + 2uv^2 + vu^4 + vu^2 - v^2 - v^3 = 0\}.$$

We will follow the steps that are described in section 6 in order to check the existence of stable constrained dynamics.

Step 1. Since $g_5(0,v) = -2$, the conditions of Lemma 5.7 are satisfied.

Step 2. Using Maple[3], we obtain the following Sturm sequence:

$$\begin{aligned}
f_0 &= -2u^5 - 2u^3 + 2uv + 2uv^2 + vu^4 + vu^2 - v^2 - v^3,\\
f_1 &= -10u^4 - 6u^2 + 2v + 2v^2 + 4vu^3 + 2vu,\\
f_2 &= -\left(-\frac{4}{5} + \frac{2}{25}v^2\right)u^3 - \frac{12}{25}vu^2 - \left(\frac{41}{25}v^2 + \frac{8}{5}v\right)u + \frac{24}{25}(v^2+v^3),\\
f_3 &= -25\frac{(-24 + 7v^4 + 8v^3 - 80v - 82v^2)u^2}{(-10+v^2)^2} + 50\frac{v(-15v^2 + 4v^3 + 4v^4 - 16v - 4)u}{(-10+v^2)^2}\\
&\quad -50\frac{4v^3 + v^4 + v^5 + 4 + 4v + 4v^2}{(-10+v^2)^2},\\
f_4 &= -[8(12800v + 41680v^2 + 68240v^3 + 52516v^4 + 7268v^5 - 10960v^6\\
&\quad -3152v^7 + 449v^8 + 133v^9 + 8v^{10} + 4v^{11} + 1600)]/[25(-24 + 7v^4\\
&\quad +8v^3 - 80v - 82v^2)^2] + [v(161600v + 548160v^2 + 923680v^3\\
&\quad +727392v^4 + 113716v^5 - 142400v^6 - 41100v^7 + 4456v^8 + 1033v^9\\
&\quad +196v^{10} + 100v^{11} + 19200)u]/[25(-24 + 7v^4 + 8v^3 - 80v - 82v^2)^2],\\
f_5 &= [50(49v^{15} + 161v^{14} - 2148v^{13} - 8948v^{12} + 27908v^{11} + 175332v^{10}\\
&\quad +5760v^9 - 1338048v^8 - 2333952v^7 + 1619072v^6 + 10299904v^5 + 15313920v^4\\
&\quad +11967488v^3 + 5407744v^2 + 1351680v + 147456)v]/[(25v^5 + 24v^4\\
&\quad +728v^3 + 1360v^2 + 848v + 192)^2(-10+v^2)^2].
\end{aligned} \tag{8.6}$$

[3]Copyright ©1981–1992 by the University of Waterloo.

From the Sturm sequence we find the leading coefficient functions:

$$
\begin{aligned}
&-2,\ -10,\ -\left(-\frac{4}{5}+\frac{2}{25}v^2\right),\\
&-25\frac{(-24+7v^4+8v^3-80v-82v^2)}{(-10+v^2)^2},\\
&[v(161600v+548160v^2+923680v^3+727392v^4+113716v^5\\
&-142400v^6-41100v^7+4456v^8+1033v^9\\
&+196v^{10}+100v^{11}+19200)]/[25(-24+7v^4+8v^3-80v-82v^2)^2],\\
&[50(49v^{15}+161v^{14}-2148v^{13}-8948v^{12}+27908v^{11}+175332v^{10}\\
&+5760v^9-1338048v^8-2333952v^7+1619072v^6+10299904v^5+15313920v^4\\
&+11967488v^3+5407744v^2+1351680v+147456)v]/[(25v^5+24v^4\\
&+728v^3+1360v^2+848v+192)^2(-10+v^2)^2].
\end{aligned}
\tag{8.7}
$$

Using the formula for bounds on roots [4] we find that the highest coefficient functions do not change their signs for v belonging to intervals $]-\infty,-312529.98[$ and $]312529.98,+\infty[$. In other words, the estimate of D_1 is $\hat{D}_1 = 312529.98$.

Step 3. All intersections of the variety V_z with A_1, A_2, B_1, and B_2 lie in the interval $]-4,+4[$. It is easy to check that $\hat{D}_3 = 2,\ \hat{D}_4 = 4,\ \hat{D}_5 = 2$, and $\hat{D}_6 = 3$.

Step 4. Therefore, the estimates of sets S_1 and S_2 are defined using the number $\hat{T} = 312529.98$.

Step 5. We now substitute any number v from the interval $]-\infty,-312529.98[$ into (8.5) and find all real roots. We obtain the following set of points in $\Re^2$:

$$\{(-312530,u) : (-312530,+559.04293),(-312530,-559.04293),(-312530,-156265)\}.$$

Similarly, we obtain the set of roots

$$\{(312530,u) : (+312530,559.04383),(312530,-559.04383),(312530,156265)\}$$

when we substitute $v^{**} = 312530$ that belongs to the interval $]312529.98,+\infty[$ into (8.5). All these points represent branches and hence $V_m^{S_1} \subset O_5$ and $V_M^{S_2} \subset O_1$.

Step 6. We conclude that there exists stable constrained dynamics for this system since point 1.a of Lemma 6.1 is satisfied. We could work with better bounds on the roots in order to obtain better estimates for the intervals S_1 and S_2 or better still find the exact roots of the polynomials in the Sturm sequence. However, the proposed method is able to check the existence of the constrained dynamics quickly.

We have provided a constructive method to verify the existence of a criterion of choice leading to (*globally*) stable constrained dynamics. The method of [17, 18] appears not to be able to deal with this aspect in general, as the example shows. Indeed, the feedback law required in the method of [17, 18] for this example cannot be expressed in an explicit form (this requires an analytic solution for a fifth-degree polynomial equation).

9. Conclusion. We have presented necessary and sufficient conditions for output dead beat controllability for a class of discrete time systems described by a single I-O polynomial equation. The highest exponent of the current input is assumed to be an odd integer. The output controllability test amounts to checking whether a set of polynomial divisions is satisfied or not.

We obtained necessary and sufficient conditions for the existence of stable constrained dynamics defined by a scalar implicit equation. We assumed that the constrained dynamics exist for every u_{k-1}, but the defining polynomial $f(0, u_{k-1}, u_k)$ may be even. A dead beat controller that zeros the output of the system in minimum time and which yields stable constrained dynamics is derived.

It is important to say that all algorithms that we presented are computationally expensive and that the computational complexity for systems defined by polynomials of high total degree may be prohibitive. This is an intrinsic feature of polynomial systems and not a drawback of the particular methods that we used.

REFERENCES

[1] G. Bastin, F. Jarachi, and I.M.Y. Mareels, *Dead beat control of recursive nonlinear systems*, in Proc. 32nd IEEE Conference on Decision and Control, San Antonio, TX, 1993, pp. 2965–2971.

[2] F. Jarachi, G. Bastin, and I.M.Y. Mareels, *One-step ahead control of nonlinear discrete time systems with one dimensional zero dynamics: Global stability conditions*, in Proc. NOLCOS-III, Tahoe City, CA, June 26–28, 1995, pp. 848–852.

[3] G. Bastin, F. Jarachi, and I.M.Y. Mareels, *Output deadbeat control of nonlinear discrete time systems with one dimensional zero dynamics: Global stability conditions*, submitted to IEEE Trans. Automat. Control, 1997.

[4] R. Benedetti and J.J. Risler, *Real Algebraic and Semi-Algebraic Sets*, Hermann, Paris, 1990.

[5] S.A. Billings and W.S.F. Voon, *A prediction-error and stepwise-regression estimation algorithm for non-linear systems*, Internat. J. Control, 44 (1986), pp. 803–822.

[6] D. Cox, J. Little, and D. O'Shea, *Ideals, Varieties and Algorithms*, Springer-Verlag, New York, Berlin, 1992.

[7] M.E. Evans and D.N.P. Murthy, *Controllability of a class of discrete time bilinear systems*, IEEE Trans. Automat. Control, 22 (1977), pp. 78–83.

[8] M.E. Evans and D.N.P. Murthy, *Controllability of discrete time inhomogeneous bilinear systems*, Automatica J. IFAC, 14 (1978), pp. 147–151.

[9] S.T. Glad, *Output dead-beat control for nonlinear systems with one zero at infinity*, Systems Control Lett., 9 (1987), pp. 249–255.

[10] S.T. Glad, *Dead beat control for nonlinear systems*, in Analysis and Control of Nonlinear Systems, C.I. Byrnes, C.F. Martin, and R.E. Saeks, eds., North–Holland, Amsterdam, 1988, pp. 437–442.

[11] T. Goka, T.J. Tarn, and J. Zaborszky, *On the controllability of a class of discrete bilinear systems*, Automatica J. IFAC, 9 (1973), pp. 615–622.

[12] O.M. Grasselli, A. Isidori, and F. Nicolo, *Dead-beat control of discrete-time bilinear systems*, Internat. J. Control, 32 (1980), pp. 31–39.

[13] R. Haber and H. Unbehauen, *Structure identification of nonlinear dynamic systems—A survey of input/output approaches*, Automatica, 26 (1990), pp. 651–677.

[14] R. Haber and L. Keviczky, *Identification of nonlinear dynamic systems*, in Proc. 4th IFAC Symp. on Identification and System Parameter Estimation, Tbilisi, USSR, 1978, pp. 79–126.

[15] V.L. Kocic and G. Ladas, *Global Behaviour of Nonlinear Difference Equations of Higher Order with Applications*, Kluwer Academic Publishers, Dordrecht, 1993.

[16] J. O'Reilly, *The discrete linear time invariant time-optimal control problem—An overview*, Automatica J. IFAC, 17 (1981), pp. 363–370.

[17] S. Monaco and D. Normand-Cyrot, *Minimum-phase nonlinear discrete-time systems*, in Proc. 28th Conference on Decision and Control, Los Angeles, CA, 1987, pp. 979–986.

[18] S. Monaco and D. Normand-Cyrot, *Zero dynamics of sampled nonlinear systems*, Systems Control Lett., 11 (1988), pp. 229–234.

[19] N. Rouche, P. Habets, and M. Laloy, *Stability Theory by Liapunov's Direct Method*, Springer-Verlag, Berlin, 1977.
[20] D. Nešić, I.M.Y. Mareels, R. Mahony, and G. Bastin, *ν-step controllability of polynomial scalar systems*, in Proc. 3rd ECCC, Rome, Italy, 1994, pp. 277–282.
[21] D. Nešić and I.M.Y. Mareels, *Dead beat control of polynomial scalar systems*, submitted to Automatica J. IFAC, 1996.
[22] D. Nešić, I.M.Y. Mareels, G. Bastin, and R. Mahony, *Necessary and sufficient conditions for output dead beat controllability for a class of polynomial systems*, in Proc. 34th CDC, New Orleans, LA, 1995, pp. 7–12.
[23] E.D. Sontag, *Polynomial Response Maps*, Springer-Verlag, Berlin, New York, 1979.
[24] E.D. Sontag and Y. Rouchaleau, *On discrete-time polynomial systems*, Nonlinear Anal., 1 (1976), pp. 55–64.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 1, pp. 273–289, January 1998

015

OPTIMAL CONTROL OF PROBLEMS GOVERNED BY ABSTRACT ELLIPTIC VARIATIONAL INEQUALITIES WITH STATE CONSTRAINTS*

MAÏTINE BERGOUNIOUX†

Abstract. In this paper we investigate optimal control problems governed by elliptic variational inequalities with additional state constraints. We present a relaxed formulation for the problem. With penalization methods and approximation techniques we give qualification conditions to get first-order optimality conditions.

Key words. state and control constrained optimal control problems, variational inequalities, optimality conditions

AMS subject classifications. 49J20, 49M29

PII. S0363012996302615

1. Introduction. In this paper we investigate optimal control problems governed by elliptic variational inequalities with additional state constraints. This topic has been widely studied by many authors. We mainly could mention Barbu [1, 2, 3], Friedman [11, 12], Mignot [13], Mignot and Puel [14], Tiba [15], and Bermudez and Saguez [8]. Most of these contributions (for example [1, 2, 3]) study the problem via the penalization of the state (in)equation. On the other hand Mignot and Puel [14] (for instance) give an equivalent formulation of the variational inequality via the associated Lagrange multiplier for the obstacle problem example. We have followed this point of view; our purpose is to set optimality conditions for such a problem that could easily be used from the numerical point of view. This paper is the generalization of the case of the obstacle problem that we have been studying in [5]. We deal here with quite abstract variational inequalities. Following our previous work, we first present a relaxed form of the original problem which can be considered as a good "approximation" of this problem. Then using both Moreau–Yosida approximation techniques and a penalization method we are able to set optimality conditions. We end the paper with the example of the obstacle problem.

2. Setting the problem. Let V and H be a pair of real Hilbert spaces such that V is a dense subset of H and $V \subset H \subset V'$ algebraically and topologically (V' denotes the dual of V). We suppose in addition that

$$\text{the injection } V \subset H \text{ is compact} \tag{2.1}$$

so that $H \subset V'$ is compact too. (For example, one may choose $V = H_o^1(\Omega)$ and $H = L^2(\Omega)$, where Ω is an open bounded "regular" subset of $\mathbb{R}^3$.) We denote $\langle\ ,\ \rangle$ the pairing between V and V', $(\ ,\ )_H$ the H-scalar product and $|\ |_V$ the norm of V. We call $\Lambda_V : V \to V'$ the canonical isomorphism. Let U be another Hilbert space (such that $U = U'$); we consider the variational inequality

$$Ay + \partial\Phi(y) \ni Bu + f, \tag{2.2}$$

*Received by the editors April 26, 1996; accepted for publication (in revised form) November 26, 1996.

http://www.siam.org/journals/sicon/36-1/30261.html

†URA-CNRS 1803, Université d'Orléans, U.F.R. Sciences, B.P. 6759, F-45067 Orléans cedex 2, France (maitine@univ-orleans.fr).

where $A : V \to V'$ is a linear, continuous operator satisfying the coercivity condition

$$\exists \omega > 0 \ \forall v \in V \qquad \langle Av, v\rangle \geq \omega |v|_V^2; \tag{2.3}$$

$$\begin{gathered}\Phi \text{ is a convex, proper, lower semicontinuous (LSC) function} \\ \text{from } V \text{to } \mathbb{R} \cup \{+\infty\}.\end{gathered} \tag{2.4}$$

We denote

$$\text{dom } \Phi = \{\ y \in V \mid \Phi(y) < +\infty\ \}$$

the domain of Φ (which is convex and V-closed). We recall that the subdifferential $\partial\Phi(y_o)$ of Φ at $y_o \in V$ is

$$\partial\Phi(y_o) = \{\ z^* \in V' \mid \forall y \in V \quad \Phi(y) - \Phi(y_o) - \langle z^*, y - y_o\rangle \geq 0\}$$

and that dom $\Phi = \overline{\text{dom } \partial\Phi}$.

$f \in V'$ and

$$B \text{ is a linear, compact operator from } U \text{ to } V'. \tag{2.5}$$

Let us recall some general results about solutions of (2.2) (see [2, 3] for example).

THEOREM 2.1. (Barbu [2, p. 40]). *Under assumptions* (2.3)–(2.4) *and for all* $\psi \in V'$ *the variational inequality*

$$Ay + \partial\Phi(y) \ni \psi$$

has a unique solution $y(\psi) \in V$ *and the mapping* $\psi \mapsto y(\psi)$ *is Lipschitz from* V' *to* V.

COROLLARY 2.1. (Barbu [2, p. 63]). *With the assumptions of the previous theorem, for all* $u \in U$ *there exists a unique* $y(u) \in V$ *solution of* (2.2) *and the mapping* $u \mapsto y(u)$ *is weakly strongly continuous from* U *to* V.

In order to get some regularity results, we suppose from now on that

$$f \in H \text{ and } B \in \mathcal{L}(U, H), \tag{2.6}$$

so that (2.5) is fulfilled and we may use in addition the following result (Barbu [2, p. 42]): let us denote $A_H : H \to H$ the operator

$$A_H(y) = Ay \text{ for all } y \in D(A_H) = \{\ y \in V \mid Ay \in H\ \}. \tag{2.7}$$

This operator is maximal monotone in $H \times H$ and we have Theorem 2.2.

THEOREM 2.2. *Assume* (2.3) *and suppose in addition that there exists* $z \in H$ *and* $c \in \mathbb{R}$ *such that*

$$\forall y \in V, \forall \lambda > 0 \qquad \Phi((I + \lambda A_H)^{-1}(y + \lambda z)) \leq \Phi(y) + c\lambda. \tag{2.8}$$

Then for every $\psi \in H$ *the solution* $y(\psi)$ *of* $Ay + \partial\Phi(y) \ni \psi$ *belongs to* $D(A_H)$ *and*

$$|Ay(\psi)|_H \leq c(1 + |\psi|_H).$$

From now we suppose that (2.8) is ensured. This is the case for example for the obstacle problem given as an example in the last section of this paper, where $V = H_o^1(\Omega)$, $H = L^2(\Omega)$, and $D(A_H) = H^2(\Omega) \cap H_o^1(\Omega)$.

Applying this regularity result to our case we get that for all $u \in U$, $f + Bu \in H$ so that the solution of (2.2) y belongs to $D(A_H) \subset V$; that is, $Ay \in H$.

REMARK 2.1. *One could think that this regularity assumption is not really necessary. Indeed, it is not useful to prove the results of next section. Nevertheless, when we investigate penalized problems, then we shall need some "strong" convergence for the penalized solutions, that is with the compactness assumptions "weak" convergence in the pivot space H.*

Now, we investigate the following optimal control problem:

$$(\mathcal{P}) \qquad \begin{cases} \min \quad g(y) + h(u), \\ Ay + \partial\Phi(y) \ni Bu + f, \\ (y, u) \in K \times U_{ad}, \end{cases}$$

where the following hold.

• g is convex from H to $\mathbb{R}$, finite everywhere ($\mathrm{dom}(g) = H$) and continuous. This implies ([4, Proposition 1.9, p. 85]) that

$$\exists (a_g, c_g) \in H \times \mathbb{R} \quad \text{such that} \quad \forall y \in H \quad g(y) \geq (a_g, y)_H + c_g \tag{2.9}$$

(because g is LSC) and g is everywhere subdifferentiable.

• h is convex from U to $\mathbb{R}$, finite everywhere ($\mathrm{dom}(h) = U$), continuous, and coercive:

$$\lim_{|u|_U \to +\infty} \frac{h(u)}{|u|_U} = +\infty. \tag{2.10}$$

• U_{ad} (resp., K) is a closed, convex, nonempty subset of U (resp., V). We note that $y \in \mathrm{dom}\ \partial\Phi \subset \mathrm{dom}\ \Phi$ so that one may always suppose that

$$K \subset \ \mathrm{dom}\ \Phi. \tag{2.11}$$

REMARK 2.2. *From now, we always suppose that these assumptions are satisfied. They are not of course optimal. To get more information one can refer to Barbu* [3, p. 150].

We end this section with an existence result for $(\mathcal{P})$.

THEOREM 2.3. *Under assumptions* (2.3), (2.4), (2.9), (2.10), *problem* $(\mathcal{P})$ *has (at least) one optimal solution.*

Proof. The proof is quite similar to the one given in Barbu [3, p. 151]. The main difference is the addition of the state constraint $y \in K$, which does not modify the proof. □

3. "Relaxation" of the problem. We denote $\Phi^* : V' \to \mathbb{R}$ the conjugate function of Φ; it is also convex, proper, LSC and we know that (see [4, 10])

$$z \in \partial\Phi(y) \Leftrightarrow y \in \partial\Phi^*(z) \Leftrightarrow \Phi(y) + \Phi^*(z) = \langle y, z \rangle . \tag{3.1}$$

Because of the regularity result we always have $Bu + f - Ay \in H$, so that $z = Bu + f - Ay \in \partial\Phi(y) \cap H$ and (3.1) is equivalent to

$$z \in \partial\Phi(y) \Leftrightarrow y \in \partial\Phi^*(z) \Leftrightarrow \Phi(y) + \Phi^*(z) = (y, z)_H .$$

REMARK 3.1. *In addition such an element z belongs to* $\mathrm{dom}\ \partial\Phi^* \subset \mathrm{dom}\ \Phi^*$ *so that the condition "$z \in \mathrm{dom}\ \Phi^*$" is implicitly included in relation* (3.1).

Finally, problem $(\mathcal{P})$ is equivalent to

$$(\widetilde{\mathcal{P}}) \qquad \begin{cases} \min \quad g(y)+h(u), \\ Ay = Bu + f - z \in H, \\ \Phi(y) + \Phi^*(z) - (y,z)_H = 0, \\ y \in D(A_H) \cap K,\ (u,z) \in U_{ad} \times (\text{dom } \Phi^* \cap H)\,. \end{cases}$$

$w = (u,z)$ is now considered as a new control variable. Problem $(\widetilde{\mathcal{P}})$ is a state-constrained optimal control problem with a nonconvex (because of the bilinear term) constraint coupling the state y and the control w. This constraint is quite difficult to deal with. It is not convex and the equality constraint makes the interior of the feasible domain empty in a very strong sense. So as we have done in [5] for the particular case of the obstacle problem, we had rather study a "relaxed" problem. More precisely we consider

$$(\mathcal{P}_\alpha^R) \qquad \begin{cases} \min \quad g(y)+h(u), \\ Ay = Bu + f - z, \\ \Phi(y) + \Phi^*(z) - (y,z)_H \le \alpha, \\ y \in K,\ (u,z) \in U_{ad} \times B_R^*, \end{cases}$$

where $\alpha > 0,\ R > 0,\ B_R^* = B_H(0,R) \cap \text{dom } \Phi^*$, and $B_H(0,R)$ is the H-ball of radius R. B_R^* is convex and H-closed (since dom Φ^* is convex and V'-closed).

REMARK 3.2. *Let us comment on this "relaxed" form for problem $(\widetilde{\mathcal{P}})$. First we know that $\Phi(y) + \Phi^*(z) - (y,z)_H$ is always nonnegative. So*

$$\Phi(y) + \Phi^*(z) - (y,z)_H \le \alpha \tag{3.2}$$

is equivalent to $|\Phi(y) + \Phi^*(z) - (y,z)_H\,| \le \alpha$. This is the relaxed term: we have replaced the equality "$= 0$" with the inequality "$\le \alpha$," where α may be as small as wanted. This is quite realistic from the numerical point of view where equalities are indeed inequalities up to α.

On the other hand, if we do not add the constraint "$z \in B_H(0,R)$" the relaxed problem is not coercive and so in general it has no solution. Moreover, by virtue of assumptions (2.8) *and* (2.10) *the optimal solutions (y,u) and Ay remain in a bounded set of $H \times U$ and H and the constant R has to be chosen accordingly (that is large enough); in particular, R is greater than $|A\bar{y} - f - B\bar{u}|_H$ for any $(\bar{y},\bar{u})$ solution of $(\mathcal{P})$, so that the feasible domain of $(\mathcal{P}_\alpha^R)$ is nonempty.*

Anyway, this additional condition is not very restrictive. One could instead add a regularization term (as $|z|_H^2/R$) to the cost functional, which would have exactly the same effect. One may also add adapted penalization terms to this cost functional as $|y-\bar{y}|_V^2$ or $|z-\bar{z}|_H^2$.

From now we fix R so that we always omit the index R in the notations and $(\mathcal{P}_\alpha^R)$ becomes $(\mathcal{P}_\alpha)$.

THEOREM 3.1. *For every $\alpha > 0$, $(\mathcal{P}_\alpha)$ has at least one optimal solution denoted $(y_\alpha, u_\alpha, z_\alpha)$. Moreover, when $\alpha \to 0$, y_α strongly converges to $\bar{y}$ in V, u_α weakly converges to $\bar{u}$ in U, and z_α weakly converges to $\bar{z}$ in H where $(\bar{y},\bar{u})$ is a solution of $(\mathcal{P})$ and $\bar{z} = A\bar{y} - B\bar{u} - f \in H$.*

Proof. Let $\alpha > 0$; we have chosen R such that the feasible domain of $(\mathcal{P}_\alpha)$ is always nonempty. We first prove that $d_\alpha = \inf\ (\mathcal{P}_\alpha) \in \mathbb{R}$. The coercivity and continuity assumptions on A yield that A is an isomorphism from V to V'. Let

(y_n, u_n, z_n) be a minimizing sequence: $y_n = A^{-1}(Bu_n + f - z_n)$, $|z_n|_H \leq R$, $u_n \in U_{ad}$, $\Phi(y_n) + \Phi^*(z_n) - (y_n, z_n)_H \leq \alpha$, and $g(y_n) + h(u_n) \to d_\alpha$. Because of (2.9) we have

$$\begin{aligned} g(y_n) + h(u_n) &\geq \left(a_g, A^{-1}(Bu_n + f - z_n)\right)_H + h(u_n) + c_g \\ &\geq \left(a_g, A^{-1}Bu_n\right)_H + h(u_n) - \left(a_g, A^{-1}z_n\right)_H + c_g + \left(a_g, A^{-1}f\right)_H. \end{aligned}$$

As $|z_n|_H \leq R$, then $-\left(a_g, A^{-1}z_n\right)_H + c_g + \left(a_g, A^{-1}f\right)_H$ is bounded from below.

If $d_\alpha = -\infty$, then $\left(a_g, A^{-1}Bu_n\right)_H + h(u_n) \to -\infty$. If (u_n) were bounded in U, then (extracting a subsequence) u_n would be weakly convergent to some $\tilde{u}$ in U; as B is continuous Bu_n would be convergent to $B\tilde{u}$ weakly in H and strongly in V'. Therefore, $A^{-1}Bu_n$would be strongly convergent to $A^{-1}B\tilde{u}$ in V and $\left(a_g, A^{-1}Bu_n\right)_H \to \left(a_g, A^{-1}B\tilde{u}\right)_H$. Moreover h is LSC, so that $-\infty < h(\tilde{u}) \leq \lim_{n\to+\infty} \inf\, h(u_n)$, so we get a contradiction. This means that (u_n) is unbounded. Coercivity assumption (2.10) implies that

$$\lim_{n\to+\infty} \frac{h(u_n)}{|u_n|_U} = +\infty.$$

Moreover the Cauchy–Schwarz inequality shows that

$$\left| \frac{\left(a_g, A^{-1}Bu_n\right)_H}{|u_n|_U} \right| \leq c_o\ \frac{|u_n|_U}{|u_n|_U},$$

so

$$\lim_{n\to+\infty} \left(a_g, A^{-1}Bu_n\right)_H + h(u_n) = |u_n|_U \left[\frac{\left(a_g, A^{-1}Bu_n\right)_H}{|u_n|_U} + \frac{h(u_n)}{|u_n|_U} \right] = +\infty,$$

and we get a contradiction.

• As $|z_n|_H \leq R$ one may extract a subsequence (still denoted z_n) weakly convergent in H to $z_\alpha \in B_R^*$ (since B_R^* is weakly closed). As $d_\alpha > -\infty$, $h(u_n)$ is bounded, and by coercivity (u_n) is bounded in U; so (extracting a subsequence) u_n weakly converges to $u_\alpha \in U_{ad}$ (U_{ad} is weakly closed in U). The continuity of B yields that Bu_n converges to Bu_α weakly in H. So $Ay_n = Bu_n + f - z_n$ converges to $Bu_\alpha + f - z_\alpha$ weakly in H and strongly in V'. As A is an isomorphism from V to V', y_n converges to $y_\alpha = A^{-1}(Bu_\alpha + f - z_\alpha)$ strongly in V. Moreover $y_\alpha \in K$ since K is closed.

• Let us prove that $(y_\alpha, u_\alpha, z_\alpha)$ is feasible for $(\mathcal{P}_\alpha)$. It remains to show that $\Phi(y_\alpha) + \Phi^*(z_\alpha) - (y_\alpha, z_\alpha)_H \leq \alpha$. Φ and Φ^* are convex and LSC so they are weakly LSC and we have

$$\Phi(y_\alpha) + \Phi^*(z_\alpha) \leq \liminf_{n\to+\infty} \Phi(y_n) + \liminf_{n\to+\infty} \Phi^*(z_n) \leq \liminf_{n\to+\infty} \left[\Phi(y_n) + \Phi^*(z_n)\right].$$

Moreover the strong convergence of y_n to y_α in H and the weak convergence of z_n to z_α in H give

$$\lim_{n\to+\infty} (y_n, z_n)_H = (y_\alpha, z_\alpha)_H.$$

Finally

$$\Phi(y_\alpha) + \Phi^*(z_\alpha) - (y_\alpha, z_\alpha)_H \leq \liminf_{n\to+\infty} \left[\Phi(y_n) + \Phi^*(z_n) - (y_n, z_n)_H\right] \leq \alpha.$$

Therefore $(y_\alpha, u_\alpha, z_\alpha)$ is feasible for $(\mathcal{P}_\alpha)$ and $g(y_\alpha) + h(u_\alpha) \geq d_\alpha$. As g and h are LSC, we also have

$$g(y_\alpha) + h(u_\alpha) \leq \liminf_{n\to+\infty} g(y_n) + \liminf_{n\to+\infty} h(u_n) \leq \liminf_{n\to+\infty} [g(y_n) + h(u_n)] = d_\alpha.$$

Finally $g(y_\alpha) + h(u_\alpha) = d_\alpha$ and $(y_\alpha, u_\alpha, z_\alpha)$ is an optimal solution for $(\mathcal{P}_\alpha)$.

• It remains to prove the convergence of $(y_\alpha, u_\alpha, z_\alpha)$ to an optimal solution of $(\mathcal{P})$. Let (y_o, u_o, z_o) be an optimal solution of $(\mathcal{P})$ such that $z_o \in B_R^*$ (remember that we have chosen R to ensure the existence of such a solution). It is also a feasible triple for $(\mathcal{P}_\alpha)$ for any $\alpha > 0$. So

$$\forall \alpha > 0 \quad -\infty < d_\alpha = g(y_\alpha) + h(u_\alpha) \leq g(y_o) + h(u_o) = d_o.$$

So d_α is bounded from above in $\mathbb{R}$. If it were not bounded from below, then we could find a sequence $\alpha_n \to 0$ such that $d_{\alpha_n} \to -\infty$. The same proof as before shows that it is impossible. So $h(u_\alpha)$ is bounded (independently of α) and by coercivity (u_α) is bounded in U as well. Similarly (z_α) is bounded in H ($z_\alpha \in B_R^*$). Then one can show (as we have proved the existence of $(y_\alpha, u_\alpha, z_\alpha)$) that

$$y_\alpha \to \bar{y} \text{ strongly in } V,\ u_\alpha \rightharpoonup \bar{u} \text{ weakly in } U, \text{ and } z_\alpha \rightharpoonup \bar{z} \text{ weakly in } H,$$

where $(\bar{y}, \bar{u})$ is a solution of $(\mathcal{P})$ with $\bar{z} = A\bar{y} - B\bar{u} - f$ and that

$$\lim_{\alpha\to 0} [g(y_\alpha) + h(u_\alpha)] = g(\bar{y}) + h(\bar{u}). \qquad \square$$

4. Penalization of $(\mathcal{P}_\alpha)$.

4.1. The approximated-penalized problem. From now on, we fix also $\alpha > 0$ as small as we want and we shall omit the index α most of time. We are going to approximate and penalize the state equation of $(\mathcal{P}_\alpha)$ to get an approximated problem $(\mathcal{P}_\alpha^\varepsilon)$. Then we shall derive optimality conditions for this problem and set qualification conditions allowing us to pass to the limit with respect to ε. For $\varepsilon > 0$, we consider the following problem:

$$(\mathcal{P}_\alpha^\varepsilon) \qquad \begin{cases} \min J_\varepsilon(y, u, z), \\ (y, u, z) \in K \times U_{ad} \times B_R^*, \end{cases}$$

where

$$\begin{aligned} J_\varepsilon(y, u, z) = \quad & g_\varepsilon(y) + h_\varepsilon(u) \\ & + \frac{1}{2\varepsilon} |Ay - Bu - f + z|_{V'}^2 + \frac{1}{2\varepsilon} [\Phi_\varepsilon(y) + \Phi_\varepsilon^*(z) - (y, z)_H - \alpha]_+^2 \\ & + \frac{1}{2} |y - y_\alpha|_V^2 + \frac{1}{2} |u - u_\alpha|_U^2 + \frac{1}{2} |z - z_\alpha|_H^2. \end{aligned}$$

Here $g_+ = \max(0, g)$, and $g_\varepsilon,\ h_\varepsilon,\ \Phi_\varepsilon$, and Φ_ε^* are the Moreau–Yosida approximations of $g,\ h,\ \Phi$, and Φ^*.

First, we briefly recall some useful properties of the Moreau–Yosida approximation of convex functions. Let φ be a convex, proper, LSC function from $\mathcal{H}$ to $\mathbb{R} \cup \{+\infty\}$ where $\mathcal{H}$ is a Hilbert space (not necessarily identified with its dual). The Moreau–Yosida approximation of φ is defined by

$$\varphi_\varepsilon(x) = \inf \left\{ \frac{|x - y|_{\mathcal{H}}^2}{2\varepsilon} + \varphi(y), y \in \mathcal{H} \right\}$$

and we have the following properties [3, pp. 49–55] in Theorem 4.1.

THEOREM 4.1. *Let us call* $I_\varepsilon = (\Lambda_{\mathcal{H}} + \varepsilon D)^{-1}$ *the proximal mapping with* $D = \partial\varphi$, $\Lambda_{\mathcal{H}}$ *the canonical isomorphism from* $\mathcal{H}$ *to* $\mathcal{H}'$, *and* $D_\varepsilon = -\varepsilon^{-1}\Lambda_{\mathcal{H}}(I_\varepsilon - I)$.

i. I_ε *is single valued and nonexpansive.*

ii. $D_\varepsilon x \in \partial\varphi(I_\varepsilon x)$ *for all* $x \in \mathcal{H}$, *and for all* $x \in \mathrm{dom}(\partial\varphi)$, $\lim_{\varepsilon\to 0} D_\varepsilon x = D^o x \in \partial\varphi(x)$ *(strongly in* $\mathcal{H}'$*), where* $D^o(x)$ *is the element of minimal norm of* $\partial\varphi(x)$.

iii. *For all* $x \in \mathrm{dom}\ \varphi$, $I_\varepsilon x$ *converges strongly in* $\mathcal{H}$ *toward* x.

iv. *If* $\varepsilon_n \to 0$, $x_{\varepsilon_n} \to x_o$ *strongly in* $\mathcal{H}$, *and* $D_{\varepsilon_n} x_{\varepsilon_n} \rightharpoonup y_o$ *weakly in* $\mathcal{H}'$, *then* $y_o \in \partial\varphi(x_o)$.

v. φ_ε *is Fréchet differentiable and* $\varphi'_\varepsilon = D_\varepsilon$ *is Lipschitz (so that* φ_ε *is* $\mathcal{C}^1$*).*

vi. *For all* $x \in \mathcal{H}$ *and* $\varepsilon > 0$ $\varphi(I_\varepsilon x) \le \varphi_\varepsilon(x) \le \varphi(x)$.

Moreover, for all $x \in \mathcal{H}$, $\lim_{\varepsilon\to 0} \varphi_\varepsilon(x) = \varphi(x)$.

In addition, as we need sharper convergence results, we set some further assumptions about the function φ and we suppose that

$$\begin{aligned} &\forall (x_\varepsilon) \in \mathrm{dom}\ \varphi \text{ strongly convergent (in } \mathcal{H}\text{) to } x \in \mathrm{dom}\ \varphi \\ &\qquad \text{then } \varphi'_\varepsilon(x_\varepsilon) \text{ is bounded in } \mathcal{H}' \text{ (with respect to } \varepsilon). \end{aligned} \tag{4.1}$$

Then we have the following useful theorem

THEOREM 4.2. *For any convex, proper, LSC function* φ,

i. *if* x_ε *strongly converges to some* x *in* $\mathcal{H}$, *then* $\lim_{\varepsilon\to 0} I_\varepsilon x_\varepsilon = x$ *(strongly in* $\mathcal{H}$*);*

ii. *if* x_ε *weakly converges to some* $x \in \mathrm{dom}\ \partial\varphi$ *in* $\mathcal{H}$, *then* $\lim_{\varepsilon\to 0} \inf\ \varphi_\varepsilon(x_\varepsilon) \ge \varphi(x)$;

iii. *if* φ *satisfies condition* (4.1) *and if* $x_\varepsilon \in \mathrm{dom}\ \varphi$ *strongly converges to some* $x \in \mathrm{dom}\ \varphi$, *then* $\lim_{\varepsilon\to 0} \varphi_\varepsilon(x_\varepsilon) = \varphi(x)$ *and* $x \in \mathrm{dom}\ \partial\varphi$.

Proof. i and ii are direct consequences of Theorem 4.1. To prove iii, we use the relation (2.18) given in Barbu [3, p. 66]:

$$\forall z, y \in \mathcal{H}, \forall \varepsilon > 0, \quad \varphi_\varepsilon(y) - \varphi_\varepsilon(z) \le \langle \varphi'_\varepsilon(y), y - z\rangle_{\mathcal{H}',\mathcal{H}}$$

where $\langle\ ,\ \rangle_{\mathcal{H}',\mathcal{H}}$ denotes the pairing between $\mathcal{H}$ and $\mathcal{H}'$.

We use it first with $z = x_\varepsilon$ and $y = x$ and then with $y = x_\varepsilon$ and $y = x$; this gives

$$|\varphi_\varepsilon(x_\varepsilon) - \varphi_\varepsilon(x)| \le \max(|\varphi'_\varepsilon(x_\varepsilon)|_{\mathcal{H}'}, |\varphi'_\varepsilon(x)|_{\mathcal{H}'})|x - x_\varepsilon|_{\mathcal{H}}.$$

With Theorem 4.1 ii, assumption (4.1), and the strong convergence of x_ε to x, we get the strong convergence of $\varphi_\varepsilon(x_\varepsilon)$ to $\varphi_\varepsilon(x)$. We conclude with Theorem 4.1 vi. □

REMARK 4.1. *The above property is satisfied for any convex, proper, LSC function* φ *as soon as* $x \in \mathrm{int}(\mathrm{dom}\ \varphi)$ *since* $\partial\varphi(x)$ *is locally bounded in this case (see* [4, p. 60]*). Anyway, here it may happen that* $\mathrm{int}(\mathrm{dom}\ \varphi)$ *is empty and this result cannot be used.*

Now we make precise the hypotheses on functions Φ and Φ^*; from now we assume that

$$\Phi \text{ satisfies (4.1) with } \mathcal{H} = V \text{ and } \Phi^* \text{ satisfies (4.1) with } \mathcal{H} = V'\ . \tag{4.2}$$

This assumption is not so restrictive since it allows us to consider a wide class of convex functions; let us give some examples.

EXAMPLE 4.1. (*convex functions satisfying* (4.1)).

• Any continuous, convex function defined on the whole space V satisfies (4.1) since $\mathrm{int}(\mathrm{dom}\ \varphi) = V$ so that $\partial\varphi(x)$ is locally bounded for any x (we use also Theorem 4.2 i, Theorem 4.1 ii, and a result of Barbu and Precupanu [4, p. 60]).

• Any indicator function $\varphi = 1_C$ of a convex, closed, nonempty subset C of $\mathcal{H}$ satisfies (4.1) also; we recall that

$$1_C(y) = \begin{cases} 0 & \text{if } y \in C, \\ +\infty & \text{else,} \end{cases}$$

and $I_\varepsilon(x) = P_C(x)$ where P_C is the $\mathcal{H}$-projection on C. Then

$$\varphi_\varepsilon(x) = \frac{|x - P_C(x)|^2_{\mathcal{H}}}{2\varepsilon} \text{ and } \varphi'_\varepsilon(x) = \Lambda_{\mathcal{H}} \left(\frac{x - P_C(x)}{\varepsilon} \right).$$

If x_ε strongly converges to x in C =dom φ, then $\varphi'_\varepsilon(x_\varepsilon) = 0$ for all ε and so remains bounded in $\mathcal{H}'$.

EXAMPLE 4.2. (*convex functions satisfying* (4.2)).

• If Φ is the indicator function of a convex closed cone C of V, then $\Phi^* = 1_{C^*}$, where C^* is the polar cone of C in V'; so with Example 4.1 we see that (4.2) is ensured.

This case involves the obstacle problem or the Signorini problem.

• If $\Phi(x) = |x|_V$ is continuous and dom $(\Phi) = H$ then Φ^* is the indicator function of the unit ball of V'.

• $\Phi(x) = \frac{1}{p}|x|^p_V$ then $\Phi^*(x) = \frac{1}{p'}|x|^{p'}_{V'}$ where $p, p' \in]1, +\infty[$ are conjugate numbers (see Ekeland–Temam [10]). This leads to a *semilinear* state equation.

REMARK 4.2. *The approximation process concerns the functions g, h, Φ, and Φ^* which are not necessarily Fréchet differentiable and are replaced by their Moreau–Yosida approximations. This method provides $\mathcal{C}^1$ functions.*

We have also added two kinds of penalization terms: the state equation and the inequality (nonconvex) constraint are penalized in a standard way. The other terms are adapted penalization terms which ensure the strong convergence of the penalized solution toward the desired solution (when uniqueness does not hold).

First we have an existence and convergence result for $(\mathcal{P}^\varepsilon_\alpha)$.

THEOREM 4.3. *For all $\varepsilon > 0$, problem $(\mathcal{P}^\varepsilon_\alpha)$ has (at least) a solution $(y_\varepsilon, u_\varepsilon, z_\varepsilon)$. Moreover, when $\varepsilon \to 0$, $(y_\varepsilon, u_\varepsilon, z_\varepsilon) \to (y_\alpha, u_\alpha, z_\alpha)$ strongly in $V \times U \times H$.*

Proof. We first prove the existence of a solution for $(\mathcal{P}^\varepsilon_\alpha)$. We notice that $(y_\alpha, u_\alpha, z_\alpha)$ is a feasible triple for $(\mathcal{P}^\varepsilon_\alpha)$ so that the feasible domain of $(\mathcal{P}^\varepsilon_\alpha)$ is nonempty, and we may find a minimizing sequence $(y^n_\varepsilon, u^n_\varepsilon, z^n_\varepsilon)$ converging to $d_\varepsilon = \inf(\mathcal{P}^\varepsilon_\alpha)$. Setting $I_{\varepsilon,g} = (I_H + \varepsilon\partial g)^{-1}$ and $I_{\varepsilon,h} = (I_U + \varepsilon\partial h)^{-1}$ we get

$$J_\varepsilon(y, u, z) \geq g_\varepsilon(y) + h_\varepsilon(u) \geq g(I_{\varepsilon,g}(y)) + h(I_{\varepsilon,h}(u))$$

and

$$\inf_{V\times U\times H} J_\varepsilon(y, u, z) \geq \inf_{V\times U} g(I_{\varepsilon,g}(y)) + h(I_{\varepsilon,h}(u)) \geq \gamma > -\infty$$

because of the properties of g and h. So $d_\varepsilon \in \mathbb{R}$ and the end of the proof is standard (see Theorem 3.1).

Now we prove the convergence result. Since $(y_\alpha, u_\alpha, z_\alpha)$ is a feasible triple for $(\mathcal{P}^\varepsilon_\alpha)$ we have

$$d_\varepsilon = J_\varepsilon(y_\varepsilon, u_\varepsilon, z_\varepsilon) \leq g(y_\alpha) + h(u_\alpha) = d_\alpha. \tag{4.3}$$

We have just seen that d_ε is lower bounded (with respect to ε) so that y_ε, u_ε, and z_ε are bounded in V, U, and H. Extracting a subsequence, we get the weak convergence of $(y_\varepsilon, u_\varepsilon, z_\varepsilon)$ to $(\tilde{y}, \tilde{u}, \tilde{z})$ in $V \times U \times H$; in particular, this yields that $Ay_\varepsilon - Bu_\varepsilon - f + z_\varepsilon$ converges to $A\tilde{y} - B\tilde{u} - f + \tilde{z}$ weakly in V'.

Moreover $Ay_\varepsilon - Bu_\varepsilon - f + z_\varepsilon$ converges to zero strongly in V' so that $A\tilde{y} - B\tilde{u} - f + \tilde{z} = 0$. In addition, as U_{ad}, K and B^*_R are weakly closed we get $\tilde{u} \in U_{ad}$, $\tilde{y} \in K$, and $\tilde{z} \in B^*_R$. The injection of V in H is compact, so $y_\varepsilon \to \tilde{y}$ strongly in H; as $z_\varepsilon \rightharpoonup \tilde{z}$

weakly in H we get the convergence of $(y_\varepsilon, z_\varepsilon)_H$ to $(\tilde{y}, \tilde{z})_H$. Moreover Theorem 4.2 gives

$$\liminf_{\varepsilon\to 0} \ \Phi_\varepsilon(y_\varepsilon) \geq \Phi(\tilde{y}) \text{ and } \liminf_{\varepsilon\to 0} \Phi_\varepsilon^*(z_\varepsilon) \geq \Phi^*(\tilde{z}).$$

So we get

$$[\Phi(\tilde{y}) + \Phi^*(\tilde{z}) - (\tilde{y}, \tilde{z})_H - \alpha]_+ \leq \liminf_{\varepsilon\to 0}[\Phi_\varepsilon(y_\varepsilon) + \Phi_\varepsilon^*(z_\varepsilon) - (y_\varepsilon, z_\varepsilon)_H - \alpha]_+.$$

Since $\lim_{\varepsilon\to 0}[\Phi_\varepsilon(y_\varepsilon) + \Phi_\varepsilon^*(z_\varepsilon) - (y_\varepsilon, z_\varepsilon)_H - \alpha]_+^2 = 0$, this yields

$$[\Phi(\tilde{y}) + \Phi^*(\tilde{z}) - (\tilde{y}, \tilde{z})_H - \alpha]_+ = 0.$$

So $(\tilde{y}, \tilde{u}, \tilde{z})$ is feasible for $(\mathcal{P}_\alpha)$. Now relation (4.3) gives

$$g_\varepsilon(y_\varepsilon) + h_\varepsilon(u_\varepsilon) + \frac{1}{2}|y_\varepsilon - y_\alpha|_V^2 + \frac{1}{2}|u_\varepsilon - u_\alpha|_U^2 + \frac{1}{2}|z_\varepsilon - z_\alpha|_H^2 \leq g(y_\alpha) + h(u_\alpha).$$

Passing to the inf-limit in the above relation we get

$$g(\tilde{y}) + h(\tilde{u}) + \frac{1}{2}|\tilde{y} - y_\alpha|_V^2 + \frac{1}{2}|\tilde{u} - u_\alpha|_U^2 + \frac{1}{2}|\tilde{z} - z_\alpha|_H^2 \leq g(y_\alpha) + h(u_\alpha) \leq g(\tilde{y}) + h(\tilde{u}),$$

since $(\tilde{y}, \tilde{u}, \tilde{z})$ is feasible for $(\mathcal{P}_\alpha)$. So $\tilde{y} = y_\alpha$, $\tilde{u} = u_\alpha$ and $\tilde{z} = z_\alpha$. Furthermore $\lim_{\varepsilon\to 0} |y_\varepsilon - y_\alpha|_V = 0$, $\lim_{\varepsilon\to 0} |u_\varepsilon - u_\alpha|_U = 0$, and $\lim_{\varepsilon\to 0} |z_\varepsilon - z_\alpha|_H = 0$ and we get the strong convergence. □

COROLLARY 4.1. *There exists* $(y^*, z^*) \in \partial\Phi^*(z_\alpha) \times \partial\Phi(y_\alpha)$ *such that* $\Phi_\varepsilon'(y_\varepsilon) \rightharpoonup z^*$ *weakly in* V' *and* $\Phi_\varepsilon^{*'}(z_\varepsilon) \rightharpoonup y^*$ *weakly in* V *(and strongly in* H*). Moreover*

$$\lim_{\varepsilon\to 0} \langle \Phi_\varepsilon'(y_\varepsilon), y_\varepsilon\rangle = \langle z^*, y_\alpha\rangle \ \text{ and } \lim_{\varepsilon\to 0} \left(\Phi_\varepsilon^{*'}(z_\varepsilon), z_\varepsilon\right)_H = (y^*, z_\alpha)_H. \tag{4.4}$$

Proof. As $y_\varepsilon \in K \subset \mathrm{dom}(\Phi)$ strongly converges to y_α in V, we use assumption (4.2) to infer that $\Phi_\varepsilon'(y_\varepsilon)$ is bounded in V'. So we may extract a subsequence (denoted similarly) such that $\Phi_\varepsilon'(y_\varepsilon)$ weakly converges in V' to z^*. Theorem 4.1 iv implies that $z^* \in \partial\Phi(y_\alpha)$. Similarly, we may prove that $\Phi_\varepsilon^{*'}(z_\varepsilon) \rightharpoonup y^* \in \partial\Phi^*(z_\alpha)$ weakly in V and strongly in H, since z_ε strongly converges to z_α in V'. Relations (4.4) are obvious. □

4.2. Optimality conditions for $(\mathcal{P}_\alpha^\varepsilon)$. Now, we want to derive optimality conditions for $(\mathcal{P}_\alpha^\varepsilon)$. J_ε is $\mathcal{C}^1$ and the feasible domain of $(\mathcal{P}_\alpha^\varepsilon)$ is convex, so using convex variations we have

$$\forall (y, u, z) \in K \times U_{ad} \times B_R^*, \quad \nabla J_\varepsilon(y_\varepsilon, u_\varepsilon, z_\varepsilon)(y - y_\varepsilon, u - u_\varepsilon, z - z_\varepsilon) \geq 0. \tag{4.5}$$

This leads to the following penalized optimality system in Theorem 4.4.

THEOREM 4.4. *For all* $\varepsilon > 0$ *(small enough), there exist* $q_\varepsilon \in V$ *and* $\lambda_\varepsilon \in \mathbb{R}^+$ *such that*

$$\begin{array}{l} \forall y \in K \\ (g_\varepsilon'(y_\varepsilon), y - y_\varepsilon)_H + (y_\varepsilon - y_\alpha, y - y_\varepsilon)_V + \langle A^* q_\varepsilon + \lambda_\varepsilon[\Phi_\varepsilon'(y_\varepsilon) - z_\varepsilon], y - y_\varepsilon\rangle \geq 0, \end{array} \tag{4.6}$$

$$\forall u \in U_{ad} \quad (h_\varepsilon'(u_\varepsilon) - B^* q_\varepsilon + u_\varepsilon - u_\alpha, u - u_\varepsilon)_U \geq 0, \tag{4.7}$$

$$\forall z \in B_R^* \quad \Big(q_\varepsilon + \lambda_\varepsilon[\Phi_\varepsilon^{*'}(z_\varepsilon) - y_\varepsilon] + z_\varepsilon - z_\alpha, z - z_\varepsilon\Big)_H \geq 0, \tag{4.8}$$

where A^* and B^* are the adjoint operators of A and B.

Proof. Relation (4.5) may be decoupled to obtain

$$\forall y \in K, \quad \nabla_y J_\varepsilon(y_\varepsilon, u_\varepsilon, z_\varepsilon)(y - y_\varepsilon) \geq 0, \tag{4.9}$$

$$\forall u \in U_{ad}, \quad \nabla_u J_\varepsilon(y_\varepsilon, u_\varepsilon, z_\varepsilon)(u - u_\varepsilon) \geq 0, \tag{4.10}$$

$$\forall z \in B_R^*, \quad \nabla_z J_\varepsilon(y_\varepsilon, u_\varepsilon, z_\varepsilon)(z - z_\varepsilon) \geq 0. \tag{4.11}$$

Let us make precise these relations: setting $q_\varepsilon = \Lambda_V^{-1}(s_\varepsilon) \in V$ and

$$s_\varepsilon = \frac{Ay_\varepsilon - Bu_\varepsilon - f + z_\varepsilon}{\varepsilon} \in H \subset V', \ \lambda_\varepsilon = \frac{[\Phi_\varepsilon(y_\varepsilon) + \Phi_\varepsilon^*(z_\varepsilon) - (y_\varepsilon, z_\varepsilon)_H - \alpha]_+}{\varepsilon} \in \mathbb{R}^+ ,$$

relation (4.9) gives for all $y \in K$

$$\begin{aligned} &(g_\varepsilon'(y_\varepsilon), y - y_\varepsilon)_H + (y_\varepsilon - y_\alpha, y - y_\varepsilon)_V \\ +\ &\langle \lambda_\varepsilon[\Phi_\varepsilon'(y_\varepsilon) - z_\varepsilon], y - y_\varepsilon\rangle + \langle q_\varepsilon, A(y - y_\varepsilon)\rangle \geq 0; \end{aligned}$$

introducing the adjoint operator A^* of A we get (4.6). The other relations are obtained similarly. □

REMARK 4.3. *Equation* (4.8) *(and* (4.7) *as well) can be reformulated using the normal cone to B_R^*. Indeed, as B_R^* is convex this normal cone is characterized with*

$$N_{B_R^*}(z_\varepsilon) = \{\ \xi \in H \mid\ (\xi, z_\varepsilon - z)_H \geq 0 \quad \forall z \in B_R^*\ \}$$

(see for instance Clarke [9, p. 53]*), so that relation* (4.8) *is equivalent to*

$$-[q_\varepsilon + \lambda_\varepsilon(\Phi_\varepsilon^{*'}(z_\varepsilon) - y_\varepsilon)] \in z_\varepsilon - z_\alpha + N_{B_R^*}(z_\varepsilon). \tag{4.12}$$

5. Optimality conditions for ($\mathcal{P}_\alpha$). In order to pass to the limit (with respect to ε) in the previous relations we need further estimations on the multipliers q_ε and λ_ε.

5.1. Estimations of the penalized multipliers. Let $(y, u, z) \in K \times U_{ad} \times B_R^*$ and let us add relations (4.6)–(4.8). This gives

$$\begin{aligned} &(g_\varepsilon'(y_\varepsilon), y - y_\varepsilon)_H + (h_\varepsilon'(u_\varepsilon), u - u_\varepsilon)_U \\ +\ &(y_\varepsilon - y_\alpha, y - y_\varepsilon)_V + (u_\varepsilon - u_\alpha, u - u_\varepsilon)_U + (z_\varepsilon - z_\alpha, z - z_\varepsilon)_H \\ +\ &\langle q_\varepsilon, Ay - Bu + z - f - (Ay_\varepsilon - Bu_\varepsilon + z_\varepsilon - f)\rangle \\ +\ &\lambda_\varepsilon \Big[\langle \Phi_\varepsilon'(y_\varepsilon) - z_\varepsilon, y - y_\varepsilon\rangle + \Big(\Phi_\varepsilon^{*'}(z_\varepsilon) - y_\varepsilon, z - z_\varepsilon\Big)_H\Big] \geq 0\ . \end{aligned}$$

Using the definition of q_ε and that $\varepsilon\,\langle q_\varepsilon, \Lambda_V q_\varepsilon\rangle \geq 0$ we get

$$\begin{aligned} \langle -q_\varepsilon, Ay - Bu + z - f\rangle - \lambda_\varepsilon\Big[\langle \Phi_\varepsilon'(y_\varepsilon) - z_\varepsilon, y - y_\varepsilon\rangle + \Big(\Phi_\varepsilon^{*'}(z_\varepsilon) - y_\varepsilon, z - z_\varepsilon\Big)_H\Big] \\ \leq (g_\varepsilon'(y_\varepsilon), y - y_\varepsilon)_H + (h_\varepsilon'(u_\varepsilon), u - u_\varepsilon)_U \\ + (y_\varepsilon - y_\alpha, y - y_\varepsilon)_V + (u_\varepsilon - u_\alpha, u - u_\varepsilon)_U + (z_\varepsilon - z_\alpha, z - z_\varepsilon)_H\,. \end{aligned}$$

The right-hand-side term is bounded since $(y_\varepsilon, u_\varepsilon, z_\varepsilon) \to (y_\alpha, u_\alpha, z_\alpha)$ strongly in $V \times U \times H$, and $(g_\varepsilon'(y_\varepsilon), h_\varepsilon'(u_\varepsilon)) \rightharpoonup (y_\alpha^g, u_\alpha^h) \in \partial g(y_\alpha) \times \partial h(u_\alpha)$ weakly in $H \times U$ (because

of Theorem 4.2 iii and the continuity of g and h). The bounding constant σ depends only on (y, u, z). So we have, for all $\varepsilon > 0$ small enough,

$$-\lambda_\varepsilon \left[\langle \Phi_\varepsilon'(y_\varepsilon) - z_\varepsilon, y - y_\varepsilon \rangle + \left(\Phi_\varepsilon^{*'}(z_\varepsilon) - y_\varepsilon, z - z_\varepsilon \right)_H \right] - \langle q_\varepsilon, Ay - Bu + z - f \rangle \leq \sigma(y, u, z). \tag{5.1}$$

We first estimate the real number λ_ε. If the solution $(y_\alpha, u_\alpha, z_\alpha)$ is such that the nonconvex constraint is inactive, i.e.,

$$\Phi(y_\alpha) + \Phi^*(z_\alpha) - (y_\alpha, z_\alpha)_H - \alpha = G(y_\alpha, z_\alpha) < 0,$$

then convergence results yield

$$\exists \varepsilon_o > 0 \ \forall \varepsilon \leq \varepsilon_o, \ \Phi_\varepsilon(y_\varepsilon) + \Phi_\varepsilon^*(z_\varepsilon) - (y_\varepsilon, z_\varepsilon)_H - \alpha < 0$$

as well and $\lambda_\varepsilon = 0$; hence the limit $\lambda_\alpha = 0$.

Now, we investigate the case when the constraint is active; that is,

$$\Phi(y_\alpha) + \Phi^*(z_\alpha) - (y_\alpha, z_\alpha)_H - \alpha = G(y_\alpha, z_\alpha) = 0.$$

Let us assume the following condition (H_1):

$$\begin{gathered} \forall \alpha \text{ such that } G(y_\alpha, z_\alpha) = 0 \ \forall (y^*, z^*) \in \partial\Phi^*(z_\alpha) \times \partial\Phi(y_\alpha), \\ \exists (\tilde{y}, \tilde{u}, \tilde{z}) \in K \times U_{ad} \times B_R^* \text{ such that } A\tilde{y} = B\tilde{u} + f - \tilde{z}, \text{ and} \end{gathered} \tag{H_1}$$

$$[\Phi(y_\alpha) - \Phi(y^*) - \langle y_\alpha - y^*, \tilde{z} \rangle] + [\Phi^*(z_\alpha) - \Phi^*(z^*) - \langle z_\alpha - z^*, \tilde{y} \rangle] < 2\alpha. \tag{5.2}$$

REMARK 5.1. *Relation* (5.2) *is indeed equivalent to*

$$\langle y_\alpha - y^*, \tilde{z} - z_\alpha \rangle + \langle z_\alpha - z^*, \tilde{y} - y_\alpha \rangle > 0,$$

as we shall prove later. Moreover, in our case, $\langle y_\alpha - y^*, \tilde{z} \rangle = (y_\alpha - y^*, \tilde{z})_H$ *since* $\tilde{z} \in H$.

THEOREM 5.1. *Assume* (H_1)*; then* λ_ε *is bounded by a constant independent of* ε*, and we may extract a subsequence converging to* $\lambda_\alpha \in \mathbb{R}^+$.

Proof. If α is such that $G(y_\alpha, z_\alpha) < 0$, we have already seen that $\lambda_\alpha = 0$.

If $G(y_\alpha, z_\alpha) = 0$, we use (H_1). Let $(y^*, z^*) \in \partial\Phi^*(z_\alpha) \times \partial\Phi(y_\alpha) \subset V \times V'$ be given by Corollary 4.1. Let us apply relation (5.1) with the triple $(\tilde{y}, \tilde{u}, \tilde{z})$ given by (H_1). We get

$$\lambda_\varepsilon \left[\langle z_\varepsilon - \Phi_\varepsilon'(y_\varepsilon), \tilde{y} - y_\varepsilon \rangle + \left(y_\varepsilon - \Phi_\varepsilon^{*'}(z_\varepsilon), \tilde{z} - z_\varepsilon \right)_H \right] \leq \tilde{C} \tag{5.3}$$

and

$$\Phi(y_\alpha) - \Phi(y^*) - (y_\alpha - y^*, \tilde{z})_H + \Phi^*(z_\alpha) - \Phi^*(z^*) - \langle z_\alpha - z^*, \tilde{y} \rangle < 2\alpha.$$

As $y^* \in \partial\Phi^*(z_\alpha)$ and $z^* \in \partial\Phi(y_\alpha)$, we have $\Phi(y^*) + \Phi^*(z_\alpha) = (y^*, z_\alpha)_H$ and $\Phi^*(z^*) + \Phi(y_\alpha) = \langle z^*, y_\alpha \rangle$.

Moreover we are in the case where $\Phi(y_\alpha) + \Phi^*(z_\alpha) = (y_\alpha, z_\alpha)_H + \alpha$, so that (5.2) is equivalent to

$$\rho = (y_\alpha - y^*, \tilde{z} - z_\alpha)_H + \langle z_\alpha - z^*, \tilde{y} - y_\alpha \rangle > 0,$$

as mentioned in Remark 5.1.

Convergence results given in Theorem 4.3 and Corollary 4.1 imply that

$$\lim_{\varepsilon\to 0}\Big(y_\varepsilon-\Phi_\varepsilon^{*'}(z_\varepsilon),\tilde z-z_\varepsilon\Big)_H+\langle z_\varepsilon-\Phi_\varepsilon'(y_\varepsilon),\tilde y-y_\varepsilon\rangle=\rho>0.$$

So, there exists $\varepsilon_o>0$ such that for all $0<\varepsilon<\varepsilon_o$ we have

$$\Big(y_\varepsilon-\Phi_\varepsilon^{*'}(z_\varepsilon),\tilde z-z_\varepsilon\Big)_H+\langle z_\varepsilon-\Phi_\varepsilon'(y_\varepsilon),\tilde y-y_\varepsilon\rangle\geq\frac{\rho}{2}.$$

Then relation (5.3) gives

$$\forall\varepsilon<\varepsilon_o\qquad 0\leq\frac{\rho}{2}\lambda_\varepsilon\leq\tilde C.$$

So λ_ε is bounded and converges to some $\lambda_\alpha\in\mathbb{R}^+$ (extracting a subsequence). □

It remains to bound q_ε. Following [7] we assume the (qualification) condition (H_2):

$$(H_2)\qquad\begin{cases}\exists W \text{ separable Banach subspace such that}\\ W\subset V' \text{ continuously and densely,}\\ \exists\mathcal{M}\subset K\times U_{ad}\times B_R^* \text{ bounded in } V\times U\times H, \text{ such that}\\ 0\in \operatorname{Int}_W\mathcal{T}(\mathcal{M}) \text{ in } W\text{-topology,}\\ \text{where } \mathcal{T}(y,u,z)=Ay-Bu-f+z\ .\end{cases}$$

More precisely, $0\in\operatorname{Int}_W\mathcal{T}(\mathcal{M})$ means the existence of $\rho>0$ such that

$$\forall\xi\in W,\ \ |\xi|_W\leq 1,\ \exists(y_\xi,u_\xi,z_\xi)\in\mathcal{M}\ \text{ such that } Ay_\xi=Bu_\xi+f-z_\xi+\rho\xi.$$

THEOREM 5.2. *Assume (H_1) and (H_2); then q_ε is bounded in W', and one may extract a subsequence converging weak* to q_α in W'.*

Proof. Let $\rho>0$ be given by (H_2) and $\xi\in W$ such that $|\xi|_W\leq 1$. We use relation (5.1) with (y_ξ,u_ξ,z_ξ) and we get

$$\langle -q_\varepsilon,\rho\xi\rangle\leq C_1+C_2\lambda_\varepsilon,$$

where C_1 and C_2 are constants dependent only on (y_ξ,u_ξ,z_ξ). Assumption (H_1) provides a bound for λ_ε and $\mathcal{M}$ is bounded. So there exists a constant C (depending only on $\mathcal{M}$) such that

$$\forall\xi\in W,\ \ |\xi|_W\leq 1,\langle q_\varepsilon,\xi\rangle_{W',W}\leq C$$

(as $W\subset V'$ and $q_\varepsilon\in V$ then $q_\varepsilon\in W'$). Thus q_ε is bounded in W'. □

Now, we are able to pass to the limit in the penalized optimality system with respect to ε.

5.2. Optimality conditions for $(\mathcal{P}_\alpha)$.

THEOREM 5.3. *Let be $\alpha>0$ and assume (H_1) and (H_2); then there exists $(y_\alpha^g,u_\alpha^h,z_\alpha^*,y_\alpha^*)\in\partial g(y_\alpha)\times\partial h(u_\alpha)\times\partial\Phi(y_\alpha)\times\partial\Phi^*(z_\alpha)\subset H\times U\times V'\times V$ and $(q_\alpha,\lambda_\alpha)\in W'\times\mathbb{R}^+$ such that*

$$(5.4)\qquad\begin{array}{l}\forall y\in K \text{ such that } A(y-y_\alpha)\in W,\\ (y_\alpha^g,y-y_\alpha)_H+\langle q_\alpha,A(y-y_\alpha)\rangle_{W',W}+\lambda_\alpha\langle z_\alpha^*-z_\alpha,y-y_\alpha\rangle_{V',V}\geq 0\ ,\end{array}$$

$$(5.5)\qquad\begin{array}{l}\forall u\in U_{ad} \text{ such that } B(u-u_\alpha)\in W,\\ \big(u_\alpha^h,u-u_\alpha\big)_U-\langle q_\alpha,B(u-u_\alpha)\rangle_{W',W}\geq 0,\end{array}$$

$$\begin{aligned}&\forall z \in B_R^* \text{ such that } z - z_\alpha \in W,\\ &\langle q_\alpha, z - z_\alpha\rangle_{W',W} + \lambda_\alpha \left(y_\alpha^* - y_\alpha, z - z_\alpha\right)_H \geq 0,\end{aligned} \tag{5.6}$$

$$\lambda_\alpha \left[\Phi(y_\alpha) + \Phi^*(z_\alpha) - \left(y_\alpha, z_\alpha\right)_H - \alpha\right] = 0. \tag{5.7}$$

Proof. Let $y \in K$ be such that $A(y - y_\alpha) \in W,\ u \in U_{ad}$ such that $B(u - u_\alpha) \in W$ and $z \in B_R^*$ such that $z - z_\alpha \in W$. We use relations (4.6)–(4.8) with these test functions and add them to get

$$\begin{aligned}&\left(g_\varepsilon'(y_\varepsilon), y - y_\varepsilon\right)_H + \left(h_\varepsilon'(u_\varepsilon), u - u_\varepsilon\right)_U\\ &+\ \left(y_\varepsilon - y_\alpha, y - y_\varepsilon\right)_V + \left(u_\varepsilon - u_\alpha, u - u_\varepsilon\right)_U + \left(z_\varepsilon - z_\alpha, z - z_\varepsilon\right)_H\\ &+\ \langle q_\varepsilon, Ay - Bu + z - f - (Ay_\varepsilon - Bu_\varepsilon + z_\varepsilon - f)\rangle\\ &+\ \lambda_\varepsilon \left[\langle \Phi_\varepsilon'(y_\varepsilon) - z_\varepsilon, y - y_\varepsilon\rangle + \left(\Phi_\varepsilon^{*'}(z_\varepsilon) - y_\varepsilon, z - z_\varepsilon\right)_H\right] \geq 0;\end{aligned}$$

that is, as subsection 5.1,

$$\begin{aligned}&\left(g_\varepsilon'(y_\varepsilon), y - y_\varepsilon\right)_H + \left(h_\varepsilon'(u_\varepsilon), u - u_\varepsilon\right)_U\\ &+\ \left(u_\varepsilon - u_\alpha, u - u_\varepsilon\right)_U + \left(z_\varepsilon - z_\alpha, z - z_\varepsilon\right)_H\\ &+\ \langle q_\varepsilon, A(y - y_\alpha) - B(u - u_\alpha) + z - z_\alpha\rangle_{W',W}\\ &+\ \lambda_\varepsilon \left[\langle \Phi_\varepsilon'(y_\varepsilon) - z_\varepsilon, y - y_\varepsilon\rangle + \left(\Phi_\varepsilon^{*'}(z_\varepsilon) - y_\varepsilon, z - z_\varepsilon\right)_H\right] \geq 0.\end{aligned}$$

As g and h are continuous and finite everywhere, $(g_\varepsilon'(y_\varepsilon), h_\varepsilon'(u_\varepsilon))$ converges toward some $(y_\alpha^g, u_\alpha^h) \in \partial g(y_\alpha) \times \partial h(y_\alpha)$. Then we may pass to the limit in the above relation to infer

$$\begin{aligned}&\left(y_\alpha^g, y - y_\alpha\right)_H + \left(u_\alpha^h\right), u - u_\alpha\big)_U\\ &+\ \langle q_\alpha, A(y - y_\alpha) - B(u - u_\alpha) + z - z_\alpha\rangle_{W',W}\\ &+\ \lambda_\alpha \left[\langle z_\alpha^* - z_\alpha, y - y_\alpha\rangle + \left(y^*\alpha - y_\alpha, z - z_\alpha\right)_H\right] \geq 0\ .\end{aligned}$$

Taking in turn $y = y_\alpha,\ u = u_\alpha$, and $z = z_\alpha$ we obtain relations (5.4)–(5.6).

Finally if $\Phi(y_\alpha) + \Phi^*(z_\alpha) - \left(y_\alpha, z_\alpha\right)_H - \alpha < 0$ we have seen that $\lambda_\alpha = 0$. So relation (5.7) is satisfied. □

REMARK 5.2. *As we already mentioned in Remark* 4.3*, equation* (5.6) *is equivalent to*

$$-q_\alpha - \lambda_\alpha(y_\alpha^* - y_\alpha) \in N_{B_R^* \cap (z_\alpha + W)}(z_\alpha). \tag{5.8}$$

COROLLARY 5.1. *With assumptions of Theorem* 5.3*, there exist* $(y_\alpha^g, u_\alpha^h) \in \partial g(y_\alpha) \times \partial h(u_\alpha) \subset H \times U$ *and* $(q_\alpha, \lambda_\alpha) \in W' \times \mathbb{R}^+$ *such that*

$$\begin{aligned}&\forall y \in K \text{ s.t. } A(y - y_\alpha) \in W,\\ &\left(y_\alpha^g, y - y_\alpha\right)_H + \langle q_\alpha, A(y - y_\alpha)\rangle_{W',W}\\ &+\ \lambda_\alpha \left[\Phi(y) + \Phi^*(z_\alpha) - \left(y, z_\alpha\right)_H - \alpha\right] \geq 0\end{aligned} \tag{5.9}$$

$$\forall u \in U_{ad} \text{ s.t. } B(u - u_\alpha) \in W \qquad \left(u_\alpha^h, u - u_\alpha\right)_U - \langle q_\alpha, B(u - u_\alpha)\rangle_{W',W} \geq 0,$$

$$\begin{aligned}&\forall z \in B_R^* \text{ s.t. } z - z_\alpha \in W,\\ &\langle q_\alpha, z - z_\alpha\rangle_{W',W} + \lambda_\alpha \left[\Phi(y_\alpha) + \Phi^*(z) - \left(y_\alpha, z\right)_H - \alpha\right] \geq 0,\end{aligned} \tag{5.10}$$

$$\lambda_\alpha \left[\Phi(y_\alpha) + \Phi^*(z_\alpha) - \left(y_\alpha, z_\alpha\right)_H - \alpha\right] = 0.$$

Proof. Theorem 5.3 gives $(z_\alpha^*, y_\alpha^*) \in \partial\Phi(y_\alpha) \times \partial\Phi^*(z_\alpha) \subset V' \times V$ such that relations (5.4) and (5.6) are satisfied.

As $z_\alpha^* \in \partial\Phi(y_\alpha)$ we get, for all $y \in K$ such that $A(y - y_\alpha) \in W$,

$$\Phi(y) - \Phi(y_\alpha) \geq \langle z_\alpha^*, y - y_\alpha \rangle ,$$

so that relation (5.4) becomes

$$(y_\alpha^g, y - y_\alpha)_H + \langle q_\alpha, A(y - y_\alpha)\rangle_{W',W} + \lambda_\alpha \left[\Phi(y) - \Phi(y_\alpha) - (z_\alpha, y - y_\alpha)_H\right] \geq 0.$$

Using (5.7) we obtain relation (5.9). Similarly, we can show relation (5.10). ∎

COROLLARY 5.2. *With assumptions of the previous theorem and if g and h are Gâteaux differentiable, there exists $(q_\alpha, \lambda_\alpha) \in W' \times \mathbb{R}^+$ such that*

$$\begin{array}{l} \forall y \in K \ \textit{s.t.}\ A(y - y_\alpha) \in W, \\ (g'(y_\alpha), y - y_\alpha)_H + \langle q_\alpha, A(y - y_\alpha)\rangle_{W',W} + \lambda_\alpha \left[\Phi(y) + \Phi^*(z_\alpha) - (z_\alpha, y)_H - \alpha\right] \geq 0, \end{array}$$

$$\forall u \in U_{ad} \ \textit{s.t.}\ B(u - u_\alpha) \in W \qquad (h'(u_\alpha), u - u_\alpha)_U - \langle q_\alpha, B(u - u_\alpha)\rangle_{W',W} \geq 0,$$

$$\begin{array}{l} \forall z \in B_R^* \ \textit{s.t.}\ z - z_\alpha \in W, \\ \langle q_\alpha, z - z_\alpha\rangle_{W',W} + \lambda_\alpha \left[\Phi(y_\alpha) + \Phi^*(z) - (z, y_\alpha)_H - \alpha\right] \geq 0, \end{array}$$

$$\lambda_\alpha \left[\Phi(y_\alpha) + \Phi^*(z_\alpha) - (y_\alpha, z_\alpha)_H - \alpha\right] = 0.$$

REMARK 5.3. *The natural idea would now be to study the asymptotic behavior of the previous optimality system when $\alpha \to 0$. Unfortunately, we would have to set an "$(H1)$-like" assumption with $\alpha = 0$, to be able to pass to the limit in the α-optimality system. This is impossible since the interior of the feasible domain of $\mathcal{P}$ is empty because of the nonconvex equality constraint and an assumption like $(H1)$ with $\alpha = 0$ would never be ensured. However, as we have already mentioned, this relaxed approach is sufficient for numerical applications.*

6. Example of the obstacle problem. In this section we study an example where the variational inequality leads to an obstacle problem.

Let Ω be an open, bounded subset of $\mathbb{R}^n$ with a smooth boundary $\partial\Omega$. We consider a bilinear form $a(.,.)$ defined on $H_o^1(\Omega) \times H_o^1(\Omega)$ and A the continuous linear operator from $H_o^1(\Omega)$ to $H^{-1}(\Omega)$ associated with a such that

$$\left\{ \begin{array}{c} Ay = -\displaystyle\sum_{i,j=1}^{n} \partial_{x_i}(a_{ij}(x)\partial_{x_j} y) + a_0(x)y \ \text{ with} \\ a_{ij}, a_0 \in \mathcal{C}^2(\bar{\Omega}) \text{ for } i,j = 1,\ldots,n, \ \inf\ \{a_0(x) \mid x \in \bar{\Omega}\} > 0, \\ \displaystyle\sum_{ij=1}^{n} a_{ij}(x)\xi_i\xi_j \geq \delta \sum_{i=1}^{n} \xi_i^2 \forall x \in \bar{\Omega} \forall \xi \in \mathbb{R}^n, \delta > 0. \end{array} \right. \tag{6.1}$$

We shall denote $\|\ \|$, the $L^2(\Omega)$-norm, $(\ ,\ )$ the $L^2(\Omega)$-scalar product, and $\langle\ ,\ \rangle$ any duality product. We set

$$V = H_o^1(\Omega)\ ,\ H = L^2(\Omega)\ ,\ D_H(A) = H^2(\Omega) \cap H_o^1(\Omega)\ ,\ U = L^2(\Omega), \text{ and } B = Id_{L^2(\Omega)}.$$

Let us set also

$$K = V \text{ and } C^+ = \{y \mid y \in H_o^1(\Omega),\ y \geq 0 \text{ a.e. in } \Omega\}.$$

The convex function Φ is the indicator function I_+ of C^+ Then Φ^* is the indicator function I_- of the negative cone C^- of H^{-1}, and we have already mentioned that Φ and Φ^* satisfy condition (4.2). Then we get as a state equation

$$Ay = f + v - z \text{ in } \Omega, y = 0 \text{ on } \Gamma, \tag{6.2}$$

with $f,\ v$, and z belonging to $L^2(\Omega)$ (because of the regularity result mentioned in section 1). The constraint $z \in \partial\Phi(y)$ becomes

$$y \geq 0\ ,\ z \leq 0\ ,\ (y,z) = 0,$$

and the α-inequality constraint $\Phi(y) + \Phi^*(z) - (y,z) \leq \alpha$ gives:

$$y \geq 0\ ,\ z \leq 0\ ,\ (y,-z) \leq \alpha.$$

We set $\xi = -z$, so that the original control problem is defined as follows (see [5]):

$$\min\left\{J(y,v) = \frac{1}{2}\int_\Omega (y - z_d)^2\,dx + \frac{M}{2}\int_\Omega v^2\,dx\right\}, \tag{$\mathcal{P}$}$$

$$Ay = f + v + \xi \quad \text{in} \quad \Omega,\ y = 0 \quad \text{on} \quad \Gamma, \tag{6.3}$$

$$(y,v,\xi) \in \mathcal{D}, \tag{6.4}$$

where

$$\mathcal{D} = \{(y,v,\xi) \in H_o^1(\Omega) \times L^2(\Omega) \times L^2(\Omega) \mid v \in U_{ad},\ y \geq 0,\ \xi \geq 0, (y,\xi) = 0\}$$

and $z_d \in L^2(\Omega)$. The relaxed problem is

$$\min J(y,v), \tag{$\mathcal{P}^\alpha$}$$

$$Ay = v + \xi \quad \text{in} \quad \Omega,\ y \in H_o^1(\Omega), \tag{6.5}$$

$$(y,v,\xi) \in \mathcal{D}_\alpha^R, \tag{6.6}$$

where

$$\mathcal{D}_\alpha^R = \{(y,v,\xi) \in H_o^1(\Omega) \times L^2(\Omega) \times L^2(\Omega) \mid v \in U_{ad},\ y \geq 0,\ \xi \geq 0, \|\xi\| \leq R, (y,\xi) \leq \alpha\}.$$

The results of the previous section may be applied with $W = L^p(\Omega)$ and we get Theorem 6.1.

THEOREM 6.1. *Assume*

$$\begin{array}{c} \forall \alpha \text{ such that } \langle y_\alpha, \xi_\alpha\rangle = \alpha\ , \\ \exists(\tilde{y},\tilde{v},\tilde{\xi}) \in C^+ \times U_{ad} \times B_R^* \quad \text{such that} \\ A\tilde{y} = \tilde{v} + \tilde{\xi} \quad \text{and} \quad (\tilde{y},\xi_\alpha) + \left(y_\alpha,\tilde{\xi}\right) < 2\alpha \end{array} \tag{$\mathcal{H}_1$}$$

and

$$(\mathcal{H}_2)\qquad \begin{array}{c}\exists p\in[1,+\infty[,\ \exists\ \rho>0\quad \forall\ \chi\in L^p(\Omega),\ \|\chi\|_{L^p(\Omega)}\leq 1,\\ \exists(y_\chi,v_\chi,\xi_\chi)\ \textit{bounded in } C^+\times U_{ad}\times B_R^*\ \textit{(independently of } \chi)\\ \textit{such that } Ay_\chi=v_\chi+\xi_\chi+\rho\chi\quad \textit{in } \Omega,\end{array}$$

and let $(y_\alpha,v_\alpha,\xi_\alpha)$ *be a solution of* $(\mathcal{P}^\alpha)$*; then a Lagrange multiplier* $(q_\alpha,\lambda_\alpha)\in L^{p'}(\Omega)\times\mathbb{R}^+$ *exists such that*

$$\begin{array}{l}\forall y\in C^+\quad \textit{such that } A(y-y_\alpha)\in L^p(\Omega),\\ (y_\alpha-z_d,y-y_\alpha)+\langle q_\alpha,A(y-y_\alpha)\rangle+\lambda_\alpha\,(\xi_\alpha,y-y_\alpha)\geq 0\ ,\end{array}\tag{6.7}$$

$$\forall v\in U_{ad},\ v-v_\alpha\in L^p(\Omega),\quad \langle Mv_\alpha-q_\alpha,v-v_\alpha\rangle\ \geq 0,\tag{6.8}$$

$$\forall\xi\in B_R^*,\quad \xi-\xi_\alpha\in L^p(\Omega),\quad \langle\lambda_\alpha y_\alpha-q_\alpha,\xi-\xi_\alpha\rangle\ \geq 0,\tag{6.9}$$

$$\lambda_\alpha\,((y_\alpha,\xi_\alpha)-\alpha)=0.\tag{6.10}$$

For more details one can refer to [5]. We just mention that assumptions $(\mathcal{H}_1)$ and $(\mathcal{H}_2)$ are satisfied, for instance, if $U_{ad}=L^2(\Omega)$ or $U_{ad}=\{v\in L^2(\Omega)\mid v\geq\psi\geq 0$ a.e. in $\Omega\}$.

7. Conclusion. As already mentioned at the beginning of this paper, we have in mind the numerical aspects of the question: that is, why we have underlined that the "relaxed" problem $\mathcal{P}_\alpha$ is a good approximation of the original problem. Now, we think that the main tool for a good numerical approach for such problems is the (necessary) optimality conditions that we have obtained in Theorem 5.3. They allow us to interpret the optimal solution as the first argument of the saddle point of a linearized Lagrangian function, although the problem is not convex. We have developed this point of view and presented some algorithms in [6] for the case of the obstacle problem. The numerical behavior of these methods is quite nice.

On the other hand, though we have not tested methods using Yosida approximation, we believe that the use of penalization is not helpful for numerics. It seems to be too unstable (because of the suitable choice of the parameter ε), and we think it is only a theoretical tool.

REFERENCES

[1] V. Barbu, *Necessary conditions for non convex distributed control problems governed by elliptic variational inequalities*, J. Math. Anal. Appl., 80 (1981), pp. 566–598.
[2] V. Barbu, *Optimal Control of Variational Inequalities*, Res. Notes Math. 100, Pitman, Boston, 1984.
[3] V. Barbu, *Analysis and Control of Infinite Dimensional Systems*, Math. Sci. Engrg. 190, Academic Press, New York, 1993.
[4] V. Barbu and T. Precupanu, *Convexity and Optimization in Banach Spaces*, Sijthoff and Noordhoff, Leyden, 1978.
[5] M. Bergounioux, *Optimal control of an obstacle problem*, Appl. Math. Optim., 36 (1997), pp. 147–172.
[6] M. Bergounioux, *On the Use of Augmented Lagrangian Methods for Optimal Control of Obstacle Problems*, J. Optim. Theory Appl., 95 (1997).
[7] M. Bergounioux and D. Tiba, *General optimality conditions for constrained convex control problems*, SIAM J. Control Optim., 34 (1996), pp. 698–711.

[8] A. Bermudez and C. Saguez, *Optimal control of variational inequalities*, Control Cybernet., 14 (1985), pp. 9–30.
[9] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Classics in Applied Mathematics 5, SIAM, Philadelphia, 1990.
[10] I. Ekeland and R. Temam, *Analyse Convexe et Problèmes Variationnels*, Dunod-Gauthier-Villars, Paris, 1974.
[11] A. Friedman, *Variational Principles and Free-Boundary Problems*, Wiley, New York, 1982.
[12] A. Friedman, *Optimal control for variational inequalities*, SIAM J. Control Optim., 24 (1986), pp. 439–451.
[13] F. Mignot, *Contrôle dans les inéquations variationnelles elliptiques*, J. Funct. Anal., 22 (1976), pp. 130–185.
[14] F. Mignot and J. P. Puel, *Optimal control in some variational inequalities*, SIAM J. Control Optim., 22 (1984), pp. 466–476.
[15] D. Tiba, *Optimal Control of Nonsmooth Distributed Parameter Systems*, Lecture Notes in Math. 1459, Springer-Verlag, Berlin, 1990.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 1, pp. 290–312, January 1998

016

ON THE ATTAINABLE SET FOR SCALAR NONLINEAR CONSERVATION LAWS WITH BOUNDARY CONTROL*

FABIO ANCONA† AND ANDREA MARSON‡

Abstract. We consider the initial value problem with boundary control for a scalar nonlinear conservation law

$$(*) \qquad u_t + [f(u)]_x = 0, \qquad u(0,x) = 0, \qquad u(\cdot,0) = \tilde{u} \in \mathcal{U},$$

on the domain $\Omega = \{(t,x) \in \mathbb{R}^2 : \ t \geq 0, \ x \geq 0\}$. Here $u = u(t,x)$ is the state variable, $\mathcal{U}$ is a set of bounded boundary data regarded as controls, and f is assumed to be strictly convex. We give a characterization of the set of attainable profiles at a fixed time $T > 0$ and at a fixed point $\bar{x} > 0$:

$$\begin{aligned} &\mathcal{A}(T,\mathcal{U}) = \{u(T,\cdot) : \ u \text{ is a solution of } (*)\}, \\ &\mathcal{A}(\bar{x},\mathcal{U}) = \{u(\cdot,\bar{x}) : \ u \text{ is a solution of } (*)\}, \end{aligned} \qquad \mathcal{U} = L^\infty(\mathbb{R}^+).$$

Moreover we prove that $\mathcal{A}(T,\mathcal{U})$ and $\mathcal{A}(\bar{x},\mathcal{U})$ are compact subsets of L^1 and $L^1{}_{loc}$, respectively, whenever $\mathcal{U}$ is a set of controls which pointwise satisfy closed convex constraints, together with some additional integral inequalities.

Key words. conservation laws, boundary control, attainable set

AMS subject classifications. 35B37, 35L65

PII. S0363012996304407

1. Introduction.

The paper is concerned with the initial boundary value problem for a scalar nonlinear conservation law in one space dimension:

$$u_t + [f(u)]_x = 0, \tag{1.1}$$

$$u(0,x) = 0, \qquad t,x \geq 0, \tag{1.2}$$

$$u(t,0) = \tilde{u}(t), \tag{1.3}$$

where $u = u(t,x)$ is the state variable, $\tilde{u}$ is a measurable bounded boundary data, and f is assumed to be a strictly convex function. Following [14] we shall consider only weak entropic solutions of (1.1)–(1.2) which satisfy the boundary condition (1.3) in a weak sense.

Here we study the system (1.1)–(1.3) from the point of view of control theory [8], regarding the boundary data $\tilde{u}$ as a control. Given a set $\mathcal{U} \subset L^\infty(\mathbb{R}^+)$ of admissible controls, we study the set of attainable profiles at a fixed time T

$$\mathcal{A}(T,\mathcal{U}) = \Big\{u(T,\cdot) : \ u \text{ is a solution to (1.1)–(1.3) with } \tilde{u} \in \mathcal{U}\Big\}.$$

We will give a precise characterization of the attainable set when $\mathcal{U} = L^\infty(\mathbb{R}^+)$ by using the theory of generalized characteristics developed by Dafermos [5]. Applications to calculus of variations and problems of optimization motivate the study of topological properties of $\mathcal{A}(T,\mathcal{U})$. Here closure and compactness of the attainable set will

*Received by the editors May 28, 1996; accepted for publication (in revised form) November 27, 1996. This research was partially supported by TMR project HCL ERBFMRXCT360033.
http://www.siam.org/journals/sicon/36-1/30440.html

†Dipartimento di Matematica and CIRAM, Università di Bologna, Piazza P.S. Donato n. 5, Bologna 40127, Italy (ancona@ciram3.ing.unibo.it).

‡Dipartimento di Elettronica per l'Automazione, Università di Brescia, Via Branze N. 38, Brescia 25123, Italy (marson@bsing.ing.unibs.it).

be established in connection with classes of boundary controls which are measurable selections of a bounded multifunction with closed convex values and satisfy certain integral inequalities. In the proof of such results a key role will be played by the weak* compactness of the set of fluxes $\{f(\tilde{u}) : \ u \in \mathcal{U}\}$ of admissible boundary controls.

Results concerning the set of attainable profiles at a fixed point in space $\bar{x} > 0$,

$$\mathcal{A}(\bar{x}, \mathcal{U}) = \Big\{ u(\cdot, \bar{x}) : \ u \text{ is a solution to (1.1)–(1.3) with } \tilde{u} \in \mathcal{U} \Big\},$$

can be derived by similar arguments.

The compactness of the attainable sets allows us to prove the existence of solutions for a class of optimization problems, where the cost functional depends on the profiles of the solutions at some time T or at a fixed point $\bar{x}$. In section 5 we apply these results to a model of traffic flow where one wants to minimize the average time spent by cars travelling through a given stretch of highway. The controller acts by varying the density of cars entering the highway.

2. Preliminaries and statements of main results.

2.1. Formulation of the problem. On the domain $\Omega = \{(t, x) \in \mathbb{R}^2 : \ t \geq 0, \ x \geq 0\}$ consider the mixed initial boundary value hyperbolic problem

$$u_t + [f(u)]_x = 0, \tag{2.1}$$

$$u(0, x) = \bar{u}(x), \qquad t, x \geq 0, \tag{2.2}$$

$$u(t, 0) = \tilde{u}(t), \tag{2.3}$$

where $\tilde{u} \in L^\infty(\mathbb{R}^+)$, $\bar{u} \in L^\infty(\mathbb{R}^+) \cap L^1(\mathbb{R}^+)$, and $f : \mathbb{R} \to \mathbb{R}$ is a twice continuously differentiable strictly convex function. Denote $b(x) = (f')^{-1}(x)$ whenever $x \in \text{Range } (f')$ and $b(0) = -\infty$ if $0 \notin \text{Range } (f')$.

We recall that problems of this type do not possess classical solutions since discontinuities arise in finite time even if the initial and boundary data are smooth (see [4], [15]). Hence it is natural to consider weak solutions in the sense of distributions satisfying the usual *entropy conditions* [11], [13]

$$u(t, x-) \geq u(t, x+), \qquad t, x > 0. \tag{2.4}$$

As pointed out in [3], [6], and [14], in general the Dirichlet condition (2.3) may not be fulfilled pointwise a.e.; thus following [14] we require that an entropic solution u to (2.1)–(2.3) satisfies the above condition in a weaker sense which is motivated by the classical vanishing viscosity method (see [3], [14], and Definition 1). In [3] an entropic solution to (2.1)–(2.3) is obtained as the limit of solutions of suitable approximating parabolic problems, while in [14] Le Floch generalizes a result of Lax for the Cauchy problem for the scalar conservation law (see [12]), expressing a solution in terms of the pointwise minimum of a function $y \mapsto \Psi(t, x, y)$ for any $(t, x) \in \mathbb{R}^+ \times \mathbb{R}^+$ (see also Remark 2.1). Concerning uniqueness, in [14] an L^1-semigroup property in the class of piecewise regular solutions is established (see Remark 2.2).

As observed in [14], any solution of (2.1)–(2.3) with boundary data $\tilde{u}$ such that $f'(\tilde{u}(t)) < 0$ on a subset I of $\mathbb{R}^+$ of positive measure can be obtained with the boundary data

$$\tilde{u}'(t) = \begin{cases} b(0) & \text{if} \quad t \in I, \\ \tilde{u}(t) & \text{otherwise.} \end{cases}$$

Hence it is not restrictive to assume that the characteristics at the boundary are always entering the domain, i.e., $f'(\tilde{u}(t)) \geq 0$ for a.e. t: this hypothesis will be adopted in the rest of the paper. We recall here the definition of the solution to (2.1)–(2.3) as stated in [14].

DEFINITION 1. *A function $u \in L^1(\Omega;\mathbb{R})$ is a solution of (2.1)–(2.3) if*

(i) *it is a weak entropic solution of (2.1) in the interior of Ω;*

(ii) *there exists a set $\mathcal{E} \subset \mathbb{R}^+$ with zero measure such that*

$$\lim_{\substack{t\to 0^+ \\ t\notin\mathcal{E}}} \int_0^x u(t,\xi)\ d\xi = \int_0^x \bar{u}(\xi)\ d\xi, \qquad x \geq 0; \tag{2.5}$$

(iii) *the boundary condition is satisfied in the following weak sense: there exist a set $\mathcal{F} \subset \mathbb{R}^+$ with zero measure and two functions $\Upsilon : \mathbb{R}^+ \to \mathbb{R}$ and $\mu : \mathbb{R}^+ \to \{-1, 0, 1\}$ such that*

$$\lim_{\substack{x\to 0^+ \\ x\notin\mathcal{F}}} \int_0^t f(u(s,x))\ ds = \int_0^t \Upsilon(s)\ ds, \qquad t \geq 0, \tag{2.6}$$

$$\lim_{\substack{x\to 0^+ \\ x\notin\mathcal{F}}} sgn\ f'(u(t,x)) = \mu(t), \qquad a.e.\ t \geq 0, \tag{2.7}$$

and

$$\begin{cases} \Upsilon(t) = f(\tilde{u}(t)) & \text{if } \mu(t) \geq 0, \\ \Upsilon(t) \geq f(\tilde{u}(t)) & \text{if } \mu(t) = -1 \end{cases} \qquad a.e.\ t > 0. \tag{2.8}$$

Remark 2.1. In [14] Le Floch proves that under the above assumptions there exists a solution u to (2.1)–(2.3), having right and left limits in t and x at every point in the interior of Ω and such that for any fixed $t \geq 0$ $u(t,\cdot)$ has at most countably many discontinuities. Moreover it satisfies the bounds

(2.9)
$$\|u(\cdot,\cdot)\|_\infty \leq \max\left\{\|\bar{u}(\cdot)\|_\infty, \|\tilde{u}(\cdot)\|_\infty\right\},$$
$$\min\left\{f(u) :\ |u| \leq \|\tilde{u}\|_\infty, \|\bar{u}(\cdot)\|_\infty\right\} \leq\ \Upsilon(t)\ \leq \max\left\{\|f(\bar{u}(\cdot))\|_\infty,\ \|f(\tilde{u})(\cdot)\|_\infty\right\}$$

for a.e. $t > 0$. Such a solution admits the following explicit representation inside the domain:

$$u(t,x) = b\left(\frac{x - y(t,x)}{t}\right), \qquad t > 0, \quad x > 0, \tag{2.10}$$

where $y(t,x)$ denotes a point of minimum value for the function

$$y \mapsto \Psi_\Upsilon(t,x,y) = \begin{cases} \displaystyle\int_0^y \bar{u}(s)\ ds + t\ g\left(\frac{x-y}{t}\right) & \text{if } y \geq 0, \\ \displaystyle -\int_0^\tau \Upsilon(s)\ ds + (t-\tau)\ g\left(\frac{x}{t-\tau}\right) & \text{if } y \leq 0, \end{cases} \tag{2.11}$$

with g denoting the Legendre transform of a superlinear convex map $\tilde{f}$ which coincides with f on the closed ball $\{u \in \mathbb{R} :\ |u| \leq \|\tilde{u}\|_\infty\}$ and τ satisfying

$$\frac{x-y}{t} = \frac{x}{t-\tau}, \qquad y \leq 0.$$

Notice that in [11] it is shown that for any given $t \in [0,T]$ the function $y \mapsto \Psi_\Upsilon(t,x,y)$ attains its minimum at a single point for all but at most countably many $x > 0$. Furthermore the existence of the traces at $x = 0$ in the sense of (2.6)–(2.7) for the functions $f(u)$, sgn $f'(u)$ holds in general for any map u admitting a representation as in (2.10) with Ψ_Υ defined by (2.11) in connection with some L^∞ function Υ.

Remark 2.2. Regarding uniqueness in [14], the following L^1-semigroup property is established: if u and v are piecewise continuously differentiable solutions of (2.1)–(2.3) associated with initial and boundary data $\bar u$, $\tilde u$ and $\bar v$, $\tilde v$, respectively ($\tilde u$, $\tilde v \geq b(0)$), then

(2.12)
$$\int_0^{+\infty} |u(t,x) - v(t,x)|\, dx \leq \int_0^{+\infty} |\bar u(x) - \bar v(x)|\, dx + \int_0^t |f(\tilde u(s)) - f(\tilde v(s))|\, ds$$

holds for any $t > 0$. This property can be extended to all the solutions associated with an L^∞ boundary condition (for details see the Appendix), and hence any solution to (2.1)–(2.3) admits a representation of the form (2.10) for a.e. $(t,x) \in \text{int}\,\Omega$.

In this paper we are interested only in solution of (2.1)–(2.3) with null initial data $\bar u$. From now on we will adopt the semigroup notation $S_t\tilde u$ for the unique solution of (1.1)–(1.3) at time t. We shall be concerned with basic properties of the attainable sets for (1.1)–(1.2):

$$\mathcal{A}(T,\mathcal{U}) \doteq \{S_T\tilde u :\ \tilde u \in \mathcal{U}\}, \tag{2.13}$$

$$\mathcal{A}(\bar x,\mathcal{U}) \doteq \left\{S_{(\cdot)}\tilde u(\bar x) :\ \tilde u \in \mathcal{U}\right\}, \tag{2.14}$$

which consist of all profiles that can be attained at a fixed time $T > 0$ and at a fixed point $\bar x > 0$ by solutions of (1.1)–(1.2) with boundary data that varies inside a given class $\mathcal{U} \subseteq L^\infty$ of admissible boundary controls. In particular we give a characterization of

$$\mathcal{A}(T) \doteq \left\{S_T\bar u :\ \tilde u \in L^\infty(\mathbb{R}^+),\ \tilde u \geq b(0)\right\}, \tag{2.15}$$

$$\mathcal{A}(\bar x) \doteq \left\{S_{(\cdot)}\tilde u(\bar x) :\ \tilde u \in L^\infty(\mathbb{R}^+),\ \tilde u \geq b(0)\right\}, \tag{2.16}$$

and we establish the compactness of (2.13), (2.14) in connection with a special class of admissible boundary controls.

2.2. Statements of the main results. We present here the statements of the main results. Throughout the following,

$$D^-w(x) = \liminf_{h\to 0} \frac{w(x+h)-w(x)}{h}, \qquad D^+w(x) = \limsup_{h\to 0} \frac{w(x+h)-w(x)}{h}$$

will denote, respectively, the lower and upper Dini derivatives of a function w at x.

THEOREM 1. *In connection with problem* (1.1)–(1.2)*, for any fixed* $T > 0$, $\mathcal{A}(T)$ *is the set of all bounded functions* w *which satisfy the following conditions:*

$$w(x) \neq 0 \Longrightarrow f'(w(x)) \geq \frac{x}{T}, \tag{2.17}$$

$$w(x-) \neq 0 \quad \textit{and} \quad w(y) = 0\ \forall\ y > x \Longrightarrow f'(w(x-)) > \frac{x}{T}, \tag{2.18}$$

$$D^+w(x) \leq \frac{f'(w(x))}{xf''(w(x))} \tag{2.19}$$

for every $x > 0$.

Remark 2.3. By definition an element $\tilde{w} \in \mathcal{A}(T) \subseteq L^\infty(\mathbb{R}^+)$ is an equivalence class of essentially bounded measurable functions. Hence the above characterization must be interpreted in the sense that $\tilde{w} \in \mathcal{A}(T)$ iff there exists a representative w in the class $\tilde{w}$ satisfying (2.17)–(2.19).

Notice that if a bounded function w satisfies (2.17), then there exists $a > 0$ such that $w(x) = 0$ if $x \geq a$. Therefore, the boundedness of w together with (2.17), (2.19) imply that w has finite total increasing variation (and hence finite total variation as well) on subsets of $\mathbb{R}^+$ bounded away from the origin. Thus we may assume that w admits left limit in any point and (2.18) makes sense. Moreover from (2.19) it follows that $w(x-) > w(x+)$ at every point of discontinuity.

Remark 2.4. Having in mind the extension of the above result to attainable sets for classes of admissible boundary controls in $L^1(\mathbb{R}^+)$ (see [1]), it is useful to rewrite condition (2.19) in the following form:

$$w(y) \leq w(x) + \int_x^y \frac{f'(w(\xi))}{\xi f''(w(\xi))}\, d\xi \qquad \forall\ x, y > 0, \quad y \geq x, \tag{2.19$'$}$$

which is shown to be equivalent to (2.19) at the end of section 3.

THEOREM 2. *In connection with problem* (1.1)–(1.2)*, for any fixed* $\bar{x} > 0$*,* $\mathcal{A}(\bar{x})$ *is the set of all bounded functions* ρ *which satisfy the following conditions:*

$$\rho(t) \neq 0 \Longrightarrow f'(\rho(t)) \geq \frac{\bar{x}}{t}, \tag{2.20}$$

$$\rho(\tau+) \neq 0 \quad \textit{and} \quad \rho(t) = 0\ \forall\ t < \tau \Longrightarrow f'(\rho(\tau+)) > \frac{\bar{x}}{\tau}, \tag{2.21}$$

$$D^-\rho(t) \geq \frac{f'(\rho(t))}{tf''(\rho(t))} \tag{2.22}$$

for every $t > 0$.

The proof of Theorem 1 is given in section 3; the proof of Theorem 2 is entirely similar so it is omitted.

In order to achieve the closure of the attainable sets for (1.1)–(1.2) we need to restrict the class of admissible boundary controls by means of a suitable multifunction G.

THEOREM 3. *Let* $G : \mathbb{R}^+ \hookrightarrow [b(0), +\infty)$ *be a measurable uniformly bounded multifunction with convex closed values,* $q_i : \mathbb{R}^+ \times \mathbb{R} \to \mathbb{R}$*,* $i = 1, \ldots, N$*, measurable maps convex w.r.t. the second variable,* $g_i : \mathbb{R}^+ \to \mathbb{R}$*,* $i = 1, \ldots, N$*, measurable maps and let* J *be a possibly empty subset of* $\mathbb{R}^+$*. Denote*

$$\mathcal{U} = \Big\{ \tilde{u} \in L^\infty(\mathbb{R}^+) : \tilde{u}(t) \in G(t), \quad \textit{for a.e. } t, \\ \int_0^t q_i\big(s, f(\tilde{u}(s))\big)\, ds \leq g_i(t) \quad \forall\ t \in J,\ \forall\ i = 1, \ldots, N \Big\}. \tag{2.23}$$

Then $\mathcal{A}(T,\mathcal{U})$*,* $T > 0$*, and* $\mathcal{A}(\bar{x},\mathcal{U})$*,* $\bar{x} > 0$ *are compact subsets of* $L^1(\mathbb{R}^+)$ *and* ${L^1}_{loc}(\mathbb{R}^+)$*, respectively.*

The proof of Theorem 3 is given in section 4. (For references on the multifunction G see [2].)

Remark 2.5. The convexity assumption on the multifunction G cannot be relaxed in order to ensure the closure of the attainable set, as shown by the following example.

Example. Consider the problem (1.1)–(1.2) associated with the Burgers equation

$$u_t + \left(\frac{u^2}{2}\right)_x = 0, \tag{2.24}$$

and assume that the admissible boundary controls are all the measurable functions taking values in $\{0, 2\}$. We claim that the corresponding attainable set at time $T = 1$ is not closed in the topology of L^1. Indeed, define

$$\tilde{u}^\nu(t) = \begin{cases} 2 & \text{if} \quad \dfrac{k}{2^\nu} \le t \le \dfrac{k+1}{2^\nu} \quad k \text{ even}, \\ \\ 0 & \text{if} \quad \dfrac{k}{2^\nu} \le t \le \dfrac{k+1}{2^\nu} \quad k \text{ odd}, \end{cases} \qquad 0 \le k \le 2^\nu - 1. \tag{2.25}$$

Observe that $f(\tilde{u}^\nu)$ converges weakly in L^1 to $f(\tilde{u})$, with $\tilde{u}(t) \equiv \sqrt{2}$. Hence by the same arguments of section 4 it can be shown that $S_{(\cdot)}\tilde{u}^\nu(\cdot)$ converges in the L^1-norm to a solution of (2.24), (1.2), (1.3) with boundary data $\tilde{u}$: then

$$S_1\tilde{u}(x) = \begin{cases} \sqrt{2} & \text{if } 0 < x < \sqrt{2}/2, \\ 0 & \text{otherwise.} \end{cases} \tag{2.26}$$

It can be easily seen that such a profile cannot be obtained with a boundary data $\tilde{u}'$ which takes values in $\{0, 2\}$. Indeed, by tracing the backward generalized characteristics [5] and recalling (2.8), one gets

$$\tilde{u}'(t) = \sqrt{2} \quad \forall\, t \in [1/2, 1]. \tag{2.27}$$

Remark 2.6. The convexity assumption on the functions q_i cannot be relaxed too. Indeed, consider the Burgers equation (2.24) with admissible boundary data $\tilde{u}$ taking values in $[0, 2]$ and satisfying the inequality

$$\int_{1/2}^{1} \tilde{u}(s)\, ds \le \frac{1}{2}, \tag{2.28}$$

which is an integral constraint of the type given in (2.23) with

$$q(s, v) \doteq \begin{cases} 0 & \text{if} \quad 0 \le s < 1/2, \\ \operatorname{sgn}(v)\sqrt{2|v|} & \text{otherwise.} \end{cases}$$

Observe that the same sequence defined by (2.25) fulfills such a constraint. On the other hand, from (2.27) it follows that the profile in (2.26) cannot be attained by using any boundary control satisfying (2.28).

As stated in the introduction, the compactness of the attainable sets guarantees the existence of optimal controls for a class of minimization problems.

COROLLARY 1. *Let $F_1 : L^1(\mathbb{R}^+) \to \mathbb{R}$, $F_2 : L^1([0, \tau]) \to \mathbb{R}$, $\tau > 0$, be lower semicontinuous functionals and let $\mathcal{U}$ be defined as in* (2.23). *Then for every fixed $T, \bar{x} > 0$ the optimal control problems*

$$\min_{\tilde{u} \in \mathcal{U}} F_1\left(S_T\tilde{u}(\cdot)\right), \qquad \min_{\tilde{u} \in \mathcal{U}} F_2\left(S_{(\cdot)}\tilde{u}(\bar{x})\right)$$

admit a solution.

3. Proof of Theorem 1. The proof will be divided into two steps:

Step 1. Show that any element $S_T\tilde{u} \in L^\infty(\mathbb{R}^+)$ of the attainable set satisfies (2.17)–(2.19).

Step 2. Show that if $w \in \mathrm{BV}([\alpha, +\infty))\ \forall\ \alpha > 0$ is a bounded function satisfying (2.17)–(2.19), then there exists $\tilde{u} \in L^\infty([0,T])$, $\tilde{u} \geq b(0)$ such that $S_T\tilde{u} = w$.

3.1. Step 1. A technical result will be proved first.

LEMMA 3.1. *Let $w : \mathbb{R} \to \mathbb{R}$, $x > 0$, be a bounded right continuous function having right and left limits in any point. Then $\varphi : x \mapsto \frac{f'(w(x))}{x}$ is nonincreasing iff* (2.19) *holds.*

Proof. Observe that nonincreasing monotonicity of φ is equivalent to

$$D^+\varphi(x) \leq 0 \qquad \forall\ x > 0. \tag{3.1}$$

Suppose first that $x > 0$ is a point of continuity for w. Hence $f'' > 0$,

$$\begin{aligned} &\limsup_{h\to 0} \frac{\varphi(x+h) - \varphi(x)}{h} \\ &= \limsup_{h\to 0} \left[\frac{f'(w(x+h)) - f'(w(x))}{(w(x+h) - w(x))}\ \frac{w(x+h) - w(x)}{(x+h)h} - \frac{f'(w(x))}{x(x+h)} \right] \qquad (3.2) \\ &= \frac{f''(w(x))}{x} \limsup_{h\to 0} \frac{w(x+h) - w(x)}{h} - \frac{f'(w(x))}{x^2}, \end{aligned}$$

which shows that (3.1) and (2.19) are equivalent.

In the case when w is not continuous at x, assume (3.1) holds: then $w(x-) > w(x)$. Indeed, if it is false, then $f'(w(x-)) < f'(w(x))$ by convexity of f; hence there exists $y < x$ such that $\varphi(y) < \varphi(x)$ which contradicts the monotonicity assumption on φ. There follows that

$$D^+w(x) = \limsup_{h\to 0^+} \frac{w(x+h) - w(x)}{h};$$

thus (2.19) follows taking in (3.2) the lim sup as $h \to 0^+$. Conversely, if (2.19) holds then still $w(x-) > w(x)$. Since w and hence φ are right continuous it follows that $\varphi(x-) > \varphi(x)$, due to the monotonicity of f'. Thus it is sufficient to prove (3.1) for $h \to 0^+$. This follows immediately from (3.2) using the same arguments as before. □

Recalling Remark 2.1 we can choose a representative function w of $S_T\tilde{u}$ which is right continuous. Assume that $f'(w(x)) < x/T$ and let $\xi(\cdot)$ denote the maximal backward generalized characteristic through (T, x). Observe that $\xi(\cdot)$ is a genuine characteristic (see [5, Theorem 3.2]) and hence, by Theorem 3.3 in [5], $S_{(\cdot)}\tilde{u}(\xi(\cdot)) = v$ a.e. on $[0,T]$ for some constant v such that $\dot{\xi} = f'(v)$. Since Theorem 4.1 in [5] implies $v(0) = w(x)$, it follows that $\xi(t) = x + f'(w(x))(t-T)$ for all $t \in [0,T]$. Hence $\xi(0) = x - Tf'(w(x)) > 0$, which implies $w(x) = S_0\tilde{u}(\xi(0)) = 0$ thus proving (2.17).

Next, suppose that there exists $x > 0$ such that $f'(w(x-)) \leq x/T$. If $w(x-) = 0$ there's nothing to prove. Otherwise $f'(w(x-)) = x/T$. If $w(x+) = w(x-)$, again there's nothing to prove, otherwise, from arguments similar to the previous ones and since genuine characteristics do not intersect in the interior of Ω, it follows that $w(y) = 0\ \forall y > x$ and hence $w(x-) > 0$. Observe now that the values of the solution in the

interior of the funnel confined between minimal and maximal backward characteristics through (T,x) depend only on the values of the solution at $t=0$. Thus $S_t\tilde{u}(x)=0$ for any $0<t<T$ and $x>f'(w(x-))t$. There follows that the minimal characteristic is not genuine, which gives a contradiction, proving (2.18).

To prove (2.19) by Lemma 3.1 it is sufficient to show that the function $\varphi: x\mapsto f'(w(x))/x$ is nonincreasing. Let $0<x_1<x_2$ be given and trace the maximal backward characteristics $\xi_1(\cdot)$, $\xi_2(\cdot)$ through (T,x_1) and (T,x_2), respectively. By the same arguments as above they have the form

$$\xi_i(t)=x_i+f'(w(x_i))(t-T),\qquad i=1,2 \tag{3.3}$$

as long as they exist. Assume that $f'(w(x_1))<f'(w(x_2))$ (otherwise the result is obvious) and let $\tau\in\mathbb{R}$ be such that $\xi_1(\tau)=\xi_2(\tau)$ where, with an abuse of notation, $\xi_i(\cdot)$ denote the functions in (3.3) defined for all $t\in\mathbb{R}$. Since ξ_1 and ξ_2 are genuine characteristics and hence do not intersect in the interior of Ω (see [5]), we deduce that $\xi_i(\tau)\le 0$. Otherwise it should be $\tau<0$ which implies, by arguments as above, $f'(w(x_1))=f'(w(x_2))=f'(0)$. Therefore,

$$1+\frac{f'(w(x_1))}{x_1}(\tau-T)=\frac{\xi_1(\tau)}{x_1}\le\frac{\xi_2(\tau)}{x_2}=1+\frac{f'(w(x_2))}{x_2}(\tau-T)$$

showing $\varphi(x_1)\ge\varphi(x_2)$.

3.2. Step 2. Choose $w\in L^\infty(\mathbb{R}^+)$ satisfying (2.17)–(2.19). By Remark 2.3 we can assume that w is right continuous. Observe first that if $w\equiv 0$ then the boundary control

$$\tilde{u}\equiv\begin{cases}0 & \text{if}\quad f'(0)\ge 0,\\ b(0) & \text{if}\quad f'(0)<0\end{cases}$$

clearly produces the null solution. Next we prove the result in the case when w is made up of two constant states.

PROPOSITION 3.1. *Let $\omega,r>0$ be given with $f'(\omega)>r/T$. Then there exists $\tilde{u}\in L^\infty([0,T])$, $\tilde{u}\ge b(0)$, such that*

$$S_T\tilde{u}(x)=\begin{cases}\omega & \textit{if } x<r,\\ 0 & \textit{otherwise.}\end{cases} \tag{3.4}$$

Proof. If $c\doteq[f(\omega)-f(0)]/\omega\ge r/T$, set $t_1=T-r\,\omega/[f(\omega)-f(0)]\ge 0$. Then

$$\tilde{u}(t)=\begin{cases}\omega & \text{if}\quad t_1<t<T,\\ 0 & \text{if}\quad 0<t<t_1 \text{ and } f'(0)\ge 0,\\ b(0) & \text{if}\quad 0<t<t_1 \text{ and } f'(0)<0\end{cases}$$

produces the solution

$$S_t\tilde{u}(x)=\begin{cases}\omega & \text{if}\quad 0<x<r+\dfrac{f(\omega)-f(0)}{\omega}(t-T),\\ 0 & \text{otherwise}\end{cases}$$

which satisfy (3.4).

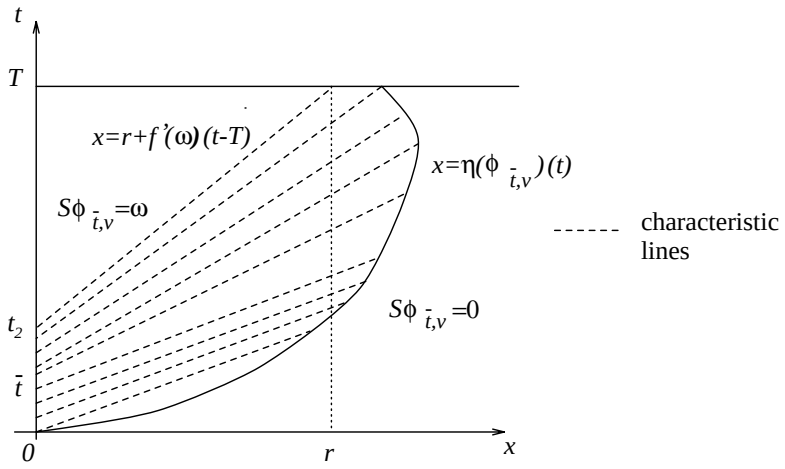


FIG. 1.

Now assume $c < r/T$ and call $t_2 = T - r/f'(\omega) > 0$. For any $\bar{t} \in [0, t_2)$ and $v \geq \omega$ define the function $\phi_{\bar{t},v} : [0,T] \to [\omega, +\infty)$ by setting

$$\phi_{\bar{t},v}(t) = \begin{cases} v & \text{if } 0 \leq t < \bar{t}, \\ b\left[f'(v) + \dfrac{t-\bar{t}}{t_2-\bar{t}}\,(f'(\omega) - f'(v))\right] & \text{if } \bar{t} \leq t < t_2, \\ \omega & \text{if } t \geq t_2. \end{cases} \tag{3.5}$$

If $v \geq \omega$ satisfies $f(v) > f(0)$, since $t \mapsto \phi_{\bar{t},v}(t)$ is decreasing on $[0, t_2]$ it can be easily seen that $S_{(\cdot)}\phi_{\bar{t},v}$ has a single shock curve $t \mapsto \eta(\phi_{\bar{t},v})(t)$ departing from the origin such that $S_t\phi_{\bar{t},v}(x) = 0$ for $x > \eta(\phi_{\bar{t},v})(t)$ as long as $\eta(\phi_{\bar{t},v})(\cdot)$ exists (see Figure 1).

We claim that there exist ω_0, $\omega_1 > \omega$ and $0 \leq \tau_0$, $\tau_1 < t_2$ such that $\eta(\phi_{\tau_0,\omega_0})(\cdot)$ and $\eta(\phi_{\tau_1,\omega_1})(\cdot)$ are defined on $[0,T]$ and

$$\eta(\phi_{\tau_0,\omega_0})(T) < r \leq \eta(\phi_{\tau_1,\omega_1})(T). \tag{3.6}$$

First we prove the existence of τ_1 and ω_1. To this end we show that there exist $v > \omega$ and $s \in (0,T)$ such that

$$\frac{f(v) - f(0)}{v}\; s > r + |c|(T - s), \tag{3.7}$$

$$0 < s - \frac{1}{f'(v)}\;\frac{f(v) - f(0)}{v}\; s < t_2. \tag{3.8}$$

Indeed, if $\lim_{v\to+\infty} f'(v) = +\infty$ then choose $s = t_2/2$ and $v > \omega$ satisfying (3.7). Otherwise, $f'(\omega) > r/T$ and hence

$$\frac{r}{T} < \lim_{v\to+\infty} f'(v) = \lim_{v\to+\infty} \frac{f(v) - f(0)}{v}, \tag{3.9}$$

there exists $\bar{v} > \omega$ such that $T\left[f(\bar{v}) - f(0)\right]/\bar{v} > r$. Then, using the continuity of the map

$$t \mapsto \frac{f(\bar{v}) - f(0)}{\bar{v}}\; t - r - |c|(T - t),$$

we find some $s \in (0,T)$ satisfying (3.7) with $v = \bar v$. But (3.9) and the convexity of f guarantee that there exists $v \geq \bar v$ satisfying (3.7)–(3.8) as well. Now set

$$\omega_1 = v, \qquad \tau_1 = s - \frac{1}{f'(v)}\ \frac{f(v)-f(0)}{v}\ s. \tag{3.10}$$

It follows that

$$\begin{aligned}\eta(\phi_{\tau_1,\omega_1})(T) &= \int_0^s \dot\eta(\phi_{\tau_1,\omega_1})(t)\ dt + \int_s^T \dot\eta(\phi_{\tau_1,\omega_1})(t)\ dt\\ &\geq \frac{f(\omega_1)-f(0)}{\omega_1}\ s + c(T-s)\\ &> r + (|c|+c)(T-s) \geq r.\end{aligned} \tag{3.11}$$

Now we set $\tau_0 = 0$ and prove the existence of ω_0. If $c > 0$, take $\omega_0 = \omega$. Otherwise set

$$\bar v = \sup\left\{v \geq \omega :\ S_T\phi_{0,v} \equiv 0\right\}. \tag{3.12}$$

By the previous analysis, $\bar v < +\infty$. Moreover, since the map $v \mapsto \phi_{0,v}$ is continuous from $[\omega, +\infty)$ into $L^\infty([0,T])$ w.r.t. the L^1-norm, from Remark 2.2 it follows that $S_T\phi_{0,\bar v} \equiv 0$. If $v > \bar v$, then $\eta(\phi_{0,v})(\cdot)$ is defined on $[0,T]$ and $\eta(\phi_{0,v})(T) > 0$. Indeed, if not, then there exists $\tau < T$ such that $\eta(\phi_{0,v})(\tau) = 0$. There follows that $S_\tau\phi_{0,v} \equiv 0$ and that $f(\phi_{0,v}(t)) \leq f(\phi_{0,v}(\tau)) < f(0)\ \ \forall t \geq \tau$. Hence $S_t\phi_{0,v} \equiv 0\ \ \forall t \geq \tau$, which contradicts (3.12). Moreover, if $0 < x < \eta(\phi_{0,v})(T)$, then $S_T\phi_{0,v}(x) \geq \omega$. In fact, due to (2.18), the minimal backward characteristic through $(T, \eta(\phi_{0,v})(T))$ reaches the t-axis in positive time. Since genuine characteristics do not intersect, all maximal backward characteristics through (T,x), $0 < x < \eta(\phi_{0,v})(T)$, intersect the t-axis. Since $\phi_{0,v}(t) \geq \omega$ for any $t \in [0,T]$, by arguments similar to the ones used in Step 1 we deduce that $S_T\phi_{0,v}(x) \geq \omega$. There exists $\delta > 0$ such that if $\bar v < v < \bar v + \delta$ then $\eta(\phi_{0,v})(T) < r$. Indeed assume by contradiction that there exists a decreasing sequence $(v_n)_{n\in\mathbb{N}}$ converging to $\bar v$ such that $\eta(\phi_{0,v_n})(T) \geq r\ \ \forall n$. Then

$$\|S_T\phi_{0,\bar v} - S_T\phi_{0,v_n}\|_{L^1} \geq \int_0^r |S_T\phi_{0,v_n}(x)|\ dx \geq \omega r,$$

which contradicts the continuity of the map $v \mapsto S_T\phi_{0,v}$, proving the existence of ω_0 with the required property. Consider now the continuous map $\phi : [0,1] \to L^\infty([0,T])$ defined by

$$\phi(\lambda) = \lambda\phi_{\tau_1,\omega_1} + (1-\lambda)\phi_{\tau_0,\omega_0}. \tag{3.13}$$

Set $\eta(\phi(\lambda))(T) = 0$ if $S_T\phi(\lambda) \equiv 0$. Then from the continuity of $\lambda \mapsto S_T\phi(\lambda)$, it follows that the map $\lambda \mapsto \eta(\phi(\lambda))(T)$ is continuous. Indeed, by the previous analysis, $S_T\phi(\lambda)(x) \geq \omega$ whenever $x < \eta(\phi(\lambda))(T)$. Hence

$$\begin{aligned}|\eta(\phi(\lambda_1))(T) - \eta(\phi(\lambda_2))(T)| &\leq \frac{1}{\omega}\left|\int_{\eta(\phi(\lambda_1))(T)}^{\eta(\phi(\lambda_2))(T)} |S_T\phi(\lambda_1)(x) - S_T\phi(\lambda_2)(x)|\ dx\right|\\ &\leq \frac{1}{\omega}\|S_T\phi(\lambda_1) - S_T\phi(\lambda_2)\|_{L^1},\end{aligned}$$

which approaches zero as $\lambda_1 - \lambda_2 \to 0$. It follows that there exists $\bar\lambda \in [0,1]$ such that $\eta(\phi(\bar\lambda))(T) = r$. We claim that $S_T\phi(\bar\lambda)$ satisfies (3.4). Indeed if $x < r$ let $t \mapsto \theta(t)$ be the maximal backward characteristic through (T,x). Then by (2.17) there exists $\tau \geq 0$ such that $\theta(\tau) = 0$. Actually $\tau \geq t_2$. If not, then

$$\dot\theta(t) = f'(S_T\phi(\bar\lambda)(x)) = \frac{x}{T-\tau} < \frac{r}{T-t_2} = f'(\omega),$$

which gives a contradiction since f' is increasing and $S_T\phi(\bar\lambda)(x) \geq \omega$. Thus $\tau \geq t_2$, from which it follows that $S_T\phi(\bar\lambda)(x) = \omega$. □

Throughout the following we denote by $\psi(\omega, r) \in L^\infty([0,T])$ a boundary control such that $S_T\psi(\omega,r)$ satisfies (3.4). In order to prove Step 2 in the general case we shall adopt the following procedure.

1. For every $x > 0$ we trace the lines θ_x^-, θ_x^+ through (T,x) with slope $f'(w(x-))$ and $f'(w(x+))$, respectively. These will be the minimal and maximal backward characteristics through (T,x) of the candidate solution. Due to (2.17), if $w(x) \neq 0$ they reach the t-axis in positive time. Assumption (2.19) guarantees that the lines $\{\theta_x^\pm : \ x > 0\}$ do not intersect each other in the interior of Ω.

2. Since a solution is constant along minimal and maximal backward characteristics [5], for every $t \in [0,T]$ for which there exists $x > 0$ such that $\theta_x^\pm(t) = 0$, we define $\tilde u(t) = w(x)$. The set of the remaining t is a disjoint union of open intervals. On any of such intervals $\tilde u$ is defined so as to produce a compression wave which generates a discontinuity at time T.

3. By using the fact that a solution is constant along genuine characteristics, we define a function $u : (0,T) \times \mathbb{R}^+ \to \mathbb{R}$, which is candidate, to be $S_{(\cdot)}\tilde u$ and we prove that u is a weak entropic solution of (1.1)–(1.2) in the interior of Ω.

4. We show that u satisfies the boundary condition related to the boundary control $\tilde u$ in the sense of Definition 1 (iii) and that $u(T-,\cdot) = w$.

1. For each $x > 0$ consider the lines

$$\theta_x^- : t \mapsto x + f'(w(x-))(t-T), \tag{3.14}$$

$$\theta_x^+ : t \mapsto x + f'(w(x))(t-T), \tag{3.15}$$

defined for $t \leq T$. By Remark 2.3 and convexity of f one has $\theta_x^-(t) \leq \theta_x^+(t) \ \ \forall\, t$. We claim that for any $0 < x < y$ the lines $\theta_x^\pm$ and $\theta_y^\pm$ do not intersect in the interior of Ω. By the previous argument it suffices to prove that $\theta_x^+(t) > \theta_y^-(t)$ in the interior of Ω. If $f'(w(x)) \geq f'(w(y-))$ the claim is obvious. Otherwise since $w(x) \neq w(y-)$, one of the two is nonzero. Hence due to (2.17) one of the two holds: $f'(w(x)) \geq x/T$ or $f'(w(y-)) \geq x/T$. Let $\tau < T$ be such that $\theta_x^+(\tau) = \theta_y^-(\tau) \doteq \xi$. Then $\tau \geq 0$ or $\xi \leq 0$. Actually $\xi \leq 0$. Indeed, let φ be as in Lemma 3.1. Then $\varphi(y-) \leq \varphi(x)$. Hence

$$\frac{\xi}{x} = 1 + \varphi(x)(\tau - T) \leq 1 + \varphi(y-)(\tau - T) = \frac{\xi}{y}$$

and since $x < y$ it follows that $\xi \leq 0$, which proves the claim.

2. Define

$$x_0 \doteq \inf\{x > 0 : \ w(y) = 0 \ \forall\ y \geq x\}. \tag{3.16}$$

To get a boundary control $\tilde u$ that produces a solution of (1.1)–(1.3) that attains w,

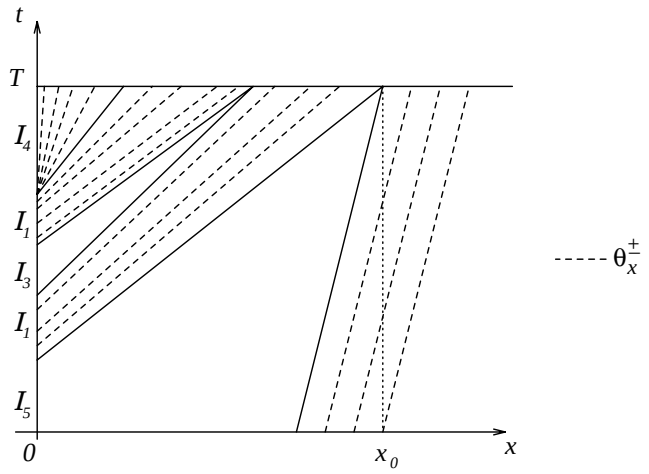


FIG. 2.

we consider the following partition of the interval $[0,T]$ (see Figure 2):

$$I_1 \doteq \left\{t\in[0,T]:\ \exists!\ x>0:\ \theta_x^-(t)=0 \text{ or } \theta_x^+(t)=0\right\}, \tag{3.17}$$

$$I_2 \doteq \left\{t\in[0,T]:\ \exists\ 0<x<y:\ \theta_x^+(t)=\theta_y^-(t)=0\right\}, \tag{3.18}$$

$$\begin{aligned} I_3 \doteq \Big\{t\in[0,T]:\ \ &\nexists\ x>0:\ \theta_x^-(t)=0 \text{ or } \theta_x^+(t)=0,\\ &\exists\ t'\in(0,t)\cap[I_1\cup I_2]\,,\ \exists\ t''\in(t,T)\cap[I_1\cup I_2]\,\Big\}, \end{aligned} \tag{3.19}$$

$$I_4 \doteq \left\{t\in[0,T]:\ \forall t'\geq t\ \ \nexists\ x>0:\ \theta_x^-(t')=0 \text{ or } \theta_x^+(t')=0\right\}, \tag{3.20}$$

$$I_5 \doteq \left\{t\in[0,T]:\ \forall t'\leq t\ \ \nexists\ x>0:\ \theta_x^-(t')=0 \text{ or } \theta_x^+(t')=0\right\}. \tag{3.21}$$

Here any of these sets could be empty. The above sets, whenever nonempty, satisfy the following properties:

(i) I_2 contains at most countably many points;

(ii) I_3 is the disjoint union of at most countably many open intervals $(\mathcal{I}_\nu)_{\nu\in\mathbb{N}}$ of the form

$$\mathcal{I}_\nu=(\tau_\nu^1,\tau_\nu^2),\qquad \theta^+_{x_\nu}(\tau_\nu^1)=\theta^-_{x_\nu}(\tau_\nu^2)=0\qquad \exists\ x_\nu>0, \tag{3.22}$$

where x_ν is a point of discontinuity for w.

(iii) I_4 is an interval of the form $I_4=(\tau^4,T]$ with $\tau^4\in I_1\cup I_2$.

(iv) I_5 is an interval of the form $I_5=[0,\tau^5)$ with $\theta^-_{x_0}(\tau^5)=0$.

To show (i) it is sufficient to observe that, since the lines $\{\theta_x^\pm\}_{x>0}$ do not intersect in the interior of Ω, for each $t\in I_2$ the set

$$J_t \doteq \left\{x>0:\ \theta_x^-(t)=0 \text{ or } \theta_x^+(t)=0\right\} \tag{3.23}$$

is an interval and $J_s\cap J_t=\emptyset$ for any $s,t\in I_2,\ s\neq t$.

Regarding (ii)–(iv), we first show that $I_3\cup I_4\cup I_5$ is open in $[0,T]$. Indeed, let $t\in I_3\cup I_4\cup I_5$ and assume by contradiction that $(t_\nu)_{\nu\in\mathbb{N}}\subseteq I_1\cup I_2$ is a sequence converging to t. Then there exists a sequence $(y_\nu)_{\nu\in\mathbb{N}}\subseteq\mathbb{R}^+$ such that $\theta^\pm_{y_\nu}(t_\nu)=0$. By eventually taking a subsequence, we shall assume $\theta^+_{y_\nu}(t_\nu)=0$, the other case being entirely similar. Since w is bounded, from (2.17) it follows that $(y_\nu)_{\nu\in\mathbb{N}}$ is bounded, so

it admits a converging subsequence which is still denoted by $(y_\nu)_{\nu\in\mathbb{N}}$. Call $\bar y$ its limit point. Again, up to a subsequence we can assume that $f'(w(y_\nu)) \to f'(w(\bar y))$. Then $0 = \theta^+_{y_\nu}(t_\nu) \to \theta^+_{\bar y}(t)$, which gives a contradiction. Observe now that $\inf I_4 \in I_1 \cup I_2$. Indeed, if $\inf I_4 = 0$, then it belongs to $I_1 \cup I_2$ by (2.17) since $w \not\equiv 0$. Otherwise, since $I_3 \cup I_4 \cup I_5$ is open, if $\inf I_4 \notin I_1 \cup I_2$, then there exists $t' < \inf I_4$ such that $(t', \inf I_4) \subseteq I_3 \cup I_4 \cup I_5$ which clearly gives a contradiction. Since by definition I_4 is an interval and $\sup I_4 = T$, this suffices to prove (iii).

Concerning (iv), in a similar way it can be proved that $\tau^5 = \sup I_5 \in I_1 \cup I_2$. Set

$$z \doteq \sup\left\{x > 0 :\ \theta^-_x(\tau^5) = 0 \text{ or } \theta^+_x(\tau^5) = 0\right\}. \tag{3.24}$$

Let $y > z$ and suppose that $w(y) \neq 0$. Then by (2.17) and (3.21), $\theta_y(\tau^5) = 0$, which contradicts (3.24). Thus it must be $z \geq x_0$. If $z > x_0$, then $0 = \theta^\pm_z(\tau^5) = z + f'(0)(\tau^5 - T)$. Hence there exists $y > z$ and $t \in (0, \tau^5)$ such that $\theta^\pm_y(t) = y + f'(0)(t - T) = 0$, which gives a contradiction by the definition of I_5. Thus $z = x_0$ and hence $\theta^-_{x_0}(\tau^5) = 0$ proving (iv).

Regarding (ii), since $\inf I_4, \sup I_5 \notin I_3$, I_3 is open; hence it is a disjoint union of at most countably many open intervals $\mathcal{I}_\nu = (\tau^1_\nu, \tau^2_\nu)$. Moreover $\tau^1_\nu, \tau^2_\nu \in I_1 \cup I_2$ since $I_3 \cup I_4 \cup I_5$ is open. Call

$$\begin{aligned} x^1_\nu &\doteq \inf\left\{x > 0 :\ \theta^-_x(\tau^1_\nu) = 0 \text{ or } \theta^+_x(\tau^1_\nu) = 0\right\}, \\ x^2_\nu &\doteq \sup\left\{x > 0 :\ \theta^-_x(\tau^2_\nu) = 0 \text{ or } \theta^+_x(\tau^2_\nu) = 0\right\}. \end{aligned}$$

Then $x^1_\nu = x^2_\nu \doteq x_\nu$. In fact $x^2_\nu \leq x^1_\nu$ since the lines $\{\theta^\pm_x\}_{x>0}$ do not intersect in the interior of Ω. If $x^2_\nu < x^1_\nu$, then choose $y \in (x^2_\nu, x^1_\nu)$. Then there exists $\tau \in (\tau^1_\nu, \tau^2_\nu)$ such that $\theta^\pm_y(\tau) = 0$, which is a contradiction. Since by (2.19) w satisfies (2.8), the conclusion of (ii) follows immediately.

Now we are ready to define the boundary data which produces the given profile:

$$\tilde u(t) = \begin{cases} w(x-) & \text{if} \quad t \in I_1,\ \theta^-_x(t) = 0, \\ w(x) & \text{if} \quad t \in I_1,\ \theta^+_x(t) = 0, \\ w\left((\sup J_t)-\right) & \text{if} \quad t \in I_2, \\ b\left(\dfrac{x_\nu}{T-t}\right) & \text{if} \quad t \in \mathcal{I}_\nu \subseteq I_3, \\ b(0) & \text{if} \quad t \in I_4, \\ \psi(w(x_0-), x_0)(t) & \text{if} \quad t \in I_5. \end{cases} \tag{3.25}$$

Notice that if $t \in \mathcal{I}_\nu \subseteq I_3$, then

$$f'(w(x_\nu)) < \frac{x_\nu}{T-t} < f'(w(x_\nu-)),$$

and hence $x_\nu/(T-t) \in \text{Range } f'$. Moreover, if $I_4 \neq \emptyset$, then $b(0) > -\infty$. Indeed, fix $\varepsilon > 0$. Then for any $x \in (0, \varepsilon(t - \tau^4))$ we have $0 < f'(w(x)) \leq \varepsilon$. In fact let $\xi > 0$ be such that $\theta^\pm_\xi(\tau^4) = 0$. If $f'(w(x)) > \varepsilon$, then there exists $\tau > \tau^4$ such that $\theta^\pm_\zeta(\tau) = 0$, thus contradicting (3.20). If $f'(w(x)) \leq 0$, then $\theta^\pm_\zeta$ and $\theta^\pm_\xi$ would intersect in the interior of Ω. Hence $\lim_{x\to 0^+} f'(w(x)) = 0$. Due to the boundedness of w, this implies $0 \in \text{Range } f'$. Thus (3.25) is well defined.

3. For each $s \in \mathcal{I}_\nu \subseteq I_3$ define the line

$$\theta_s : t \mapsto f'(\tilde u(s))(t - s) = \frac{x_\nu}{T-s}(t - s), \qquad s < t < T, \tag{3.26}$$

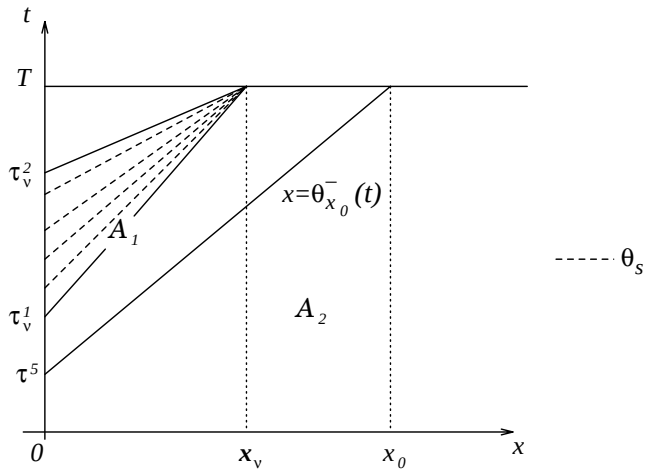


FIG. 3.

which is entirely contained in the open set $\left\{(t,x):\ s<t<T,\ \theta^-_{x_\nu}(t)<x<\theta^+_{x_\nu}(t)\right\}$. Observe that any of the θ_s cannot intersect one of the $\theta^\pm_x$ in the interior of Ω, otherwise $\theta^\pm_x$ would intersect $\theta^-_{x_\nu}$ or $\theta^+_{x_\nu}$ too. Denote (see Figure 3)

$$\begin{aligned}\mathcal{A}_1 &\doteq \left\{(\tau,\xi)\in \text{int }\Omega:\ \xi\le\theta^-_{x_0}(\tau)\right\},\\ \mathcal{A}_2 &\doteq \left\{(\tau,\xi)\in \text{int }\Omega:\ \xi>\theta^-_{x_0}(\tau))\right\}.\end{aligned}\tag{3.27}$$

We claim that for any $(\tau,\xi)\in\mathcal{A}_1$ there exists a unique line through (τ,ξ) belonging to the family $\Theta\doteq\{\theta^\pm_x:\ x>0\}\cup\{\theta_s:\ s\in I_3\}$. The uniqueness of such a line follows from the previous remark and from the fact that the lines of each family $\{\theta^\pm_x:\ x>0\}$ and $\{\theta_s:\ s\in I_3\}$ do not intersect in the interior of Ω. Regarding the existence observe that if $\xi\neq\theta^\pm_x(\tau)$ for any $x>0$, then there exists $s\in I_3$ such that $\theta_s(\tau)=\xi$. Indeed, the set

$$\mathcal{B}(\tau)\doteq\left\{0<x<\theta^-_{x_0}(\tau):\quad \nexists\ y>0:\ \theta^\pm_y(\tau)=x\right\}\tag{3.28}$$

is open. In fact, let $x\in\mathcal{B}(\tau)$ and assume by contradiction that there exists in $(0,\theta^-_{x_0}(\tau))$ a sequence $x_\nu=\theta^\pm_{y_\nu}(\tau),\ y_\nu>0$, converging to x. By eventually taking a subsequence, we shall assume that $x_\nu=\theta^+_{y_\nu}(\tau)$, the other case being entirely similar. Since w is bounded, from (2.17) it follows that $(y_\nu)_{\nu\in\mathbb{N}}$ is bounded. Therefore, there exists a subsequence, which we still denote by $(y_\nu)_{\nu\in\mathbb{N}}$, converging to some $\bar y>0$ and such that $f'(w(y_\nu))\to f'(w(\bar y))$. Then $\theta^+_{y_\nu}(\tau)\to\theta^+_{\bar y}(\tau)$ and hence $x=\theta^+_{\bar y}(\tau)$ which gives a contradiction.

Now, let (ξ_1,ξ_2) be the connected component of $\mathcal{B}(\tau)$ containing ξ. Then as above there exists $y>0$ such that $\theta^-_y(\tau)=\xi_1$ and $\theta^+_y(\tau)=\xi_2$. Let $t_1>t_2$ be such that $\theta^-_y(t_1)=\theta^+_y(t_2)=0$. Then clearly it must be $(t_2,t_1)=\mathcal{I}_\nu$, $y=x_\nu$, and

$$\frac{x_\nu-\xi}{T-\tau}=\frac{x_\nu}{T-s}=\dot\theta_s$$

for some $\nu\in\mathbb{N}$ and $s\in(t_2,t_1)$. Thus by (3.26) one has $\theta_s(\tau)=\xi$.

Consider now the function $u : (0,T) \times \mathbb{R}^+ \to \mathbb{R}$ defined by

$$u(\tau,\xi) = \begin{cases} w(x) & \text{if} \quad (\tau,\xi) \in \mathcal{A}_1,\ \theta_x^+(\tau) = \xi \quad \exists\ x > 0, \\ w(x-) & \text{if} \quad (\tau,\xi) \in \mathcal{A}_1,\ \theta_x^-(\tau) = \xi \quad \exists\ x > 0, \\ \tilde{u}(s) & \text{if} \quad (\tau,\xi) \in \mathcal{A}_1,\ \theta_s(\tau) = \xi \quad \exists\ s \in I_3, \\ S_\tau \psi(w(x_0-), x_0)(\xi) & \text{if} \quad (\tau,\xi) \in \mathcal{A}_2,\ w(x_0-) > 0, \\ 0 & \text{if} \quad (\tau,\xi) \in \mathcal{A}_2,\ w(x_0-) = 0. \end{cases} \tag{3.29}$$

We claim that, for every $(\tau,\xi) \in \mathcal{A}_1$, $u(\tau,\cdot)$ *is continuous on* $(0, \theta^-_{x_0}(\tau)]$ and $u(\cdot,\xi)$ *is continuous on* $[\tau, T)$. We only give the proof of the first property, the second one being derived in an entirely similar way. To this end we first show that $u(\tau,\cdot)$ satisfies the following properties on $(0, \theta^-_{x_0}(\tau)]$:

(a) if there exists $x > 0$ such that $\theta_x^-(\tau) = \xi$, then $u(\tau,\cdot)$ is left continuous at ξ;

(b) if there exists $x > 0$ such that $\theta_x^+(\tau) = \xi$, then $u(\tau,\cdot)$ is right continuous at ξ;

(c) if $\xi \in \mathcal{B}(\tau)$, then $u(\tau,\cdot)$ is continuous at ξ.

Observe first that if $\zeta \in \mathcal{B}(\tau)$, so that $\theta^-_{x_\nu}(\tau) < \zeta < \theta^+_{x_\nu}(\tau)$ for some $\nu \in \mathbb{N}$ and $\zeta = \theta_s(\tau)$ for some $s \in \mathcal{I}_\nu$, then

$$f'(w(x_\nu)) = \dot{\theta}^+_{x_\nu} = \frac{x_\nu}{T - \tau_\nu^1} < \frac{x_\nu}{T-s} < \frac{x_\nu}{T-\tau_\nu^2} = \dot{\theta}^-_{x_\nu} = f'(w(x_\nu-)).$$

Hence, since f' is strictly increasing,

$$w(x_\nu) < u(\tau,\zeta) < w(x_\nu-). \tag{3.30}$$

We now prove (a). Let $x, \xi > 0$ be such that $\theta_x^-(\tau) = \xi$. Then, by (3.29) $u(\tau,\xi) = w(x-)$. Fix $\varepsilon > 0$ and choose $\delta > 0$ such that

$$|w(y) - w(x-)| \le \varepsilon \qquad \forall y \in (x-\delta, x). \tag{3.31}$$

Let $\xi_\delta = \theta^+_{x-\delta}(\tau)$. By point 1, $\xi_\delta < \xi$. Then, for every $\zeta \in (\xi_\delta, \xi)$,

$$|u(\tau,\zeta) - u(\tau,\xi)| \le \varepsilon. \tag{3.32}$$

Indeed, using again point 1, if $\zeta = \theta_y^\pm(\tau)$ for some $y > 0$ then $y \in (x-\delta, x)$ and hence (3.32) follows from (3.31). Otherwise $\zeta \in \mathcal{B}(\tau)$ and (3.30) holds for some $x_\nu \in (x-\delta, x)$. Again (3.32) follows from (3.31).

The proof of (b) is entirely similar and (c) follows with an analogous argument by using the continuity of $\tilde{u}$ on I_3 instead of the existence of right and left limits of w.

Using (a), (b), and (c) we now derive the continuity of $u(\tau,\cdot)$ on $(0, \theta^-_{x_0}(\tau)]$. Indeed if $\xi = \theta_x^-(\tau) = \theta_x^+(\tau)$ for some $x > 0$ or $\xi \in \mathcal{B}(\tau)$ the conclusion is obvious. Otherwise, assume $\xi = \theta^-_{x_\nu}(\tau) < \theta^+_{x_\nu}(\tau)$ for some $\nu \in \mathbb{N}$. Since $\zeta \in \mathcal{B}(\tau)$ for any $\zeta \in (\xi, \theta^+_{x_\nu}(\tau))$ it follows

$$\lim_{\zeta \to \xi^+} u(\tau,\zeta) = \lim_{\zeta \to \xi^+} b\left(\frac{x_\nu - \zeta}{T-\tau}\right) = b\left(\frac{x_\nu - \xi}{T-\tau}\right) = b(f'(w(x_\nu-))) = u(\tau,\xi);$$

i.e., $u(\tau,\cdot)$ is right continuous at ξ, and hence continuous as well by (a). In a similar way it can be shown that if $\xi = \theta_x^+(\tau) > \theta_x^-(\tau)$, then $u(\tau,\cdot)$ is continuous at ξ.

In order to prove that u is a weak entropic solution of (1.1) in the region $\mathcal{A}_1$, we now show that u *is locally Lipschitz continuous*. As above we prove only that, for

every $\tau \in (0,T)$, $u(\tau,\cdot)$ is locally Lipschitz continuous on $(0,\theta^-_{x_0}(\tau))$. Observe first that, with the same arguments of Step 1, from the definition of u it follows

$$D^+_\xi u(\tau,\xi) \leq \frac{f'(u(\tau,\xi))}{\xi f''(u(\tau,\xi))}$$

for any $0<\xi<\theta^-_{x_0}(\tau)$. Hence, to derive the Lipschitz continuity of $u(\tau,\cdot)$ it suffices to show that locally there exists a constant $C_1 \leq 0$ such that

$$D^-_\xi u(\tau,\xi) \geq C_1 \qquad \forall\ \xi \in (0,\theta^-_{x_0}(\tau)). \tag{3.33}$$

If $D^-_\xi u(\tau,\xi) \geq 0$, there is nothing to prove. Otherwise let $\tau < T' < T$ be fixed. Since by construction

$$u(t,\xi + f'(u(\tau,\xi)))(t-\tau) = u(\tau,\xi) \qquad \forall\ t \in [\tau,T],\ (\tau,\xi) \in \mathcal{A}_1, \tag{3.34}$$

for every $\zeta \in (0,\theta^-_{x_0}(\tau))$ there exists a unique $z = z(\zeta) \in (0,\theta^-_{x_0}(T'))$ such that

$$\zeta = z + f'\left(u(T',z)\right)(\tau - T'), \qquad u(T',z) = u(\tau,\zeta).$$

Observe that

(3.35)

$$\begin{aligned} D^-_\xi u(\tau,\xi) &= \liminf_{z\to z(\xi)} \frac{u(T',z)-u(T',z(\xi))}{(z-z(\xi)) + [f'(u(T',z)) - f'(u(T',z(\xi)))]\,(\tau-T')} \\ &= \liminf_{z\to z(\xi)} \left(\frac{z-z(\xi)}{u(T',z)-u(T',z(\xi))} + \frac{f'(u(T',z)) - f'(u(T',z(\xi)))}{u(T',z)-u(T',z(\xi))}(\tau-T')\right)^{-1}. \end{aligned}$$

Choose a sequence $(z_\nu)_{\nu\in\mathbb{N}}$ converging to $z(\xi)$ such that

(3.36)

$$\begin{aligned} &D^-_\xi u(\tau,\xi) \\ &\quad = \lim_{\nu\to+\infty} \left(\frac{z_\nu - z(\xi)}{u(T',z_\nu)-u(T',z(\xi))} + \frac{f'(u(T',z_\nu)) - f'(u(T',z(\xi)))}{u(T',z_\nu)-u(T',z(\xi))}(\tau-T')\right)^{-1}. \end{aligned}$$

By the continuity of $u(T',\cdot)$,

$$\lim_{\nu\to+\infty} \frac{f'(u(T',z_\nu)) - f'(u(T',z(\xi)))}{u(T',z_\nu)-u(T',z(\xi))} = f''(u(T',z(\xi)))$$

and hence

$$\lim_{\nu\to+\infty} \frac{z_\nu - z(\xi)}{u(T',z_\nu)-u(T',z(\xi))}$$

does exist. Call ℓ its value. We observe that $\ell \leq 0$. In fact, assume by contradiction that $\ell > 0$. For ν sufficiently large

$$\frac{u(T',z_\nu)-u(T',z(\xi))}{z_\nu - z(\xi)} > 0. \tag{3.37}$$

Let $\xi_\nu = z_\nu + f'(u(T', z_\nu))(\tau - T')$. Hence $\xi_\nu \to \xi$ as $\nu \to +\infty$. Since f' is increasing, (3.34) and (3.37) imply

$$\frac{u(\tau, \xi_\nu) - u(\tau, \xi)}{\xi_\nu - \xi} = \frac{u(T', z_\nu) - u(T', z(\xi))}{\xi_\nu - \xi} > 0,$$

which contradicts the assumption on $D^-_\xi u(\tau, \xi)$. By (3.36)

$$D^-_\xi u(\tau, \xi) \geq \frac{1}{f''\left(u(T', z(\xi))\right)(\tau - T')},$$

proving (3.33).

Since u is locally Lipschitz continuous, then it is a.e. differentiable on $\mathcal{A}_1$ and by construction it satisfies $u_t + f'(u)u_x = 0$ a.e. Moreover by definition it is a weak entropic solution to (1.1) in $\mathcal{A}_2$. Now observe that, for any $t \in [0, T]$, $u(t, \theta^-_{x_0}(t)-) = w(x_0-)$ since $u(t, \cdot)$ is left continuous at $\theta^-_{x_0}(t)$. On the other hand, if $w(x_0-) > 0$ then one has $w(x_0-) = S_t\psi(w(x_0-), x_0)(\theta^-_{x_0}(t)-) = S_t\psi(w(x_0-), x_0)(\theta^-_{x_0}(t)+)$ since $\theta_{x_0^-}$ is a minimal backward characteristic of $S_{(\cdot)}\psi(w(x_0-), x_0)$. If $w(x_0-) = 0$ then $u(t, \theta^-_{x_0}(t)+) = 0$. Thus $u(t, \theta^-_{x_0}(t)-) = u(t, \theta^-_{x_0}(t)+)$ for any $t \in (0, T)$. It follows that u is a weak entropic solution to (1.1) in the interior of Ω. Furthermore it clearly fulfills (1.2) in the sense of (ii) in Definition 1.

4. We claim that for any $t \in I_1 \cup I_3 \cup I_4$,

$$\lim_{x \to 0^+} u(t, x) = \tilde{u}(t). \tag{3.38}$$

If $t \in I_1 \cup I_3$ (3.38) follows by using the same arguments at point 3. Let $t \in I_4$ and fix $\varepsilon > 0$. For any $x \in (0, \varepsilon(t - \tau^4))$ we have $0 < f'(u(t, x)) \leq \varepsilon$. Indeed fix $\xi > 0$ such that $\theta^\pm_\xi(\tau^4) = 0$. By construction $s \in I_3$ does not exist such that $\theta_s(t) = x$. Hence $x = \theta^\pm_\zeta(t)$ for some $\zeta > 0$ and $f'(u(t, x)) = f'(w(\zeta\pm))$. If $f'(u(t, x)) > \varepsilon$, then there exists $\tau > \tau^4$ such that $\theta^\pm_\zeta(\tau) = 0$, thus contradicting (3.20). If $f'(u(t, x)) \leq 0$, then $\theta^\pm_\zeta$ and $\theta^\pm_\xi$ would intersect in the interior of Ω. Hence $\lim_{x \to 0^+} f'(u(t, x)) = 0$, so that (3.38) holds. Moreover since $f'(\tilde{u}(t)) > 0$ for every $t \in I_1 \cup I_3$, it follows that

$$\lim_{x \to 0^+} \operatorname{sgn} f'(u(t, x)) = 1 \qquad \forall\, t \in I_1 \cup I_3 \cup I_4. \tag{3.39}$$

Thus if $t \in I_1 \cup I_3 \cup I_4$, then u satisfies the boundary condition related to $\tilde{u}$ in the sense of Definition 1. If $t \in I_5$ such a boundary condition is fulfilled by construction. Hence u solves (1.1)–(1.3) with $\tilde{u}$ as in (3.25). Now we show that

$$\lim_{t \to T^-} \int_0^{+\infty} |u(t, x) - w(x)|\ dx = 0. \tag{3.40}$$

Let $(t_\nu)_{\nu \in \mathbb{N}}$ be an arbitrary increasing sequence converging to T. Then

$$\int_0^{+\infty} |u(t_\nu, x) - w(x)|\ dx = \int_0^{x_0} |u(t_\nu, x) - w(x)|\ dx + \int_{x_0}^{+\infty} |u(t_\nu, x)|\ dx. \tag{3.41}$$

Let us estimate each term in the right-hand side of (3.41). Concerning the first term we show that

$$\lim_{\nu \to +\infty} u(t_\nu, x) = w(x) \qquad \forall\, x \in (0, x_0). \tag{3.42}$$

In fact, let $\varepsilon > 0$ be given and fix $\delta > 0$ such that $|w(y) - w(x)| \le \varepsilon$ whenever $x \le y < x + \delta$. Let $\tau < T$ be such that $\theta^-_{x+\delta}(\tau) = x$ (such a τ does exist since $f'(w(x)) \ge x/T$). We claim that if $t_\nu > \tau$ then $|u(t_\nu, x) - w(x)| \le \varepsilon$. Assume first $x \in \mathcal{B}(t_\nu)$. Then $\theta^-_{x_{k(\nu)}}(t_\nu) < x < \theta^+_{x_{k(\nu)}}(t_\nu)$ for some $k(\nu) \in \mathbb{N}$, with $x \le x_{k(\nu)} < x + \delta$ since $\theta^+_x, \theta^\pm_{x_{k(\nu)}}$, and $\theta^-_{x+\delta}$ do not intersect each other in the interior of Ω. Hence from the above remark and (3.30) it follows $|u(t_\nu, x) - w(x)| \le \varepsilon$. Suppose now that $x \notin \mathcal{B}(t_\nu)$. Then with arguments similar to the previous ones we get that $x = \theta^\pm_y(t_\nu)$ with $x \le y < x + \delta$ and $u(t_\nu, x) = w(y\pm)$. The conclusion follows easily.

Furthermore there exists $C_2 > 0$ such that $|u(t_\nu, x) - w(x)| \le C_2$ for any $x \in (0, x_0)$. Hence by the dominated convergence theorem we get

$$\lim_{\nu \to +\infty} \int_0^{x_0} |u(t_\nu, x) - w(x)| \, dx = 0. \tag{3.43}$$

Concerning the second term in the right-hand side of (3.41), observe first that if $w(x_0-) = 0$, then $f'(0) \ge x/T$, due to (2.17). Hence $u(t_\nu, x) = 0$ for any $x \ge x_0$ since $x_0 + f'(0)(t_\nu - T) \le x_0$. Otherwise, $t \mapsto S_t\psi(w(x_0-), x_0)$ is continuous as a map from $[0, T]$ into $L^1(\mathbb{R}^+)$ and $S_T\psi(w(x_0-), x_0)(y) = 0$ whenever $y \ge x_0$. By combining this with (3.41) and (3.43) and by the arbitrary choice of $(t_\nu)_{\nu \in \mathbb{N}}$, we obtain (3.40).

3.3. Proof of Remark 2.4. As in Remark 2.3 the boundedness of w together with (2.19′) imply that w has finite total increasing variation (and hence total increasing variation as well) on sets bounded away from the origin. Thus we can assume that w has left and right limits at every point and is right continuous. Moreover (2.19′) implies that $w(x-) \ge w(x)$. Next observe that (2.19′) holds iff the function

$$\gamma : x \mapsto w(x) - \int_c^x \frac{f'(w(\xi))}{\xi f''(w(\xi))} \, d\xi, \qquad c > 0, \tag{3.44}$$

is nonincreasing on $\mathbb{R}^+$ and hence iff

$$D^+\gamma(x) \le 0 \qquad \forall\ x > 0. \tag{3.45}$$

Now we show that

$$D^+\gamma(x) = D^+w(x) - \frac{f'(w(x))}{xf''(w(x))}. \tag{3.46}$$

If $x > 0$ is a point of continuity for w then

$$\begin{aligned} D^+\gamma(x) &= \limsup_{h \to 0} \left[\frac{w(x+h) - w(x)}{h} - \frac{1}{h} \int_x^{x+h} \frac{f'(w(\xi))}{\xi f''(w(\xi))} \, d\xi \right] \\ &= D^+w(x) - \frac{f'(w(x))}{xf''(w(x))}. \end{aligned}$$

Otherwise since w is right continuous and $w(x-) > w(x)$,

$$\limsup_{h \to 0^+} \left[\frac{w(x+h) - w(x)}{h} - \frac{1}{h} \int_x^{x+h} \frac{f'(w(\xi))}{\xi f''(w(\xi))} \, d\xi \right] = D^+w(x) - \frac{f'(w(x))}{xf''(w(x))},$$

$$\limsup_{h \to 0^-} \left[\frac{w(x+h) - w(x)}{h} - \frac{1}{h} \int_x^{x+h} \frac{f'(w(\xi))}{\xi f''(w(\xi))} \, d\xi \right] = -\infty,$$

which imply (3.46).

4. Proof of Theorem 3. We will give the proof of the statement concerning $\mathcal{A}(T,\mathcal{U})$, the one concerning $\mathcal{A}(\bar{x},\mathcal{U})$ being entirely similar. Let $(\tilde{u}_\nu)_{\nu\in\mathbb{N}} \subset \mathcal{U}$. Then, being G bounded, by (2.9) and (2.17) there exist $C, \alpha > 0$ such that

$$|S_t\tilde{u}_\nu(x)| \le \begin{cases} C & \text{if} \quad x < \alpha \\ 0 & \text{if} \quad x \ge \alpha \end{cases} \qquad \forall\, t \in [0,T] \quad \forall\, \nu \in \mathbb{N}. \tag{4.1}$$

Hence $(S_T\tilde{u}_\nu)_{\nu\in\mathbb{N}}$, $(S_{(\cdot)}\tilde{u}_\nu)_{\nu\in\mathbb{N}}$ are weak* relatively compact in $L^\infty(\mathbb{R}^+)$, $L^\infty(\Omega)$, respectively, so that we can assume

$$S_T\tilde{u}_\nu \overset{*}{\rightharpoonup} w \quad \text{in} \quad L^\infty(\mathbb{R}^+), \tag{4.2}$$

$$S_{(\cdot)}\tilde{u}_\nu \overset{*}{\rightharpoonup} u \quad \text{in} \quad L^\infty(\Omega), \tag{4.3}$$

for some functions $w \in L^\infty(\mathbb{R}^+)$, $u \in L^\infty(\Omega)$. We shall prove that $w \in \mathcal{A}(T,\mathcal{U})$ and that there exists a subsequence of $(S_T\tilde{u}_\nu)_{\nu\in\mathbb{N}}$ converging to w in $L^1(\mathbb{R}^+)$. By (4.1) and Remark 2.3 for every $a > 0$ there exists $C_a > 0$ such that

$$\mathrm{TV}\left\{S_t\tilde{u}_\nu; [a,+\infty)\right\} \le C_a \qquad \forall\, t \in [0,T] \quad \forall\, \nu. \tag{4.4}$$

Moreover there exists $L > 0$ such that if $0 < a' < a$, then

$$\int_a^{+\infty} |S_t\tilde{u}_\nu(x) - S_s\tilde{u}_\nu(x)|\ dx \le L|t-s|C_{a'} \qquad \forall\, t,s > 0 \quad \forall\, \nu. \tag{4.5}$$

By Helly's theorem for any fixed $a > 0$ there exists a subsequence $\left(S_t\tilde{u}_{\nu_j}\right)_{j\in\mathbb{N}}$ which converges to some function $v_a(t,\cdot)$ in $L^1{}_{loc}([a,+\infty))$ for every $t \in [0,T]$. But (4.3) implies that such a function must coincide with u and hence by using (4.1), for every $t \in [0,T]$, the original sequence $(S_t\tilde{u}_\nu)_{\nu\in\mathbb{N}}$ converges to $u(t,\cdot)$ in $L^1(\mathbb{R}^+)$. In particular, from the convergence of $(S_T\tilde{u}_\nu)_{\nu\in\mathbb{N}}$ to $u(T,\cdot)$ and (4.2) it follows that $u(T,\cdot) = w$. Thus to complete the proof it remains to show that u is a solution of (1.1)–(1.3) corresponding to a boundary data $\tilde{u} \in \mathcal{U}$.

By (4.1) and the regularity of f it can be assumed that, for every $t \in [0,T]$, the sequence $(f(S_t\tilde{u}_\nu))_{\nu\in\mathbb{N}}$ converges in $L^1(\mathbb{R}^+)$ to $f(u(t,\cdot))$. It follows that, for any nonnegative $\mathcal{C}^1$ function ϕ with compact support in $[0,T]\times(0,+\infty)$ and for any $k \in \mathbb{R}$, we obtain

$$\begin{aligned} &\iint \left\{|u-k|\,\phi_t + \big(f(u)-f(k)\big)\mathrm{sgn}\,(u-k)\phi_x\right\}\ dxdt \\ &\quad = \lim_{\nu\to+\infty} \iint \left\{|S_t\tilde{u}_\nu - k|\,\phi_t + \big(f(S_t\tilde{u}_\nu)-f(k)\big)\mathrm{sgn}\,(S_t\tilde{u}_\nu - k)\phi_x\right\}\ dxdt \\ &\quad \ge 0. \end{aligned} \tag{4.6}$$

Hence u is a weak entropic solution of (1.1)–(1.2) in the interior of Ω.

Next we show that the traces of the functions $f(u)$, sgn $f'(u)$ at $x = 0$ exist in the sense of (2.6)–(2.7). By Remark 2.1 it is sufficient to prove that u admits in the interior of Ω the representation (2.10). Let Υ_ν, $\nu \in \mathbb{N}$, be the traces of $f(S_{(\cdot)}\tilde{u}_\nu)$, $\nu \in \mathbb{N}$. By Remarks 2.1–2.2, for every given $t \in [0,T]$ and for any $\nu \in \mathbb{N}$, $S_t\tilde{u}_\nu(x) = b\left((x - y_\nu(t,x))/t\right)$ for a.e. $x > 0$ with $y_\nu(t,x)$ denoting the unique point where the function $y \mapsto \Psi_{\Upsilon_\nu}(t,x,y)$ defined by (2.11) attains its minimum. Since by (2.9) and (4.1) Υ_ν are uniformly bounded, there exists a subsequence still denoted

$(\Upsilon_\nu)_{\nu\in\mathbb{N}}$ which converges weak* in L^∞ to some function $\Upsilon \in L^\infty([0,T])$. Thus for every $(t,x) \in \text{int } \Omega$ the sequence of maps $(\Psi_{\Upsilon_\nu}(t,x,\cdot))_{\nu\in\mathbb{N}}$ converges uniformly to $\Psi_\Upsilon(t,x,\cdot)$ and hence for all $t \in [0,T]$ and for a.e. $x > 0$ the corresponding minimum points $y_\nu(t,x)$ being unique (see Remark 2.1) converge to the minimum point $y(t,x)$ of $\Psi_\Upsilon(t,x,\cdot)$ proving that u satisfies (2.10).

Observe now that $f(\tilde{u}_\nu)$ are uniformly bounded, and hence it can be assumed that

$$f(\tilde{u}_\nu) \overset{*}{\rightharpoonup} \Phi \quad \text{in} \quad L^\infty([0,T])$$

for some function $\Phi \in L^\infty([0,T])$. Since $f(\tilde{u}_\nu(t)) \in f(G(t))$ and by (2.8) $f(\tilde{u}_\nu(t)) \leq \Upsilon_\nu(t)$ for a.e. t, being f convex and G convex closed valued it follows that $\Phi(t) \in f(G(t))$ and $\Phi(t) \leq \Upsilon(t)$ for a.e. t. Hence there exists a measurable selection $\tilde{u}$ from G such that

$$\Phi(t) = f(\tilde{u}(t)), \quad f(\tilde{u}(t)) \in f(G(t)), \quad f(\tilde{u}(t)) \leq \Upsilon(t) \quad \text{for a.e. } t > 0.$$

Since, for any $t \in J$, on bounded subsets of L^∞ the functionals $y \mapsto \int_0^t q_i(s, y(s))\, ds$, $i = 1, \dots, N$, are sequentially lower semicontinuous w.r.t. weak convergence on L^1 (see Theorem 3 in [10]), it follows that $\tilde{u} \in \mathcal{U}$. Therefore, to prove that u fulfills (iii) in Definition 1, it remains to show that $\Upsilon(t) = f(\tilde{u}(t))$ whenever $\mu(t) \geq 0$, with μ denoting the trace of sgn $f'(u)$ at $x = 0$ as defined in (2.7). Assume that $\mu(t) = 0$. Then there exists $\delta > 0$ such that $f'(u(t,x)) = 0$ whenever $x \in (0,\delta) \setminus \mathcal{F}$, so that $\Upsilon(t) = f(b(0)) = f(\tilde{u}(t))$.

Now consider the set

$$\mathcal{P} \doteq \{t \in [0,T] : \ \mu(t) = 1\} \tag{4.7}$$

and assume that $\mathcal{P}$ has positive measure. Let μ_ν be the trace of sgn $f'(S_{(\cdot)}\tilde{u}_\nu)$ as defined in (2.7). We claim that

$$\liminf_{\nu\to+\infty} \mu_\nu(t) \geq 0 \qquad \text{for a.e. } t \in \mathcal{P}. \tag{4.8}$$

Indeed, suppose that (4.8) does not hold. Then there exists $\mathcal{P}' \subseteq \mathcal{P}$ with positive measure such that for every $t \in \mathcal{P}'$ there is a subsequence $(\mu_{\nu_k}(t))_{k\in\mathbb{N}}$ of $(\mu_\nu(t))_{\nu\in\mathbb{N}}$ such that $\mu_{\nu_k}(t) = -1$ for all k. This means that, for any such t, $f'(S_t\tilde{u}_{\nu_k}(x)) < 0$ for x sufficiently close to zero. Hence by (2.17), since genuine characteristics do not intersect in the interior of the domain, it follows that $S_t\tilde{u}_{\nu_k}(x) = 0$ for every $x > 0$ and hence $f'(0) < 0$. Fix $R > 0$ and define

$$\mathcal{R} \doteq \{(t,x) \in \mathcal{P}' \times [0,R] : \ f'(u(t,x)) > 0\}. \tag{4.9}$$

Clearly $meas(\mathcal{R}) > 0$. Let $0 < \varepsilon < meas(\mathcal{R})/2$. By Egoroff's theorem there exists $\mathcal{R}' \subset \mathcal{R}$ such that $meas(\mathcal{R} \setminus \mathcal{R}') < \varepsilon$ and $S_{(\cdot)}\tilde{u}_\nu$ converges uniformly to u on $\mathcal{R}'$. Therefore, if $(t,x) \in \mathcal{R}'$, for ν sufficiently large $S_t\tilde{u}_\nu(x) \geq b(0)$ which gives a contradiction since $f'(0) < 0$ implies $0 < b(0)$ by the convexity of f. Hence $\lim_{\nu\to\infty}(f(\tilde{u}_\nu)(t) - \Upsilon_\nu(t)) = 0$ for a.e. $t \in \mathcal{P}$. Since $f(\tilde{u}_\nu) \overset{*}{\rightharpoonup} f(\tilde{u})$ and $\Upsilon_\nu \overset{*}{\rightharpoonup} \Upsilon$ in L^∞, we get $f(\tilde{u})(t) = \Upsilon(t)$ for a.e. $t \in \mathcal{P}$.

5. An application. When modelling traffic phenomena in first approximation we find it is reasonable to treat a flow of traffic on a highway as a continuum with an observable density $u(t,x)$ equal to the number of cars per unit length and a flux

$f(t,x)$ equal to the number of cars crossing the point x per unit time. Making the assumption that at each point x the flux f is a function only of the density u at x leads to the conservation law (see [9])

$$u_t + [uv(u)]_x = 0, \tag{5.1}$$

where $v(u)$ represents the velocity of the cars as a function of their density. In practice one often takes $v(u) = a_1 \ln(a_2/u)$ for suitable constants a_1 and a_2. Consider the problem of minimizing the mean time which occurs in driving through a stretch of the highway between an entry at a point $x = 0$ and an exit at a point $x = \bar{x}$ by controlling the density $\tilde{u}(t)$ of cars entering the highway at time t equal to the value of u at the boundary $x = 0$. Suppose that at time $t = 0$ no cars are on the stretch of highway $[0, \bar{x}]$. Let $g(t)$ be the number of cars arriving at $x = 0$ per unit of time. We may assume that g is a continuous function with compact support. Let u_m be the maximum density, i.e., the value for which the cars are bumper to bumper. Then there are quite natural assumptions that can be made on the boundary data $\tilde{u}$:

(i) the net flux of cars entering the stretch of highway must be equal to the total number of cars arriving at the entry:

$$\int_0^{+\infty} \tilde{u}(s)v(\tilde{u}(s))\ ds = \int_0^{+\infty} g(s)\ ds; \tag{5.2}$$

(ii) at any time $t > 0$ the total number of cars which have entered the highway until that moment must be less than or equal to the total number of cars that have arrived at the entry in the same period of time:

$$\int_0^{t} \tilde{u}(s)v(\tilde{u}(s))\ ds \le \int_0^{t} g(s)\ ds; \tag{5.3}$$

(iii) the maximum number of cars entering the highway must be less than or equal to the maximum density of cars allowed on the highway:

$$\tilde{u}(t) \in [0, u_m]; \tag{5.4}$$

(iv) after a period of time sufficiently large no cars enter the highway:

$$\tilde{u}(t) = 0, \qquad t > \tau, \qquad \exists \tau > 0. \tag{5.5}$$

Then if $(t,x) \mapsto S_t\tilde{u}(x)$ denotes the solution to (5.1), (1.2), (1.3), we will be interested in minimizing the difference between the average incoming time of cars at $x = \bar{x}$ and at $x = 0$:

$$\left(\int_0^{+\infty} t\ S_t\tilde{u}(\bar{x})v\left(S_t\tilde{u}(\bar{x})\right)\ dt - \int_0^{+\infty} t\ g(t)\ dt\right)\left(\int_0^{+\infty} g(t)\ dt\right)^{-1}, \tag{5.6}$$

which clearly is equivalent to the minimization problem

$$\min_{\tilde{u}\in\mathcal{U}} \int_0^{+\infty} t\ S_t\tilde{u}(\bar{x})v\left(S_t\tilde{u}(\bar{x})\right)\ dt, \tag{5.7}$$

where the admissible set $\mathcal{U}$ consists of all L^∞ functions $\tilde{u}$ satisfying (5.2)–(5.5) for a.e. $t > 0$. Here we have a strictly concave flux $f(u) = uv(u)$. Since it is not restrictive to consider boundary data with characteristics entering the domain $\mathbb{R}^+ \times \mathbb{R}^+$, one can

assume that $\tilde{u} \in [0, b(0)] \subseteq [0, u_m]$ for a.e. $t > 0$ and for any admissible boundary data $\tilde{u}$. Moreover by the basic structure of a solution to (1.1)–(1.3), from (5.5) it follows that $S_t\tilde{u}(\bar{x}) = 0$ for a.e. $t > \tau + \bar{x}\ b(0)/f(b(0)) \doteq \tau'$. Therefore problem (5.7) can be restated

$$\min_{\tilde{u}\in\mathcal{U}} \int_0^{\tau'} t\ S_t\tilde{u}(\bar{x})\upsilon\big(S_t\tilde{u}(\bar{x})\big)\ dt, \tag{5.8}$$

where $\mathcal{U}$ is a set of the form (2.23), q being the identity map and G the multifunction

$$G(t) = \begin{cases} [0, b(0)] & \text{if} \quad t \le \tau', \\ \{0\} & \text{otherwise,} \end{cases}$$

with an additional constraint given by (5.2). Observe that the compactness of the attainable set $\mathcal{A}(\bar{x}, \mathcal{U})$ still holds in connection with such an admissible set of boundary controls as it follows from the proof of Theorem 3. Thus, since the map $u \mapsto \int_0^{\tau'} t\ u(t)\upsilon(u(t))\ dt$ is continuous as a functional from $\{\, u \in L^\infty([0,\tau']) :\ ||u||_\infty \le b(0)\,\}$ into $\mathbb{R}$ w.r.t. the L^1-norm, by Corollary 1 problem (5.8), admits a solution.

6. Appendix. Here we extend the L^1-contraction property (2.12) established in [14] for piecewise continuously differentiable solutions of the mixed initial boundary value problem (2.1)–(2.3) to the class of all solutions associated with every initial and boundary data in the domain

$$\mathcal{D} \doteq \big\{\, (\bar{u}, \tilde{u}) \in L^\infty(\mathbb{R}^+) \cap L^1(\mathbb{R}^+) \times L^\infty(\mathbb{R}^+) :\ \tilde{u}(t) \ge b(0) \quad \text{a.e.} \quad t \,\big\}.$$

In the following we denote $\mathcal{T}_t : L^\infty \to L^\infty,\ t > 0$, the translation operator, i.e., $\mathcal{T}_t\tilde{u}(s) \doteq \tilde{u}(t+s)\ \forall s > 0$.

THEOREM 4. *Let $f : \mathbb{R} \to \mathbb{R}$ be a continuously differentiable strictly convex function. Then there exists a continuous map $S : \mathbb{R}^+ \times \mathcal{D} \to L^\infty(\mathbb{R}^+)$ with the following properties:*

(i) $S_0(\bar{u}, \tilde{u}) = \bar{u},\ S_{s+t}(\bar{u}, \tilde{u}) = S_s\big(S_t(\bar{u}, \tilde{u}), \mathcal{T}_t\tilde{u}\big)\ \forall s, t > 0$;

(ii) $\|S_t(\bar{u}, \tilde{u}) - S_t(\bar{v}, \tilde{v})\|_{L^1(\mathbb{R}^+)} \le \|\bar{u} - \bar{v}\|_{L^1(\mathbb{R}^+)} + \|f(\tilde{u}) - f(\tilde{v})\|_{L^1([0,t])}\ \forall t > 0$;

(iii) *each trajectory $t \to S_t(\bar{u}, \tilde{u})$ yields the unique solution (in the sense of Definition 1) to the initial boundary value problem* (2.1)–(2.3).

Proof. For any given $R > 0$ consider the set

$$\mathcal{D}_R \doteq \big\{\, (\bar{u}, \tilde{u}) \in \mathcal{D} :\ \|\tilde{u}\|_\infty \le R \big\}$$

endowed with the product topology of $L^1(\mathbb{R}^+) \times {L^1}_{loc}(\mathbb{R}^+)$. Then to prove Theorem 4 it suffices to show that for any $R > 0$ there exists a continuous map $S : \mathbb{R}^+ \times \mathcal{D}_R \to L^\infty(\mathbb{R}^+)$ satisfying (i), (ii), (iii).

Let $\widehat{\mathcal{D}}_R$ be the set of couples $(\bar{u}, \tilde{u}) \in \mathcal{D}_R$ of piecewise constant functions (with finite number of discontinuities). Observe first that any solution of (2.1)–(2.3) associated with initial and boundary data in $\widehat{\mathcal{D}}_R$ is piecewise continuously differentiable. Then for every $(\bar{u}, \tilde{u}) \in \widehat{\mathcal{D}}_R$ let $\widehat{S}_t(\bar{u}, \tilde{u})$ be the value at time t of the solution to (2.1)–(2.3) which, by Remark 2.2, is unique, admits a representation of the form (2.10), and satisfies the L^1 contraction property (ii). Since $\widehat{\mathcal{D}}_R$ is a dense subset of $\mathcal{D}_R$ the continuous flow $\widehat{S} : \mathbb{R}^+ \times \widehat{\mathcal{D}}_R \to \mathcal{D}_R$ can be uniquely extended by continuity to a continuous map $S : \mathbb{R}^+ \times \mathcal{D}_R \to \mathcal{D}_R$ satisfying (ii) as well. Thus the proof will be

completed if we show that $t \to S_t(\bar{u}, \tilde{u})$ admits a representation of the form (2.10) for every $(\bar{u}, \tilde{u}) \in \mathcal{D}_R$.

Let $(\bar{u}_\nu)_{\nu\in\mathbb{N}}$, $(\tilde{u}_\nu)_{\nu\in\mathbb{N}}$, $(\bar{u}_\nu, \tilde{u}_\nu) \in \mathcal{D}_R$, be two sequences of piecewise constant functions such that

$$\bar{u}_\nu \to \bar{u} \quad \text{in} \quad L^1(\mathbb{R}^+), \tag{6.1}$$

$$f(\tilde{u}_\nu) \to f(\tilde{u}) \quad \text{in} \quad L^1{}_{loc}(\mathbb{R}^+). \tag{6.2}$$

Then by previous arguments, for every fixed $t > 0$, one has

$$S_t(\bar{u}_\nu, \tilde{u}_\nu)(x) = b\left(\frac{x - y_\nu(t,x)}{t}\right)$$

for a.e. $x > 0$, $y_\nu(t,x)$ denoting the unique minimum point for the function $y \mapsto \Psi_{\Upsilon_\nu}(t,x,y)$ defined by (2.11) in connection with the trace Υ_ν at $x = 0$ of $f\big(S_{(\cdot)}(\bar{u}_\nu, \tilde{u}_\nu)\big)$. Observe that by (2.9) Υ_ν are uniformly bounded. Thus there exists a subsequence still denoted $(\Upsilon_\nu)_{\nu\in\mathbb{N}}$ which converges weak* in L^∞ to some function $\Upsilon \in L^\infty(\mathbb{R}^+)$. Therefore, for every $x > 0$ the sequence of maps $(\Psi_{\Upsilon_\nu}(t,x,\cdot))_{\nu\in\mathbb{N}}$ converges uniformly to $\Psi_\Upsilon(t,x,\cdot)$. This implies that for a.e. $x > 0$ the corresponding minimum points $y_\nu(t,x)$ being unique (see Remark 2.1) converge to the minimum point $y(t,x)$ of $\Psi_\Upsilon(t,x,\cdot)$ and hence $\big(b\left((x - y_\nu(t,x))/t\right)\big)_{\nu\in\mathbb{N}}$ converges to $b\left((x - y(t,x))/t\right)$ for a.e. $x > 0$ proving that $S_t(\bar{u}, \tilde{u})$ satisfies (2.10). □

Acknowledgments. The authors would like to thank Prof. Alberto Bressan for suggesting the problem and for many helpful discussions.

REFERENCES

[1] F. Ancona and A. Marson, *Scalar Non-linear Conservation Laws with Integrable Boundary Data*, Nonlinear Anal., to appear.
[2] J. Aubin and A. Cellina, *Differential Inclusion*, Springer-Verlag, Berlin, New York, 1984.
[3] C. Bardos, A.Y. Leroux, and J.C. Nedelec, *First order quasilinear equations with boundary conditions*, Comm. in Partial Differential Equations, 9 (1979), pp. 1017–1034.
[4] A. Bressan, *Lectures Notes on Systems of Conservation Laws*, S.I.S.S.A., International School for Advanced Studies, Trieste, Italy, 1995.
[5] C. Dafermos, *Generalized characteristic and the structure of solutions of hyperbolic conservation laws*, Indiana Math. J., 26 (1977), pp. 1097–1119.
[6] F. Dubois and P. Le Floch, *Boundary conditions for nonlinear hyperbolic systems of conservation laws*, J. Differential Equations, 71 (1988), pp. 93–122.
[7] A.F. Filippov, *Differential equations with discontinuous right-hand side*, Amer. Math. Soc. Transl. Ser. 2, 42 (1964), pp. 199–231.
[8] A.V. Fursikov and O.Yu. Imanuvilov, *Controllability of Evolution Equations*, Lecture Notes Series 34, Seoul National University, Research Institute of Mathematics, Global Analysis Research Center, Seoul, 1996.
[9] K.E. Gustafson, *Partial Differential Equations Second Edition*, John Wiley & Sons, New York, 1987.
[10] A.D. Ioffe, *On lower semicontinuity of integral functionals*. I, SIAM J. Control Optim., 15 (1977), pp. 521–538.
[11] S.N. Kruzkov, *First order quasilinear equations in several independent variables*, Math. USSR Sbornik, 10 (1970), pp. 217–243.
[12] P.D. Lax, *Hyperbolic systems of conservation laws* II, Comm. Pure Appl. Math., 10 (1957), pp. 537–566.
[13] P.D. Lax, *The formation and decay of shock waves*, Amer. Math. Monthly, 79 (1972), pp. 227–241.
[14] P. Le Floch, *Explicit formula for scalar non-linear conservation laws with boundary condition*, Math. Methods Appl. Sci., 10 (1988), pp. 265–287.
[15] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, 2nd ed., Springer-Verlag, Berlin, New York, 1994.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 1, pp. 313–335, January 1998

017

INFINITE LINEAR PROGRAMMING AND MULTICHAIN MARKOV CONTROL PROCESSES IN UNCOUNTABLE SPACES*

ONÉSIMO HERNÁNDEZ-LERMA† AND JUAN GONZÁLEZ-HERNÁNDEZ‡

Abstract. In this paper we use infinite linear programming to study Markov control processes in Borel spaces and the average cost criterion in the "unichain" and "multichain" cases. Under appropriate assumptions we show that in both cases the associated linear programs are solvable and that there is no duality gap. Moreover, conditions are given for minimizing (respectively, maximizing) sequences for the primal (respectively, dual) programs to converge to optimal solutions.

Key words. (discrete-time) Markov control processes, average cost criterion, infinite linear programming, generalized Farkas theorem

AMS subject classifications. 93E20, 90C40

PII. S0363012995292238

1. Introduction.

In this paper we use infinite-dimensional linear programming to study discrete-time Markov control processes (MCPs) with Borel state and control spaces and the average cost (AC) criterion. Our results extend several recent works on MCPs in Borel [17, 18, 19] and countable [21, 25] spaces. Namely, we consider the so-called unichain and multichain cases and for each of them we show that the associated linear programs, both the primal and the dual, are *solvable* and that there is *no duality gap*. Furthermore, conditions are given for minimizing (respectively, maximizing) sequences for the primal (respectively, dual) programs to converge to optimal solutions.

Linear programming (LP) is a standard technique to study many different classes of control problems—see, e.g., [13, 19, 27, 28, 31] and their references. In particular, as shown in the historical remarks in [1, 4, 19, 23], LP has been used since the early 1960s to study MCPs, but except for just a few papers, all of the literature deals with *countable*—mainly *finite*—spaces. To our knowledge, this is the first paper dealing with *multichain* MCPs and *uncountable* spaces.

For *finite* state MCPs, the LP approach to the multichain case has been used for many years, beginning with Denardo and Fox [12] and further extended by several authors, including Hordijk and Kallenberg [20, 23]—see also [1, 4, 19] for additional references. Moreover, extending previous work of Kallenberg [23], Altman and Spieksma [2] have recently found an explicit relation between the measure ν in an optimal solution (μ, ν) for the corresponding linear program P (see (3.5) or (3.7)) and the so-called deviation matrix (see also [22]), but their analysis and the result itself heavily depend on the finiteness of the state space.

In the *countably infinite* case, the LP approach was initiated in an interesting recent paper by Hordijk and Lasserre [21], in which one can already appreciate the difficulties in going from the finite to the infinite-space situation. For instance, in

*Received by the editors September 22, 1995; accepted for publication (in revised form) November 27, 1996. This research was partially supported by Consejo Nacional de Ciencia y Tecnología (CONACYT, México) grants 1332–E9206 and 3115P–E9608.

http://www.siam.org/journals/sicon/36-1/29223.html

†Departamento de Matemáticas, CINVESTAV-IPN, Apartado Postal 14–740, México, D.F. 07000, Mexico (ohernand@math.cinvestav.mx).

‡Departamento de Matemáticas, Facultad de Ciencias, Universidad Nacional Autónoma de México, Ciudad Universitaria, 04510 México, D.F., Mexico.

the latter case there seems to be nothing remotely close to the "deviation matrix" interpretation in [2] and there seems to be no direct way of getting the no-duality-gap condition, which is trivial in the finite-space case. More explicitly, one of the main technical difficulties in showing that our *infinite-dimensional* linear programs EP and P (see (3.3)–(3.4) and (3.5)–(3.6)) are consistent and solvable and that there is no duality gap requires verifying that some sets are *weakly closed* (see Theorems 5.1 and 5.6). Since typically this cannot be done by direct methods, we have to use "perturbations" of EP (see (4.4)–(4.6)) and P (see (4.13)) and "subconsistency" results (Theorem 5.5).

In addition to the LP approach, one could also use (in principle) recurrence-like conditions to study the multichain case. This has been done by Schäl [30] in the *countable*-space case under a Lyapunov condition and by Kurano [24] for *compact*-space MCPs under the Doeblin hypothesis. Both papers heavily rely of course on their respective assumptions, and extending their results to uncountable noncompact spaces poses a challenging problem.

The remainder of the paper is organized as follows. In section 2 we introduce the AC problems we shall be dealing with, and in section 3 we present the associated linear programs, which are called EP and P in the unichain and multichain cases, respectively, and their corresponding duals are called EP^* and P^*. In section 4, first we present two sets of hypotheses (Assumptions 4.1 and 4.2), and then we state our main results for EP (Theorems 4.5 and 4.6) and P (Theorems 4.9 and 4.10); their proofs are given in sections 6 and 7, respectively, whereas section 5 contains several technical preliminaries for the proofs. The latter include the *generalized Farkas theorem* of Craven and Koliha [11], which is used to give necessary and sufficient conditions for EP and P to be consistent. We conclude in section 8 with some general remarks.

2. Markov control processes and AC problems. MCPs have been discussed by many authors, so our review can be brief. Except for small changes, our notation generally follows Hernández-Lerma and Lasserre [17], which provides many related references. (See also [18], [19, Chapter 6].)

We consider an MCP $(X, A, \{A(x)|x \in X\}, Q, c)$ with state space X and control (or action) set A, both assumed to be Borel spaces with Borel σ-algebras $\mathcal{B}(X)$ and $\mathcal{B}(A)$, respectively. For every state $x \in X$, $A(x) \in \mathcal{B}(A)$ is the (nonempty) set of admissible control actions in x. We assume that the set

$$\mathbb{K} := \{(x, a)|x \in X,\ a \in A(x)\} \tag{2.1}$$

is a Borel subset of $X \times A$ and that it contains the graph of a measurable function from X to A. The transition law $Q(B|x, a)$, with B in $\mathcal{B}(X)$ and (x, a) in $\mathbb{K}$, is a stochastic kernel on A given $\mathbb{K}$, and the one-stage cost $c(x, a)$ is a *nonnegative* measurable function on $\mathbb{K}$.

Additional assumptions on the MCP are imposed in the following sections.

A *control policy* $\pi = \{\pi_0, \pi_1, \ldots\}$ is a sequence of stochastic kernels $\pi_t(\cdot|h_t)$ on A, given the previous history $h_t = (x_0, a_0, \ldots, x_{t-1}, a_{t-1}, x_t)$, satisfying the constraint

$$\pi_t(A(x_t)|h_t) = 1 \qquad \forall h_t,\ t = 0, 1, \ldots, \tag{2.2}$$

where x_n and a_n denote the state and control at time n. The set of all control policies is denoted by Π.

DEFINITION 2.1. (a) $\mathbb{F}$ *denotes the set of all measurable functions* $f : X \to A$ *such that* $f(x) \in A(x)$ *for all* X, *and* Φ *stands for the set of all stochastic kernels on* A, *given* X, *such that* $\varphi(A(x)|x) = 1$ *for all* $x \in X$.

(b) *A control policy* $\pi = \{\pi_t\}$ *is said to be a* randomized stationary *(or* relaxed*) policy—also known as a* Young measure *(see, e.g., Balder [5, 6])—if there exists* $\varphi \in \Phi$ *such that* $\pi_t(\cdot|h_t) = \varphi(\cdot|x_t)$ *for every history* h_t *and* $t = 0, 1, \ldots$.

(c) $\pi = \{\pi_t\}$ *is said to be a (deterministic or nonrandomized)* stationary policy *if there exists* $f \in \mathbb{F}$ *such that* $\pi_t(\cdot|h_t)$ *is concentrated at* $f(x_t)$ *for every* h_t *and* t.

We shall identify $\mathbb{F}$ (respectively, Φ) with the set of all stationary (respectively, randomized stationary) policies.

P_γ^π denotes the induced probability measure when using the policy π, given the initial distribution γ, and E_γ^π stands for the expectation operator with respect to P_γ^π. If γ is the unit mass at the initial state $x_0 = x$, we write P_γ^π and E_γ^π as P_x^π and E_x^π, respectively.

Remark 2.2. If $v(x, a)$ is a function on $\mathbb{K}$ and $\varphi \in \Phi$ is a randomized stationary policy, we write

$$v(x, \varphi) := \int_A v(x, a)\varphi(da|x), \quad x \in X.$$

In particular, for a stationary policy $f \in \mathbb{F}$, $v(x, f) := v(x, f(x))$.

AC problems. Let $J_n(\pi, \gamma)$ be the n-stage total expected cost when using the policy π, given the initial distribution γ; that is,

$$J_n(\pi, \gamma) := E_\gamma^\pi \left[\sum_{t=0}^{n-1} c(x_t, a_t) \right], \quad n = 1, 2, \ldots,$$

and $J_0(\cdot) \equiv 0$. The long-run expected AC is then defined as

$$J(\pi, \gamma) := \limsup_{n \to \infty} J_n(\pi, \gamma)/n, \tag{2.3}$$

and the *AC-value function* is

$$J^*(\gamma) := \inf_\pi J(\pi, \gamma), \quad \gamma \in \mathcal{P}(X), \tag{2.4}$$

where $\mathcal{P}(X)$ denotes the set of all probability measures on X. Then the so-called AC problem is to find a policy π^* such that

$$J(\pi^*, \gamma) = J^*(\gamma) \quad \text{for } \textit{all} \quad \gamma \in \mathcal{P}(X). \tag{2.5}$$

A policy π^* satisfying (2.5) is called *AC optimal*.

The usual *dynamic programming* approach to solve the AC problem is to find a *canonical triplet* (g, h, f^*), which consists of two real-valued functions g and h on X, and a stationary policy $f^* \in \mathbb{F}$, such that the pair (g, h) satisfies the *AC optimality equation* (2.6)–(2.7) below, and $f^*(x) \in A(x)$ attains the minimum in (2.6)–(2.7) for every $x \in X$, which (using the notation in Remark 2.2) yields (2.8)–(2.9): $\forall x \in X$,

$$g(x) = \inf_{a \in A(x)} \int_X g(y)Q(dy|x, a), \tag{2.6}$$

$$g(x) + h(x) = \inf_{a \in A(x)} \left[c(x, a) + \int_X h(y)Q(dy|x, a) \right], \tag{2.7}$$

$$g(x) = \int_X g(y)Q(dy|x, f^*), \quad \text{and} \tag{2.8}$$

$$g(x) + h(x) = c(x, f^*) + \int_X h(y)Q(dy|x, f^*). \tag{2.9}$$

Finally, if h is such that

$$\limsup_{n\to\infty} E_x^\pi[h(x_n)]/n = 0 \quad \text{for every } \pi \in \Pi \text{ and } x \in X, \tag{2.10}$$

then f^* is AC optimal and $g(\cdot)$ is the AC value function, i.e.,

$$J(f^*, x) = g(x) = J^*(x) \quad \forall x \in X, \tag{2.11}$$

and (2.5) is immediately obtained.

The LP formulation to the AC problem is closely related to (2.6)–(2.7), except that (2.5) is solved for every *fixed* initial distribution; i.e., given the initial distribution, say γ_0, we find π^* such that

$$J(\pi^*, \gamma_0) = J^*(\gamma_0) := \inf_\pi J(\pi, \gamma_0). \tag{2.12}$$

Another related problem is to find a *minimum pair* (π^0, γ^0), i.e., a policy π^0 and an initial distribution γ^0 such that

$$J(\pi^0, \gamma^0) = \inf_\gamma J^*(\gamma) =: \rho^*. \tag{2.13}$$

Of course, if the AC value function J^* is a *constant*, as in the so-called unichain case, using dynamic programming the idea would be to solve (2.6)–(2.9) with a *constant* function $g(\cdot)$, and then, under (2.10), we would obtain (2.11) with $g(\cdot) = J^*(\cdot) = \rho^*$.

The minimum pair problem is precisely the one solved in [17] and analyzed again here from a different viewpoint. We also study (2.12), which we call the γ_0-*AC problem* or AC problem in the *multichain* case (with initial distribution γ_0).

3. Associated linear programs. In this section we informally present the linear programs associated with the AC problems (2.12) and (2.13). The hypotheses under which they are well defined, together with our main results, are given in section 4.

As in [17] (see also [3], [18], or [19]), we first introduce the dual pairs $(M(\mathbb{K}), F(\mathbb{K}))$ and $(M(X), F(X))$ in Definition 3.1 below, where we use the functions

$$w(x, a) := 1 + c(x, a) \quad \text{and} \quad w_0(x) := \inf_{a\in A(x)} w(x, a). \tag{3.1}$$

Since (by assumption) c is nonnegative, we have

$$1 \le w_0(x) \le w(x, a) \qquad \forall (x, a) \in \mathbb{K}, \tag{3.2}$$

and—under Assumptions 4.1 or 4.2—w_0 is measurable.

DEFINITION 3.1. (a) $M(\mathbb{K})$ *denotes the normed vector space of finite signed measures* μ *on* $\mathbb{K}$ *with*

$$\|\mu\|_w := \int_{\mathbb{K}} w d|\mu| < \infty,$$

where $|\mu|$ stands for the total variation of μ, and $F(\mathbb{K})$ denotes the normed vector space of real-valued measurable functions u on $\mathbb{K}$ with

$$\|u\|_w := \sup_{(x,a)} |u(x,a)|/w(x,a) < \infty.$$

$(M(\mathbb{K}), F(\mathbb{K}))$ is a dual pair with respect to the bilinear form

$$\langle \mu, u \rangle := \int_K u d\mu, \quad \mu \in M(\mathbb{K}),\ u \in F(\mathbb{K}).$$

(b) *Similarly, replacing $\mathbb{K}$ and w by X and w_0, respectively, we get the dual pair $(M(X), F(X))$.*

We shall in fact consider several topologies on $M(\mathbb{K})$ (see Remark 5.3), but unless explicitly stated otherwise we shall always consider $M(\mathbb{K})$ to be endowed with the *weak topology* $\sigma(M(\mathbb{K}), F(\mathbb{K}))$. A similar remark holds for $M(X)$.

Notation. In any vector space, the "null" or "zero" vector will be denoted by "0". Thus $\mu(\cdot) = 0$ is the trivial measure and $u(\cdot) = 0$ is the null function. $M_+(\mathbb{K})$ denotes the convex cone of nonnegative measures in $M(\mathbb{K})$ and similarly for $M_+(X)$.

Now we introduce four linear operators (which will be *weakly continuous* under Assumptions 4.1 or 4.2—see Lemma 4.4):

$$L_0, L_1 : M(\mathbb{K}) \to M(X), \quad L : M(\mathbb{K}) \to \mathbb{R} \times M(X), \quad \text{and} \quad T : M(\mathbb{K})^2 \to M(X)^2,$$

defined, for every μ and ν in $M(\mathbb{K})$ and $B \in \mathcal{B}(X)$, as

$$\begin{aligned} &L_0\mu := \mu_1 := \text{marginal (or projection) of } \mu \text{ on } X, \\ &(L_1\mu)(B) := \mu_1(B) - \int_{\mathbb{K}} Q(B|x,a)\mu(d(x,a)), \\ &L\mu := (\langle \mu, 1\rangle, L_1\mu) \quad [\text{with} \quad \langle \mu, 1\rangle = \int_{\mathbb{K}} d\mu = \mu(\mathbb{K})], \end{aligned}$$

and

$$T(\mu, \nu) := (L_0\mu + L_1\nu, L_1\mu).$$

The corresponding adjoints

$$L_0^*, L_1^* : F(X) \to F(\mathbb{K}), L^* : \mathbb{R} \times F(X) \to F(\mathbb{K}), \text{ and } T^* : F(X)^2 \to F(\mathbb{K})^2$$

are given, for every g and h in $F(X)$, $\rho \in \mathbb{R}$, and $(x,a) \in \mathbb{K}$, by

$$\begin{aligned} &(L_0^* g)(x,a) := g(x), \\ &(L_1^* g)(x,a) := g(x) - \int_X g(y) Q(dy|x,a), \\ &L^*(\rho, g) := \rho + L_1^* g, \end{aligned}$$

and

$$T^*(g,h) := (L_0^* g + L_1^* h, L_1^* g).$$

With this notation, the *primal* linear program associated with the *minimum pair* problem (2.13) is (as in [17] or [19])

EP: minimize $\langle \mu, c \rangle$
(3.3) subject to $\mu \in M_+(\mathbb{K})$, $\langle \mu, 1\rangle = 1$, and $L_1\mu = 0$ (i.e., $L\mu = (1,0)$).

(See Remark 3.2(a).) The *dual* linear program is

$$\begin{aligned} &\boldsymbol{EP}^*\text{: maximize } \rho \ [= \langle (1,0),(\rho,h)\rangle] \\ &(3.4)\ \text{ subject to } L^*(\rho,h) \le c,\ (\rho,h) \in \mathbb{R} \times F(X). \end{aligned}$$

Similarly, for the γ_0 *AC problem* (2.12) the associated *primal* program is

$$\begin{aligned} &\boldsymbol{P}\text{: minimize } \langle \mu, c\rangle \ [= \langle (\mu,\nu),(c,0)\rangle] \\ &(3.5)\ \text{ subject to } T(\mu,\nu) = (\gamma_0, 0),\ (\mu,\nu) \in M_+(\mathbb{K})^2. \end{aligned}$$

The dual of P is

$$\begin{aligned} &\boldsymbol{P}^*\text{: maximize } \langle \gamma_0, g\rangle \ [= \langle (\gamma_0,0),(g,h)\rangle] \\ &(3.6)\ \text{ subject to } T^*(g,h) \le (c,0),\ (g,h) \in F(X)^2. \end{aligned}$$

Remark 3.2. (a) Vector equalities and inequalities are understood componentwise. For instance, (3.5) means that

$$L_0\mu + L_1\nu = \gamma_0 \quad \text{and} \quad L_1\mu = 0, \tag{3.7}$$

and (3.6) is the same as writing

$$L_0^* g + L_1^* h \le c \quad \text{and} \quad L_1^* g \le 0. \tag{3.8}$$

Incidentally, note that (3.8) (or (3.6)) can be "derived" from (2.6)–(2.7), in the sense that if (g,h) satisfies (2.6)–(2.7), then it satisfies (3.8). Thus, solving P^* is basically the same as finding the "maximal subsolution" of (2.6)–(2.7); see Lemma 7.1. This is interesting to note because, historically speaking, it is the way the "LP approach" to MCPs was born: trying to solve a dynamic programming equation rewriting it as a linear program.

(b) If (μ,ν) satisfies (3.5) (= (3.7))—in other words, if (μ,ν) is *feasible* for P—then μ is a *probability measure*. Of course, a similar remark holds if μ is feasible for EP—see (3.3).

(c) A function g in $F(X)$ will be identified with the function $(L_0^* g)(x,a) := g(x)$ in $F(\mathbb{K})$, in which case we can write $F(X) \subset F(\mathbb{K})$. This can be done because, by (3.2), $\|g\|_w \le \|g\|_{w_0} < \infty$ if g is in $F(X)$. Hence, we can also write $\langle \mu, g\rangle = \langle \mu_1, g\rangle$ for every $g \in F(X)$ and $\mu \in M(\mathbb{K})$, where $\mu_1 := L_0\mu$.

(d) The definition of the weight functions w and w_0 in (3.1) is useful because it automatically yields that c is in $F(\mathbb{K})$ and that (3.2) holds. However, it is important to keep in mind that to develop the LP approach we can take *any weight functions* w and w_0 as long as $c \in F(\mathbb{K})$ and (3.2) (together with Assumption 4.1(c)=4.2(d)) hold true. Similarly, Assumption 4.1(a) below can be replaced by the following statement: c is l.s.c. (lower semicontinuous) and the weight function w is inf-compact. This condition together with the requirement on w in Assumption 4.1(d) ensures, for instance, that the set of measures μ that satisfy (3.3) and $\int c d\mu < \infty$ is *tight* (cf. Assumption 5.1(c) in [17]).

4. Main results. In this section we present two different sets of assumptions, which are briefly discussed in Remark 4.3, and then we state our main results.

We shall use the following *notation*: If S is a metric space, we denote by $C_b(S)$ the Banach space of real-valued bounded continuous functions on S, endowed with

the supremum norm ($\|u\| := \sup_s |u(s)|$), and by $C_0(S)$ the subspace of continuous functions vanishing at infinity. If S is *compact*, then $C_b(S) = C_0(S)$.

The following assumption collects the hypotheses used in [17].

Assumption 4.1. (a) The one-stage cost function $c : \mathbb{K} \to \mathbb{R}_+$ ($\mathbb{R}_+ := [0, \infty)$) is *inf-compact*; i.e., for every real number r, the set $\{(x, a) \in \mathbb{K} | c(x, a) \leq r\}$ is compact.

(b) The transition law Q is *weakly continuous*; i.e., the function

$$\int_X v(y)Q(dy|\cdot) \text{ is in } C_b(\mathbb{K}) \text{ whenever } v \in C_b(X).$$

(c) $\int_X w_0(y)Q(dy|\cdot)$ is in $F(\mathbb{K})$; i.e., there is a constant C such that

$$\int_X w_0(y)Q(dy|x, a) \leq Cw(x, a) \quad \forall (x, a) \in \mathbb{K}.$$

(d) There is a policy π such that for every initial distribution γ, the average cost $J(\pi, \gamma) < \infty$, or, equivalently (by (3.1)),

$$\limsup_{n\to\infty} E^\pi_\gamma \left[\sum_{t=0}^{n-1} w(x_t, a_t)\right] /n < \infty.$$

Assumption 4.2. (a) X and $\mathbb{K}$ are locally compact separable metric spaces.

(b) The one-stage cost c is l.s.c.

(c) Q is weakly continuous (Assumption 4.1(b)) and, in addition, for every compact subset K of X, $Q(K|\cdot)$ vanishes at infinity; i.e., for every $\varepsilon > 0$ there is a compact set $K' = K'(\varepsilon, K)$ in $\mathbb{K}$ such that

$$Q(K|x, a) \leq \varepsilon \quad \forall (x, a) \notin K'.$$

(d) This is the same as Assumption 4.1(c).

Remark 4.3. (a) It is obvious that Assumptions 4.1 and 4.2 are not comparable. For instance, in Assumption 4.1 it is implicit that X and $\mathbb{K}$ are *Borel spaces* (see the second paragraph of section 2), which is a condition weaker than Assumption 4.2(a). But on the other hand, Assumption 4.1(a) is stronger than (i.e., it implies) Assumption 4.2(b). Similarly, Assumption 4.2(c) implies 4.1(b), but 4.1(d) is *not* required in Assumption 4.2.

(b) In most applications of MCPs, the spaces X and $\mathbb{K}$—and also the control set A—are "nice" subsets of Euclidean spaces, so Assumption 4.2(a) is not really too restrictive. A *sufficient condition* for it is that X and A are both locally compact separable metric spaces (which is a necessary and sufficient condition for $X \times A$ to be locally compact separable metric—see, e.g., Dieudonné [14, p. 75]) and that $\mathbb{K}$ is either open or closed in $X \times A$ [14, p. 66]. The main reason for requiring X and $\mathbb{K}$ as in Assumption 4.2(a) is explained in Remark 5.3.

(c) Assumption 4.2(c) on Q—as well as (4.10) below (see also the Remark following Theorem 4.5)—is related to conditions given by Benes [7] to ensure the existence of invariant distributions for Feller–Markov chains—observe that weak continuity of Q is a Feller-like condition. In our present context, Assumption 4.2(c) implies in particular that, as is easily shown,

$$\int_X u(y)Q(dy|\cdot) \quad \text{is in} \quad C_0(\mathbb{K}) \quad \text{if} \quad u \in C_0(X). \tag{4.1}$$

For a general (say, $\mathbb{R}^d$-valued) discrete-time system $x_{t+1} = F(x_t, a_t, \xi_t)$, $t = 0, 1, \ldots$, with independent and identically distributed disturbances ξ_t, Assumption 4.2(c) holds if, for every s, the function $F(x, a, s)$ is continuous in $(x, a) \in \mathbb{K}$ and, moreover, $F(x, a, s) \to \infty$ as $(x, a) \to \infty$. The general discrete-time system, in particular the additive-noise case in (d), below, includes many control models found in applications; see [4, 13, 16, 19].

(d) As an *example*, consider the additive-noise system

$$x_{t+1} = G(x_t, a_t) + \xi_t, \quad t = 0, 1, \ldots,$$

with state and control spaces $X = A = \mathbb{R}$ (so that $\mathbb{K} = \mathbb{R}^2$), where the disturbances ξ_t are independent and identically distributed random variables, independent of the initial state x_0. Moreover, the one-stage cost c is supposed to be a quadratic function, say $c(x, a) = \alpha x^2 + \beta a^2$ with positive α and β, and $G(x, a)$ is a given continuous function. In this case, Assumptions 4.1(a),(b) and 4.2(a),(b) are obviously satisfied, and conditions sufficient for Assumption 4.1(c),(d) are easily determined (see, for instance, [17]). Finally, the second part of Assumption 4.2(c) is also satisfied if, for instance, $|G(x, a)| \to \infty$ as $|(x, a)| \to \infty$. In other words, under the given conditions, Assumptions 4.1 and 4.2 both hold. Similar statements hold in the vector case.

(e) (See [3, p. 37] or [11, p. 984].) Let $(\mathcal{X}, \mathcal{Y})$ and $(\mathcal{Z}, \mathcal{W})$ be two dual pairs of vector spaces and $G : \mathcal{X} \to \mathcal{Z}$ a linear mapping with adjoint G^*. Then G is weakly continuous if and only if G^* maps $\mathcal{W}$ into $\mathcal{Y}$. For instance, by Remark 3.2(c), (3.2) implies that L_0^* maps $F(X)$ into $F(\mathbb{K})$ and, therefore, $L_0 : M(\mathbb{K}) \to M(X)$ is *weakly continuous*. Similarly, Assumption 4.1(c) (=4.2(d)) yields that, for all g in $F(X)$,

$$\left| \int g(y) Q(dy|x, a) \right| \le \|g\|_{w_0} \int w_0(y) Q(dy|x, a) \le \|g\|_{w_0} C w(x, a).$$

Hence, this assumption and (3.2) imply that L_1^* maps $F(X)$ into $F(\mathbb{K})$, so that $L_1 : M(\mathbb{K}) \to M(X)$ is *weakly continuous*. Combining these facts with the definitions of L and T (see section 3), we obtain part (a) in the following lemma.

LEMMA 4.4. (a) *Each of the operators L_0, L_1, L, and T is weakly continuous.*

(b) *L_0^* maps $C_b(X)$ into $C_b(\mathbb{K})$, and so does L_1^* if Q is weakly continuous* (*see Assumption* 4.1(b) *or* 4.2(c)).

Proof. Part (a) follows from Remark 4.3(e), and part (b) is obvious. □

Plainly, the dual programs EP^* and P^* are both consistent. For instance, for any constant k the pair $(\rho, h) := (0, k)$ and $(g, h) := (0, k)$ satisfy (3.4) and (3.6), respectively. In the remainder of this section, (i) we give necessary and sufficient conditions for the primal problems EP and P to be *consistent*; (ii) we show that for each of them there is *no duality gap*, i.e.,

$$\sup(EP^*) = \inf(EP) \qquad (= \rho^*; \quad \text{see (2.13)}), \tag{4.2}$$

and similarly for P; and, finally, (iii) we show that in each case there is *strong duality*, which means that EP and EP^* are both *solvable* and their optimal values satisfy

$$\max(EP^*) = \min(EP) \tag{4.3}$$

and similarly for P and P^*. (In writing (4.3) we follow the usual convention that "min" replaces "inf" for an attained infimum and similarly for "max" and "sup.") We shall first state the results for EP and then for P; the proofs are presented in sections 6 (for EP) and 7 (for P).

Main results for *EP*. Under Assumption 4.1, it is shown in [17] that EP is solvable and that it has no duality gap. Here, first we use Assumption 4.2 to obtain a *necessary and sufficient condition for EP to be consistent* (Theorem 4.5). To do this using a *generalized Farkas theorem* (Theorem 5.1) we would need to show that the set $L(M_+(\mathbb{K})) \subset \mathbb{R} \times M(X)$ is weakly closed. Since we are unable to prove this directly, we "perturbate" EP as follows.

Let v_0 be a strictly positive function in $C_0(X)$, and consider the linear operator $\mathcal{L} : M(\mathbb{K}) \times \mathbb{R}^2 \to M(X) \times \mathbb{R}^2$ with

$$\mathcal{L}(\mu, r_1, r_2) := (L_1\mu, \langle \mu, 1\rangle + r_1, \langle \mu, v_0\rangle - r_2). \tag{4.4}$$

The adjoint $\mathcal{L}^* : F(X) \times \mathbb{R}^2 \to F(\mathbb{K}) \times \mathbb{R}^2$ is given by

$$\mathcal{L}^*(h, \rho_1, \rho_2) = (L_1^* h + \rho_1 + \rho_2 v_0, \rho_1, -\rho_2). \tag{4.5}$$

Now note that EP *is consistent* (i.e., there is a measure μ satisfying (3.3)) *if and only if the linear equation*

$$\mathcal{L}(\mu, r_1, r_2) = (0, 1, \varepsilon) \text{ has a solution } (\mu, r_1, r_2) \text{ in } M_+(\mathbb{R}) \times \mathbb{R}^2_+ \tag{4.6}$$

for some $\varepsilon > 0$. This is obvious because if μ satisfies (3.3), then $(\mu, 0, 0)$ satisfies (4.6) with $\varepsilon := \langle \mu, v_0\rangle$; and, conversely, if (μ, r_1, r_2) satisfies (4.6), then $\langle \mu, v_0\rangle \geq \varepsilon$ implies $\langle \mu, 1\rangle > 0$ and, therefore, $\mu^* := \mu/\langle \mu, 1\rangle$ satisfies (3.3). We will show that (4.6) and the generalized Farkas theorem (see Theorem 5.1 below) yield part (a) in the following result.

THEOREM 4.5. *Suppose that Assumption* 4.2 *holds. Then*

(a) EP *is consistent (equivalently,* (4.6) *holds) if and only if*

$$L_1^* h + \rho_1 + \rho_2 v_0 \geq 0 \quad \text{with} \quad h \in F(X),\ \rho_1 \geq 0, \ \text{and}\ \rho_2 \leq 0 \tag{4.7}$$

implies

$$\rho_1 + \varepsilon\rho_2 \geq 0 \qquad \text{for some} \quad \varepsilon > 0. \tag{4.8}$$

(b) *Let* v_0 *and* w_0 *be as in* (4.4) *and* (3.1)*, respectively, and suppose there exists* $\varepsilon > 0$ *and a relaxed policy* φ *such that* $\forall x \in X$

$$\liminf_{n\to\infty} E_x^\varphi[w_0(x_n)]/n = 0 \quad \text{and} \tag{4.9}$$

$$\liminf_{n\to\infty} \sum_{t=0}^{n-1} E_x^\varphi[v_0(x_t)]/n \geq \varepsilon. \tag{4.10}$$

Then EP *is consistent.*

Remark. The hypotheses of Theorem 4.5(b) imply that the stochastic kernel $Q(\cdot|\cdot, \varphi)$ satisfies Benes's [7] condition (v) (hence, the equivalent conditions (i)–(iv)) for the existence of a nontrivial invariant measure. Indeed, suppose that $\varepsilon > 0$ is such that $\varepsilon \leq \|v_0\|$, and choose $0 < \varepsilon_0 < \varepsilon$. As $v_0(\cdot) > 0$ is in $C_0(X)$, there is a compact set K_0 such that $0 < v_0(x) \leq \varepsilon_0$ for all x not in K_0. Then (writing X as the union of K_0 and its complement) for all $t = 0, 1, \ldots$ and $x \in X$,

$$E_x^\varphi[v_0(x_t)] = \int_X v_0(y) Q^t(dy|x, \varphi) \leq (\|v_0\| - \varepsilon_0) Q^t(K_0|x, \varphi) + \varepsilon_0.$$

This implies, by (4.10),

$$\liminf_{n\to\infty} \frac{1}{n}\sum_{t=0}^{n-1} Q^t(K_0|x,\varphi) \geq (\varepsilon-\varepsilon_0)/(\|v_0\|-\varepsilon) > 0.$$

Finally, take Benes's measure μ as (for instance) the initial distribution γ_0 in (2.12) to obtain the condition (v) in [7].

In Theorem 4.6 we use the following definition. A sequence of measures μ^n in $M(\mathbb{K})$ is said to be a *minimizing sequence* for EP if each μ^n is feasible for EP (see (3.3)) and $\langle \mu^n, c\rangle \downarrow \inf(EP)$. Similarly (ρ_n, h_n) is a *maximizing sequence* for the dual EP^* if each (ρ_n, h_n) is feasible for EP^* and $\rho_n \uparrow \sup(EP^*)$. Of course, minimizing and maximizing sequences for P and P^*, respectively, are defined analogously.

THEOREM 4.6. *Suppose that either* (i) *Assumption* 4.1 *holds, or* (ii) *Assumptions* 4.2 *and* 4.1(a) *hold and EP is consistent with a finite value. Then the following hold.*

(a) *EP is solvable and there is no duality gap; i.e., there exists a feasible solution μ^* for EP such that*

$$\langle \mu^*, c\rangle = \min(EP) = \sup(EP^*) \quad (=\rho^*; \quad \textit{see } (2.13)). \tag{4.11}$$

(b) *If $\{\mu^n\}$ is a minimizing sequence for EP, then there exists a subsequence $\{j\}$ of $\{n\}$ such that μ^j converges in the weak topology $\sigma(M(\mathbb{K}), C_b(\mathbb{K}))$ (see Remark* 5.3(b)) *to an optimal solution for EP.*

(c) *If (ρ_n, h_n) is a maximizing sequence for EP^* with $\{h_n\}$ bounded in the w_0-weighted norm (i.e., there is a constant k such that $\|h_n\|_{w_0} \leq k\ \forall n$), then* (i) *$EP^*$ is solvable,* (ii) *strong duality holds* (*see* (4.3)), *and* (iii) *if μ^* is optimal for EP and $\mu_1^* = L_0\mu$ is its marginal on X, then the AC optimality equation* (2.6)–(2.7) *holds μ_1^*-almost everywhere (a.e.) with $g(\cdot) = \rho^*$; in fact, there is a function $h \in F(X)$ and a policy $f^* \in \mathbb{F}$ such that for μ_1^*-almost every $x \in X$:*

$$\begin{aligned} \rho^* + h(x) &= \min_{a\in A(x)} \left[c(x,a) + \int_X h(y) Q(dy|x,a) \right] \\ &= c(x,f^*) + \int_X h(y) Q(dy|x,f^*). \end{aligned} \tag{4.12}$$

We now turn our attention to the linear program P.

Main results for *P*. In analogy with EP, we use Theorem 5.1 to obtain a *necessary and sufficient condition for P to be consistent* (Theorem 4.9). And again, as with EP, instead of (3.5), we consider an *equivalent*, "perturbated" linear equation, (4.13), as follows.

Consider the linear mapping $\tau : M(\mathbb{K})^2 \times \mathbb{R} \to M(X)^2 \times \mathbb{R}$ defined by

$$\begin{aligned} \tau(\mu,\nu,r) &:= (L_0\mu + L_1\nu, L_1\mu, \langle \mu+\nu, w\rangle + r) \\ &= (T(\mu,\nu), \langle \mu+\nu, w\rangle + r). \end{aligned}$$

The adjoint $\tau^* : F(X)^2 \times \mathbb{R} \to F(\mathbb{K})^2 \times \mathbb{R}$ is given by

$$\begin{aligned} \tau^*(g,h,\rho) &:= (L_0^* g + L_1^* h + \rho w, L_1^* g + \rho w, \rho) \\ &= (T^*(g,h), 0) + \rho(w,w,1). \end{aligned}$$

Also consider the primal linear program, with $m \geq 1$ (see Remark 4.7),

$\boldsymbol{P}_m$: minimize $\langle \mu, c\rangle$ $[= \langle (\mu,\nu,r),\ (c,0,0)\rangle]$
(4.13) subject to $\tau(\mu,\nu,r) = (\gamma_0, 0, m)$, $(\mu,\nu,r) \in M_+(\mathbb{K})^2 \times \mathbb{R}_+$.

The corresponding dual is

$\boldsymbol{P}_m^*$: maximize $\langle \gamma_0, g\rangle + m\rho$ $[= \langle (\gamma_0, 0, m),\ (g,h,\rho)\rangle]$
subject to $\tau^*(g,h,\rho) \le (c,0,0)$, $(g,h,\rho) \in F(X)^2 \times \mathbb{R}$.

Remark 4.7. If (μ,ν,r) satisfies (4.13), then μ is a probability measure, which combined with (3.1), yields $m \ge 1$ since

$$m = \langle \mu + \nu, w\rangle + r \ge \langle \mu, w\rangle \ge \langle \mu, 1\rangle = 1.$$

The following proposition shows that P and P_m, as well as the duals P^* and P_m^*, are equivalent for all m s.l. (sufficiently large).

PROPOSITION 4.8. (a) *P is consistent if and only if P_m is consistent for m s.l.; moreover,* $\inf(P_m) \le \inf(P)$ *for all* $m \ge 1$.

(b) *If P_m is consistent ($m \ge 1$), then so is $P_{m'}$ and* $\inf(P_m) \ge \inf(P_{m'})$ $\forall m' \ge m$; *hence*

$$\inf(P) = \inf(P_m) \quad \textit{for all } m \textit{ s.l.} \tag{4.14}$$

(c) *If there is no duality gap for P_m, then there is no duality gap for P, and for all m s.l.:*

$$\inf(P) = \inf(P_m) = \sup(P_m^*) = \sup(P^*). \tag{4.15}$$

THEOREM 4.9. *Suppose that Assumption* 4.2 *holds. Then the following statements hold.*

(a) *P is consistent (equivalently, P_m is consistent for some $m \ge 1$) if and only if*

$$(g,h) \in F(X)^2,\quad L_0^*g + L_1^*h + \rho w \ge 0,\quad L_1^*g + \rho w \ge 0,\quad \textit{and}\ \ \rho \ge 0 \tag{4.16}$$

imply

$$\langle \gamma_0, g\rangle + \rho m \ge 0. \tag{4.17}$$

(b) *If P_m is consistent with a finite value, then it is solvable and there is no duality gap; hence (by* (4.15)*) for all m s.l., P is solvable and*

$$\min(P) = \min(P_m) = \sup(P_m^*) = \sup(P^*). \tag{4.18}$$

THEOREM 4.10. *Suppose that Assumptions* 4.2 *and* 4.1(d) *hold. Then*

(a) *P is solvable and there is no duality gap for it.*

(b) *Let (μ^n, ν^n) be a minimizing sequence for P and suppose that either* (i) *$\langle \mu^n + \nu^n, w\rangle$ is bounded, or* (ii) *Assumption* 4.1(a) *holds and $\langle \nu^n, w\rangle$ is bounded. Then there is subsequence $\{j\}$ of $\{n\}$ such that (μ^j, ν^j) converges to an optimal solution for P in the weak* topology $\sigma(M(\mathbb{K}), C_0(\mathbb{K}))$.*

(c) *Let (g^n, h^n) be a maximizing sequence for P^*. If the sequences g^n, h^n are bounded in the w_0-weighted norm (i.e., there is a constant k such that $\|g^n\|_{w_0}$, $\|h^n\|_{w_0} \le k$ $\forall n$), then* (i) *P^* is solvable. Hence,* (ii) *there is strong duality, i.e.,* $\min(P) = \max(P^*) = \langle \gamma_0, J^*\rangle$, *and* (iii) *an optimal solution (μ,ν) for P is such that*

- (2.6) *holds* ν_1*-a.e., and*
- (2.7) *holds* μ_1*-a.e.*

We conclude with the following interesting remark whose easy proof is left to the reader.

Remark 4.11. Under Assumptions 4.1 or 4.2, the set of optimal solutions for P is a weakly closed, convex, extremal subset of the set of feasible solutions for P—where *extremal* means that if (μ, ν) is optimal for P, if (μ^i, ν^i) is feasible $(i = 1, 2)$, and if

$$(\mu, \nu) = r(\mu^1, \nu^1) + (1 - r)(\mu^2, \nu^2) \quad \text{for} \quad 0 < r < 1,$$

then (μ^i, ν^i) is optimal for P $(i = 1, 2)$ (see [9]).

5. Technical preliminaries for the proofs. In this section we collect some facts used in the proofs of our main results.

If $(\mathcal{X}, \mathcal{Y})$ is a dual pair of vector spaces, we denote by $\sigma(\mathcal{X}, \mathcal{Y})$ the *weak topology* on $\mathcal{X}$ [3, 10, 15]. If $\mathcal{Y}$ is a Banach space and $\mathcal{X} = \mathcal{Y}^*$ the topological dual, then $\sigma(\mathcal{X}, \mathcal{Y})$ is called the *weak* topology* on $\mathcal{X}$.

Let $(\mathcal{X}, \mathcal{Y})$ be a dual pair, and S a convex cone in $\mathcal{X}$. Then the *dual cone* S^* is the set $\{y \in \mathcal{Y} | \langle x, y\rangle \geq 0 \ \forall x \in S\}$.

THEOREM 5.1 (generalized Farkas theorem [11, Theorem 2]). *Let* $(\mathcal{X}, \mathcal{Y})$ *and* $(\mathcal{Z}, \mathcal{W})$ *be two real dual pairs, let* S *be a convex cone in* $\mathcal{X}$*, and let* $G : \mathcal{X} \to \mathcal{Z}$ *be a weakly continuous linear map. If* $G(S)$ *is weakly closed, then the following are equivalent conditions on* $b \in \mathcal{Z}$ *(where* $G^* : \mathcal{W} \to \mathcal{Y}$ *denotes the* adjoint *of* G*):*

(a) *The equation* $Gx = b$ *has a solution* $x \in S$*.*

(b) $G^*w \in S^* \Rightarrow \langle b, w\rangle \geq 0$.

The following result is the *Alaoglu*, or *Banach–Alaoglu–Bourbaki theorem* (see [10, 15]), which is used in sections 6 and 7 in combination with Remark 5.3.

THEOREM 5.2. *Let* $\mathcal{X}$ *be a Banach space with topological dual* $\mathcal{X}^*$ *and let* U *be the closed unit sphere in* $\mathcal{X}^*$*. Then* U *is compact in the weak* topology* $\sigma(\mathcal{X}^*, \mathcal{X})$*. Moreover, if* $\mathcal{X}$ *is separable, then the weak* topology of* U *is metrizable.*

Remark 5.3. *(a)* Let us view $M(\mathbb{K})$ (see Definition 3.1) as the Banach space of finite signed measures on $\mathbb{K}$ endowed with the *total variation* norm $\|\cdot\|_{TV}$. Note that, by (3.1)–(3.2),

$$\|\mu\|_w = \int w d|\mu| \geq \|\mu\|_{TV},$$

so that $(M(\mathbb{K}), \|\cdot\|_w)$ is a subspace of $(M(\mathbb{K}), \|\cdot\|_{TV})$. On the other hand, with $\mathbb{K}$ as in Assumption 4.2(a), $(M(\mathbb{K}), \|\cdot\|_{TV})$ is the dual of the *separable* Banach space $C_0(\mathbb{K})$.

(b) We shall consider three topologies on $M(\mathbb{K})$: the *weak topology* $\sigma(M(\mathbb{K}), F(\mathbb{K}))$, the *weak topology* $\sigma(M(\mathbb{K}), C_b(\mathbb{K}))$, and the *weak* topology* $\sigma(M(\mathbb{K}), C_0(\mathbb{K}))$. However, as already noted in section 3, by "weak topology" we always mean $\sigma(M(\mathbb{K}), F(\mathbb{K}))$, unless explicitly stated otherwise.

(c) Under Assumption 4.2(a), if $\{\mu^j\}$ is a bounded sequence of measures on $\mathbb{K}$ that converges in the weak* topology, then also the sequence of marginals $L_0\mu^j = \mu_1^j$ on X converges in the weak* topology. That is, if

$$\text{(i)} \qquad \langle \mu^j, v\rangle \longrightarrow \langle \mu, v\rangle \qquad \forall v \in C_0(\mathbb{K}),$$

then

$$\text{(ii)} \qquad \langle \mu^j, u\rangle \longrightarrow \langle \mu, u\rangle \qquad \forall u \in C_0(X),$$

where in (ii) we have used Remark 3.2(c) to write $\langle \mu^j, u\rangle$ and $\langle \mu, u\rangle$ instead of $\langle \mu_1^j, u\rangle$ and $\langle \mu_1, u\rangle$, respectively. To prove (ii) first note that, under Assumption 4.2(a), $\mathbb{K}$ is σ-compact and, therefore, there exists an increasing sequence of compact sets $K_n \uparrow \mathbb{K}$. Moreover, by Urysohn's lemma [29, p. 39], for any given $\varepsilon > 0$ and all $n = 1, 2, \ldots$, there is a function α_n in $C_0(\mathbb{K})$ such that $0 \le \alpha_n \le 1$, with $\alpha_n = 1$ on K_n and $\alpha_n(x, a) = 0$ if the distance from (x, a) to K_n is $\ge \varepsilon$. Now, given u in $C_0(X)$, define $u_n(x, a) := u(x)\alpha_n(x, a)$ on $\mathbb{K}$. Then u_n is in $C_0(\mathbb{K})$, and for all (x, a) in $\mathbb{K}$,

$$|u_n(x, a)| \le |u(x)| \le \|u\| < \infty, \qquad \text{and} \qquad u_n(x, a) \to u(x) \qquad \text{as} \qquad n \to \infty.$$

Hence, by the bounded convergence theorem, for every fixed j,

$$\text{(iii)} \qquad \langle \mu^j, u_n\rangle \longrightarrow \langle \mu^j, u\rangle \qquad \text{and} \qquad \langle \mu, u_n\rangle \longrightarrow \langle \mu, u\rangle \qquad \text{as} \qquad n \to \infty.$$

On the other hand, for every fixed n, (i) yields

$$\text{(iv)} \qquad \langle \mu^j, u_n\rangle \longrightarrow \langle \mu, u_n\rangle \qquad \text{as} \qquad j \to \infty.$$

Finally, the desired conclusion (ii) follows from (iii), (iv), and the inequality

$$\begin{aligned} |\langle \mu^j, u\rangle - \langle \mu, u\rangle| \le\ & |\langle \mu^j, u\rangle - \langle \mu^j, u_n\rangle| \\ & + |\langle \mu^j, u_n\rangle - \langle \mu, u_n\rangle| + |\langle \mu, u_n\rangle - \langle \mu, u\rangle|. \end{aligned}$$

Linear programming. Let $(\mathcal{X}, \mathcal{Y})$, $(\mathcal{Z}, \mathcal{W})$, S, G, and b be as in Theorem 5.1, and consider the linear program, with $c \in \mathcal{Y}$,

$\boldsymbol{\mathcal{P}}$: minimize $\langle x, c\rangle$
(5.1) subject to $Gx = b$, $x \in S$.

The dual of $\mathcal{P}$ is

$\boldsymbol{\mathcal{P}}^*$: maximize $\langle b, w\rangle$
subject to $c - G^*w \in S^*, w \in \mathcal{W}$.

Weak duality. If x is feasible for $\mathcal{P}$ and w is feasible for $\mathcal{P}^*$, then $\langle b, w\rangle \le \langle x, c\rangle$; hence

$$\sup(\mathcal{P}^*) \le \inf(\mathcal{P}). \tag{5.2}$$

DEFINITION 5.4. (see [3, p. 40]). *Let H be the subset of $\mathcal{Z} \times \mathbb{R}$ defined as*

$$H := \{(Gx, \langle x, c\rangle + r) | x \in S, r \ge 0\}. \tag{5.3}$$

The program $\mathcal{P}$ is said to be subconsistent *if there is some $r \in \mathbb{R}$ with (b, r) in the weak closure* $\mathrm{cl}(H)$ *of H. If $\mathcal{P}$ is subconsistent, its* subvalue *is defined as the infimum of all $r \in \mathbb{R}$ for which (b, r) is in* $\mathrm{cl}(H)$. *(Thus, the subvalue is the infimum of r for which there is a net $\{x^\alpha\}$ in S with $Gx^\alpha \to b$ and $\langle x^\alpha, c\rangle \to r$.)*

THEOREM 5.5 (see [3, Theorem 3.3]). *$\mathcal{P}$ is subconsistent with a finite value r^* if and only if $\mathcal{P}^*$ is consistent with a finite value r^*.*

THEOREM 5.6 (see [3, Theorems 3.9, 3.22]). *If $\mathcal{P}$ is consistent with a finite value and H is weakly closed, then $\mathcal{P}$ is solvable and there is no duality gap for $\mathcal{P}$.*

6. Proof of the results for *EP*.

Proof of Theorem 4.5. (a) In this proof we use Theorem 5.1 with the following identifications (see (4.4)–(4.6))]:

$$(\mathcal{X},\mathcal{Y}) := (M(\mathbb{K})\times\mathbb{R}^2, F(\mathbb{K})\times\mathbb{R}^2),\ (\mathcal{Z},\mathcal{W}) := (M(X)\times\mathbb{R}^2, F(X)\times\mathbb{R}^2),$$

$$G := \mathcal{L},\quad S := M_+(\mathbb{K})\times\mathbb{R}^2_+,\quad b := (0,1,\varepsilon). \tag{6.1}$$

Then condition (a) in Theorem 5.1 is the same as (4.6), and condition (b) becomes

$$\mathcal{L}^*(h,\rho_1,\rho_2)\geq 0 \quad\Longrightarrow\quad \langle(0,1,\varepsilon),(h,\rho_1,\rho_2)\rangle\geq 0,$$

which is precisely that condition "(4.7) implies (4.8)." Hence, to prove Theorem 4.5(a) we only need to verify the hypotheses of Theorem 5.1; namely, we need to check that (in view of (6.1))

(i) $\mathcal{L}$ is weakly continuous, and

(ii) $\mathcal{L}(S)$ is weakly closed.

The requirement (i) follows from Lemma 4.4(a) and the definition of $\mathcal{L}$—see (4.4). Therefore, to complete the proof of Theorem 4.5(a) it only remains to prove (ii).

Proof of (ii). To prove that $\mathcal{L}(S)$ is weakly closed, let $(D,\leq)$ be a directed set, and let $\{(\mu^\alpha, r_1^\alpha, r_2^\alpha), \alpha\in D\}$ be a net in S such that $\mathcal{L}(\mu^\alpha, r_1^\alpha, r_2^\alpha)$ converges weakly to $(\nu,\rho_1,\rho_2)\in M(X)\times\mathbb{R}^2$; i.e. (by (4.4)),

$$\langle L_1\mu^\alpha, u\rangle \to \langle\nu, u\rangle \quad \forall u\in F(X), \tag{6.2}$$

$$\langle\mu^\alpha, 1\rangle + r_1^\alpha \to \rho_1, \quad \text{and} \tag{6.3}$$

$$\langle\mu^\alpha, v_0\rangle - r_2^\alpha \to \rho_2. \tag{6.4}$$

We wish to show that (ν,ρ_1,ρ_2) is in $\mathcal{L}(S)$; i.e., there exists (μ^0, r_1^0, r_2^0) in S such that

$$\text{(a)}\quad L_1\mu^0 = \nu,\ \text{(b)}\ \langle\mu^0,1\rangle + r_1^0 = \rho_1,\ \text{and (c)}\ \langle\mu^0, v_0\rangle - r_2^0 = \rho_2. \tag{6.5}$$

If $\rho_1 = 0$ in (6.3), then r_1^α and $\langle\mu^\alpha,1\rangle\to 0$, and it is easily verified that necessarily $\nu = 0$ and that (6.4) holds with $(\mu^0, r_1^0, r_2^0) = (0,0,-\rho_2)$. We shall now consider the case $\rho_1 > 0$.

Suppose that $\rho_1 > 0$. Then, by (6.3), there exists $\alpha_0\in D$ such that

$$0\leq\langle\mu^\alpha,1\rangle = \mu^\alpha(\mathbb{K})\leq 2\rho_1 \quad\forall\alpha\geq\alpha_0. \tag{6.6}$$

Hence, by Theorem 5.2 and Remark 5.3(a), there exists a measure μ^0 on $\mathbb{K}$ and a sequence $\{j\}$ in D such that $\mu^j\to\mu^0$ in the weak* topology $\sigma(M(\mathbb{K}), C_0(\mathbb{K}))$; i.e.,

$$\langle\mu^j, v\rangle\to\langle\mu^0, v\rangle \quad\forall v\in C_0(\mathbb{K}). \tag{6.7}$$

Moreover, μ^0 is in $M(\mathbb{K})$, since $0\leq\mu^0(\mathbb{K})\leq\liminf_j\mu^j(\mathbb{K})\leq 2\rho_1$ (see, e.g., [10, Proposition III.12]). Now, let u be an arbitrary function in $C_0(x)$. Then (6.7) and (4.1) yield, as $j\to\infty$,

$$\left\langle\mu^j, \int_X u(y)Q(dy|\cdot)\right\rangle \longrightarrow \left\langle\mu^0, \int_X u(y)Q(dy|\cdot)\right\rangle,$$

whereas (6.7) and Remark 5.3(c) yield

$$\langle \mu^j, u\rangle \longrightarrow \langle \mu^0, u\rangle.$$

Therefore, as $u \in C_0(X)$ was arbitrary, $L_1\mu^j$ converges to $L_1\mu^0$ in the weak* topology; i.e.,

$$\langle L_1\mu^j, u\rangle \longrightarrow \langle L_1\mu^0, u\rangle \qquad \forall u \in C_0(X). \tag{6.8}$$

By (6.2) and (6.8), we obtain $L_1\mu^0 = \nu$, which is condition (6.5)(a), and, finally, (b) and (c) in (6.5) hold with $r_1^0 := \rho_1 - \langle \mu^0, 1\rangle$ and $r_2^0 := \rho_2 - \langle \mu^0, v_0\rangle$. This proves (ii), which, as already noted, concludes the proof of Theorem 4.5(a).

(b) In view of part (a), to prove (b) it suffices to show that (4.7), together with (4.9) and (4.10), implies (4.8). So, let φ be as in (4.9) and let us rewrite (4.7) as

$$h(x) \geq \int h(y)Q(dy|x,a) - \rho_1 - \rho_2 v_0(x),\ h \in F(X),\ \rho_1 \geq 0,\ \rho_2 \leq 0. \tag{6.9}$$

Note that (4.9) obviously implies

$$\liminf_n E_x^\varphi[h(x_n)]/n = 0 \quad \forall h \in F(X),\ x \in X. \tag{6.10}$$

On the other hand, since (6.9) holds for all (x,a) in $\mathbb{K}$, integration with respect to $\varphi(\cdot|x)$ yields

$$h(x) \geq \int h(y)Q(dy|x,\varphi) - \rho_1 - \rho_2 v_0(x) \quad \forall x \in X,$$

which in turn (by iteration) implies, $\forall x \in X$ and $n = 1, 2, \ldots$,

$$h(x) \geq E_x^\varphi h(x_n) - n\rho_1 - \rho_2 \sum_{t=0}^{n-1} E_x^\varphi v_0(x_t);$$

i.e.,

$$h(x) + n\rho_1 + \rho_2 \sum_{t=0}^{n-1} E_x^\varphi v_0(x_t) \geq E_x^\varphi h(x_n).$$

Thus, multiplying by $1/n$ and taking lim inf as $n \to \infty$, (6.10) and (4.10) yield (4.8). Therefore, by part (a), EP is consistent. □

In the proof of Theorem 4.6 we shall use the following lemma in which (a) is a well-known result on disintegration of measures [32], and (b) is a result of Blackwell [8]—see also, e.g., [16, pp. 88, 97] or [19, Proposition D.8].

LEMMA 6.1. (a) *If μ is a finite measure on $\mathbb{K}$, then there exists a randomized stationary policy φ such that $\mu(d(x,a)) = \varphi(da|x)\mu_1(dx)$, where $\mu_1 = L_0\mu$ is the marginal of μ on X; i.e.,*

$$\mu(B \times C) = \int_B \varphi(C|x)\mu_1(dx) \quad \forall B \in \mathcal{B}(X),\ C \in \mathcal{B}(A). \tag{6.11}$$

(b) *If φ is a randomized stationary policy and $v : \mathbb{K} \to \mathbb{R}$ is a measurable function such that $v(\cdot,\varphi)$ (recall Remark 2.2) is a finite-valued function on X, then there exists a stationary policy $f \in \mathbb{F}$ satisfying*

$$v(x,\varphi) \geq v(x,f) \quad \forall x \in X.$$

Also note that if μ is feasible for EP and (ρ, h) is feasible for EP^*, then

$$\langle L_1\mu, h\rangle = 0 \quad \textit{and} \quad \langle L\mu, (\rho, h)\rangle = \rho. \tag{6.12}$$

Proof of Theorem 4.6. (a) This part is proved in [17] under condition (i), namely, Assumption 4.1, and in fact the proof works in exactly the same way under condition (ii).

(b) Let $\{\mu^n\}$ be a minimizing sequence for EP; that is, μ^n is feasible for EP, which (by (3.1)) means that

$$\text{(i)} \quad \langle \mu^n, 1\rangle = 1, \quad \text{and} \quad \text{(ii)} \quad L_1\mu^n = 0 \qquad \forall n, \tag{6.13}$$

and

$$\langle \mu^n, c\rangle \;\downarrow\; \min(EP). \tag{6.14}$$

In particular, (6.14) implies that for any given $\varepsilon > 0$ there exists $n(\varepsilon)$ such that

$$\min(EP) \le \langle \mu^n, c\rangle \le \min(EP) + \varepsilon \quad \forall n \ge n(\varepsilon).$$

The right-hand side inequality and Assumption 4.1(a) imply (see, for instance, [5, 6, 7, 17, 19]) the existence of a probability measure μ^* on $\mathbb{K}$ and a subsequence $\{j\}$ of $\{n\}$ such that μ^j converges to μ^* in the weak topology $\sigma(M(\mathbb{K}), C_b(\mathbb{K}))$; i.e.,

$$\langle \mu^j, v\rangle \to \langle \mu^*, v\rangle \quad \forall v \in C_b(\mathbb{K}). \tag{6.15}$$

Hence, as c is l.s.c.,

$$\langle \mu^*, c\rangle \le \liminf_j \langle \mu^j, c\rangle \le \min(EP) + \varepsilon.$$

As $\varepsilon > 0$ was arbitrary, the latter inequality shows that μ^* satisfies (4.11). Thus, to complete the proof of part (b) it only remains to show that μ^* is indeed feasible for EP (see (3.3)). But this, however, follows directly from (6.15) and (6.13)—in particular, note that Assumption 4.1(b), which is part of 4.2(c), implies that L_1^* maps $C_b(X)$ into $C_b(\mathbb{K})$ (see Lemma 4.4(b)) and, therefore, $\forall u \in C_b(X)$,

$$\langle L_1\mu^*, u\rangle = \langle \mu^*, L_1^*u\rangle = \lim_j \langle \mu^j, L_1^*u\rangle = \lim_j \langle L_1\mu^j, u\rangle = 0;$$

that is, $L_1\mu^* = 0$.

(c) Let (ρ_n, h_n) be a maximizing sequence for EP^*, i.e., (ρ_n, h_n) is feasible for EP^*, so that (by (3.4)) $\forall\ n = 1, 2, \ldots, (x, a) \in \mathbb{K}$,

$$\rho_n + h_n(x) \le c(x, a) + \int_X h_n(y)Q(dy|x, a), \tag{6.16}$$

and

$$\rho_n = \langle (1, 0), (\rho_n, h_n)\rangle \uparrow \sup(EP^*) = \rho^* \ \text{(by (4.11))}.$$

Define $h(x) := \limsup_n h_n(x)$, $x \in X$. Since, by hypotheses, $\|h_n\|_{w_0}$ is bounded, h is in $F(X)$ and, moreover, applying Fatou's lemma in (6.16) we see that

$$\rho^* + h(x) \le c(x, a) + \int h(y)Q(dy|x, a) \quad \forall (x, a) \in \mathbb{K}; \tag{6.17}$$

that is, (ρ^*, h) is feasible—hence *optimal* (as $\rho^* = \sup(EP^*)$)—for EP^*. This fact, combined with (4.11), yields *strong duality*: there exists an optimal solution μ^* for EP, and optimal solution (ρ^*, h) for EP^*, and $\langle \mu^*, c\rangle = \rho^*$. On the other hand, by (6.12),

$$\langle \mu^*, L^*(\rho^*, h)\rangle = \langle L\mu^*, (\rho^*, h)\rangle = \rho^*.$$

This yields

$$\langle \mu^*, c - L^*(\rho^*, h)\rangle = 0,$$

or, equivalently, writing $\mu^*(d(x,a)) = \varphi^*(da|x)\mu_1^*(dx)$ as in Lemma 6.1(a),

$$\begin{aligned} 0 &= \int \left[c(x,a) - L^*(\rho^*, h)(x,a)\right] \mu^*(d(x,a)) \\ &= \int \left[c(x,\varphi^*) - \rho^* - h(x) + \int h(y)Q(dy|x,\varphi^*)\right] \mu_1^*(dx). \end{aligned}$$

This equation and (6.17) yield, for μ_1^*-a.e. $x \in X$:

$$\begin{aligned} \rho^* + h(x) &= \min_{a\in A(x)} \left[c(x,a) + \int_X h(y)Q(dy|x,a)\right] \\ &= c(x,\varphi^*) + \int_X h(y)Q(dy|x,\varphi^*) \quad \text{(see (2.6)–(2.9))}. \end{aligned}$$

Finally, by Lemma 6.1(b), there is a stationary policy f^* such that

$$\rho^* + h(x) \geq c(x,f^*) + \int_X h(y)Q(dy|x,f^*)$$

for μ_1^*-a.e. $x \in X$, which combined with (6.17) yields (4.12). This completes the proof of Theorem 4.6. □

7. Proof of the results for *P*.

Proof of Proposition 4.8. (a) If (μ,ν) satisfies (3.5), then, in particular, $\langle \mu + \nu, w\rangle \leq m$ for some $1 \leq m < \infty$; hence (μ,ν,r) with $r := m - \langle \mu+\nu, w\rangle$ satisfies (4.13). Conversely, if (μ,ν,r) satisfies (4.13) for *any* $m \geq 1$, then obviously (μ,ν) satisfies (3.5). Moreover, the latter fact implies $\inf(P_m) \leq \inf(P)$ for all $m \geq 1$.

(b) If (μ,ν,r) is feasible for P_m and $m' \geq m$, then (μ,ν,r') with $r' := r + m' - m$ is feasible for $P_{m'}$, and, therefore, $\inf(P_m) \geq \inf(P_{m'})$. On the other hand, by part (a), if P is consistent then P_m is consistent for some $m \geq 1$ and, therefore, $P_{m'}$ is consistent for all $m' \geq m$. Combining this fact with part (a), there is some m s.l. such that

$$\inf(P) \geq \inf(P_m) \geq \inf(P_{m'}) \geq \inf(P) \quad \forall m' \geq m,$$

which yields (4.14).

(c) If (g,h) is feasible for P^* (i.e., (3.6) holds true), then (g,h,ρ) is feasible for P_m^* for all $\rho \leq 0$, and $\langle \gamma_0, g\rangle \geq \langle \gamma_0, g\rangle + m\rho$ $\forall m \geq 1$ and $\rho \leq 0$. Therefore,

$$\sup(P^*) \geq \sup(P_m^*). \tag{7.1}$$

Now suppose that there is no duality gap for P_m so that, by (4.14),

$$\inf(P) = \inf(P_m) = \sup(P_m^*). \tag{7.2}$$

If $\inf(P_m) = +\infty$, then (4.15) follows (by (7.1)). Now suppose that $\inf(P_m) < \infty$. Then if $\sup(P^*) > \sup(P_m^*)$, (7.1) and (7.2) yield $\sup(P^*) > \inf(P)$, which contradicts the weak duality property (see (5.2)). Hence $\sup(P^*) = \sup(P_m^*)$ and we obtain (4.15). □

Proof of Theorem 4.9. (a) The proof of this part is very similar to the proof of Theorem 4.5(a), so we only mention the main steps and leave the details to the reader.

As in the proof of Theorem 4.5(a), the idea is to use the *generalized Farkas theorem*, Theorem 5.1, for which we make the following identifications:

$$(\mathcal{X}, \mathcal{Y}) := (M(\mathbb{K})^2 \times \mathbb{R}, F(\mathbb{K})^2 \times \mathbb{R}),\ (\mathcal{Z}, \mathcal{W}) := (M(X)^2 \times \mathbb{R}, F(X)^2 \times \mathbb{R}),$$

$$S := M_+(\mathbb{K})^2 \times \mathbb{R}_+, \quad b := (\gamma_0, 0, m), \text{ and } G := \tau. \tag{7.3}$$

With this notation, part (a) in Theorem 5.1 turns out to be the same as (4.13), and part (b) becomes

$$\tau^*(g, h, \rho) \geq 0 \quad \Longrightarrow \quad \langle (\gamma_0, 0, m), (g, h, \rho) \rangle \geq 0,$$

which is the same as "(4.16) $\Longrightarrow$ (4.17)." Hence, Theorem 4.9(a) will be proved if we can show that (i) τ is weakly continuous, and (ii) $\tau(S)$ is weakly closed. Since (i) follows from Lemma 4.4(a), we only need to prove (ii). To do this, let $\{(\mu^\alpha, \nu^\alpha, r^\alpha), \alpha \in D\}$ be a net is S such that (recalling the notation (7.3))

$$\tau(\mu^\alpha, \nu^\alpha, r^\alpha) \to (\theta_1, \theta_2, \rho) \in \mathcal{Z} \text{ in the weak topology } \sigma(\mathcal{Z}, \mathcal{W}), \tag{7.4}$$

so we need to show that $(\theta_1, \theta_2, \rho)$ is in $\tau(S)$; in other words, there is (μ^0, ν^0, r^0) in S with

$$\tau(\mu^0, \nu^0, r^0) = (\theta_1, \theta_2, \rho). \tag{7.5}$$

This is done exactly, mutatis mutandis, as in (6.2)–(6.8).

(b) By Theorem 5.6 (see also (5.3)), we only need to show that (using notation (7.3)) the set

$$H := \{(\tau(\mu, \nu, r), \langle \mu, c \rangle + s) | (\mu, \nu, r) \in S,\ s \geq 0\}$$

is weakly closed in $\mathcal{Z} \times \mathbb{R} := M(\mathbb{K})^2 \times \mathbb{R}^2$; that is, if $(\mu^\alpha, \nu^\alpha, r^\alpha, s^\alpha)$ is a net in $S \times \mathbb{R}$ for which we have the weak convergence

$$(\tau(\mu^\alpha, \nu^\alpha, r^\alpha), \langle \mu^\alpha, c \rangle + s^\alpha) \to (\theta_1, \theta_2, \rho_1, \rho_2) \in \mathcal{Z} \times \mathbb{R},$$

then $(\theta_1, \theta_2, \rho_1, \rho_2)$ is in H. The latter means that

$$\tau(\mu^0, \nu^0, r^0) = (\theta_1, \theta_2, \rho_1) \text{ and } \langle \mu^0, c \rangle + s^0 = \rho_2$$

for some (μ^0, ν^0, r^0, s^0) in $S \times \mathbb{R}$. The first equality is obtained exactly as in (7.4)–(7.5), and the second holds with $s^0 := \rho_2 - \langle \mu^0, c \rangle$. □

In the proof of Theorem 4.10 we use the following lemma.

LEMMA 7.1. *If (g, h) is feasible for P^*, then $g(x) \leq J(\pi, x)$ for every initial state x and every policy π; hence $g(x) \leq J^*(x)$ $\forall x \in X$. Therefore, under Assumption* 4.1(d)*,*

$$\langle \gamma_0, g \rangle \leq \langle \gamma_0, J^* \rangle < \infty, \tag{7.6}$$

and so P^* is consistent with a finite value

$$\sup(P^*) \le \langle \gamma_0, J^* \rangle. \tag{7.7}$$

Proof. Standard calculations [4, 17, 18, 19] show that the second inequality in (3.6) (or (3.8)) implies

$$E_x^\pi g(x_n) \ge g(x),$$

and, similarly, the first inequality in (3.6) (or (3.8)) yields

$$\begin{aligned} J_n(\pi, x) + E_x^\pi h(x_n) &\ge h(x) + \sum_{t=0}^{n-1} E_x^\pi g(x_t) \\ &\ge h(x) + ng(x). \end{aligned}$$

Finally, since (2.10) holds for every policy π as in Assumption 4.1(d), we obtain $J(\pi, x) \ge g(x)$, which in turn yields the desired conclusion. □

Proof of Theorem 4.10. (a) Since P^* is consistent with a finite value $\sup(P^*)$ (Lemma 7.1), then (by (7.1)) so is P_m^* for every $m \ge 1$, with finite value

$$\sup(P_m^*) \le \sup(P^*) \le \langle \gamma_0, J^* \rangle. \tag{7.8}$$

Therefore (Theorem 5.5), P_m is subconsistent with subvalue $v = \sup(P_m^*)$. In other words (by Definition 5.4), there is a net $(\mu^\alpha, \nu^\alpha, r^\alpha)$ in S (see (7.3)) such that

$$\tau(\mu^\alpha, \nu^\alpha, r^\alpha) \to (\gamma_0, 0, m) \text{ weakly}, \tag{7.9}$$

and

$$\langle (\mu^\alpha, \nu^\alpha, r^\alpha), (c, 0, 0) \rangle = \langle \mu^\alpha, c \rangle \to v. \tag{7.10}$$

Moreover, (7.9) implies (see (7.4)–(7.5)) the existence of $(\mu^0, \nu^0, r^0) \in S$ such that

$$\tau(\mu^0, \nu^0, r^0) = (\gamma_0, 0, m),$$

whereas (7.10) yields $\langle \mu^0, c \rangle \le v$. This implies that (μ^0, ν^0) is feasible for P and

$$\begin{aligned} \langle \mu^0, c \rangle \le v = \sup(P_m^*) &\le \inf(P_m) \qquad \text{(by (5.2))} \\ = \inf(P) &\le \langle \mu^0, c \rangle \qquad \text{(by (4.14))}. \end{aligned} \tag{7.11}$$

Therefore, since equality holds throughout (7.11), (μ^0, ν^0) is an optimal solution for (P) and there is no duality gap for P_m, hence for P, by Proposition 4.8(c).

(b) Let (μ^n, ν^n) be a minimizing sequence for P and suppose that (i) holds; that is, $\langle \mu^n + \nu^n, w \rangle$ is bounded. Then, for every n, (μ^n, ν^n) satisfies (3.5), so that

$$L_0\mu^n + L_1\nu^n = \gamma_0,\ L_1\mu^n = 0, \quad \mu^n \text{ and } \nu^n \in M_+(\mathbb{K}), \tag{7.12}$$

and

$$\langle \mu^n, c \rangle = \langle (\mu^n, \nu^n), (c, 0) \rangle \downarrow \inf(P), \tag{7.13}$$

and

$$\langle \mu^n, 1 \rangle + \langle \nu^n, 1 \rangle \le \langle \mu^n + \nu^n, w \rangle \le k \tag{7.14}$$

for some constant k. By (7.14), Remark 5.3(a), and Theorem 5.2, there exist measures μ^*, ν^*, and a subsequence $\{j\}$ of $\{n\}$ such that

$$\mu^j \to \mu^* \text{ and } \nu^j \to \nu^* \text{ in the weak}^* \text{ topology } \sigma(M(\mathbb{K}), C_0(\mathbb{K})). \tag{7.15}$$

Therefore (as in (6.7)–(6.8)), (7.15), (7.14), and (7.12) yield that (μ^*, ν^*) is feasible for P, i.e.,

$$L_0\mu^* + L_1\nu^* = \gamma_0,\ L_1\mu^* = 0, \quad \mu^*, \nu^* \in M_+(\mathbb{K}),$$

whereas (7.13) and the fact that c is l.s.c. yield

$$\inf(P) = \lim_j \langle \mu^j, c\rangle \geq \langle \mu^*, c\rangle.$$

This completes the proof of (b), under condition (i), as $\langle \mu^*, c\rangle \geq \inf(P)$.

Let us now suppose (ii). Then (7.13) and Assumption 4.1(a) imply (as in (6.14)–(6.15)) the existence of a subsequence $\{i\}$ of $\{n\}$ and a probability measure μ^* in $M_+(\mathbb{K})$ such that $\mu^i \to \mu^*$ in the weak topology $\sigma(M(\mathbb{K}), C_b(\mathbb{K}))$—hence in the weak* topology $\sigma(M(\mathbb{K}), C_0(\mathbb{K}))$—and $\langle \mu^*, c\rangle = \inf(P)$. On the other hand, as $\langle \nu^i, w\rangle$ is bounded, there is a subsequence $\{j\}$ of $\{i\}$ and a measure ν^* such that (μ^j, ν^j) satisfies (7.15). Now (b) is concluded as in the previous paragraph.

(c) Let us first note the following obvious fact: *If* (g_1, h) *and* (g_2, h) *are two feasible solutions for* P^* (see (3.6) or (3.8)) *and* $g := \max(g_1, g_2)$, *then* (g, h) *is also feasible for* P.

Now let (g^n, h^n) be a maximizing sequence for P^*, with g^n and h^n bounded in the w_0-weighted norm; that is, for every $(x, a) \in \mathbb{K}$ and $n = 1, 2, \ldots$

$$g^n(x) + h^n(x) \leq c(x,a) + \int h^n(y) Q(dy|x,a), \tag{7.16}$$

$$g^n(x) \leq \int g^n(y) Q(dy|x,a), \tag{7.17}$$

$$\langle \gamma_0, g^n\rangle = \int g^n(x)\gamma_0(dx) \uparrow \sup(P^*), \tag{7.18}$$

and, for some $k \geq 0$,

$$|g^n(x)|,\ |h^n(x)| \leq k w_0(x). \tag{7.19}$$

By the remark in the previous paragraph (and recalling that $(0, N)$ is feasible for P^* for any constant N) we may assume that $\{g^n\}$ is a *nonnegative increasing* sequence in $F(X)$. Finally, define

$$g(\cdot) := \lim_n g^n(\cdot) \text{ and } h(\cdot) := \limsup_n h^n(\cdot), \tag{7.20}$$

and let $n \to \infty$ in (7.16) and (7.17). Then, by Fatou's lemma (which is indeed applicable, by (7.19) and Assumption 4.1(d)) and the monotone convergence theorem we obtain that (g, h) is feasible for P^* and $\langle \gamma_0, g\rangle = \sup(P^*)$; that is, (g, h) is an optimal solution for P^*. This proves (i), which combined with part (a) yields (ii).

To prove (iii) let (μ, ν) be an optimal solution for P and (g, h) an optimal solution for P^*. Then by strong duality

$$\langle (\mu, \nu), (c, 0)\rangle = \langle \mu, c\rangle = \langle \gamma_0, g\rangle.$$

On the other hand (for any (μ, ν) and (g, h) that are feasible for P and P^*, respectively),

$$\langle (\mu, \nu), T^*(g, h) \rangle = \langle T(\mu, \nu), (g, h) \rangle = \langle \gamma_0, g \rangle.$$

Hence

$$\langle (\mu, \nu), (c, 0) - T^*(g, h) \rangle = 0,$$

which is equivalent to

$$\int_{\mathbb{K}} (c - L_0^* g - L_1^* h) d\mu = \int_{\mathbb{K}} (L_1^* g) d\nu.$$

As the left-hand side is ≥ 0 and the right-hand side is ≤ 0 (see (3.6) or (3.8)), we obtain that each side equals 0; i.e.,

$$\int_{\mathbb{K}} \left[c(x, a) + \int_X h(y) Q(dy|x, a) - g(x) - h(x) \right] \mu(d(x, a)) = 0,$$

and

$$\int_{\mathbb{K}} \left[g(x) - \int_X g(y) Q(dy|x, a) \right] \nu(d(x, a)) = 0$$

Thus, disintegrating μ and ν as in Lemma 6.1(a) (see the proof of Theorem 4.6(c)), we conclude part (iii). □

8. Closing remarks. In the previous sections we have presented an LP approach to studying "unichain" and "multichain" AC problems for MCPs on uncountable spaces. Our results include necessary and sufficient conditions for the related linear programs to be consistent (Theorems 4.5 and 4.9) and conditions for solvability and strong duality (Theorems 4.6 and 4.10). There are, on the other hand, many questions that come to mind: (i) Why *LP* (why not one of the more standard approaches, such as dynamic programming or the "vanishing discount" approach)? (ii) Why do the assumptions look so "restrictive"? (iii) Can we actually compute a solution of the primal and/or the dual programs?

Concerning question (i) (and (iii)), at least for finite-state, finite-action MCPs, LP is the most widely used technique, and even in the countable-state case it has been proved to be very useful to study *constrained* and *adaptive* MCPs; see, for instance, [1] for an extensive list of related references. Hence, in the first place, it seems natural to try to extend the LP techniques to MCPs on more general spaces. But more importantly, *in the multichain uncountable-space case LP seems to be* (to the best of our knowledge) *the only "reasonable" approach*. Namely, all the other techniques (dynamic programming, "vanishing discount," etc.) assume, explicitly or implicitly, the *unichain* setting (see [4, 16, 19]); and even under very strong probabilistic-like hypotheses, *multichain* results are *not* ensured (see, for instance, [24]). These remarks are also related to question (ii): a comparison of our assumptions with those required by the other techniques would hardly yield that ours are more "restrictive." As to the nature of these assumptions, it should be noted that basically they are designed to obtain the "weak closedness" that makes applicable background results such as Theorems 5.1 and 5.6.

Finally, we have the computational question (iii), which is in fact a *common difficulty to all the known solution methods*. An approach we are currently investigating is based on the well-known fact that *in a Polish space S the class of probability measures with finite support is weakly dense in $\mathcal{P}(S)$*, that is, in the weak topology $\sigma(M(S), C_b(S))$. Now assume that X and A are both Polish spaces, and let D_X and D_A be countable dense sets in X and A, respectively. Next, consider MCPs of the form (X_n, A_n, Q_n, c), where X_n and A_n are suitably chosen *finite* subsets of D_X and D_A, respectively. And, finally, the idea is to exploit the denseness of D_X and D_A to show that, as $n \to \infty$, the values of the *finite* linear programs corresponding to (X_n, A_n, Q_n, c) converge (or at least give a good approximation) to the original linear programs EP or P in section 3. In view of Theorems 4.6(b), (c) and 4.10(b), (c), we expect to obtain that the finite programs will yield *monotone approximation schemes*, decreasing for EP and P, and increasing for the duals EP^* and P^*. More generally, from a computational viewpoint, it would be important to investigate whether some of the available approximation schemes for LP (see, for instance, [1, 27, 28]) can be extended to uncountable-state MCPs.

Acknowledgment. We would like to thank one of the anonymous reviewers for calling our attention to an embarrassing mistake in a previous statement of Lemma 4.4.

Note Added in Proof. The LP approximations mentioned in the last paragraph of section 8 have been recently developed in [33].

REFERENCES

[1] E. Altman, *Constrained Markov Decision Processes*, Rapport RR–2574, INRIA, Centre Sophia-Antipolis, 1995.
[2] E. Altman and F. Spieksma, *The linear program approach in multi-chain Markov decision processes revisited*, Z. Oper. Res., 42 (1995), pp. 169–188.
[3] E. J. Anderson and P. Nash, *Linear Programming in Infinite-Dimensional Spaces*, Wiley, Chichester, 1987.
[4] A. Arapostathis, V. S. Borkar, E. Fernández-Gaucherand, M. K. Ghosh, and S. I. Marcus, *Discrete-time controlled Markov processes with average cost criterion: A survey*, SIAM J. Control Optim., 31 (1993), pp. 282–344.
[5] E. Balder, *On equivalence of strong and weak convergence in L_1-spaces under extreme point conditions*, Israel J. Math., 75 (1991), pp. 21–47.
[6] E. Balder, *Lectures on Young Measures*, Cahiers de Mathématiques de la Décision, CEREMADE, Université Paris IX-Dauphine, 1995.
[7] V. E. Benes, *Finite regular invariant measures for Feller processes*, J. Appl. Probab., 5 (1967), pp. 203–209.
[8] D. Blackwell, *Memoryless strategies in finite-stage dynamic programming*, Ann. Math. Statist., 35 (1964), pp. 863–865.
[9] J. M. Borwein, *On the existence of Pareto efficient points*, Math. Oper. Res., 8 (1983), pp. 64–73.
[10] H. Brezis, *Analyse Fonctionnelle: Théorie et Applications*, 4^e tirage, Masson, Paris, 1993.
[11] B. D. Craven and J. J. Koliha, *Generalizations of Farkas' theorem*, SIAM J. Math Anal. Appl., 8 (1977), pp. 983–997.
[12] E. V. Denardo and B. L. Fox, *Multichain Markov renewal programs*, SIAM J. Appl. Math., 16 (1968), pp. 468–487.
[13] G. Deodhare and M. Vidyasagar, *Infinite control synthesis and infinite linear programming*, in Mathematical Control Theory, M. C. Joshi and A. V. Balakrishnan, eds., Lecture Notes in Pure and Appl. Math. 142, Marcel Dekker, New York, 1993, pp. 87–107.
[14] J. Dieudonné, *Foundations of Modern Analysis*, Academic Press, New York, 1969.
[15] N. Dunford and J. T. Schwartz, *Linear Operators, Part* I: *General Theory*, Interscience-Wiley, New York, 1957.

[16] E. B. Dynkin and A. A. Yushkevich, *Controlled Markov Processes*, Springer-Verlag, Berlin, 1979.
[17] O. Hernández-Lerma and J. B. Lasserre, *Linear programming and average optimality of Markov control processes on Borel spaces—unbounded costs*, SIAM J. Control Optim., 32 (1994), pp. 480–500.
[18] O. Hernández-Lerma and J. B. Lasserre, *Average optimality in Markov control processes via discounted cost problems and linear programming*, SIAM J. Control Optim., 34 (1996), pp. 295–310.
[19] O. Hernández-Lerma and J. B. Lasserre, *Discrete-Time Markov Control Processes: Basic Optimality Criteria*, Springer, New York, 1996.
[20] A. Hordijk and L. C. M. Kallenberg, *Linear programming and Markov decision chains*, Manage. Sci., 25 (1979), pp. 352–362.
[21] A. Hordijk and J. B. Lasserre, *Linear programming formulation of MDPs in countable state space: The multichain case*, Z. Oper. Res., 40 (1994), pp. 91–108.
[22] A. Hordijk and F. Spieksma, *A new formula for the deviation matrix*, in Probability, Statistics and Optimization, F. P. Kelly, ed., Wiley, New York, 1994, pp. 497–507.
[23] L. C. M. Kallenberg, *Linear Programming and Finite Markovian Control Problems*, Mathematical Centre Tracts 148, Mathematisch Centrum, Amsterdam, 1983.
[24] M. Kurano, *The existence of a minimum pair of state and policy for Markov decision processes under the hypothesis of Doeblin*, SIAM J. Control Optim., 27 (1989), pp. 296–307.
[25] J. B. Lasserre, *Average optimal stationary policies and linear programming in countable space MDPs*, J. Math. Anal. Appl., 183 (1994), pp. 233–249.
[26] U. Rieder, *Measurable selection theorems for optimization problems*, Manuscripta Math., 24 (1978), pp. 115–131.
[27] J. Rubio, *Control and Optimization: The Linear Treatment of Nonlinear Problems*, Manchester University Press, Manchester, and John Wiley, New York and London, 1986.
[28] J. Rubio, *The global control of nonlinear diffusion equations*, SIAM J. Control Optim., 33 (1995), pp. 308–322.
[29] W. Rudin, *Real and Complex Analysis*, 3rd ed., McGraw-Hill, New York, 1986.
[30] M. Schäl, *On the second optimality equation for semi-Markov decision models*, Math. Oper. Res., 17 (1992), pp. 470–486.
[31] M. Taksar, *Infinite dimensional linear programming approach to singular stochastic control*, SIAM J. Control Optim., 35 (1997), pp. 604–625.
[32] M. Valadier, *Désintégration d'une mesure sur un produit*, C.R. Acad. Sci. Paris Sér. A, 276 (1973), pp. 33–35.
[33] O. Hernández-Lerma and J. B. Lasserre, *Linear programming approximations for Markov control processes in metric spaces*, Acta Appl. Math., to appear.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 1, pp. 336–378, January 1998

018

SPECTRAL APPROACH TO DUALITY IN NONCONVEX GLOBAL OPTIMIZATION*

ALEXEY S. MATVEEV†

Abstract. A nonconvex problem of constrained optimization is analyzed in terms of its ordinary Lagrangian function. New sufficient conditions are obtained for the duality gap to vanish. Among them, the main condition is that the objective and constraint functions be the sums of convex functionals and nonconvex quadratic forms with certain specific spectral properties. The proofs are related to extensions of the classic Toeplitz–Hausdorff theorem, which states that a continuous quadratic mapping $(y_1, y_2) = [\mathcal{B}_1(z), \mathcal{B}_2(z)]$ from a complex Hilbert space $H = \{z\}$ into $\mathbb{R}^2 = \{(y_1, y_2)\}$ transforms the unit sphere $|z| = 1$ into a convex set. The extensions deal with a quadratic mapping $[\mathcal{B}_1(z), \dots, \mathcal{B}_k(z)]$ from a real Hilbert space into $\mathbb{R}^k$ with k being arbitrary. Applications to linear-quadratic optimal control theory are considered.

Key words. global optimization, nonconvex problems, method of duality

AMS subject classifications. 49N15, 49K27, 49N10, 47A12

PII. S0363012995277731

1. Introduction and preliminaries.

In this paper we consider problems of global optimization whose abstract presentation is as follows:

$$\mathcal{F}(z) \to \inf \quad \text{subject to} \quad z \in Z \subset H, \quad \mathcal{G}(z) \leq 0. \tag{1.1}$$

Here Z is an affine subspace of a real linear space H, the functions $\mathcal{F} : H \to \mathbb{R}$ and $\mathcal{G} : H \to Y$ are given, and Y is a finite-dimensional real linear space. This space is assumed to be ordered with a convex cone $K_+ \subset Y = \{y\}, K_+ \ni 0$. Accordingly, the inequalities $y_1 \leq y_2$ and $y_2 \geq y_1$ signify the inclusion $y_2 - y_1 \in K_+$. The interior $\mathrm{int}K_+$ of the cone K_+ is supposed to be nonempty. An example of the constraint $\mathcal{G}(z) \leq 0$ under consideration is the system of scalar inequalities

$$\mathcal{G}_1(z) \leq 0, \dots, \mathcal{G}_k(z) \leq 0. \tag{1.2}$$

In this case, $Y = \mathbb{R}^k, \mathcal{G}(z) = \|\mathcal{G}_i(z)\|$, and $K_+ = \left\{y = \|y_i\| \in \mathbb{R}^k : y_1 \geq 0, \dots, y_k \geq 0\right\}$.

In what follows, our interest will be focused on the special case when, in (1.1),

$$\mathcal{F}(z) = B_{\mathcal{F}}(z, z) + \Phi_{\mathcal{F}}(z), \quad \mathcal{G}(z) = B_{\mathcal{G}}(z, z) + \Phi_{\mathcal{G}}(z). \tag{1.3}$$

Here $B_{\mathcal{F}} : H \times H \to \mathbb{R}, B_{\mathcal{G}} : H \times H \to Y$ are bilinear symmetric mappings and $\Phi_{\mathcal{F}} : H \to \mathbb{R}, \Phi_{\mathcal{G}} : H \to Y$ are convex mappings. We do not impose assumptions that imply positivity or convexity of the forms $B_{\mathcal{F}}(z, z)$ and $B_{\mathcal{G}}(z, z)$. So, in (1.1), the objective and constraint functions may be nonconvex. As is well known, this nonconvexity involves a series of troubles both in the analysis of the problem and in the computation of its global solution.

The purpose of this paper is to investigate the validity of the duality in the Arrow–Hurwicz sense [1], which makes use of the ordinary Lagrangian function

$$S(\tau^*, z) := \mathcal{F}(z) + \tau^* \mathcal{G}(z). \tag{1.4}$$

*Received by the editors March 3, 1995; accepted for publication (in revised form) December 3, 1996. This research was supported in part by Russian Foundation for Basic Researches grant G93-01-124 and by International Science Foundation grants NW9000 and NW9300.

http://www.siam.org/journals/sicon/36-1/27773.html

†Department of Mathematics and Mechanics, St. Petersburg University, Bybliotechnaya 2, Petrodvoretz, St. Petersburg, 198904, Russia (almat@niimm.spb.su).

(Here $\tau^* \in Y^*$ is a Lagrange multiplier.) So, throughout the paper, the term "method of duality" will denote the following specific rule (I)–(IV) to solve the problem (1.1).[1] Further the inequality $\tau^* \geq 0$ expresses that $\tau^* y \geq 0$ for all $y \geq 0$.

METHOD OF DUALITY

(I) *For any $\tau^* \geq 0$, solve the problem*

$$S(\tau^*, z) \to \inf \quad \textit{subject to} \quad z \in Z. \tag{1.5}$$

More exactly, it suffices to find only the value of the infimum

$$S_0(\tau^*) := \inf_{z \in Z} S(\tau^*, z). \tag{1.6}$$

(II) *Determine some solution τ_0^* of the dual problem*

$$S_0(\tau^*) \to \max \quad \textit{subject to} \quad \tau^* \geq 0, \tag{1.7}$$

where the maximum must be attained.

(III) *Find all solutions z of the problem* (1.5) *with $\tau^* = \tau_0^*$ and omit those not satisfying at least one of the following relations:*

$$\mathcal{G}(z) \leq 0, \qquad \tau_0^* \mathcal{G}(z) = 0. \tag{1.8}$$

The resultant set $\{z\}$ must coincide with the set of all solutions of the primal problem (1.1).

(IV) *Let $\inf\{\mathcal{F}(z) : z \in D\} > -\infty$, where $D := \{z \in Z : \mathcal{G}(z) \leq 0\}$ is the admissible domain in the problem* (1.1). *A sequence $\{z_n\} \subset Z$ is minimizing in this problem*

$$\mathcal{F}(z_n) \to \inf_{z \in D} \mathcal{F}(z) \quad \textit{as} \quad n \to \infty, \quad z_n \in D \quad \forall n \tag{1.9}$$

if and only if it is minimizing in the problem (1.5) *with $\tau^* := \tau_0^*$*

$$S(\tau_0^*, z_n) \to \inf_{z \in Z} S(\tau_0^*, z) \quad \textit{as} \quad n \to \infty \tag{1.10}$$

and

$$\mathcal{G}(z_n) \leq 0 \quad \forall n\,, \quad \tau_0^* \mathcal{G}(z_n) \to 0 \quad \textit{as} \quad n \to \infty. \tag{1.11}$$

The last item is of particular interest if the primal problem (1.1) has no solution and so the operations (I)–(III) result in the empty set.

In general, the method formulated fails to be correct and may produce a wrong result. Its validity is known to be equivalent to the *duality relation* [2, 13]

$$\inf_{\substack{z \in Z \\ \mathcal{G}(z) \leq 0}} \mathcal{F}(z) = \max_{\tau^* \geq 0} \inf_{z \in Z} S(\tau^*, z). \tag{1.12}$$

This relation is valid for convex problems (1.1) (i.e., if the functions $\mathcal{F}$ and $\mathcal{G}$ are convex) provided $\mathcal{G}(z_*) \in -\mathrm{int} K_+$ for some $z_* \in Z$ [1]. The same is true for specific variational problems and problems of optimal control that can be converted into

[1] The statement of this rule follows [11, 12].

convex ones by means of relaxation [3, Chapters IX, X]. Furthermore, there is known a number of other results on the validity of (1.12), which also utilize a specific nature of the problem under consideration ([4, 5] and others; see [6] for detailed survey).

On the whole, problems for which relation (1.12) is known to be true constitute a relatively small subclass in the class of all mathematical programming problems. In connection with this, a considerable number of generalized duality schemes has been developed. (See [7, 8, 9, 10] and many others. For an excellent survey of them, the reader may consult [6].) Proceeding from various ideas, these schemes replace some constructions in the duality method (I)–(IV) (such as the Lagrangian function, the dual problem, etc.) by generalized ones. However, there are applications where the use of the nongeneralized constructions is preferable and has the advantage of the considerable simplification of the method. Some examples of such applications were given in [11, 12, 13, 14]. A number of other examples will be considered in this paper. They deal with linear-quadratic optimal control problems whose statements differ from traditional ones by the presence of additional quadratic constraints.

In this paper we indicate a new class of problems to which the method (I)–(IV) is applicable. This class not only covers all convex problems but also includes an essential supplement in the region of nonconvex ones. In the description of this class, the main point is the decomposition (1.3) and certain assumptions on the spectral properties of the quadratic forms

$$\mathcal{B}_{\tau^*}(z) := B_{\mathcal{F}}(z,z) + \tau^* B_{\mathcal{G}}(z,z) \tag{1.13}$$

over the linear subspace $\mathfrak{M} := Z - z_0$ $(z_0 \in Z)$ that is a displacement of Z. To illustrate in outline what kind of properties is meant, here we adduce a particular, but suitable for immediate formulation, consequence of the results of this paper.

PROPOSITION 1.1. *In* (1.1) *let H be a real Hilbert space and the subspace Z be closed. Assume that the decomposition* (1.3) *is valid where the mappings $B_{\mathcal{F}}(\cdot,\cdot)$, $B_{\mathcal{G}}(\cdot,\cdot)$ are continuous with respect to the norm $|\cdot|_H$ of H and $\Phi_{\mathcal{F}}(\cdot), \Phi_{\mathcal{G}}(\cdot)$ are continuous with respect to the weak topology of the space H. Also let there exist an element $z_* \in Z$ such that $\mathcal{G}(z_*) \in -\mathrm{int}K_+$. Consider the bounded self-adjoint linear operator $A_{\tau^*} : \mathfrak{M} \to \mathfrak{M}$ that corresponds to the quadratic form* (1.13) $\mathcal{B}_{\tau^*}(z) = \langle A_{\tau^*} z, z\rangle$ $(\forall z \in \mathfrak{M})$. *(The symbol $\langle\cdot,\cdot\rangle$ denotes the inner product in H.)*

If for any $\tau^* \geq 0$ this operator either has no negative isolated eigenvalues of finite geometrical multiplicity at all or, at least, the minimal point of its spectrum is not such an eigenvalue, then relation (1.12) *is true and the method* (I)–(IV) *is valid.*

We recall that an eigenvalue is called *isolated* if some of its neighborhoods has no common points with the spectrum of the operator except for this eigenvalue. The *geometrical multiplicity* of an eigenvalue is merely the dimension of the space of all its eigenvectors.

Both this assertion and the further, more general, results of the paper cannot be applied to problems with a finite-dimensional subspace Z. Nevertheless these results have a series of useful applications to problems of optimal control of dynamical systems (see section 5). To deal with them, it is important that, under the assumptions of this paper, the Lagrangian function $S(\tau^*, z)$ proves to be convex on Z whenever $\tau^* \geq 0$ and $S_0(\tau^*) > -\infty$. This means that *the method of duality* (I)–(IV) *converts the nonconvex problem of global optimization* (1.1) *into two convex problems* (1.5) *and* (1.7).

Indeed, the function $S_0(\tau^*)$ is concave as the infimum (1.6) of the functions (1.4), which are linear in τ^*. So the dual problem (1.7) is convex. (More precisely, it becomes

convex by switching the sign in (1.7) $-S_0(\tau^*) \to \min,\ \ \tau^* \geq 0$.) In (1.7), one obviously can seek the maximum only on the convex domain $\mathcal{R} := \{\tau^* \geq 0 : S_0(\tau^*) > -\infty\}$. In doing so, one has to calculate the value $S_0(\tau^*)$ for $\tau^* \in \mathcal{R}$ by solution of the corresponding problem (1.5), which is also convex as it was remarked above.

The reduction of the primal problem (1.1) to two convex ones brings the possibility to solve it with the aid of the highly developed methods of convex programming. Furthermore, the resultant problems (1.5) and (1.7) are simpler than the original one (1.1) not only for their convexity. While the primal problem (1.1) is infinite dimensional and has a quite complicated admissible domain $D := \{z \in Z : \mathcal{G}(z) \leq 0\}$, the dual problem (1.7) is finite dimensional and the admissible domain Z in the problem (1.5) is quite simple. (It is an affine subspace.) Moreover, in many important applications, the problem (1.5) belongs to a thoroughly investigated class of problems and can be solved easily.

As an example, consider the stationary infinite-horizon linear-quadratic optimal control problem with quadratic inequality constraints. It presents some typical features of problems to be covered by the theory of this paper and is stated as follows:

$$\mathcal{G}_0 \to \min \quad \text{subject to} \tag{1.14}$$

$$\dot{x} = Ax + Bu, \quad x = x(t) \in \mathbb{R}^l, \quad u = u(t) \in \mathbb{R}^m, \quad 0 \leq t < \infty\ , \tag{1.15}$$

$$x(0) = a, \qquad |x(\cdot)| + |u(\cdot)| \in L_2\ , \tag{1.16}$$

$$\mathcal{G}_1 \leq 0, \ldots, \mathcal{G}_k \leq 0, \tag{1.17}$$

where

$$\mathcal{G}_i := \int_0^\infty g_i(x,u)dt - \gamma_i \qquad (i = 0, \ldots, k). \tag{1.18}$$

Here $g_i(x,u) = x^*G_ix + 2x^*Q_iu + u^*\Gamma_iu$ is a quadratic form, the asterisk stands for transposition, A, B, G_i, Q_i, Γ_i are constant matrices, and γ_i are given reals $\gamma_0 = 0$.

Omitting the constraints (1.17), we get the problem (1.14)–(1.16), which was thoroughly investigated in linear-quadratic optimal control theory ([15, 16, 17, 18] and others). This theory places at our disposal quite efficient methods of solution. They include simply verified criteria for the infimum $\mathcal{G}_0^{\inf}$ of the objective functional to be finite. If $\mathcal{G}_0^{\inf} > -\infty$, then solution of the problem ultimately looks like computation of $l \times l$ matrix P and $l \times m$ matrix r [15, 16, 17, 18]. Namely, $\mathcal{G}_0^{\inf} = a^*Pa$ and the optimal process is generated by the closed-loop controller $u = r^*x$ in the so-called regular case (see [19] for the definition). Otherwise, the matrix r is used to construct a minimizing sequence of admissible processes [19]. There are known quite simple and efficient methods to calculate the matrices P and r ([15, 16, 17, 18] and others).

As for the problem (1.14)–(1.18), the above theory did not directly deal with the constraints (1.17), which, however, are of interest for many applications. It was first discovered in [11] that the method (I)–(IV) not only is valid for the problem (1.14)–(1.18)[2] but also *permits us to harness classic linear-quadratic optimal control theory for solution of the problem with the constraints* (1.17).

[2]The same follows from Proposition 1.1; see section 5 for details.

Indeed, rewrite first the problem (1.14)–(1.18) in the form (1.1) with the constraints (1.2)

$$z = [x(\cdot), u(\cdot)], \quad H := L_2\left\{[0, +\infty) \to \mathbb{R}^l\right\} \times L_2\left\{[0, +\infty) \to \mathbb{R}^m\right\},$$

$Z := \{z \in Z : \text{(1.15) and (1.16) are true }\}$, $\mathcal{F} := \mathcal{G}_0$, and $\mathcal{G}_i$ is defined by (1.18). The point to note is that now (1.5) is an ordinary problem of linear-quadratic optimal control theory: *minimize*

$$S[\tau^*, x(\cdot), u(\cdot)] = \int_0^\infty g_\tau(x, u)dt - \gamma_\tau \tag{1.19}$$

subject to the constraints (1.15) *and* (1.16). Here $\tau^* = \tau = \|\tau_i\| \in \mathbb{R}^k$, the function $g_\tau := g_0 + \tau_1 g_1 + \cdots + \tau_k g_k$ is a quadratic form, and $\gamma_\tau := \tau_1\gamma_1 + \cdots + \tau_k\gamma_k$ is a constant. So the problem (1.5) can be solved easily by calculating the corresponding matrices P_τ and r_τ. Then the dual problem (1.7) takes the form $a^*P_\tau a - \gamma_\tau \to \max$. It remains to find its optimum τ^0 and either to generate the solution of the original problem (1.14)–(1.18) by the closed-loop controller $u = r^*_{\tau^0}x$ or to construct a minimizing sequence.

This example underscores the *advantage of the ordinary Lagrangian function* (1.4). This function inherits the quadratic and integral nature of the objective and constraint functionals, and it is for this reason that the effective methods of linear-quadratic optimal control theory can be drawn in solution.

Studies on the validity of the duality method are often related to revealing the convexity or some neighboring properties of the set

$$C_+ := \{\xi = (t, y) \in \mathfrak{Y} := \mathbb{R} \times Y : t \geq \mathcal{F}(z) \text{ and } y \geq \mathcal{G}(z) \text{ for some } z \in Z\}. \tag{1.20}$$

To illustrate their significance, assume that there exists a point $z_* \in Z$ for which $\mathcal{G}(z_*) \in -\mathrm{int}K_+$. Then relation (1.12) is valid if either (A) *the set* (1.20) *is convex* or (B) *its closure* $\overline{C_+}$ *is convex* [2, 13].

For the convex problem (1.1), assertion (A) is apparently true for the convexity of $\mathcal{F}$ and $\mathcal{G}$. There is another research trend, which proves the same assertion by reasons that may ultimately be boiled down to the so-called effect of Lyapunov [2, Chapter 2], [3, pp. 367–373], [20, p. 24]. This paper represents the third research trend, which does not appeal to the above reasons but utilizes the quadratic structure of the functionals under consideration. Its origins may be traced back to the following classic result [21, p. 166].

THEOREM 1.2 (Toeplitz–Hausdorff). *Let H be a complex Hilbert space and $\mathcal{G}_1, \mathcal{G}_2$ be continuous Hermitian forms on H.*

Then the image of the unit sphere $S := \{h \in H : |h| = 1\}$ under the mapping $\mathcal{G} := [\mathcal{G}_1, \mathcal{G}_2]$ is convex.

The following are basic known and quite general facts on the validity of the method (I)–(IV) that are based upon the quadratic structure of the functionals. This method is valid for the problem (1.1) with the constraints (1.2) if either (1) *H is a real linear space, $k = 1$, and $\mathcal{F}$, $\mathcal{G}_1$ are quadratic functionals*[3] [22] or (2) *H is a complex linear space, $k = 2$, and $\mathcal{F}$, $\mathcal{G}_1$, $\mathcal{G}_2$ are quadratic functionals* [23]. In (1) and (2), the number of constraints k cannot be increased because it would lead to the assertions,

[3] A functional $G : H \to \mathbb{R}$ is called *quadratic* if it can be written as the sum $G(h) = \mathcal{B}(h) + \mathrm{Re}l^*z + c$ where $B(h, h) \in \mathbb{R}$ is a quadratic (Hermitian in the case of a complex space H) form, $l^* \in H^*$ is a linear functional, and $c \in \mathbb{R}$.

which are wrong in general [23]. In [11, 14], the method (I)–(IV) was justified for the problem (1.14)–(1.18) with $\mathcal{G}_i$ being the sum of the integral (1.18) and a linear continuous functional on $L_2 \times L_2$. The backbone of the approach taken in [11, 14] was a result of [24] on the convexity of the set $\overline{\mathcal{G}(\mathfrak{M})} \subset R^k$. Here $\mathfrak{M}$ is the collection of all pairs $[x(\cdot),\ u(\cdot)]$ satisfying (1.15), (1.16) with $a := 0$ and $\mathcal{G} := [\mathcal{G}_1, \dots, \mathcal{G}_k]$ where $\mathcal{G}_i$ is defined by (1.18). In [11, 14], the method (I)–(IV) was also justified for certain specific abstract linear-quadratic problems with a finite [11] or infinite [14] number of inequality constraints. A generalization on the case when there also is a finite number of equality constraints was considered in [13]. Applications of the foregoing theory to stationary optimal control problems were considered in [11, 12, 13, 14].

All the results mentioned either impose very strong restrictions on the dimension of the space Y [22, 23] or appeal explicitly or implicitly to the periodicity or the stationarity of the problem [11, 12, 13, 14, 24]. This paper presents a more general approach and enlarges the class of problems to which the method (I)–(IV) is proven to be applicable. This approach does not appeal to the periodicity of the problem and it does not involve restrictions on $\dim Y$. Instead, this approach is actually related to new extensions of the founding Toeplitz–Hausdorff theorem. Since we do not need the explicit formulation of these extensions to prove our results, we state an example of such an extension here only to reveal the underlying ideas.

Let V be a linear space. A set $C \subset V$ is called *almost convex* if there exists a convex set $C_0 \subset V$ such that $C_0 \subset C \subset \overline{C_0}$.

THEOREM 1.3. *Let H be a real Hilbert space and $\mathcal{G}_1, \dots, \mathcal{G}_k$ be continuous quadratic forms on H. Given $\tau = \|\tau_i\| \in \mathbb{R}^k$, consider the quadratic form $\mathcal{G}_\tau(z) := \tau_1\mathcal{G}_1(z) + \dots + \tau_k\mathcal{G}_k(z) = \langle A_\tau z, z\rangle$, where A_τ is the corresponding self-adjoint bounded linear operator. Denote by $\sigma(A_\tau)$ its spectrum.*

If for any $\tau \in \mathbb{R}^k$ the extreme (i.e., the minimal and the maximal) points of the spectrum $\sigma(A_\tau)$ are not isolated eigenvalues of finite geometrical multiplicity, then the image of the unit sphere $S := \{z \in H : |z| = 1\}$ under the mapping $\mathcal{G}(z) := [\mathcal{G}_1(z), \dots, \mathcal{G}_k(z)]$ is almost convex.

The proof of this theorem will be given in section 3 below.[4]

The body of the paper is organized as follows. In section 2, we state our main results. They deal with the abstract problem (1.1). The proof of these results is given in section 4, which is prefaced with the study of vector-valued quadratic forms in section 3. The approach taken in this section develops some ideas and constructions from [27] as well as from [11, 13, 24]. Section 5 is devoted to applications of the general theory developed in the paper. We indicate there a series of infinite-horizon nonstationary and nonconvex optimal control problems to which the method (I)–(IV) is applicable. An example of its application is given in section 6.

Note in conclusion that, according to Lemma 1.1 of [13], results on the applicability of the method (I)–(IV) can be interpreted as criteria for equivalence of the following two conditions (A) and (B). (A) $\mathcal{F}(z) \geq 0$ *in the domain* $z \in Z$, $\mathcal{G}(z) \leq 0$. (B) *There exists a functional* $\tau^* \geq 0$ *such that* $S(\tau^*, z) \geq 0$ *for all* $z \in Z$, *where* $S(\tau^*, z)$ *is the Lagrangian function* (1.4). Called the *S-procedure*, the substitution (B) in place of (A) is used in the theory of stability [22], in the theory of H_∞-optimization, in the theory of robustness of uncertain systems [24], and also in some other branches

[4] In [25, 26], the reader may find a result that is close to another extension of the Toeplitz–Hausdorff theorem that also deals with an arbitrary number of forms but differs from Theorem 1.3 in assumptions.

of systems and control theory [28]. Thus, this paper indicates new cases when the S-procedure is applicable.

2. Statement of basic results. Given $y_1, y_2 \in Y$, the strict inequalities $y_1 < y_2$ and $y_2 > y_1$ denote the inclusion $y_2 - y_1 \in \mathrm{int}K_+$.

DEFINITION 2.1. *The problem* (1.1) *is called regular if there exists an element* $z_* \in Z$ *such that*

$$\mathcal{G}(z_*) < 0. \tag{2.1}$$

Given a topological space X, the *limit inferior* of a function $f : X \to \mathbb{R}$ at a point $x_0 \in X$ is defined as

$$\varliminf_{x \to x_0} f(x) := \liminf_{x \to x_0} f(x) := \sup_{V \in \mathcal{O}(x_0)} \inf_{x \in V} f(x). \tag{2.2}$$

Here $\mathcal{O}(x_0)$ is the collection of all neighborhoods of the point x_0. The following theorem offers a criterion for the method (I)–(IV) to be applicable.

THEOREM 2.2. *In* (1.1) *let* Z *be an affine subspace of a real locally convex topological linear space* H*, and let the finite-dimensional linear space* Y *be ordered with a convex cone* K_+*, which contains an interior point. Let also the functions* $\mathcal{F} : H \to \mathbb{R}$ *and* $\mathcal{G} : H \to Y$ *be given. Assume that*

(A) *the decomposition* (1.3) *is valid where* $B_{\mathcal{F}} : H \times H \to \mathbb{R}$ *and* $B_{\mathcal{G}} : H \times H \to Y$ *are bilinear symmetric mappings and the functions* $\Phi_{\mathcal{F}} : H \to \mathbb{R}, \Phi_{\mathcal{G}} : H \to Y$ *are continuous and convex on* Z*;*

(B) *given* $z \in Z$*, the linear operators* $B_{\mathcal{F}}(z, \cdot)$ *and* $B_{\mathcal{G}}(z, \cdot)$ *are continuous on the linear subspace* $\mathfrak{M} := Z - z_0 \quad (z_0 \in Z)$ *that is a displacement of* Z*;*

(C) *given* $\tau^* \in Y^*, \tau^* \geq 0$*, the quadratic form* (1.13) *has the following property:*

$$\mathcal{B}_{\tau^*}(h) < 0 \quad \textit{for some} \quad h \in \mathfrak{M} \Longrightarrow \varliminf_{\substack{h \to 0 \\ h \in \mathfrak{m}}} \mathcal{B}_{\tau^*}(h) < 0. \tag{2.3}$$

If, in addition, the problem (1.1) *is regular, then relation* (1.12) *is true and the method* (I)–(IV) *is valid. Furthermore, the Lagrangian function* (1.4) *is convex on* Z *and the quadratic form* (1.13) *is nonnegative on* $\mathfrak{M}$ *provided that* $\tau^* \geq 0$ *and the infimum* (1.6) *is finite.*

The proof of this theorem will be given in section 4 below.

It easily follows from (2.2) that $\varliminf_{x \to 0} f(x) = \varliminf_{x \to 0} f(\rho x)$ for any $\rho > 0$ and also that $\varliminf_{x \to 0} f(x) \leq 0$ provided $f(0) = 0$. Picking here $X := \mathfrak{M}, f(x) := \mathcal{B}_{\tau^*}(x)$, we get

$$\sigma_{\lim}(\tau^*) := \varliminf_{\substack{h \to 0 \\ h \in \mathfrak{m}}} \mathcal{B}_{\tau^*}(h) = \varliminf_{\substack{h \to 0 \\ h \in \mathfrak{m}}} \mathcal{B}_{\tau^*}(\rho h) = \rho^2 \sigma_{\lim}(\tau^*) \leq 0 \qquad \forall \rho > 0.$$

This implies that either $\sigma_{\lim}(\tau^*) = 0$ or $\sigma_{\lim}(\tau^*) = -\infty$. So the inequality $\sigma_{\lim}(\tau^*) < 0$ from (2.3) is equivalent to the equality $\sigma_{\lim}(\tau^*) = -\infty$.

The following two lemmas are useful to verify assumption (C) of Theorem 2.2.

LEMMA 2.3. *Any of the following assumptions* (C.1)–(C.3) *implies assumption* (C) *of Theorem* 2.2.

(C.1) *If* $\tau^* \in Y^*$, $\tau^* \geq 0$, $h \in \mathfrak{M}$*, and* $\mathcal{B}_{\tau^*}(h) < 0$*, then there exists a sequence* $\{h_n\}_{n=0}^{\infty} \subset \mathfrak{M}$ *such that*

$$h_n \to 0 \quad \textit{as} \quad n \to \infty, \quad \varliminf_{n \to \infty} \mathcal{B}_{\tau^*}(h_n) < 0. \tag{2.4}$$

(C.2) *For any* $\tau^* \in Y^*,\ \tau^* \geq 0$, *and* $h \in \mathfrak{M}$, *there exists a sequence* $\{h_n\}_{n=0}^{\infty} \subset \mathfrak{M}$ *such that*

$$h_n \to 0 \quad as \quad n \to \infty, \quad \liminf_{n\to\infty} \mathcal{B}_{\tau^*}(h_n) \leq \mathcal{B}_{\tau^*}(h). \tag{2.5}$$

(C.3) *There exists a sequence of mappings* $T_n : H \to H,\ n = 0, 1, \ldots$ *such that* $T_n\mathfrak{M} \subset \mathfrak{M}$ *for all* n *and*

$$B_{\mathcal{P}}(T_n h, T_n h) \to B_{\mathcal{P}}(h, h),\ T_n h \to 0 \ \ as \ \ n \to \infty \ \ (\forall h \in \mathfrak{M},\ \mathcal{P} := \mathcal{F}, \mathcal{G}). \tag{2.6}$$

Here $B_{\mathcal{F}}$ *and* $B_{\mathcal{G}}$ *are the quadratic summands in the decomposition* (1.3).

Proof. The proof comes from the chain of obvious implications (C.3) $\Rightarrow$ (C.2) $\Rightarrow$ (C.1) $\Rightarrow$ (C). □

In [11, 13, 14], the validity of the method (I)–(IV) was proved under assumptions, which included (C.3) with T_n being linear continuous operators.

In many applications, H is introduced to be a Hilbert space equipped with the corresponding weak topology. In this case, the limit inferior from (2.3) can be calculated explicitly. To do this, we recall some notions.

Let X be a real Hilbert space and $\mathcal{B} : X \to \mathbb{R}$ be a continuous quadratic form. A linear subspace $L \subset X$ is called $\mathcal{B}$-*negative* iff $\mathcal{B}(x) < 0$ whenever $x \in L$ and $x \neq 0$. The *(negative) index of inertia* $n_-[\mathcal{B}(\cdot)]$ of the form $\mathcal{B}(\cdot)$ is defined to be the least upper bound of $\dim L$ over all $\mathcal{B}$-negative linear subspaces $L \subset X$. Consider the self-adjoint continuous linear operator $A : X \to X$ that corresponds to the form $\mathcal{B}(x) = \langle Ax, x\rangle$. If its spectrum $\sigma(A)$ contains no negative points, then $n_-[\mathcal{B}(\cdot)] = 0$. If any point $\lambda \in \sigma(A) \cap (-\infty, 0)$ is an eigenvalue, then $n_-[\mathcal{B}(\cdot)] = \sum \deg\lambda$ where the sum is over $\lambda \in \sigma(A) \cap (-\infty, 0)$ and $\deg\lambda$ is the geometrical multiplicity of the eigenvalue λ. In general, $n_-[\mathcal{B}(\cdot)] = \dim \mathrm{Im} P\{(-\infty, 0)\}$, where $P(d\lambda)$ is the resolution of the identity for the operator A [29, p. 889].

LEMMA 2.4. *Let* X *be a real Hilbert space and* $\mathcal{B} : X \to \mathbb{R}$ *be a scalar quadratic form. Assume that the form* $\mathcal{B}$ *is continuous with respect to the norm of* X. *Then*

$$\sigma_{\lim} := \liminf_{x \rightharpoonup 0} \mathcal{B}(x) = \begin{cases} 0 & if \ \ n_-[\mathcal{B}(\cdot)] < \infty, \\ -\infty & if \ \ n_-[\mathcal{B}(\cdot)] = \infty. \end{cases} \tag{2.7}$$

Here the arrow $\rightharpoonup$ *denotes the convergence with respect to the weak topology.*

Proof. Since $\sigma_{\lim} = 0, -\infty$, it suffices to show that $\sigma_{\lim} = 0 \Leftrightarrow n_-[\mathcal{B}(\cdot)] < \infty$.

Let $\sigma_{\lim} = 0$. By (2.2), there exists a weak neighborhood V of the origin such that $c_-(V) := \inf_{x\in V} \mathcal{B}(x) > -\infty$. By the definition of the weak topology, we can pick linear functionals $m_1^*, \ldots, m_p^* \in X^*$ and a real $\epsilon > 0$ such that $V \supset \{x : |m_i^* x| < \epsilon,\ i = 1, \ldots, p\}$. Introducing the linear subspace $M := \{x : m_1^* x = 0, \ldots, m_p^* x = 0\}$, we have $M \subset V$ and, therefore, $\mathcal{B}(x) \geq c_-(V)$ for all $x \in M$. Here, putting $x := \rho x$ and letting $\rho \to \infty$ result in the inequality $\mathcal{B}(x) \geq 0$. Thus,

$$\boxed{m_1^* x = 0, \ldots, m_p^* x = 0} \Longrightarrow \mathcal{B}(x) \geq 0. \tag{2.8}$$

Show that $n_-[\mathcal{B}(\cdot)] \leq p$. Suppose to the contrary that $n_-[\mathcal{B}(\cdot)] \geq p + 1$. Then there exists a linear subspace $L \subset X$ such that $\dim L = p + 1$ and $\mathcal{B}(x) < 0$ whenever $x \in L$ and $x \neq 0$. Choose a basis of this subspace $e_1, \ldots, e_{p+1}$. The system of p linear algebraic equations

$$\sum_{i=1}^{p+1} \alpha_i \left(m_j^* e_i\right) = 0 \quad (\forall j = 1, \ldots, p)$$

has a nonzero solution $\alpha_1, \ldots, \alpha_{p+1}$. Then $x := \alpha_1 e_1 + \cdots + \alpha_{p+1} e_{p+1} \neq 0,\ x \in L$, and so $\mathcal{B}(x) < 0$ by the choice of L. But, on the other hand, $m_j^* x = 0$ for $j = 1, \ldots, p$ and, hence, $\mathcal{B}(x) \geq 0$ due to (2.8). This contradiction proves that $n_-[\mathcal{B}(\cdot)] \leq p < \infty$. Thus $\sigma_{\lim} = 0 \Rightarrow n_-[\mathcal{B}(\cdot)] < \infty$.

Conversely, let $n_-[\mathcal{B}(\cdot)] < \infty$. Consider the self-adjoint continuous linear operator $A : X \to X$ that corresponds to $\mathcal{B}$. Let $P(d\lambda)$ be its resolution of the identity [29, p. 889]. Denoting $P_- := P\{(-\infty, 0)\}$, we have $\dim \operatorname{Im} P_- = n_-[\mathcal{B}(\cdot)] < \infty$. From this it follows that P_- is a continuous operator from the space X endowed with the weak topology into the same space equipped with the norm topology. So the set $V := \{x : |P_- x| < 1\}$ is a weak neighborhood of the origin. Denote $-\lambda_- := \min_{\lambda \in \sigma(A)} \lambda$ and $-\lambda_-^0 := \min\{-\lambda_-,\ 0\}$, where $\sigma(A)$ is the spectrum of A. Given $x \in V$, we have [29, pp. 893, 899]

$$\mathcal{B}(x) = \int_{-\infty}^{+\infty} \lambda \ \ \langle P(d\lambda)x, x\rangle = \int_{(-\infty,0)} \lambda \ \ \langle P(d\lambda)x, x\rangle + \int_0^{\infty} \lambda \ \ \langle P(d\lambda)x, x\rangle$$

$$\geq \int_{\left[-\lambda_-^0, 0\right)} \lambda \ \ \langle P(d\lambda)x, x\rangle \geq -\lambda_-^0 \ \ \langle P_- x, x\rangle = -\lambda_-^0 \, |P_- x|^2 \geq -\lambda_-^0 > -\infty.$$

Passing to the infimum over $x \in V$ and taking into account equation (2.2), we get $\sigma_{\lim} := \liminf_{x \to 0} \mathcal{B}(x) \geq -\lambda_-$. Since $\sigma_{\lim} = 0, -\infty$, we have the equality desired: $\sigma_{\lim} = 0$. □

An immediate consequence of Lemma 2.4 is the following useful corollary.

COROLLARY 2.5. *Let the assumptions of Theorem* 2.2 *be fulfilled except for* (B), (C)*, and let H be a real Hilbert space endowed with the weak topology. Also let the mappings $B_{\mathcal{F}}(\cdot,\cdot), B_{\mathcal{G}}(\cdot,\cdot)$ be continuous with respect to the norm of H and let Z be closed. Then assumption* (B) *of Theorem* 2.2 *is satisfied, and assumption* (C) *of this theorem is equivalent to the following assertion.*

Given $\tau^* \in Y^*,\ \tau^* \geq 0$, the next implication takes place for the quadratic form (1.13)

$$n_-[\mathcal{B}_{\tau^*}|_{\mathfrak{m}}] \neq 0 \Longrightarrow n_-[\mathcal{B}_{\tau^*}|_{\mathfrak{m}}] = \infty. \tag{2.9}$$

In particular, if this assertion is valid, then the conclusion of Theorem 2.2 *is true.*

Proof. Since $\dim Y < \infty$ and the operators $B_{\mathcal{F}}(z,\cdot), B_{\mathcal{G}}(z,\cdot)$ are continuous with respect to the norm of H, they are also continuous with respect to the weak topology; i.e., assumption (B) of Theorem 2.2 is valid. Note that $n_-[\mathcal{B}_{\tau^*}|_{\mathfrak{m}}] \neq 0 \Leftrightarrow \mathcal{B}_{\tau^*}(h) < 0$ for some $h \in \mathfrak{M}$. So (2.9) implies (2.3) by (2.7). Thus, assumption (C) of Theorem 2.2 is true. □

Let the affine subspace Z be closed. Denote by $\sigma(A_{\tau^*})$ the spectrum of the bounded self-adjoint linear operator $A_{\tau^*} : \mathfrak{M} \to \mathfrak{M}$ that corresponds to the quadratic form (1.13) $\mathcal{B}_{\tau^*}(h) = \langle A_{\tau^*} h, h\rangle\ (\forall h \in \mathfrak{M})$. It is well known that the inequality $n_-[\mathcal{B}_{\tau^*}|_{\mathfrak{m}}] < \infty$ is true if and only if the negative part of the spectrum $\sigma(A_{\tau^*}) \cap (-\infty, 0)$ either is empty or consists of a finite number of eigenvalues each having finite geometrical multiplicity. Consequently, *the implication* (2.9) *means that the second case does not take place for any $\tau^* \geq 0$.* Thus, Proposition 1.1 readily follows from Corollary 2.5.

3. Upper limitrophe cones of quadratic forms. This section provides preliminary studies to be used further in the demonstration of Theorem 2.2.

Throughout the section, $\mathfrak{M}$ is a real locally convex linear topological space and Y is a real finite-dimensional linear space ordered with a nonempty convex cone $K_+ \subset Y$. Now we do not insist on its interior to be nonempty. A function $\mathcal{B} : \mathfrak{M} \to Y$ is called *(vector) quadratic form* iff it can be represented as follows $\mathcal{B}(h) = B(h,h)$ with $B(\cdot,\cdot)$ being a bilinear symmetric mapping $B : \mathfrak{M} \times \mathfrak{M} \to Y$. This mapping is determined uniquely by the simply verified formula $B(h_1,h_2) = 1/4\,[\mathcal{B}(h_1+h_2) - \mathcal{B}(h_1-h_2)]$. It shows that the continuity of the form $\mathcal{B}(\cdot)$ is equivalent to the continuity of the mapping $B(\cdot,\cdot)$.

The *upper image* of a set $V \subset \mathfrak{M}$ under a mapping $f : \mathfrak{M} \to Y$ is defined by $f(V)^+ := \{y \in Y : y \geq f(h) \text{ for some } h \in V\} = f(V) + K_+$. Let $\mathcal{B} : \mathfrak{M} \to Y$ be a quadratic form. Denote by $\mathcal{O}$ the collection of all neighborhoods $V \subset \mathfrak{M}$ of the origin. The set

$$\mathcal{K}^+(\mathcal{B}) := \bigcap_{V \in \mathcal{O}} \overline{\mathcal{B}(V)^+} \tag{3.1}$$

is called the *upper limitrophe cone* of the form $\mathcal{B}$. If $K_+ = \{0\}$, the adjective "upper" and the index $^+$ are dropped.

The usage of the term "cone" with respect to the set (3.1) is justified by the following lemma.

LEMMA 3.1. *Let $\mathcal{B} : \mathfrak{M} \to Y$ be a quadratic form.*

(a) *The set* (3.1) *is a cone; i.e., $\rho\mathcal{K}^+(\mathcal{B}) \subset \mathcal{K}^+(\mathcal{B})$ for all $\rho \geq 0$.*

(b) *The upper limitrophe cone* (3.1) *is closed and $K_+ \subset \mathcal{K}^+(\mathcal{B})$.*

(c) *If the form $\mathcal{B}$ is bounded on some neighborhood $V_0 \in \mathcal{O}$, then $\mathcal{K}^+(\mathcal{B}) = \overline{K_+}$.*

Proof. (a) Let $\rho > 0$. By (3.1), we have

$$\rho\mathcal{K}^+(\mathcal{B}) = \bigcap_{V \in \mathcal{O}} \rho\left[\overline{\mathcal{B}(V)^+}\right] = \bigcap_{V \in \mathcal{O}} \overline{\rho\mathcal{B}(V)^+} = \bigcap_{V \in \mathcal{O}} \overline{\rho\left[\mathcal{B}(V) + K_+\right]}$$

$$= \bigcap_{V \in \mathcal{O}} \overline{[\rho\mathcal{B}(V) + \rho K_+]} = \bigcap_{V \in \mathcal{O}} \overline{\mathcal{B}(\sqrt{\rho}V) + K_+} = \bigcap_{V \in \mathcal{O}} \overline{\mathcal{B}(\sqrt{\rho}V)^+}.$$

Here $\sqrt{\rho}V$ runs over $\mathcal{O}$ provided that V does so. This means that the last intersection coincides with (3.1) and $\rho\mathcal{K}^+(\mathcal{B}) = \mathcal{K}^+(\mathcal{B})$. If $\rho = 0$, then $\rho\mathcal{K}^+(\mathcal{B}) = \{0\} \subset \mathcal{K}^+(\mathcal{B})$.

(b) The proof is obvious.

c) Since (b) implies the inclusion $\overline{K_+} \subset \mathcal{K}^+(\mathcal{B})$, it remains to prove the opposite one, $\mathcal{K}^+(\mathcal{B}) \subset \overline{K_+}$. Let $y \in \mathcal{K}^+(\mathcal{B})$. Choose $\rho > 0, \epsilon > 0$. Since $\rho V_0 \in \mathcal{O}$, we have, by (3.1), $y \in \overline{\mathcal{B}\,(\rho V_0)^+}$ and so $y = \mathcal{B}(\rho h_{\rho,\epsilon}) + y^+_{\rho,\epsilon} + \delta y_{\rho,\epsilon}$ for some $h_{\rho,\epsilon} \in V_0, y^+_{\rho,\epsilon} \in K_+, |\delta y_{\rho,\epsilon}| < \epsilon$. Here $|\mathcal{B}\,(\rho h_{\rho,\epsilon})\,| \leq \rho^2 c$ with $c := \sup\{|\mathcal{B}(h)| : h \in V_0\}$ being finite by the assumption. Consequently, $\mathcal{B}\,(\rho h_{\rho,\epsilon}) \to 0, \delta y_{\rho,\epsilon} \to 0$ as $\rho \to +0, \epsilon \to +0$, and the above decomposition of y implies that $y^+_{\rho,\epsilon} \to y$ as $\rho \to +0, \epsilon \to +0$; i.e., $y \in \overline{K_+}$. □

The boundedness of the form $\mathcal{B}$ on some neighborhood V_0 follows from the continuity of $\mathcal{B}$. So, by Lemma 3.1(c), nontrivial upper limitrophe cones are associated with discontinuous forms only. A widespread situation to produce such a cone is the following: $\mathfrak{M}$ is a Hilbert space endowed with the weak topology, and the form $\mathcal{B}$ is continuous with respect to the strong topology of $\mathfrak{M}$ but is not continuous with respect to the weak one. In this case, the form $\mathcal{B}$ apparently has the following important property.

ASSUMPTION 3.1. *Consider the bilinear symmetric mapping $B : \mathfrak{M} \times \mathfrak{M} \to Y$ associated with the quadratic form $\mathcal{B}$. Given $h \in \mathfrak{M}$, the linear operator $B(h, \cdot) : \mathfrak{M} \to Y$ is continuous.*

The usefulness of the notion of upper limitrophe cone is predetermined in part by the following fact.

LEMMA 3.2. *Let Assumption* 3.1 *be fulfilled. Then the upper limitrophe cone* (3.1) *is convex. Furthermore,*

$$\overline{\mathcal{B}(V)^+} + \mathcal{K}^+(\mathcal{B}) \subset \overline{\mathcal{B}(V)^+} \tag{3.2}$$

for any convex neighborhood $V \subset \mathfrak{M}$ of the origin.

Proof. Note first that (3.2) ensures the convexity of the cone (3.1). Indeed, denote by $\mathcal{O}_{conv}$ the collection of all convex neighborhoods of the origin. Since the topology of $\mathfrak{M}$ is locally convex, one can obviously substitute $\mathcal{O}_{conv}$ for $\mathcal{O}$ in (3.1). Then (3.2) immediately results in the inclusion $\mathcal{K}^+(\mathcal{B}) + \mathcal{K}^+(\mathcal{B}) \subset \mathcal{K}^+(\mathcal{B})$ where $\mathcal{K}^+(\mathcal{B})$ is a cone by (a) of Lemma 3.1. This implies that the cone $\mathcal{K}^+(\mathcal{B})$ is convex [8, p. 14].

Thus, it suffices to prove (3.2). Let $V \in \mathcal{O}_{conv}$ and $y \in \overline{\mathcal{B}(V)^+} + \mathcal{K}^+(\mathcal{B}) = \overline{\mathcal{B}(V) + K_+} + \mathcal{K}^+(\mathcal{B})$. Choose $\epsilon > 0$ and $\kappa > 0$. Then $y = \mathcal{B}(z) + y^+ + y_0 + \delta y$ for some $z \in V, y^+ \in K_+, y_0 \in \mathcal{K}^+(\mathcal{B})$, and $|\delta y| < \epsilon$. By Assumption 3.1, the operator $B(z, \cdot)$ is continuous and so the set $V' := \{z' \in \kappa V : |B(z, z')| < \epsilon\}$ is a neighborhood of the origin. By (3.1), $y_0 \in \overline{\mathcal{B}(V')^+}$ and we have $y_0 = \mathcal{B}(z') + y_0^+ + \delta y_0$ where $z' \in V', y_0^+ \in K_+$, and $|\delta y_0| < \epsilon$. Hence

$$\begin{aligned} y &= \mathcal{B}(z) + y^+ + \delta y + y_0 = \mathcal{B}(z) + \mathcal{B}(z') + y^+ + y_0^+ + \delta y + \delta y_0 \\ &= \mathcal{B}(\underbrace{z + z'}_{z_{\epsilon,\kappa}}) + \underbrace{y^+ + y_0^+}_{y^+_{\epsilon,\kappa}} + \underbrace{\delta y - 2B(z, z') + \delta y_0}_{\delta y_{\epsilon,\kappa}}. \end{aligned} \tag{3.3}$$

Here $|\delta y_{\epsilon,\kappa}| \le |\delta y| + |\delta y_0| + 2|B(z, z')| \le 4\epsilon$. Let $\epsilon \to +0$. Then (3.3) means that $\mathcal{B}(z_{\epsilon,\kappa}) + y^+_{\epsilon,\kappa} \to y$ where $y^+_{\epsilon,\kappa} \in K_+$ and $z_{\epsilon,\kappa} = z + z' \in V + \kappa V = (1 + \kappa)V$. So $y \in \overline{\mathcal{B}\left[(1 + \kappa)V\right]^+} = (1 + \kappa)^2 \overline{\mathcal{B}(V)^+}$. Dividing by $(1 + \kappa)^2$ and letting $\kappa \to 0$, we get the inclusion $y \in \overline{\mathcal{B}(V)^+}$ where the vector $y \in \overline{\mathcal{B}(V)^+} + \mathcal{K}^+(\mathcal{B})$ is arbitrary. Thus, (3.2) is true. □

Given a normed space X and a set $Q \subset X$, the symbol ri$Q$ denotes the *relative interior* of the set Q, i.e., its interior in the affine hull aff$Q$ of the set Q.

The next lemma offers an important dual characterization of upper limitrophe cones.

LEMMA 3.3. *Let $\mathcal{B} : \mathfrak{M} \to Y$ be a quadratic form and Assumption* 3.1 *be fulfilled. Denote*

$$\mathcal{E}^*_+(\mathcal{B}) := \left\{\tau^* \in Y^* : \tau^* \ge 0 \ \text{ and } \ \inf_{h \in V} \tau^* \mathcal{B}(h) > -\infty \ \text{ for some } \ V \in \mathcal{O}\right\}, \tag{3.4}$$

where $\mathcal{O}$ is the collection of all neighborhoods of the origin. Then

$$\mathcal{K}^+(\mathcal{B}) = \left\{y \in Y : \tau^* y \ge 0 \ \text{ for all } \ \tau^* \in \mathcal{E}^*_+(\mathcal{B})\right\}, \tag{3.5}$$

where $\mathcal{K}^+(\mathcal{B})$ is the upper limitrophe cone (3.1).

Proof. If $y \in Y \Rightarrow y = 0$, the lemma is obvious. Let Y contain a nonzero vector. By Lemmas 3.1 and 3.2, $\mathcal{K}^+(\mathcal{B})$ is a closed convex cone. So the positive conjugate cone

$$\mathcal{P}^*(\mathcal{B}) := \left\{\tau^* \in Y^* : \tau^* y \ge 0 \ \text{ for all } \ y \in \mathcal{K}^+(\mathcal{B})\right\} \tag{3.6}$$

restores $\mathcal{K}^+(\mathcal{B})$ by the formula [8, p. 121]

$$\mathcal{K}^+(\mathcal{B}) = \{y \in Y : \tau^* y \geq 0 \ \text{ for all } \ \tau^* \in \mathcal{P}^*(\mathcal{B})\}.$$

Here the cone $\mathcal{P}^*(\mathcal{B})$ obviously can be replaced by any set $\mathcal{E}_+^* \subset Y^*$ such that $\overline{\mathcal{E}_+^*} = \mathcal{P}^*(\mathcal{B})$. Putting $\mathcal{E}_+^* := \mathcal{E}_+^*(\mathcal{B})$ entails (3.5) and so it suffices to prove the relation $\overline{\mathcal{E}_+^*(\mathcal{B})} = \mathcal{P}^*(\mathcal{B})$. Since $\mathcal{P}^*(\mathcal{B}) = \overline{\mathrm{ri}\,\mathcal{P}^*(\mathcal{B})}$ [8, p. 46], the equality desired results from the following inclusions to be demonstrated in the remainder of the proof:

$$\mathrm{ri}\mathcal{P}^*(\mathcal{B}) \subset \mathcal{E}_+^*(\mathcal{B}) \subset \mathcal{P}^*(\mathcal{B}). \tag{3.7}$$

We start with the second inclusion. Let $\tau^* \in \mathcal{E}_+^*(\mathcal{B})$. Then, by (3.4), $\tau^* \geq 0$ and $c(\tau^*, V) := \inf_{h \in V} \tau^* \mathcal{B}(h) > -\infty$ for some $V \in \mathcal{O}$. Choose $\epsilon > 0$ and consider $y \in \mathcal{B}(\epsilon V)^+$. It is clear that $y = \epsilon^2 \mathcal{B}(h) + y^+$ for some $h \in V$ and $y^+ \in K_+$. So

$$\tau^* y = \epsilon^2 \tau^* \mathcal{B}(h) + \tau^* y^+ \geq \epsilon^2 \tau^* \mathcal{B}(h) \geq \epsilon^2 \inf_{h' \in V} \tau^* \mathcal{B}(h') = \epsilon^2 c(\tau^*, V)$$

or finally $\tau^* y \geq \epsilon^2 c(\tau^*, V)$ for all $y \in \mathcal{B}(\epsilon V)^+$. By continuity, this inequality spreads on all $y \in \overline{\mathcal{B}(\epsilon V)^+}$ where $\overline{\mathcal{B}(\epsilon V)^+} \supset \mathcal{K}^+(\mathcal{B})$ due to (3.1). Letting $\epsilon \to +0$ and taking into account (3.6), we get the second inclusion in (3.7).

To demonstrate the first inclusion in (3.7), we first assume that *the cone* $\mathcal{K}^+(\mathcal{B})$ *includes no lines.* Then $\mathrm{aff}\mathcal{P}^*(\mathcal{B}) = Y^*$ [8, p. 126] and so $\mathrm{int}\mathcal{P}^*(\mathcal{B}) = \mathrm{ri}\mathcal{P}^*(\mathcal{B}) \neq \emptyset$. Let $\tau^* \in \mathrm{int}\mathcal{P}^*(\mathcal{B})$ and $\tau^* \neq 0$. There exists $\epsilon > 0$ such that $\delta\tau^* \in Y^*$ and $|\delta\tau^*| \leq \epsilon \Rightarrow \tau^* - \delta\tau^* \in \mathcal{P}^*(\mathcal{B})$. Given $y \in \mathcal{K}^+(\mathcal{B})$, by (3.6), we have $0 \leq (\tau^* - \delta\tau^*)\, y = \tau^* y - \delta\tau^* y$ and so $\tau^* y \geq \delta\tau^* y$. By passing to the maximum over $\delta\tau^* \in Y^*$ with $|\delta\tau^*| \leq \epsilon$, we get

$$\tau^* y \geq \epsilon |y| \qquad (\forall y \in \mathcal{K}^+(\mathcal{B})). \tag{3.8}$$

This and (b) of Lemma 3.1 ensure, in particular, that $\tau^* \geq 0$.

Assume that $\tau^* \overline{\in} \mathcal{E}_+^*(\mathcal{B})$. Given a convex neighborhood V of the origin, $c(\tau^*, V) = -\infty$ due to (3.4). So $\tau^* \mathcal{B}(h) \leq -\|\tau^*\|$ for some $h \in V$. Here $\tau^* \mathcal{B}(h) \geq -\|\tau^*\| |\mathcal{B}(h)|$. Hence $-\|\tau^*\| \geq -\|\tau^*\| |\mathcal{B}(h)|$ and so $|y'| \geq 1$ for $y' := \mathcal{B}(h)$. Putting $y := y'|y'|^{-1}$, we have $y = \mathcal{B}(|y'|^{-\frac{1}{2}} h) \in \overline{\mathcal{B}(V)^+}$, $|y| = 1$, and $\tau^* y = |y'|^{-1} \tau^* \mathcal{B}(h) \leq 0$. This means that the following compact set $C(V)$ is not empty

$$C(V) := \left\{ y \in Y : |y| = 1, \tau^* y \leq 0, y \in \overline{\mathcal{B}(V)^+} \right\}. \tag{3.9}$$

Given a finite number of neighborhoods of the origin $V_1, \ldots, V_N \in \mathcal{O}$, there exists a convex neighborhood $V_0 \in \mathcal{O}$ such that $V_0 \subset V_1 \cap \cdots \cap V_N$. It is easy to see that $C(V_0) \subset C(V_1) \cap \cdots \cap C(V_N)$ where $C(V_0) \neq \emptyset$ by the foregoing. Then, in accordance with the generalized principle of Cantor, $C_\infty := \bigcap_{V \in \mathcal{O}} C(V) \neq \emptyset$. Choose $y \in C_\infty$. By (3.9), $\tau^* y \leq 0, |y| = 1$, and $y \in \bigcap_{V \in \mathcal{O}} \overline{\mathcal{B}(V)^+} = \mathcal{K}^+(\mathcal{B})$. These relations apparently contradict (3.8).

So we are forced to reject the assumption $\tau^* \overline{\in} \mathcal{E}_+^*(\mathcal{B})$ and to recognize that $\tau^* \in \mathcal{E}_+^*(\mathcal{B})$ whenever $\tau^* \in \mathrm{ri}\mathcal{P}^*(\mathcal{B})$. Let us proceed to the case when *the cone* $\mathcal{K}^+(\mathcal{B})$ *includes lines.* Reduce this case to the previous one. To this end, consider the linear subspace $L := \mathcal{K}^+(\mathcal{B}) \cap [-\mathcal{K}^+(\mathcal{B})]$, the quotient space $\widehat{Y} := Y/L$, and the canonical projection $\pi : Y \to \widehat{Y}$. Let us order the space $\widehat{Y}$ with the cone $\widehat{K}_+ := \pi(K_+)$. The quadratic form $\widehat{\mathcal{B}}(\cdot) := \pi \circ \mathcal{B}$ obviously satisfies all the assumptions of Lemma 3.3. We are going to show first that

$$\pi^{-1} \mathcal{K}^+(\widehat{\mathcal{B}}) = \mathcal{K}^+(\mathcal{B}). \tag{3.10}$$

Indeed, let $V \in \mathcal{O}$. Note that

$$\widehat{\mathcal{B}}(V)^+ = \widehat{\mathcal{B}}(V) + \widehat{K}_+ = \pi\mathcal{B}(V) + \pi K_+ = \pi\left[\mathcal{B}(V) + K_+\right] = \pi\left[\mathcal{B}(V)^+\right]. \tag{3.11}$$

From this it follows that $\mathcal{B}(V)^+ \subset \pi^{-1}\widehat{\mathcal{B}}(V)^+ \subset \pi^{-1}\overline{\widehat{\mathcal{B}}(V)^+}$, where the set $\pi^{-1}\overline{\widehat{\mathcal{B}}(V)^+}$ is closed by the continuity of π. So $\pi^{-1}\overline{\widehat{\mathcal{B}}(V)^+} \supset \overline{\mathcal{B}(V)^+} \supset \mathcal{K}^+(\mathcal{B})$ and, by (3.1),

$$\pi^{-1}\mathcal{K}^+(\widehat{\mathcal{B}}) = \pi^{-1}\bigcap_{V\in\mathcal{O}}\overline{\widehat{\mathcal{B}}(V)^+} = \bigcap_{V\in\mathcal{O}}\pi^{-1}\overline{\widehat{\mathcal{B}}(V)^+} \supset \mathcal{K}^+(\mathcal{B}).$$

Conversely, given $V \in \mathcal{O}$, there exists a convex neighborhood $V_{aux} \in \mathcal{O}$ such that $V_{aux} \subset V$. By definition, $L \subset \mathcal{K}^+(\mathcal{B})$. So (3.2) implies the inclusion $\overline{\mathcal{B}(V_{aux})^+} + L \subset \overline{\mathcal{B}(V_{aux})^+}$. It is easy to see that $\overline{\mathcal{B}(V_{aux})^+} + L = \pi^{-1}\pi[\mathcal{B}(V_{aux})^+]$. Thus, $\overline{\mathcal{B}(V_{aux})^+} = \pi^{-1}[\pi(\overline{\mathcal{B}(V_{aux})^+})]$ and so $\pi^{-1}[\widehat{Y} \setminus \pi(\overline{\mathcal{B}(V_{aux})^+})] = Y \setminus \overline{\mathcal{B}(V_{aux})^+}$, i.e., $\widehat{Y} \setminus \pi[\mathcal{B}(V_{aux})^+] = \pi[Y \setminus \overline{\mathcal{B}(V_{aux})^+}]$. Here the set $Y \setminus \overline{\mathcal{B}(V_{aux})^+}$ is open, and so too is the image $\pi[Y \setminus \overline{\mathcal{B}(V_{aux})^+}]$ because the operator π transforms open sets into open ones [30, p. 20]. Consequently, the set $\pi[\overline{\mathcal{B}(V_{aux})^+}]$ is closed where, by (3.11), $\widehat{\mathcal{B}}(V_{aux})^+ = \pi[\mathcal{B}(V_{aux})^+] \subset \pi[\overline{\mathcal{B}(V_{aux})^+}]$. Therefore, $\overline{\widehat{\mathcal{B}}(V_{aux})^+} \subset \pi[\overline{\mathcal{B}(V_{aux})^+}]$. Furthermore, due to (3.1), $\overline{\widehat{\mathcal{B}}(V_{aux})^+} \supset \mathcal{K}^+(\widehat{\mathcal{B}})$. Hence $\pi^{-1}[\mathcal{K}^+(\widehat{\mathcal{B}})] \subset \overline{\mathcal{B}(V_{aux})^+} \subset \overline{\mathcal{B}(V)^+}$ for any $V \in \mathcal{O}$. Passing here to the intersection over all $V \in \mathcal{O}$ and taking into account (3.1), we get the inclusion desired: $\pi^{-1}[\mathcal{K}^+(\widehat{\mathcal{B}})] \subset \mathcal{K}^+(\mathcal{B})$. Thus, (3.10) is true.

The cone $\mathcal{K}^+(\widehat{\mathcal{B}})$ includes no lines $l \neq \{0\}$ because, otherwise, we would have, by (3.10), $\pi^{-1}(l) = -\pi^{-1}(l) \subset \pi^{-1}[\mathcal{K}^+(\widehat{\mathcal{B}})] = \mathcal{K}^+(\mathcal{B}), \pi^{-1}(l) \supset \pi^{-1}(0) = L \neq \pi^{-1}(l)$ that would contradict the definition $L = \mathcal{K}^+(\mathcal{B}) \cap [-\mathcal{K}^+(\mathcal{B})]$ of L. Thus, as it has been proven,

$$\mathrm{ri}\mathcal{P}^*(\widehat{\mathcal{B}}) = \mathrm{int}\mathcal{P}^*(\widehat{\mathcal{B}}) \subset \mathcal{E}^*_+(\widehat{\mathcal{B}}). \tag{3.12}$$

Formula (3.10) implies the following relationship between the positive conjugate cones (3.6) [8, p. 143] $\mathcal{P}^*(\mathcal{B}) = \overline{\pi^*\mathcal{P}^*(\widehat{\mathcal{B}})}$, where $\pi^* : \widehat{Y}^* \to Y^*$ is the adjoint operator. Since $\mathrm{Im}\pi = \widehat{Y}$, this operator maps isomorphically the space $\widehat{Y}^*$ onto $\mathrm{Im}\pi^*$. As a result, on the one hand, the image $\pi^*\mathcal{P}^*(\widehat{\mathcal{B}})$ of the closed set $\mathcal{P}^*(\widehat{\mathcal{B}})$ is also closed and we have $\mathcal{P}^*(\mathcal{B}) = \pi^*\mathcal{P}^*(\widehat{\mathcal{B}})$ and, on the other hand, $\mathrm{ri}\mathcal{P}^*(\mathcal{B}) = \pi^*\mathrm{int}\mathcal{P}^*(\widehat{\mathcal{B}})$.

To conclude the proof, consider $\tau^* \in \mathrm{ri}\mathcal{P}^*(\mathcal{B})$. The last equality means that $\tau^* = \pi^*\theta^*$ for some $\theta^* \in \mathrm{int}\mathcal{P}^*(\widehat{\mathcal{B}})$, i.e., $\tau^* = \theta^* \circ \pi$. By (3.12), $\theta^* \in \mathcal{E}^*_+(\widehat{\mathcal{B}})$ and (3.4) yields that $\theta^* \geq 0$ and $\widehat{c}(\theta^*, V) := \inf\{\theta^*\widehat{\mathcal{B}}(z) : z \in V\} > -\infty$ for some $V \in \mathcal{O}$. By definition, $\theta^* \geq 0 \Leftrightarrow \theta^*\widehat{y} \geq 0$ for all $\widehat{y} \in \widehat{K}_+ = \pi K_+ \Leftrightarrow \theta^* \circ \pi y \geq 0$ for all $y \in K_+ \Leftrightarrow \tau^* \geq 0$. In addition, $\widehat{c}(\theta^*, V) = \inf\{\theta^*\pi\mathcal{B}(z) : z \in V\} = \inf\{\tau^*\mathcal{B}(z) : z \in V\} > -\infty$ and, by (3.4), $\tau^* \in \mathcal{E}^*_+(\mathcal{B})$. Thus, $\tau^* \in \mathcal{E}^*_+(\mathcal{B})$ whenever $\tau^* \in \mathrm{ri}\mathcal{P}^*(\mathcal{B})$, and the first inclusion in (3.7) does hold. □

The following lemma characterizes the set (3.4) in terms of the limit inferior (2.2).

LEMMA 3.4. *Let $\mathcal{B} : \mathfrak{M} \to Y$ be a quadratic form. Then*

$$\mathcal{E}^*_+(\mathcal{B}) = \left\{\tau^* \in Y^* : \tau^* \geq 0 \quad \textit{and} \quad \liminf_{h\to 0}\tau^*\mathcal{B}(h) = 0\right\}. \tag{3.13}$$

Proof. Denote by $\mathcal{E}^*_{lim}$ the set on the right. By (2.2) and (3.4), $\mathcal{E}^*_{lim} \subset \mathcal{E}^*_+(\mathcal{B})$. Conversely, let $\tau^* \in \mathcal{E}^*_+(\mathcal{B})$. For each $V \in \mathcal{O}$, denote $I_-(V) := \inf\{\tau^*\mathcal{B}(h) : h \in V\}$.

By (3.4), $\tau^* \geq 0$ and $I_-(V_0) > -\infty$ for some $V_0 \in \mathcal{O}$. Given $\rho > 0$, we have $\rho V_0 \in \mathcal{O}$ and then, by (2.2),

$$l_- := \liminf_{h \to 0} \tau^* \mathcal{B}(h) \geq I_-(\rho V_0) = \inf_{h \in V_0} \tau^* \mathcal{B}(\rho h) = \rho^2 I_-(V_0).$$

By letting $\rho \to +0$, we get $l_- \geq 0$. Furthermore, $I_-(V) \leq \tau^* \mathcal{B}(0) = 0$ for any $V \in \mathcal{O}$ because $0 \in V$. Then equation (2.2) implies $l_- \leq 0$. Thus $l_- = 0, \tau^* \geq 0$ and so $\tau^* \in \mathcal{E}^*_{lim}$. □

We recall that the limitrophe cone $\mathcal{K}(\mathcal{B})$ is the upper limitrophe cone (3.1) corresponding to the trivial positive cone $K_+ = \{0\}$. In other words, $\mathcal{K}(\mathcal{B}) = \bigcap_{V \in \mathcal{O}} \overline{\mathcal{B}(V)}$ where $\mathcal{B}(V) := \{y : y = \mathcal{B}(h) \text{ for some } h \in V\}$ is the ordinary image and $\mathcal{O}$ is the collection of all neighborhoods of the origin.

LEMMA 3.5. *Let $\mathcal{B} : \mathfrak{M} \to Y$ be a quadratic form, Assumption* 3.1 *be fulfilled, and the cone $\mathcal{K}(\mathcal{B})$ does not contain vectors y such that $y \in -\overline{K_+}$ and $y \neq 0$. Then*

$$\mathcal{K}^+(\mathcal{B}) = \overline{\mathcal{K}(\mathcal{B}) + K_+}. \tag{3.14}$$

Proof. If $y \in Y \Rightarrow y = 0$, the lemma is obvious. Let Y contain a nonzero vector. Given $k \subset Y^*$, denote by k° the positive conjugate cone $k^\circ := \{y \in Y : \tau^* y \geq 0 \text{ for all } \tau^* \in k\}$ and put $K_+^* := \{\tau^* \in Y^* : \tau^* \geq 0\}$. We are going to show first that

$$\operatorname{ri} K_+^* \cap \operatorname{ri} \mathcal{E}^*(\mathcal{B}) \neq \emptyset, \tag{3.15}$$

where $\mathcal{E}^*(\mathcal{B})$ is the set (3.4) corresponding to the trivial positive cone $K_+ = \{0\}$, i.e.,

$$\mathcal{E}^*(\mathcal{B}) = \left\{ \tau^* \in Y^* : \inf_{h \in V} \tau^* \mathcal{B}(h) > -\infty \quad \text{for some} \quad V \in \mathcal{O} \right\}.$$

Suppose to the contrary that formula (3.15) violates $\operatorname{ri} K_+^* \cap \operatorname{ri} \mathcal{E}^*(\mathcal{B}) = \emptyset$. Here K_+^* and $\mathcal{E}^*(\mathcal{B})$ are obviously convex cones. Therefore, they are separable with a hyperplane; i.e., there exists a vector $y \in Y$ such that $y \neq 0, \tau^* y \geq 0$ for all $\tau^* \in \mathcal{E}^*(\mathcal{B})$, and $\tau^* y \leq 0$ for all $\tau^* \in K_+^*$. By Lemma 3.3, the second inequality yields that $y \in \mathcal{K}(\mathcal{B})$. In its turn, the third one means that $-y \in \left(K_+^*\right)^\circ$ where $\left(K_+^*\right)^\circ = \overline{K_+}$ [8, p. 125]. Thus, $y \in -\overline{K_+}$ and we have a contradiction to the assumption of the lemma. Therefore, (3.15) does hold.

By (3.15) and Corollary 16.4.2 [8, p. 146],

$$\left[K_+^* \cap \mathcal{E}^*(\mathcal{B})\right]^\circ = \overline{\left[\left(K_+^*\right)^\circ + (\mathcal{E}^*(\mathcal{B}))^\circ\right]}. \tag{3.16}$$

It is clear that $K_+^* \cap \mathcal{E}^*(\mathcal{B}) = \mathcal{E}^*_+(\mathcal{B})$, where the set $\mathcal{E}^*_+(\mathcal{B})$ is given by (3.4). Then (3.5) means that the left-hand side in (3.16) is equal to $\mathcal{K}^+(\mathcal{B})$ and $(\mathcal{E}^*(\mathcal{B}))^\circ = \mathcal{K}(\mathcal{B})$. Furthermore, $\left(K_+^*\right)^\circ = \overline{K_+}$ [8, p. 125]. So (3.16) takes the form $\mathcal{K}^+(\mathcal{B}) = \overline{\mathcal{K}(\mathcal{B}) + \overline{K_+}}$ where the inside closure sign can obviously be omitted. □

Note that, in general, formula (3.14) fails to be true.

As a principal tool in the justification of the method (I)–(IV), we shall use the following key result. To state it, we recall that a set $C \subset Y$ is called *almost convex* if there exists a convex set $C_0 \subset Y$ such that $C_0 \subset C \subset \overline{C_0}$. Thus, the set C is almost convex iff it differs from some convex set C_0 at most by boundary details.

THEOREM 3.6. *Let $\mathfrak{M}$ be a real locally convex linear topological space, and let Y be a real finite-dimensional linear space ordered with a nonempty convex cone $K_+ \subset Y$. Also let $\mathcal{B} : \mathfrak{M} \to Y$ be a quadratic form and Assumption* 3.1 *be fulfilled.*

Assume that, for any $\tau^* \in Y^*, \tau^* \geq 0$, the following implication takes place

$$\tau^*\mathcal{B}(h) < 0 \quad \textit{for some} \quad h \in \mathfrak{M} \Longrightarrow \liminf_{h\to 0} \tau^*\mathcal{B}(h) = -\infty. \tag{3.17}$$

Then the upper image $\mathcal{B}(\mathfrak{M})^+ = \{y \in Y : y \geq \mathcal{B}(h)$ for some $h \in \mathfrak{M}\}$ of the space $\mathfrak{M}$ is almost convex and its closure coincides with the upper limitrophe cone (3.1).

Moreover, given a neighborhood V of the origin, its upper image $\mathcal{B}(V)^+ := \{y \in Y : y \geq \mathcal{B}(h)$ for some $h \in V\}$ is almost convex and

$$\overline{\mathcal{B}(V)^+} = \overline{\mathcal{B}(\mathfrak{M})^+} = \mathcal{K}^+(\mathcal{B}), \tag{3.18}$$

$$\mathrm{ri}\mathcal{B}(V)^+ = \mathrm{ri}\mathcal{B}(\mathfrak{M})^+ = \mathrm{ri}\mathcal{K}^+(\mathcal{B}), \tag{3.19}$$

where $\mathcal{K}^+(\mathcal{B})$ is the upper limitrophe cone (3.1).

In the case of a Hilbert space $\mathfrak{M}$ equipped with the weak topology, the meaning of the implication (3.17) was discussed in section 2 (see Lemma 2.4 and the neighboring considerations).

We break up the proof of Theorem 3.6 into a string of three lemmas.

LEMMA 3.7. *Let the assumptions of Theorem* 3.6 *be valid. Then relations* (3.18) *are true for any $V \in \mathcal{O}$.*

Proof. The inclusions

$$\mathcal{K}^+(\mathcal{B}) \subset \overline{\mathcal{B}(V)^+} \subset \overline{\mathcal{B}(\mathfrak{M})^+} \tag{3.20}$$

result from (3.1). Given $\tau^* \in \mathcal{E}^*_+(\mathcal{B})$, it follows from (3.13) and (3.17) that $\tau^*\mathcal{B}(h) \geq 0$ for all $h \in \mathfrak{M}$ and $\tau^* y^+ \geq 0$ for all $y^+ \in K_+$. Summing up, we have $\tau^*\left[\mathcal{B}(h) + y^+\right] \geq 0$ for all $h \in \mathfrak{M}$ and $y^+ \in K_+$, i.e., $\tau^* y \geq 0$ for all $y \in \mathcal{B}(\mathfrak{M})^+, \tau^* \in \mathcal{E}^*_+(\mathcal{B})$. By Lemma 3.3, this implies that $y \in \mathcal{K}^+(\mathcal{B})$ whenever $y \in \mathcal{B}(\mathfrak{M})^+$. Thus, $\mathcal{B}(\mathfrak{M})^+ \subset \mathcal{K}^+(\mathcal{B})$ where the cone $\mathcal{K}^+(\mathcal{B})$ is closed by Lemma 3.1. So $\overline{\mathcal{B}(\mathfrak{M})^+} \subset \mathcal{K}^+(\mathcal{B})$ and, thanks to (3.20), we get $\overline{\mathcal{B}(\mathfrak{M})^+} = \mathcal{K}^+(\mathcal{B})$. Then the inclusions (3.20) come to relations (3.18). □

To prove (3.19), we need a topological technique, which is developed in the following lemma.

LEMMA 3.8. *Denote by $\mathcal{S}_r$ the standard $(r-1)$-dimensional simplex*

$$\mathcal{S}_r := \{\theta = \|\theta_i\| \in \mathbb{R}^r : \theta_i \geq 0 \quad \textit{for all } i, \quad \theta_1 + \cdots + \theta_r = 1\}.$$

Let a set $C \subset Y$ be given. Assume that

(i) *its closure $\overline{C}$ is convex,*

(ii) *given $r = 1, 2, \ldots$ and elements $y_1, \ldots, y_r \in C$, there exists an infinite sequence $f_1(\cdot), f_2(\cdot), \ldots$ of continuous functions $f_m : \mathcal{S}_r \to C$ such that*

$$f_m(\theta) \to y(\theta) := \sum_{i=1}^{r} \theta_i y_i \quad \textit{as} \quad m \to \infty \tag{3.21}$$

uniformly over $\theta = \|\theta_i\| \in \mathcal{S}_r$.

Then $\mathrm{ri}\,\overline{C} \subset C$.

Proof. It needs to be proven that $y_0 \in C$ whenever $y_0 \in \mathrm{ri}\overline{C}$. Without loss of generality, we can assume that $y_0 = 0$. Then $L := \mathrm{aff}C \ni 0$ is a linear subspace. Choose a real $\epsilon > 0$ such that $\boxed{y \in L, |y| \leq \epsilon} \Rightarrow y \in \mathrm{ri}\overline{C}$ and also choose a basis $y'_1, \ldots, y'_{r-1}$ of

L such that $|y'_1| = \cdots = |y'_{r-1}| = \epsilon/(r-1)$. The vectors $y'_1, \dots, y'_{r-1}$ obviously belong to $\mathrm{ri}\overline{C} \subset \overline{C}$, as does the vector $y'_r := -\left(y'_1 + \cdots + y'_{r-1}\right)$. Consequently, any vector y'_i can be approximated by an element $y_i \in C$. Choose so close approximations that the vectors $y_1, \dots, y_{r-1}$ constitute a basis of L and all the coefficients of the vector y_r with respect to this basis are strictly negative. Then the convex hull Q of the vectors $y_1, \dots, y_r$ is apparently a neighborhood of the origin in the subspace L. Denote by $\theta(y) = [\theta_1(y), \dots, \theta_r(y)] \in \mathcal{S}_r$ the row of the barycentric coordinates of a point $y \in Q$ with respect to the apices $y_1, \dots, y_r$, i.e., $y = \sum_{i=1}^r \theta_i(y) y_i, \theta_i(y) \geq 0$ for all i, and $\sum_{i=1}^r \theta_i(y) = 1$. Consider a sequence $f_1(\cdot), f_2(\cdot), \dots$ that corresponds to the elements $y_1, \dots, y_r \in C$ by Assumption (ii). Then introduce the continuous mappings

$$I_m(y) := y\left[\theta(y)\right] - f_m\left[\theta(y)\right] \in L, \quad y \in Q, \quad m = 1, 2, 3, \dots,$$

where $y(\theta)$ was defined in (3.21). Due to (3.21), the continuous function $I_m(\cdot)$ maps the compact convex neighborhood of the origin $Q \subset L$ into itself provided that the index m is sufficiently large. By the Brouwer's fixed-point theorem, this implies the existence of a fixed point $y_m = I_m(y_m) = y\left[\theta(y_m)\right] - f_m\left[\theta(y_m)\right] \in Q$. Since, by (3.21), $y\left[\theta(y)\right] = \sum_{i=1}^r \theta_i(y) y_i = y$ for all $y \in Q$, we have $y_m = y_m - f_m\left[\theta(y_m)\right]$. Hence $f_m\left[\theta(y_m)\right] = 0$, where $f_m(\theta) \in C$ for all θ by the assumptions of the lemma. Thus, $0 \in C$. □

To prove relations (3.19), we shall apply Lemma 3.8 to the set $C := \mathcal{B}(V)^+$. Then Assumption (i) of this lemma follows from Lemmas 3.2 and 3.7. So we need to demonstrate only Assumption (ii). This gap is filled by the following lemma.

LEMMA 3.9. *Let the assumptions of Theorem* 3.6 *be valid. Consider a continuous mapping $y : \mathcal{S}_r \to \mathcal{K}^+(\mathcal{B})$ where $\mathcal{K}^+(\mathcal{B})$ is the upper limitrophe cone* (3.1).

Given $\epsilon > 0$ and a neighborhood V of the origin, the function $y(\cdot)$ may be decomposed as follows:

$$y(\theta) = \mathcal{B}\left[z(\theta)\right] + y^+(\theta) + \Delta y(\theta) \qquad \forall \theta \in \mathcal{S}_r, \tag{3.22}$$

where $z : \mathcal{S}_r \to V, y^+ : \mathcal{S}_r :\to K_+, \Delta y : \mathcal{S}_r \to Y$ are continuous functions and $|\Delta y(\theta)| \leq \epsilon$ for all $\theta \in \mathcal{S}_r$.

Proof. Due to the compactness of the simplex $\mathcal{S}_r$, we can pick a real $\kappa > 0$ such that

$$\boxed{\theta', \theta'' \in \mathcal{S}_r, |\theta' - \theta''| < \kappa} \Rightarrow |y(\theta') - y(\theta'')| < \frac{\epsilon}{2}. \tag{3.23}$$

Choose a finite collection of nonempty open sets $O_1, \dots, O_{\mathfrak{n}} \subset \mathcal{S}_r$ such that $\mathcal{S}_r = O_1 \cup \cdots \cup O_{\mathfrak{n}}$ and $\sup_{\theta, \vartheta \in O_i} |\theta - \vartheta| < \kappa$. Given $\theta \in \mathcal{S}_r$ and $i = 1, \dots, \mathfrak{n}$, we put $\zeta_i(\theta) := \min_{\vartheta \in \mathcal{S}_r \setminus O_i} |\theta - \vartheta|$. Then $\zeta_i(\theta) > 0$ if $\theta \in O_i$ and $\zeta_i(\theta) = 0$ otherwise. Hence $\zeta(\theta) := \zeta_1(\theta) + \cdots + \zeta_{\mathfrak{n}}(\theta) > 0$ for all $\theta \in \mathcal{S}_r$. So the function $\rho_i(\theta) := \zeta(\theta)^{-1} \zeta_i(\theta)$ is well defined on $\theta \in \mathcal{S}_r$ and continuous. It is also easy to see that

$$\rho_1(\theta) + \cdots + \rho_{\mathfrak{n}}(\theta) = 1, \ \ \rho_i(\theta) \geq 0 \qquad \forall \theta \in \mathcal{S}_r, i = 1, \dots, \mathfrak{n}, \tag{3.24}$$

$$\mathrm{supp}\rho_i := \overline{\{\theta : \rho_i(\theta) \neq 0\}} \neq \emptyset, \ \max_{\theta', \theta'' \in \mathrm{supp}\rho_i} |\theta' - \theta''| < \kappa \quad \forall i = 1, \dots, \mathfrak{n}. \tag{3.25}$$

Choose an element $\theta^{(i)} \in \mathrm{supp}\rho_i$ for each $i = 1, \dots, \mathfrak{n}$, and also fix a convex neighborhood of the origin $V_{\mathrm{c}} \subset V$. Denote $\nu := \epsilon \left(2\mathfrak{n}^2\right)^{-1}$ and consider the bilinear symmetric mapping $B : \mathfrak{M} \times \mathfrak{M} \to Y$ associated with the form $\mathcal{B}(\cdot)$.

As a first step, we are going to choose vectors $x_1,\dots,x_{\mathfrak{n}}\in\mathfrak{M}, y_1^+,\dots,y_{\mathfrak{n}}^+$, and $\Delta y_1,\dots,\Delta y_{\mathfrak{n}}\in Y$ such that

$$y\left(\theta^{(i)}\right)=\mathcal{B}\left(x_i\right)+y_i^+ +\Delta y_i \qquad \forall i\,, \tag{3.26}$$

$$x_i\in\mathfrak{n}^{-1}V_{\mathrm{c}},\quad y_i^+\in K_+,\quad |\Delta y_i|<\nu \qquad \forall i\,, \tag{3.27}$$

$$|B\left(x_i,x_j\right)|<\nu \qquad \forall i\neq j\,. \tag{3.28}$$

Do this in a consecutive order. Namely, choose first x_1, y_1^+, Δy_1, then x_2, y_2^+, Δy_2, and so on. Denote $V_0:=\mathfrak{n}^{-1}V_{\mathrm{c}}$. Since $y\left(\theta^{(1)}\right)\in\mathcal{K}^+(\mathcal{B})$, we have by (3.1) $y\left(\theta^{(1)}\right)\in\overline{\mathcal{B}\left(V_0\right)^+}$ that apparently implies (3.26) and (3.27) for $i=1$ with appropriate $x_1, y_1^+, \Delta y_1$.

Consider $m=1,\dots,\mathfrak{n}-1$, and assume that the vectors $x_1,\dots,x_m, y_1^+,\dots,y_m^+$, $\Delta y_1,\dots,\Delta y_m$ have already been chosen. To construct $x_{m+1}, y_{m+1}^+, \Delta y_{m+1}$, note first that, by Assumption 3.1, the operator $B(x_i,\cdot):\mathfrak{M}\to Y$ is continuous for any $i\le m$. So the set $\widehat{V}:=\left\{h\in\mathfrak{n}^{-1}V_{\mathrm{c}}:|B\left(x_i,h\right)|<\nu \text{ for all } i=1,\dots,m\right\}$ is a convex neighborhood of the origin. This permits us to repeat the above considerations with respect to $V_0:=\widehat{V}, y\left(\theta^{(m+1)}\right)$. As a result, we conclude that (3.26) and (3.27) are true for $i=m+1$ with appropriate $x_{m+1}\in\widehat{V}, y_{m+1}^+$, and Δy_{m+1}. If $\max\{i,j\}\le m$, inequality (3.28) is valid by assumption. If $\max\{i,j\}=m+1$, it follows from the inclusion $x_{m+1}\in\widehat{V}$.

Thus, the requisite vectors $x_i, y_i^+, \Delta y_i$ do exist. Show that the functions

$$\begin{gathered} z(\theta):=\textstyle\sum_{i=1}^{\mathfrak{n}}\sqrt{\rho_i(\theta)}x_i,\ \ y^+(\theta):=\sum_{i=1}^{\mathfrak{n}}\rho_i(\theta)y_i^+,\\ \Delta y(\theta):=y(\theta)-\mathcal{B}\left[z(\theta)\right]-y^+(\theta)\end{gathered} \tag{3.29}$$

have all the properties desired. Indeed, it is clear that the decomposition (3.22) takes place and that the functions $z(\cdot)$ and $y^+(\cdot)$ are continuous, as is the function $\Delta y(\cdot)$ by the following concretization of the term $\mathcal{B}\left[z(\theta)\right]$ in its definition:

$$\mathcal{B}\left[z(\theta)\right]=B\left[z(\theta),z(\theta)\right]=\sum_{i=1}^{\mathfrak{n}}\rho_i(\theta)\mathcal{B}\left(x_i\right)+2\sum_{i<j}\sqrt{\rho_i(\theta)}\sqrt{\rho_j(\theta)}B\left(x_i,x_j\right).$$

Taking into account both this relation and (3.24), (3.26), (3.29), we get

$$|\Delta y(\theta)|=\left|y(\theta)-\mathcal{B}\left[z(\theta)\right]-y^+(\theta)\right|$$

$$=\left|\sum_{i=1}^{\mathfrak{n}}\rho_i(\theta)y(\theta)-\sum_{i=1}^{\mathfrak{n}}\rho_i(\theta)\mathcal{B}\left(x_i\right)-\sum_{i=1}^{\mathfrak{n}}\rho_i(\theta)y_i^+-2\sum_{i<j}\sqrt{\rho_i(\theta)}\sqrt{\rho_j(\theta)}B\left(x_i,x_j\right)\right|$$

$$\le\sum_{i=1}^{\mathfrak{n}}\rho_i(\theta)\underbrace{\left|y\left(\theta^{(i)}\right)-\mathcal{B}\left(x_i\right)-y_i^+\right|}_{=\Delta y_i}+\sum_{i=1}^{\mathfrak{n}}\rho_i(\theta)\left|y\left(\theta^{(i)}\right)-y(\theta)\right|$$

$$+2\sum_{i<j}\sqrt{\rho_i(\theta)}\sqrt{\rho_i(\theta)}\left|B\left(x_i,x_j\right)\right|.$$

Here $\theta^{(i)} \in \operatorname{supp}\rho_i$ by choice. So (3.25) implies that $\rho_i(\theta) \neq 0 \Rightarrow |\theta - \theta^i| < \kappa$ and, by (3.23), $|y(\theta^{(i)}) - y(\theta)| \leq \epsilon/2$. Consequently, $\rho_i(\theta)|y(\theta^{(i)}) - y(\theta)| \leq \rho_i(\theta)\epsilon/2$ for all $\theta \in \mathcal{S}_r$ where $\rho_i(\theta) \leq 1$ by (3.24). This and (3.26)–(3.28) permit us to continue the estimation

$$|\Delta y(\theta)| \leq \sum_{i=1}^{\mathfrak{n}} \rho_i(\theta)\nu + \frac{\epsilon}{2}\sum_{i=1}^{\mathfrak{n}} \rho_i(\theta) + 2\sum_{i<j} \nu = \frac{\epsilon}{2} + \mathfrak{n}^2\nu,$$

where $\nu = \epsilon\left(2\mathfrak{n}^2\right)^{-1}$ by choice. Thus, we get the inequality desired: $|\Delta y(\theta)| \leq \epsilon$. The inclusion $y^+(\theta) \in K_+$ results from (3.24), (3.27), and (3.29) because K_+ is a convex cone. Due to (3.24), $\rho_i(\theta) \leq 1$ and so (3.29) and the first inclusion in (3.27) yield that $z(\theta) \in V_{\mathrm{c}} \subset V$; i.e., the last property to be proved does take place. □

Now we are ready to prove Theorem 3.6. Relations (3.18) are true by Lemma 3.7. To prove (3.19), consider a neighborhood of the origin V and apply Lemma 3.8 to $C := \mathcal{B}(V)^+$. Lemma 3.8(i) follows from Lemmas 3.2 and 3.7. To prove assumption (ii), consider $y_1, \ldots, y_r \in C = \mathcal{B}(V)^+$. By (3.18), $y_1, \ldots, y_r \in \mathcal{K}^+(\mathcal{B})$ where $\mathcal{K}^+(\mathcal{B})$ is a convex cone by Lemma 3.2. So $y(\theta) := \theta_1 y_1 + \cdots + \theta_r y_r \in \mathcal{K}^+(\mathcal{B})$ for all $\theta = \|\theta_i\| \in \mathcal{S}_r$. Given a natural m, the application of Lemma 3.9 to the mapping $y(\cdot)$ and to the real $\epsilon := m^{-1}$ results in the corresponding continuous functions $z : \mathcal{S}_r \to V, y^+ : \mathcal{S}_r \to K_+$, and $\Delta y : \mathcal{S}_r \to Y$. It is clear that $f_m(\theta) := \mathcal{B}\left[z(\theta)\right] + y^+(\theta) \in \mathcal{B}(V)^+ = C$. Since $f_m(\theta) = y(\theta) - \Delta y_m(\theta)$, the function $f_m(\cdot)$ is continuous. From (3.22), we have $|y(\theta) - f_m(\theta)| = |\Delta y_m(\theta)| \leq \epsilon = m^{-1}$, which proves (3.21).

Thus, all the assumptions of Lemma 3.8 are valid. By this lemma, $\mathrm{ri}\mathcal{K}^+(\mathcal{B}) \subset \mathcal{B}(V)^+$, where, due to (3.18), $\mathcal{B}(V)^+ \subset \overline{\mathcal{B}(V)^+} = \mathcal{K}^+(\mathcal{B})$ and so $\mathrm{ri}\mathcal{B}(V)^+ = \mathrm{ri}\mathcal{K}^+(\mathcal{B})$. Here choosing $V := \mathfrak{M}$, we get (3.19).

Consider the convex set $C_0 := \mathrm{ri}\mathcal{K}^+(\mathcal{B})$. By (3.18) and (3.19), $C_0 = \mathrm{ri}\mathcal{B}(V)^+ \subset \mathcal{B}(V)^+ \subset \overline{\mathcal{B}(V)^+} = \mathcal{K}^+(\mathcal{B}) = \overline{C_0}$ or, in brief, $C_0 \subset \mathcal{B}(V)^+ \subset \overline{C_0}$. This means that the set $\mathcal{B}(V)^+$ is almost convex. So is the set $\mathcal{B}(\mathfrak{M})^+$ because $V := \mathfrak{M}$ is a particular case of a neighborhood of the origin. Thus, the proof of Theorem 3.6 is completed.

We conclude the section with the demonstration of Theorem 1.3 (see section 1).

LEMMA 3.10. *Let Y be a real linear finite-dimensional space and C_1, C_2 be almost convex sets.*

If $\mathrm{ri}\, C_1 \cap \mathrm{ri}\, C_2 \neq \emptyset$, then the intersection $C_1 \cap C_2$ is almost convex too.

Proof. By the definition of an almost convex set, there exist convex sets $C_1^0, C_2^0 \subset Y$ such that $C_i^0 \subset C_i \subset \overline{C_i^0}$ for $i = 1, 2$. These inclusions imply that, first, $C^0 := C_1^0 \cap C_2^0 \subset C_1 \cap C_2 \subset \overline{C_1^0} \cap \overline{C_2^0}$ with the set C^0 being convex and, second, $\mathrm{ri}C_i^0 \subset \mathrm{ri}C_i \subset \mathrm{ri}\overline{C_i^0}$, where $\mathrm{ri}C_i^0 = \mathrm{ri}\overline{C_i^0}$ [8, p. 46]. Therefore, $\mathrm{ri}C_i^0 = \mathrm{ri}C_i$ and we have $\mathrm{ri}C_1^0 \cap \mathrm{ri}C_2^0 \neq \emptyset$. From this it follows that $\overline{C^0} = \overline{C_1^0 \cap C_2^0} = \overline{C_1^0} \cap \overline{C_2^0}$ [8, p. 47]. Thus, $C^0 \subset C_1 \cap C_2 \subset \overline{C^0}$; i.e., the set $C_1 \cap C_2$ is almost convex. □

LEMMA 3.11. *Let H be a real Hilbert space and $\mathcal{G}_1, \ldots, \mathcal{G}_k$ be continuous scalar quadratic forms on H. Define the quadratic form $\mathcal{B} : H \to \mathbb{R}^{k+1}$ by $\mathcal{B}(z) := \left\{\mathcal{G}_1(z), \ldots, \mathcal{G}_k(z), |z|^2\right\}$, where $|z| = \sqrt{\langle z, z\rangle}$ is the norm in H.*

If the image $\mathcal{B}(H)$ of the space H is almost convex, then the image of the unit sphere $S := \{z \in H : |z| = 1\}$ under the mapping $\mathcal{G}(z) := [\mathcal{G}_1(z), \ldots, \mathcal{G}_k(z)]$ is almost convex too.

Proof. The function $J(y_1, \ldots, y_k) := (y_1, \ldots, y_k, 1)$ maps isomorphically $\mathbb{R}^k$ onto the affine subspace $C_2 := \left\{y = \|y_i\| \in \mathbb{R}^{k+1} : y_{k+1} = 1\right\} \subset \mathbb{R}^{k+1}$. Furthermore, it is easy to see that $J\left[\mathcal{G}(S)\right] = \mathcal{B}(H) \cap C_2$. So it suffices to prove that the intersection

$\mathcal{B}(H) \cap C_2$ is almost convex. By Lemma 3.10, this follows from the relation $\mathrm{ri}\mathcal{B}(H) \cap \mathrm{ri}C_2 \neq \emptyset$ to be demonstrated in the remainder of the proof.

Assume the opposite $\mathrm{ri}\mathcal{B}(H) \cap \mathrm{ri}C_2 = \emptyset$. By the definition of an almost convex set, there exists a convex set C such that $C \subset \mathcal{B}(H) \subset \overline{C}$. Then $\mathrm{ri}C \subset \mathrm{ri}\mathcal{B}(H) \subset \mathrm{ri}\overline{C}$ where $\mathrm{ri}\overline{C} = \mathrm{ri}C$ [8, p. 46]. So $\mathrm{ri}C = \mathrm{ri}\mathcal{B}(H)$ and we have $\mathrm{ri}C \cap \mathrm{ri}C_2 = \emptyset$ with both sets being convex. Consequently, they are separable with a hyperplane; i.e., there exists $\tau = \|\tau_i\| \in \mathbb{R}^{k+1}$ such that $\tau \neq 0$ and $\tau^* y' \geq \alpha \geq \tau^* y''$ for all $y' \in \mathrm{ri}C$, $y'' \in C_2$ where the asterisk stands for transposition. By continuity, these inequalities spread on all $y' \in \overline{\mathrm{ri}C} = \overline{C} \supset \mathcal{B}(H)$. Picking $y'' := (y_1, \ldots, y_k, 1)$ and letting successively $y_1 \to \pm\infty, \ldots, y_k \to \pm\infty$, we get $\tau_1 = 0, \ldots, \tau_k = 0$, $\alpha \geq \tau_{k+1} \neq 0$. Picking $y' := \mathcal{B}(z) \in \mathcal{B}(H)$, we get $\tau^* y' = \tau_{k+1}|z|^2 \geq \alpha \geq \tau_{k+1}$, $\tau_{k+1}(|z|^2 - 1) \geq 0$ for all $z \in H$. Hence $\tau_{k+1} = 0$, which contradicts the above inequality $\tau_{k+1} \neq 0$. Thus, we are forced to recognize that $\mathrm{ri}\mathcal{B}(H) \cap \mathrm{ri}C_2 \neq \emptyset$. □

Proof of Theorem 1.3. Define the quadratic form $\mathcal{B} : H \to Y := \mathbb{R}^{k+1}$ as in Lemma 3.11. We are going to apply Theorem 3.6. To this end, denote by $\mathfrak{M}$ the space H equipped with the weak topology and pick $K_+ := \{0\}$. Then Assumption 3.1 is clearly true. By Lemma 2.4, the implication (3.17) now takes the form

$$n_- \left[\tau^* \mathcal{B}(\cdot)\right] \neq 0 \Longrightarrow n_- \left[\tau^* \mathcal{B}(\cdot)\right] = +\infty, \tag{3.30}$$

where $\tau = \|\tau_i\| \in \mathbb{R}^{k+1}$, the asterisk stands for transposition, and $n_-(\cdot)$ is the (negative) index of inertia of the scalar quadratic form. Denoting $\widehat{\tau} := (\tau_1, \ldots, \tau_k) \in \mathbb{R}^k$, we obviously have $\tau^* \mathcal{B}(z) = \langle (A_{\widehat{\tau}} + \tau_{k+1} I)\, z, z \rangle$, where the operator $A_{\widehat{\tau}}$ was introduced in Theorem 1.3 and I is the identity operator on H. Consider the resolution of the identity $P(d\lambda)$ for the operator $A_{\widehat{\tau}}$. Then the last equality looks like the following [29, pp. 893, 899]:

$$\tau^* \mathcal{B}(z) = \int_{\lambda_-}^{\lambda_+} (\lambda + \tau_{k+1}) \quad \langle P(d\lambda) z, z \rangle\,. \tag{3.31}$$

Here $\lambda_- := \min\{\lambda : \lambda \in \sigma(A_{\widehat{\tau}})\}$ and $\lambda_+ := \max\{\lambda : \lambda \in \sigma(A_{\widehat{\tau}})\}$.

Let $n_- \left[\tau^* \mathcal{B}(\cdot)\right] \neq 0$. Then (3.31) implies that, on the one hand, $\lambda_- + \tau_{k+1} < 0$ and, on the other hand, $\boxed{z \in \mathrm{Im}P\left\{[\lambda_-, -\tau_{k+1})\right\},\ z \neq 0} \Rightarrow \tau^* \mathcal{B}(z) < 0$. So $n_- \left[\tau^* \mathcal{B}(\cdot)\right] \geq \mu := \dim \mathrm{Im}P\left\{[\lambda_-, -\tau_{k+1})\right\}$. Here λ_- is not an isolated eigenvalue of finite geometrical multiplicity by the assumptions of Theorem 1.3. So $\mu = \infty$ and (3.30) does hold.

Thus, the assumptions of Theorem 3.6 are valid. By this theorem, the image $\mathcal{B}(H)$ is almost convex. Then Lemma 3.11 completes the proof. □

4. Correctness of the method (I)–(IV). This section is devoted to the proof of Theorem 2.2. We recall first the following fact [2, 13, 14].

LEMMA 4.1. *Let, in* (1.1), *Z be a subset of a set H and $\mathcal{F} : H \to \mathbb{R}$, $\mathcal{G} : H \to Y$ be given functions. Assume that the finite-dimensional linear space Y is ordered with a convex cone $K_+ \subset Y$, which contains an interior point. Let also the problem* (1.1) *be regular (see Definition* 2.1*). Denote by D the admissible domain in this problem $D := \{z \in Z : \mathcal{G}(z) \leq 0\}$ and define the set C_+ by* (1.20).

If either $\inf_{z \in D} \mathcal{F}(z) = -\infty$ or the closure $\overline{C_+}$ of the set (1.20) *is convex, then relation* (1.12) *is true and the method of duality* (I)–(IV) *is valid.*

In the remainder of the section, the assumptions of Theorem 2.2 are assumed to be fulfilled. We shall consider the linear space $\mathfrak{Y} := \mathbb{R} \times Y$ and the mapping

$G(z) := [\mathcal{F}(z), \mathcal{G}(z)] \in \mathfrak{Y}$. The space $\mathfrak{Y}$ is assumed to be ordered with the convex cone $\mathfrak{K}_+ := \{\xi = (t, y) \in \mathfrak{Y} : t \geq 0, y \geq 0\}$. By (1.3), we obviously have

$$G(z) = \mathcal{B}(z) + \Phi(z), \tag{4.1}$$

where $\mathcal{B}(z) := B(z, z), B(z_1, z_2) := [B_{\mathcal{F}}(z_1, z_2), B_{\mathcal{G}}(z_1, z_2)]$, and $\Phi(z) := [\Phi_{\mathcal{F}}(z), \Phi_{\mathcal{G}}(z)]$. The functions $B(\cdot, \cdot)$ and $\Phi(\cdot)$ apparently have the same properties that the functions $B_{\mathcal{F}}, B_{\mathcal{G}}$ and, respectively, $\Phi_{\mathcal{F}}, \Phi_{\mathcal{G}}$ have by Assumptions (A) and (B) of Theorem 2.2. Here and throughout, $\mathfrak{M}$ denotes the linear subspace $\mathfrak{M} := Z - z_0 \quad (z_0 \in Z)$ that is a displacement of Z. It is endowed with the topology induced on $\mathfrak{M}$ by H.

To prove Theorem 2.2, it suffices to show that the closure of the set (1.20) is convex provided $\inf_{z \in D} \mathcal{F}(z) > -\infty$. We preface the study of this set with four technical lemmas. Further, the problem (1.1) is assumed to be regular.

LEMMA 4.2. *For the quadratic form* $\mathcal{B}|_{\mathfrak{m}}$, *define the set* $\mathcal{E}_+^* := \mathcal{E}_+^*(\mathcal{B}|_{\mathfrak{m}}) \subset \mathfrak{Y} = \mathbb{R} \times Y$ *in accordance with* (3.13).

If $\inf_{z \in D} \mathcal{F}(z) > -\infty$, *then there exists an element* $\overline{\tau}^* = (\tau_0, \tau^*) \in \mathcal{E}_+^*$ *with* $\tau_0 > 0$.

Proof. Suppose to the contrary that $\mathcal{E}_+^* \subset L^* := \{\overline{\tau}^* \in \mathfrak{Y}^* : \tau_0 = 0\}$. Let $\xi \in L := \{(t, y) \in \mathfrak{Y} : \ y = 0\}$. For any $\overline{\tau}^* \in \mathcal{E}_+^*$, we have $\overline{\tau}^* \in L^*$ and so $\overline{\tau}^* \xi = 0$. Then, by Lemma 3.3, $\xi \in \mathcal{K} := \mathcal{K}^+(\mathcal{B}|_{\mathfrak{m}})$, i.e., $L \subset \mathcal{K}$.

Consider the element $z_* \in Z$ from (2.1), and choose $\epsilon > 0$ such that $|\mathcal{G}(z_*) - y| < \epsilon \Rightarrow y < 0$. Denote by V the set of all $h \in \mathfrak{M}$ such that

$$\begin{array}{ll} |\Phi_{\mathcal{F}}(z_* + h) - \Phi_{\mathcal{F}}(z_*)| < 1, & |B_{\mathcal{F}}(z_*, \ h)| < 1, \\ |\Phi_{\mathcal{G}}(z_* + h) - \Phi_{\mathcal{G}}(z_*)| < \epsilon/4, & |B_{\mathcal{G}}(z_*, \ h)| < \epsilon/4. \end{array} \tag{4.2}$$

By Assumptions (A) and (B) of Theorem 2.2, the set $V \subset \mathfrak{M}$ is a neighborhood of the origin in the subspace $\mathfrak{M}$.

Given $\rho \in \mathbb{R}$, we have $(\rho, 0) \in L \subset \mathcal{K}^+(\mathcal{B}|_{\mathfrak{m}})$ and, by (3.1), $(\rho, 0) \in \overline{\mathcal{B}(V)^+}$. Hence, $(\rho, 0) = [B_{\mathcal{F}}(h, h), \ B_{\mathcal{G}}(h, h)] + (\rho^+, \ y^+) + (\Delta\rho, \ \Delta y)$ for some $h \in V, \ \rho^+ \geq 0, \ y^+ \geq 0, \ |\Delta\rho| < 1, \ |\Delta y| < \epsilon/4$. Letting $z := z_* + h \in Z$, we have by (1.3) and (4.2)

$$\left|\mathcal{G}(z) + y^+ - \mathcal{G}(z_*)\right|$$

$$= \left|B_{\mathcal{G}}(z_* + h, \ z_* + h) + \Phi_{\mathcal{G}}(z_* + h) + y^+ - B_{\mathcal{G}}(z_*, \ z_*) - \Phi_{\mathcal{G}}(z_*)\right|$$

$$\leq \Big| \underbrace{B_{\mathcal{G}}(h, h) + y^+}_{=-\Delta y} \Big| + 2\,|B_{\mathcal{G}}(z_*, \ h)| + |\Phi_{\mathcal{G}}(z_* + h) - \Phi_{\mathcal{G}}(z_*)| < \epsilon,$$

and, consequently, $\mathcal{G}(z) + y^+ < 0 \Rightarrow \mathcal{G}(z) \leq 0, \ z \in Z \Rightarrow z \in D$. Likewise,

$$\left|\mathcal{F}(z) - \rho + \rho^+ - \mathcal{F}(z_*)\right| \leq \Big| \underbrace{B_{\mathcal{F}}(h, h) - \rho + \rho^+}_{=-\Delta\rho} \Big| + 2\,|B_{\mathcal{F}}(z_*, \ h)|$$

$$+ |\Phi_{\mathcal{F}}(z_* + h) - \Phi_{\mathcal{F}}(z_*)| \leq 4.$$

Hence, $\mathcal{F}(z) \leq \rho + \mathcal{F}(z_*) + 4$, where $z \in D$. Thus, $f := \inf_{z \in D} \mathcal{F}(z) \leq \rho + \mathcal{F}(z_*) + 4$. By letting $\rho \to -\infty$, we get the contradiction to the assumption $f > -\infty$ of the lemma. This proves the lemma by contraposition. ∎

Given $Q \subset \mathfrak{Y}^*$, denote by kon$Q$ the minimal cone containing Q.

LEMMA 4.3. *Let $\mathcal{C} \subset \mathfrak{Y}^*$ be a convex set such that $\mathcal{E}_+^* \subset \overline{kon\ \mathcal{C}}$ and* ri$\mathcal{C} \cap \mathcal{E}_+^* \neq \emptyset$ *where the set $\mathcal{E}_+^*$ is defined in Lemma* 4.2.

Assume that, for any $\overline{\tau}^* \in \mathcal{C},\ \overline{\tau}^* \geq 0$, we have

$$\boxed{\overline{\tau}^*\mathcal{B}(h) < 0 \ \text{ for some } \ h \in \mathfrak{M}} \Longrightarrow \sigma\left(\overline{\tau}^*\right) := \liminf_{h \to 0, h \in \mathfrak{m}} \overline{\tau}^*\mathcal{B}(h) < 0. \tag{4.3}$$

Then this implication remains valid provided $\overline{\tau}^* \geq 0$ only.

Proof. $\boxed{\overline{\tau}^* \in \mathcal{C},\ \overline{\tau}^* \geq 0,\ \sigma\left(\overline{\tau}^*\right) = 0} \Rightarrow \boxed{\overline{\tau}^*\mathcal{B}(h) \geq 0 \quad \text{for all} \quad h \in \mathfrak{M}}$ due to (4.3). Furthermore, by (3.13), $\overline{\tau}^* \in \mathcal{E}_+^* \Leftrightarrow \boxed{\overline{\tau}^* \geq 0,\ \sigma\left(\overline{\tau}^*\right) = 0}$. Thus, $\mathcal{C} \cap \mathcal{E}_+^* \subset \mathcal{N}^* := \{\overline{\tau}^* \in \mathfrak{Y}^* : \overline{\tau}^*\xi \geq 0 \text{ for all } \xi \in \mathcal{B}(\mathfrak{M})\}$. Since $\mathcal{E}_+^*$ and $\mathcal{N}$ are cones, we have kon$\mathcal{C} \cap \mathcal{E}_+^* \subset \mathcal{N}^*$. Choose $\overline{\tau}_0^* \in$ ri$\mathcal{C} \cap \mathcal{E}_+^*$ and $\overline{\tau}^* \in \mathcal{E}_+^*$. Then $\overline{\tau}^* \in \overline{\text{kon}\mathcal{C}},\ \overline{\tau}_0^* \in$ ri$\mathcal{C} \subset$ ri(kon$\mathcal{C}$) $\Rightarrow \overline{\tau}_\epsilon^* := (1-\epsilon)\overline{\tau}^* + \epsilon\overline{\tau}_0^* \in$ kon$\mathcal{C}$ for any $\epsilon \in (0, 1]$ [8, p. 45], and, by (3.4), $\overline{\tau}^*,\ \overline{\tau}_0^* \in \mathcal{E}_+^* \Rightarrow \overline{\tau}_\epsilon^* \in \mathcal{E}_+^*$. In brief, $\overline{\tau}_\epsilon^* \in$ kon$\mathcal{C} \cap \mathcal{E}_+^* \subset \mathcal{N}^*$. Letting $\epsilon \to +0$, we get $\overline{\tau}^* \in \mathcal{N}^*$ where the element $\overline{\tau}^* \in \mathcal{E}_+^*$ is arbitrary. Thus, $\mathcal{E}_+^* \subset \mathcal{N}^*$.

Recalling (3.13) and the definition of the set $\mathcal{N}^*$, we can rewrite the last inclusion as follows $\boxed{\overline{\tau}^* \geq 0,\ \sigma\left(\overline{\tau}^*\right) = 0} \Rightarrow \boxed{\overline{\tau}^*\mathcal{B}(h) \geq 0 \text{ for all } h \in \mathfrak{M}}$ or, in other words, $\boxed{\overline{\tau}^*\mathcal{B}(h) < 0 \text{ for some } h \in \mathfrak{M}} \Rightarrow \sigma\left(\overline{\tau}^*\right) \neq 0$ provided $\overline{\tau}^* \geq 0$. To complete the proof, it suffices to recall that $\sigma\left(\overline{\tau}^*\right)$ can take only two values 0 and $-\infty$. □

LEMMA 4.4. *Let $\inf_{z \in D} \mathcal{F}(z) > -\infty$. Then the quadratic form $\mathcal{B}|_{\mathfrak{m}} : \mathfrak{M} \to \mathfrak{Y}$ satisfies the assumptions of Theorem* 3.6 *and, therefore,*

$$\overline{\mathcal{B}(\mathfrak{M})^+} = \mathcal{K}^+\left(\mathcal{B}|_{\mathfrak{m}}\right) \tag{4.4}$$

where the upper limitrophe cone $\mathcal{K}^+\left(\mathcal{B}|_{\mathfrak{m}}\right)$ is defined in accordance with (3.1).

Proof. Assumption 3.1 follows from assumption (B) of Theorem 2.2. Relation (4.4) is merely a part of (3.18). So it suffices to demonstrate only the implication (3.17). Now it takes the form (4.3) and has to be proven for any $\overline{\tau}^* \in \mathfrak{Y}^*,\ \overline{\tau}^* \geq 0$. To this end, we shall apply Lemma 4.3.

Denote $\mathcal{C} := \{\overline{\tau}^* = (\tau_0,\ \tau^*) \in \mathfrak{Y}^* = \mathbb{R} \times Y^* : \tau_0 = 1\}$ and define the set $\mathcal{E}_+^* = \mathcal{E}_+^*\left(\mathcal{B}|_{\mathfrak{m}}\right)$ in accordance with (3.13). Since kon $\mathcal{C} = \{\overline{\tau}^* : \tau_0 > 0\}$, we have $\overline{\text{kon}\ \mathcal{C}} = \{\overline{\tau}^* : \tau_0 \geq 0\}$ and, by (3.13), $\mathcal{E}_+^* \subset \overline{\text{kon}\ \mathcal{C}}$. By Lemma 4.2, there exists an element $\overline{\tau}^* = (\tau_0,\ \tau^*) \in \mathcal{E}_+^*$ with $\tau_0 > 0$. Then $\widetilde{\tau}^* := \tau_0^{-1}\overline{\tau}^* \in \mathcal{E}_+^*$ due to (3.13) and, obviously, $\widetilde{\tau}^* \in \mathcal{C}$ where $\mathcal{C} =$ ri$\mathcal{C}$. Thus, ri$\mathcal{C} \cap \mathcal{E}_+^* \neq \emptyset$. Given $\overline{\tau}^* = (1,\ \tau^*) \in \mathcal{C},\ \overline{\tau}^* \geq 0$, the implication (4.3) apparently has the form (2.3) and is thereby true. Then, by Lemma 4.3, the implication (4.3) is valid for all $\overline{\tau}^* \in \mathfrak{Y},\ \overline{\tau}^* \geq 0$. □

To state the next lemma, we recall that the space $\mathfrak{Y}$ is ordered with the cone $\mathfrak{K}_+ := \{\xi = (t, y) \in \mathfrak{Y} : t \geq 0,\ y \geq 0\}$.

LEMMA 4.5. *Let the vectors $z_1, z_2 \in Z, \delta z \in \mathfrak{M}, \xi^+, \Delta\xi \in \mathfrak{Y}$ and the reals $\theta_1 \in [0, 1], \epsilon > 0$ be given. Denote $\Delta z := z_2 - z_1$ and $\theta_2 := 1 - \theta_1$. Assume that*

$$\mathcal{B}(\Delta z) = \mathcal{B}(\delta z) + \xi^+ + \Delta\xi, \ \ \xi^+ \geq 0, \ \ |\Delta\xi| < \epsilon, \tag{4.5}$$

$$\delta z \in \mathfrak{M}, \ \ |B\left(z_1, \delta z\right)| < \epsilon, \ \ |B(\Delta z, \delta z)| < \epsilon, \tag{4.6}$$

$$\left|\Phi\left(\theta_1 z_1 + \theta_2 z_2 + \sqrt{\theta_1\theta_2}\delta z\right) - \Phi\left(\theta_1 z_1 + \theta_2 z_2\right)\right| < \epsilon. \tag{4.7}$$

Then the vector

$$z := \theta_1 z_1 + \theta_2 z_2 + \sqrt{\theta_1\theta_2}\delta z \tag{4.8}$$

belongs to Z. *Furthermore, there exists a vector* $\Delta \in \mathfrak{Y}$ *such that*

$$G(z) \le \theta_1 G(z_1) + \theta_2 G(z_2) + \Delta, \qquad |\Delta| < 6\epsilon. \tag{4.9}$$

Proof. Since Z is an affine subspace and $\mathfrak{M} = Z - Z$, we have

$$\boxed{\theta_1, \theta_2 \ge 0, \theta_1 + \theta_2 = 1, z_1, z_2 \in Z, \delta z \in \mathfrak{M}} \Rightarrow z \in Z.$$

In light of (4.1) and (4.8), it is straightforward to compute that

$$\begin{aligned} G(z) &= B\left(z_1 + \theta_2\Delta z + \sqrt{\theta_1\theta_2}\delta z\ ,\ z_1 + \theta_2\Delta z + \sqrt{\theta_1\theta_2}\delta z\right) + \Phi(z) \\ &= \mathcal{B}(z_1) + 2\theta_2 B(z_1, \Delta z) + \theta_2^2\mathcal{B}(\Delta z) + \theta_1\theta_2\mathcal{B}(\delta z) + \Phi(\theta_1 z_1 + \theta_2 z_2) \\ &+ \underbrace{2\sqrt{\theta_1\theta_2}B(z_1, \delta z) + 2\theta_1^{1/2}\theta_2^{3/2}B(\Delta z, \delta z) + \Phi(z) - \Phi(\theta_1 z_1 + \theta_2 z_2)}_{\Delta_1}. \end{aligned} \tag{4.10}$$

By (4.6)–(4.8) and the inequalities $0 \le \theta_1, \theta_2 \le 1$, we get

$$|\Delta_1| \le 2\sqrt{\theta_1\theta_2}\epsilon + 2\theta_1^{1/2}\theta_2^{3/2}\epsilon + \epsilon \le 5\epsilon. \tag{4.11}$$

Calculate $\mathcal{B}(\delta z)$ from (4.5) and put the result into (4.10)

$$G(z) = \underbrace{\mathcal{B}(z_1) + 2\theta_2 B(z_1, \Delta z) + \theta_2^2\mathcal{B}(\Delta z) + \theta_1\theta_2\mathcal{B}(\Delta z)}_{b}$$

$$-\theta_1\theta_2\left[\xi^+ + \Delta\xi\right] + \Phi(\theta_1 z_1 + \theta_2 z_2) + \Delta_1.$$

Here

$$b = (\theta_1 + \theta_2)\mathcal{B}(z_1) + 2\theta_2 B(z_1, \Delta z) + \theta_2(\theta_1 + \theta_2)\mathcal{B}(\Delta z)$$

$$= \theta_1\mathcal{B}(z_1) + \theta_2\left[\mathcal{B}(z_1) + 2B(z_1, \Delta z) + \mathcal{B}(\Delta z)\right] = \theta_1\mathcal{B}(z_1) + \theta_2\mathcal{B}\underbrace{(z_1 + \Delta z)}_{=z_2}.$$

Thus, we have

$$\begin{aligned} G(z) &= \theta_1\mathcal{B}(z_1) + \theta_2\mathcal{B}(z_2) + \Phi(\theta_1 z_1 + \theta_2 z_2) - \theta_1\theta_2\xi^+ \\ &+ \underbrace{\Delta_1 - \theta_1(1 - \theta_1)\Delta\xi}_{\Delta}. \end{aligned} \tag{4.12}$$

The third relation in (4.5) and (4.11) imply the second inequality in (4.9). By assumption (A) of Theorem 2.2, the function $\Phi = [\Phi_{\mathcal{F}}, \Phi_{\mathcal{G}}]$ is convex on Z, i.e., $\Phi(\theta_1 z_1 + \theta_2 z_2) \le \theta_1\Phi(z_1) + \theta_2\Phi(z_2)$. Recalling that $\xi^+ \ge 0$ by (4.5) and taking into account (4.1), we can continue (4.12) and complete the proof

$$G(z) \le \theta_1\mathcal{B}(z_1) + \theta_2\mathcal{B}(z_2) + \theta_1\Phi(z_1) + \theta_2\Phi(z_2) + \Delta$$

$$= \theta_1\left[\Phi(z_1) + \mathcal{B}(z_1)\right] + \theta_2\left[\Phi(z_2) + \mathcal{B}(z_2)\right] + \Delta = \theta_1 G(z_1) + \theta_2 G(z_2) + \Delta. \qquad \square$$

Now we are ready to study the set (1.20).

LEMMA 4.6. *Let* $\inf_{z\in D}\mathcal{F}(z) > -\infty$. *Define the set* C_+ *by* (1.20). *Then the closure* $\overline{C_+}$ *is convex.*

Proof. In terms of the mapping $G(z) := [\mathcal{F}(z), \mathcal{G}(z)] \in \mathfrak{Y} := \mathbb{R}\times Y$ and the cone $\mathfrak{K}_+ := \{\xi = (t,y) \in \mathfrak{Y} : t \geq 0, y \geq 0\}$, formula (1.20) takes the form $C_+ = G(Z) + \mathfrak{K}_+$ and implies that $C_+ + \mathfrak{K}_+ \subset C_+$, i.e., $C_+ + \xi^+ \subset C_+$ for any $\xi^+ \in \mathfrak{K}_+$. By passing to the closure, we get $\overline{C_+} + \xi^+ \subset \overline{C_+}$. In other words,

$$\overline{C_+} + \mathfrak{K}_+ \subset \overline{C_+}. \tag{4.13}$$

To prove the lemma, it suffices to demonstrate that

$$\theta_1 G(z_1) + \theta_2 G(z_2) \in \overline{C_+} \tag{4.14}$$

whenever $z_1, z_1 \in Z, \theta_1, \theta_2 \geq 0$, and $\theta_1 + \theta_2 = 1$. Indeed, (4.14) means that $\theta_1 G(Z) + \theta_2 G(Z) \subset \overline{C_+}$. Then, taking into account (4.13), we have

$$\theta_1 C_+ + \theta_2 C_+ = \theta_1 [G(Z) + \mathfrak{K}_+] + \theta_2 [G(Z) + \mathfrak{K}_+]$$

$$= [\theta_1 G(Z) + \theta_2 G(Z)] + [\theta_1 \mathfrak{K}_+ + \theta_2 \mathfrak{K}_+] \subset \overline{C_+} + \mathfrak{K}_+ \subset \overline{C_+}.$$

This immediately implies the inclusion $\overline{C_+} \supset \overline{\theta_1 C_+ + \theta_2 C_+}$, where clearly $\overline{\theta_1 C_+ + \theta_2 C_+} \supset \theta_1\overline{C_+} + \theta_2\overline{C_+}$. Thus $\theta_1\overline{C_+} + \theta_2\overline{C_+} \subset \overline{C_+}$ whenever $\theta_1, \theta_2 \geq 0$ and $\theta_1 + \theta_2 = 1$ that means the convexity of the set $\overline{C_+}$ and proves the lemma.

Turn to demonstration of (4.14). Let $z_1, z_2 \in Z, \theta_1, \theta_2 \geq 0$, and $\theta_1 + \theta_2 = 1$. Denote $\Delta z := z_2 - z_1 \in \mathfrak{M}$, and choose $\epsilon > 0$. Assumptions (A) and (B) of Theorem 2.2 yield that the set

$$\begin{aligned} V := \{h \in \mathfrak{M} : |B(z_1, h)| < \epsilon, |B(\Delta z, h)| < \epsilon, \\ |\Phi(\theta_1 z_1 + \theta_2 z_2 + h) - \Phi(\theta_1 z_1 + \theta_2 z_2)| < \epsilon\} \end{aligned} \tag{4.15}$$

is a neighborhood of the origin in $\mathfrak{M}$. Choose a convex neighborhood of the origin $V_{\mathrm{c}} \subset V$. Taking into account (4.4), we have $\mathcal{B}(\Delta z) \in \mathcal{B}(\mathfrak{M}) \subset \overline{\mathcal{B}(\mathfrak{M})^+} = \mathcal{K}^+(\mathcal{B}|_{\mathfrak{m}})$. By (3.1), this implies $\mathcal{B}(\Delta z) \in \overline{\mathcal{B}(V_{\mathrm{c}})^+}$. Consequently, there exist vectors $\delta z \in V_{\mathrm{c}}, \xi^+ \in \mathfrak{Y}$, and $\Delta\xi \in \mathfrak{Y}$ such that (4.5) is true. Since the set V_{c} is convex, $\rho\delta z \in V_{\mathrm{c}}$ for all $\rho \in [0,1]$ and, in particular, $\delta z, \sqrt{\theta_1\theta_2}\delta z \in V_{\mathrm{c}} \subset V$. Then, due to (4.15), we have (4.6) and (4.7). Thus, the assumptions of Lemma 4.5 are fulfilled.

By this lemma, the vector (4.8) belongs to Z and relations (4.9) are valid with an appropriate vector $\Delta \in \mathfrak{Y}$. By (1.20), the first relation in (4.9) means that $\theta_1 G(z_1) + \theta_2 G(z_2) + \Delta \in C_+$. Taking into account the second relation in (4.9) and letting $\epsilon \to +0$, we get (4.14). □

Proof of Theorem 2.2. By Lemma 4.1, formula (1.12) and the applicability of the method (I)–(IV) follow from Lemma 4.6.

It remains to prove the second assertion of Theorem 2.2. Namely, we have to show that the Lagrangian function (1.4) is convex on Z and the quadratic form (1.13) is nonnegative on $\mathfrak{M}$ whenever $\tau^* \geq 0$ and $S_0(\tau^*) := \inf_{z\in Z} S(\tau^*, z) > -\infty$.

Let $\tau^* \geq 0$ and $S_0(\tau^*) > -\infty$. Choose $z_0 \in Z$. Given $h \in \mathfrak{M}$, we have $z := z_0 + h \in Z$. Then, by (1.3), (1.4), and (1.13),

$$S_0(\tau^*) \leq S(\tau^*, z) = \mathcal{B}_{\tau^*}(z) + \underbrace{\Phi_{\mathcal{F}}(z) + \tau^*\Phi_{\mathcal{G}}(z)}_{\phi(z)}$$

$$= \mathcal{B}_{\tau^*}(z_0) + 2\underbrace{[B_{\mathcal{F}}(z_0, h) + \tau^* B_{\mathcal{G}}(z_0, h)]}_{b} + \mathcal{B}_{\tau^*}(h) + \phi(z_0 + h).$$

Let $h \to 0,\ h \in \mathfrak{M}$. Then $b \to 0$ by assumption (B) of Theorem 2.2 and $\phi(z_0 + h) \to \phi(z_0)$ due to assumption (A) of this theorem. Hence

$$-\infty < S_0(\tau^*) \le \mathcal{B}_{\tau^*}(z_0) + \phi(z_0) + \overbrace{\liminf_{h \to 0,\ h \in \mathfrak{m}} \mathcal{B}_{\tau^*}(h)}^{\sigma},$$

i.e., $\sigma > -\infty$. It was shown in section 2 that either $\sigma = 0$ or $\sigma = -\infty$. Therefore, $\sigma = 0$ and (2.3) implies that $\mathcal{B}_{\tau^*}(h) \ge 0$ for all $h \in \mathfrak{M}$; i.e., the quadratic form (1.13) is nonnegative on $\mathfrak{M}$. Then, as it is well known, this form is convex on Z, and so is the function $\phi(z)$ by assumption (A) of Theorem 2.2. As a result, the Lagrangian function $S(\tau^*, z) = \mathcal{B}_{\tau^*}(z) + \phi(z)$ is also convex on Z. □

5. Method of duality for nonconvex problems of optimal control with inequality constraints. In this section we apply Theorem 2.2 to indicate a number of nonconvex optimal control problems to which the method (I)–(IV) is applicable. It is only for definiteness that we shall confine our remarks to consideration of systems described by ordinary differential equations. Analogous examples can be given for other systems (discrete-time, distributed, and so on).

Consider the following problem of optimal control:

$$\mathcal{G}_0 \to \min \quad \text{subject to} \quad \mathcal{G}_1 \le 0, \dots, \mathcal{G}_k \le 0, \tag{5.1}$$

$$\dot{x} = A(t)x + B(t)u, \quad x = x(t) \in \mathbb{R}^l, \quad u = u(t) \in \mathbb{R}^m, \quad 0 \le t < \infty\,, \tag{5.2}$$

$$x(0) = a, \quad |x(\cdot)| + |u(\cdot)| \in L_2\,, \tag{5.3}$$

$$\mathcal{G}_i := \int_0^\infty g_i(t,x,u)\,dt + \int_0^\infty \phi_i(t,x,u)\,dt - \gamma_i \quad (i = 0, \dots, k)\,. \tag{5.4}$$

Here $x = x(t)$ is the state and $u = u(t)$ is the control, $A(t)$ and $B(t)$ are matrices of respective sizes $l \times l$ and $l \times m$, $g_i(t,x,u) = x^* G_i(t) x + 2x^* Q_i(t) u + u^* \Gamma_i(t) u$ is a quadratic form in x and u, the function $\phi_i(t,x,u)$ is assumed to be at least convex in $x,\ u$, and $\gamma_0, \dots, \gamma_k$ are given reals, $\gamma_0 = 0$. The matrix-valued functions $A(\cdot), B(\cdot), G_i(\cdot) = G_i(\cdot)^*, Q_i(\cdot), \Gamma_i(\cdot) = \Gamma_i(\cdot)^*$ are measurable, so are the functions $\phi_i(\cdot, x, u)$ for all x, u. Further we shall impose additional assumptions on $g_i(\cdot)$ and $\phi_i(\cdot)$ to ensure the convergence of the integrals in (5.4) for all $x(\cdot) \in L_2,\ u(\cdot) \in L_2$. The problem under consideration is a particular case of the abstract problem (1.1) with the constraints (1.2): now $H := L_2\left([0, +\infty) \to \mathbb{R}^l\right) \times L_2\left([0, +\infty) \to \mathbb{R}^m\right),\ Z := \{z = [x(\cdot),\ u(\cdot)] \in H : (5.2) \text{ and } (5.3) \text{ are true}\},\ \mathcal{F} := \mathcal{G}_0$, and $\mathcal{G}_1, \dots, \mathcal{G}_k$ are defined by (5.4).

It is worthy to note first that, in general, the method (I)–(IV) fails to be valid for the problem (5.1)–(5.4). (See the example in subsection 5.5 at the end of the section.) Nevertheless, we shall indicate a number of particular cases of this problem for which

($\mathfrak{A}$) *the method* (I)–(IV) *is valid and the duality relation* (1.12) *is true* and

($\mathfrak{B}$) *the Lagrangian function* (1.4) *is convex on the domain* Z *of all processes* $[x(\cdot), u(\cdot)]$ *that satisfy* (5.2) *and* (5.3) *whenever* $\tau^* \ge 0$ *and the infimum of the Lagrangian function on* Z *is finite.*

In each of these cases, the admissible domain and the objective function can be nonconvex due to the absence of assumptions on $g_i(t,x,u)$ that imply the convexity of the first summand in (5.4).

Interpreting Definition 2.1, we see that *the problem* (5.1)–(5.4) *is regular if and only if there exists a process* $[x(\cdot),\ u(\cdot)]$ *such that relations* (5.2) *and* (5.3) *are true and* $\mathcal{G}_1 < 0, \dots, \mathcal{G}_k < 0$ *with* $\mathcal{G}_i$ *being defined by* (5.4).

5.1. Stationary linear-quadratic problem.

LEMMA 5.1. *Let* $A(t) = A,\ B(t) = B,\ G_i(t) = G_i,\ Q_i(t) = Q_i,$ *and* $\Gamma_i(t) = \Gamma_i$ *be constant matrices. Assume that the pair* (A, B) *is stabilizable and* $\phi_i(t, x, u) = r_i(t)^* x + \rho_i(t)^* u$ *with* $|r_i(\cdot)| + |\rho_i(\cdot)| \in L_2$.

Then assertions $(\mathfrak{A})$ *and* $(\mathfrak{B})$ *are true provided that the problem* (5.1)–(5.4) *is regular.*

Proof. Rewrite the problem (5.1)–(5.4) in the form (1.1) as it was done above. Equip H with the weak topology. Given $z_i = [x_i(\cdot),\ u_i(\cdot)] \in H,\ i = 1, 2,\ z = [x(\cdot),\ u(\cdot)] \in H$, we put

$$B_i(z_1, z_2) := \int_0^{+\infty} \begin{bmatrix} x_1(t) \\ u_1(t) \end{bmatrix}^* \begin{pmatrix} G_i(t) & Q_i(t) \\ Q_i(t)^* & \Gamma_i(t) \end{pmatrix} \begin{bmatrix} x_2(t) \\ u_2(t) \end{bmatrix} dt\ , \tag{5.5}$$
$$B_{\mathcal{F}}(\cdot,\cdot) := B_0(\cdot,\cdot), \qquad B_{\mathcal{G}}(\cdot,\cdot) := [B_1(\cdot,\cdot), \dots, B_k(\cdot,\cdot)]\ ,$$

$$\Phi_i(z) := \int_0^{+\infty} \phi_i\left[t, x(t), u(t)\right]\ dt - \gamma_i\ , \tag{5.6}$$
$$\Phi_{\mathcal{F}}(\cdot) := \Phi_0(\cdot), \qquad \Phi_{\mathcal{G}}(\cdot) = [\Phi_1(\cdot), \dots, \Phi_k(\cdot)]\ .$$

Then the decomposition (1.3) is valid and the mappings $B_{\mathcal{F}}(\cdot,\cdot),\ B_{\mathcal{G}}(\cdot,\cdot),\ \Phi_{\mathcal{F}}(\cdot)$, and $\Phi_{\mathcal{G}}(\cdot)$ are evidently continuous with respect to the norm of H with $\Phi_{\mathcal{F}}(\cdot)$ and $\Phi_{\mathcal{G}}(\cdot)$ being linear in z. This clearly implies assumptions (A) and (B) of Theorem 2.2. Its assumption (C) follows from assumption (C.3) of Lemma 2.3 by this lemma. To prove (C.3), choose a sequence $\{t_n\},\ t_n > 0,\ t_n \to \infty$, and define the operator $T_n : H \to H$ to be the right shift $T_n[x(\cdot),\ u(\cdot)] := [\widehat{x}(\cdot),\ \widehat{u}(\cdot)],\ \widehat{x}(t) := x(t - t_n),\ \widehat{u}(t) := u(t - t_n)$ if $t \geq t_n$ and $\widehat{x}(t) := 0,\ \widehat{u}(t) := 0$ otherwise. Now the space $\mathfrak{M} := Z - Z$ is obviously described by relations (5.2) and (5.3) with $a = 0$. So it is very easy to see that, due to stationarity,

$$T_n \mathfrak{M} \subset \mathfrak{M},\ B_{\mathcal{P}}(T_n h,\ T_n h) = B_{\mathcal{P}}(h, h) \quad \forall n = 1, 2, \dots,\ \mathcal{P} := \mathcal{F}, \mathcal{G}, h \in \mathfrak{M}. \tag{5.7}$$

Given $z = [x(\cdot),\ u(\cdot)] \in \mathfrak{M},\ h = [y(\cdot),\ v(\cdot)] \in L_2 \times L_2$, we have

$$\begin{aligned} |\langle T_n z, h\rangle| &= \left| \int_{t_n}^{+\infty} y(t)^* x(t - t_n)\ dt + \int_{t_n}^{+\infty} v(t)^* u(t - t_n)\ dt \right| \\ &\leq \left(\int_{t_n}^{+\infty} |y(t)|^2 \right)^{1/2} \left(\int_0^{+\infty} |x(t)|^2 \right)^{1/2} + \left(\int_{t_n}^{+\infty} |v(t)|^2 \right)^{1/2} \left(\int_0^{+\infty} |u(t)|^2 \right)^{1/2} \end{aligned} \tag{5.8}$$

and, consequently, $T_n z \rightharpoonup 0$ with respect to the weak topology as $n \to \infty$. Invoking (5.7), we see that (2.6) is true. This completes the proof of assumption (C.3).

Thus, all the assumptions of Theorem 2.2 are fulfilled and, by this theorem, Lemma 5.1 is valid. □

Remark. Lemma 5.1 was first proved in [11]. Note also that this lemma readily follows from Proposition 1.1. Indeed, consider the self-adjoint operator $A_{\tau^*} : \mathfrak{M} \to \mathfrak{M}$ that corresponds to the form (1.13) $\mathcal{B}_{\tau^*}(h) = \langle A_{\tau^*} h, h\rangle,\ h \in \mathfrak{M}$, and denote by

λ_- the minimal point of its spectrum. Then $\mathcal{B}_-(h) := \mathcal{B}_{\tau^*}(h) - \lambda_- |h|_H^2 \geq 0$ for all $h \in \mathfrak{M}$. Assume that λ_- is an eigenvalue of A_{τ^*} and consider a corresponding eigenvector $h_0 \neq 0$. Then $\mathcal{B}_-(h_0) = 0$. By (5.7), the form (1.13) is invariant, $\mathcal{B}_{\tau^*}(T_n h) = \mathcal{B}_{\tau^*}(h)$, $h \in \mathfrak{M}$, and so evidently is the norm $|T_n h|_H = |h|_H$. From this it follows that $\mathcal{B}_-(T_n h_0) = \mathcal{B}_-(h_0) = 0$ where $T_n h_0 \in \mathfrak{M}$ and $\mathcal{B}_-(h) \geq 0$ for all $h \in \mathfrak{M}$. In other words, $T_n h_0 = \arg\min_{h \in \mathfrak{m}} \mathcal{B}_-(h)$. By applying the Fermat necessary conditions, we see that $h_n := T_n h_0$ is also an eigenvector $A_{\tau^*} h_n = \lambda_- h_n$. It remains to note that the linear hull of all the shifts $h_0, h_1, \ldots$ is an infinite-dimensional subspace provided $h_0 \neq 0$. Thus, even if λ_- is an eigenvalue, its geometrical multiplicity is infinite. So the assumptions of Proposition 1.1 are valid and we prove Lemma 5.1 once more.

In what follows, we shall denote the norm of the space L_q by $|\cdot|_q$.

5.2. Almost-periodic linear-quadratic problem with vanishing convex summands in the objective and constraint functions.

LEMMA 5.2. *Assume that*

(2.i) *the functions $A(\cdot), B(\cdot), G_i(\cdot), Q_i(\cdot)$, and $\Gamma_i(\cdot)$ are almost-periodic (more precisely, each of them has an almost-periodic extension on the real line);*

(2.ii) *the functions $\phi_i(t,x,u)$ are convex in x and u for almost all t and*

$$|\phi_i(t,x,u)| \leq \alpha_i(t)\left(|x|^2 + |u|^2\right) + \beta_i(t)\left(|x| + |u|\right) + \gamma_i(t), \tag{5.9}$$

where $\alpha_0(\cdot) \geq 0, \ldots, \alpha_k(\cdot) \geq 0$ are continuous functions of $t \in [0,\infty)$ such that $\alpha_i(t) \to 0$ as $t \to \infty$ and $\beta_i(\cdot) \in L_2, \gamma_i(\cdot) \in L_1, \beta_i(\cdot) \geq 0, \gamma_i(\cdot) \geq 0$ for all $i = 0, \ldots, k$;

(2.iii) *the system* (5.2) *is stabilizable. Namely, there exists a bounded continuous $m \times l$ matrix function $C(t), 0 \leq t < \infty$ such that the solution $x(\cdot)$ of the Cauchy problem*

$$\dot{x} = (A + BC)\, x + f(t), \quad 0 \leq t < \infty, \quad x(0) = 0, \tag{5.10}$$

belongs to $L_2\left[[0,\infty) \to \mathbb{R}^l\right]$ for any $f(\cdot) \in L_2\left[[0,\infty) \to \mathbb{R}^l\right]$ and

$$|x(\cdot)|_2 \leq c\, |f(\cdot)|_2 \tag{5.11}$$

where the constant c is independent of $f(\cdot)$.

Then assertions ($\mathfrak{A}$) *and* ($\mathfrak{B}$) *are true provided that the problem* (5.1)–(5.4) *is regular.*

Proof. We recall that any collection $\{p^\nu(\cdot)\}_{\nu \in \mathcal{A}}$ of seminorms on a linear space X generates a unique locally convex topology such that a set $V \subset X$ is a neighborhood of the origin iff $V \supset \{x \in X : p^{\nu_1}(x) < \epsilon_1, \ldots, p^{\nu_s}(x) < \epsilon_s\}$ for some $\nu_1, \ldots, \nu_s \in \mathcal{A}$, $\epsilon_1 > 0, \ldots, \epsilon_s > 0$, and $s = 1, 2, \ldots$. The convergence $x_n \to x$ with respect to this topology holds iff $p^\nu(x_n - x) \to 0$ as $n \to \infty$ for all $\nu \in \mathcal{A}$.

Rewrite the problem (5.1)–(5.4) in the form (1.1) just as it was done at the beginning of the section. Denote $\alpha(t) := \alpha_0(t) + \cdots + \alpha_k(t)$, $\beta(t) := \beta_0(t) + \cdots + \beta_k(t)$, and define the seminorm $p : H \to [0, +\infty)$ as

$$\begin{aligned} p(z) := &\left[\int_0^{+\infty} \alpha(t)\left(|x(t)|^2 + |u(t)|^2\right)\, dt\right]^{1/2} \\ &+ \int_0^{\infty} \beta(t)\left(|x(t)| + |u(t)|\right)\, dt. \end{aligned} \tag{5.12}$$

Equip the space H with the topology η that is generated by the collection of seminorms $p(\cdot), \{|\langle z,\cdot\rangle|\}_{z\in H}$ where $\langle\cdot,\cdot\rangle$ is the inner product in H. It is easy to see that

$$\begin{array}{c} h_n \xrightarrow{\eta} 0 \text{ as } n\to\infty \\ \Updownarrow \\ \boxed{h_n \rightharpoonup 0 \ \text{ with respect to the weak topology of } H} \text{ and } p(h_n)\to 0 \text{ as } n\to\infty. \end{array} \tag{5.13}$$

The decomposition (1.3) is clearly valid with the summands being defined by (5.5) and (5.6). Due to (5.6), (5.9), and (5.12), the convex functions $\Phi_{\mathcal{F}}(\cdot)$ and $\Phi_{\mathcal{G}}(\cdot)$ are bounded above on the η-open set $V := \{z\in H : p(z)<1\}$. From this it follows that these functions are η-continuous [3, p. 12], and so obviously are the operators $B_{\mathcal{F}}(z,\cdot)$ and $B_{\mathcal{G}}(z,\cdot)$ for any $z\in H$. This means that assumptions (A) and (B) of Theorem 2.2 are fulfilled.

To prove (C), it suffices to demonstrate assumption (C.2) of Lemma 2.3. To this end, consider a Lagrange multiplier $\tau^* = \tau = \|\tau_i\| \in \mathbb{R}^k,\ \tau\geq 0$. The corresponding form (1.13) clearly looks as follows:

$$\mathcal{B}_\tau(z) = \int_0^{+\infty} x^*G_\tau(t)x\,dt + 2\int_0^{+\infty} x^*Q_\tau(t)u\,dt + \int_0^{+\infty} u^*\Gamma_\tau(t)u\,dt, \tag{5.14}$$

where $z=[x(\cdot),u(\cdot)],\ x=x(t),\ u=u(t)$, and $P_\tau(\cdot) := P_0(\cdot)+\tau_1P_1(\cdot)+\cdots+\tau_kP_k(\cdot)$ for $P := G,Q,\Gamma$. For $P := G,Q,\Gamma$, the function $P_\tau(\cdot)$ is almost periodic [31, p. 10], and so is the function $\Xi(t) := [A(-t),B(-t),G_\tau(t),Q_\tau(t),\Gamma_\tau(t)]$ [31, p. 10]. Applying the Bohr definition of an almost-periodic function [31, p. 3] to $\Xi(\cdot)$, we conclude that there exists a sequence $\{t_n\}_{n=1}^\infty \subset (0,+\infty)$ such that $t_n\to\infty$ as $n\to\infty$ and

$$\Delta_n^P := \sup_{t\in\mathbb{R}} |P(t) - P(t+\sigma_P t_n)| \to 0 \quad \text{as} \quad n\to\infty \tag{5.15}$$

for any function $P(\cdot) := A(\cdot),B(\cdot),G_\tau(\cdot),Q_\tau(\cdot),\Gamma_\tau(\cdot)$, where $\sigma_P=-1$ for $P(\cdot) := A(\cdot),B(\cdot)$ and $\sigma_P := 1$ otherwise.

Let $h=[x(\cdot),u(\cdot)]\in\mathfrak{M}$ where $\mathfrak{M}$ is described by relations (5.2) and (5.3) with $a=0$. We have to construct a sequence $h_n=[x_n(\cdot),u_n(\cdot)]\in\mathfrak{M}, n=1,2,\ldots,$ such that (2.5) is true. Shift the process h to the right $\widehat{h}_n := [y_n(\cdot),v_n(\cdot)],\ y_n(t) := x(t-t_n),\ v_n(t) := u(t-t_n)$ if $t\geq t_n$ and $y_n(t):=0,\ v_n(t):=0$ otherwise. Then (5.2) implies that

$$\dot{y}_n = A(t)y_n + B(t)v_n + f_n(t), \quad 0\leq t<\infty, \quad y_n(0)=0,$$

where $f_n(t) := [A(t-t_n)-A(t)]\,y_n(t) + [B(t-t_n)-B(t)]\,v_n(t)$. Taking into account (5.15), we see that $|f_n(\cdot)|_2 \leq \Delta_n^A\,|y_n(\cdot)|_2 + \Delta_n^B\,|v_n(\cdot)|_2 = \Delta_n^A\,|x(\cdot)|_2 + \Delta_n^B\,|u(\cdot)|_2 \to 0$ as $n\to\infty$. Consider the solution $x(\cdot)=\Delta x_n(\cdot)$ of the Cauchy problem (5.10) with $f(\cdot) := -f_n(\cdot)$. Denoting $\Delta u_n(\cdot) := C(\cdot)\Delta x_n(\cdot),\ x_n(\cdot) := y_n(\cdot)+\Delta x_n(\cdot),\ u_n(\cdot) := v_n(\cdot)+\Delta u_n(\cdot)$, we see that, by (5.11),

$$|\Delta x_n(\cdot)|_2 + |\Delta u_n(\cdot)|_2 \to 0 \quad \text{as} \quad n\to\infty \tag{5.16}$$

and also that $x_n(0)=0,\ \dot{x}_n = \dot{y}_n + \Delta\dot{x}_n = [Ay_n+Bv_n+f_n] + [(A+BC)\Delta x_n - f_n]$ $= A(y_n+\Delta x_n) + B(v_n+C\Delta x_n) = Ax_n+Bu_n$, i.e., $h_n := [x_n(\cdot),u_n(\cdot)]\in\mathfrak{M}$.

Thus, a sequence $\{h_n\}\subset\mathfrak{M}$ is indicated. Put $z:=h_n$ in (5.14). Then the first summand in (5.14) looks as follows

$$\int_0^\infty [y_n(t)+\Delta x_n(t)]^* G(t)\,[y_n(t)+\Delta x_n(t)]\ dt$$

$$= \underbrace{\int_0^\infty y_n(t)^* G(t) y_n(t)\ dt}_{a_1} + \underbrace{2\int_0^\infty y_n(t)^* G(t)\Delta x_n(t)\ dt}_{a_2} + \underbrace{\int_0^\infty \Delta x_n(t)^* G(t)\Delta x_n(t)\ dt}_{a_3}.$$

Denoting $g := \sup_{t\in\mathbb{R}} |G(t)|$, we have by (5.16)

$$|a_2| \le 2g\,|y_n(\cdot)|_2\,|\Delta x_n(\cdot)|_2 = 2g\,|x(\cdot)|_2\,|\Delta x_n(\cdot)|_2 \to 0,\ \ |a_3| \le g\,|\Delta x_n(\cdot)|_2^2 \to 0 \text{ as } n\to\infty.$$

In the light of (5.15), we see that

$$\left|a_1 - \int_0^\infty x(t)^* G(t) x(t)\ dt\right| = \left|\int_{t_n}^\infty x\,(t-t_n)^*\,G(t)x\,(t-t_n)\ dt - \int_0^\infty x(t)^* G(t)x(t)\ dt\right|$$

$$= \left|\int_0^\infty x(t)^*\,[G\,(t+t_n) - G(t)]\,x(t)\ dt\right| \le \Delta_n^G |x(\cdot)|_2^2 \to 0 \ \text{ as } \ n\to\infty.$$

Thus, $a_1 \to \int x(t)^* G(t) x(t)\ dt$ as $n\to\infty$.

Considering the second and the third summands in (5.14) by analogy, we get the equality

$$\lim_{n\to\infty} \mathcal{B}_\tau\,(h_n) = \mathcal{B}_\tau(h),$$

which implies the second relation in (2.5). To complete the proof of (2.5), it remains to show that $h_n \xrightarrow{\eta} 0$ as $n\to\infty$. Since, in (5.13), the assertion framed follows from (5.8) and (5.16), we have to prove only the convergence $p\,(h_n)\to 0$ as $n\to\infty$, which evidently follows from (5.12) and the estimations

$$p\,(h_n) = p\,([y_n(\cdot),\ v_n(\cdot)] + [\Delta x_n(\cdot),\ \Delta u_n(\cdot)]) \le p\,[y_n(\cdot),\ v_n(\cdot)] + p\,[\Delta x_n(\cdot),\ \Delta u_n(\cdot)]\,,$$

$$p\,[y_n(\cdot),\ v_n(\cdot)] = \left(\int_{t_n}^\infty \alpha(t)\left(|x\,(t-t_n)|^2 + |u\,(t-t_n)|^2\right)\ dt\right)^{1/2}$$

$$+ \int_{t_n}^\infty \beta(t)\,(|x\,(t-t_n)| + |u\,(t-t_n)|)\ dt$$

$$\le \left[\ \sup_{t\ge t_n} \alpha(t)\right]^{1/2} \left(|x(\cdot)|_2^2 + |u(\cdot)|_2^2\right)^{1/2} + \left(\int_{t_n}^\infty |\beta(t)|^2\ dt\right)^{1/2} (|x(\cdot)|_2 + |u(\cdot)|_2)\,,$$

$$p\,[\Delta x_n(\cdot),\ \Delta u_n(\cdot)] \le \left[\ \sup_{t\ge t_n} \alpha(t)\right]^{1/2} \left(|\Delta x_n(\cdot)|_2^2 + |\Delta u_n(\cdot)|_2^2\right)^{1/2}$$

$$+ \left(\int_{t_n}^\infty |\beta(t)|^2\ dt\right)^{1/2} (|\Delta x_n(\cdot)|_2 + |\Delta u_n(\cdot)|_2)\,.$$

Here $\sup_{t\geq t_n}\alpha(t)\to 0$ and $\int_{t_n}^{\infty}|\beta(t)|^2\,dt\to 0$ as $n\to\infty$ by (2.ii). Thus, Assumption (C.2) of Lemma 2.3 is satisfied.

Thus, all the assumptions of Theorem 2.2 are fulfilled and, by this theorem, Lemma 5.2 is valid. □

In dealing with the previous examples, the point was the use of the right shifts of the process. The subsequent examples will present another technique. It leans upon the following frequency criterion for nonnegativity of an integral quadratic form on the subspace $\mathfrak{M}$ of all processes satisfying (5.2) and (5.3) with $a=0$.

LEMMA 5.3 (see [17]). *Let, in* (5.2), (5.4), $A(t)=A$, $B(t)=B$, $G_0(t)=G_0$, $Q_0(t)=Q_0$, $\Gamma_0(t)=\Gamma_0$ *be constant matrices and* $\phi_0(t,x,u)=0$, $\gamma_0=0$. *Also let the pair* A,B *be stabilizable. Denote by* $\mathbb{C}$ *the complex plane, and extend the function* $g(\cdot):=g_0(\cdot)$ *on* $\mathbb{C}^l\times\mathbb{C}^m$ *as Hermitian form, i.e.,* $g(x'+\imath x'',\,u'+\imath u'):=g(x',u')+g(x'',u'')$ *for all* $x',x''\in\mathbb{R}^l$, $u',u''\in\mathbb{R}^m$, *where* $\imath$ *is the imaginary unity. Define* $\mathcal{G}_0$ *by* (5.4), *and denote by* I *the unit* $l\times l$*-matrix.*

Then $\mathcal{G}_0\geq 0$ *for all processes* $x(\cdot),u(\cdot)$ *satisfying* (5.2) *and* (5.3) *with* $a=0$ *if and only if*

$$g(y,v)\geq 0 \text{ for all } y\in\mathbb{C}^l,\ v\in\mathbb{C}^m,\ \text{and } \omega\in\mathbb{R} \text{ such that } \imath\omega y=Ay+Bv. \tag{5.17}$$

5.3. Linear-quadratic problem with quadratic constraints. Stationary object and nonstationary quadratic forms.

LEMMA 5.4. *Assume* that

(3.i) $g_i(t,x,u)=g_i^0(x,u)+\Delta g_i(t,x,u)$, *where* $g_i^0(x,u)=x^*G_i^0x+2x^*Q_i^0u+u^*\Gamma_i^0u$ *and* $\Delta g_i(t,x,u)=x^*\Delta G_i(t)x+2x^*\Delta Q_i(t)u+u^*\Delta\Gamma_i(t)u$ *are quadratic forms in* x,u;

(3.ii) $\Delta g_i(t,x,u)\geq 0$ *for all* x,u *and almost all* $t\geq 0$;

(3.iii) $\Delta G_i(\cdot)\in L_\infty$, $\Delta Q_i(\cdot)\in L_\infty$, $\Delta\Gamma_i(\cdot)\in L_\infty$;

(3.iv) *given* $y\in\mathbb{R}^l$ *and* $v\in\mathbb{R}^m$, *we have*

$$\frac{1}{T}\int_0^T\Delta g_i(t,y,v)\,dt\to 0\quad as\quad T\to\infty; \tag{5.18}$$

(3.v) $A(t)=A$ *and* $B(t)=B$ *are constant matrices and the pair* (A,B) *is stabilizable;*

(3.vi) $\phi_i(t,x,u)=r_i(t)^*x+\rho_i(t)^*u$ *with* $|r_i(\cdot)|+|\rho_i(\cdot)|\in L_2$.

Then assertions $(\mathfrak{A})$ *and* $(\mathfrak{B})$ *are true provided that the problem* (5.1)–(5.4) *is regular.*

Proof. Rewrite the problem (5.1)–(5.4) in the form (1.1) just as it was done at the beginning of the section. Equip H with the weak topology and define the mappings $B_{\mathcal{F}}(\cdot)$, $B_{\mathcal{G}}(\cdot)$, $\Phi_{\mathcal{F}}(\cdot)$, $\Phi_{\mathcal{G}}(\cdot)$ by (5.5) and (5.6). Then the decomposition (1.3) is obviously valid and assumptions (A) and (B) of Theorem 2.2 are fulfilled. To prove (C), it suffices to demonstrate assumption (C.1) of Lemma 2.3. To this end, consider a Lagrange multiplier $\tau^*=\tau=\|\tau_i\|\in\mathbb{R}^k$, $\tau\geq 0$. The corresponding form (1.13) now clearly looks as follows:

$$\mathcal{B}_\tau(z):=\underbrace{\int_0^{+\infty}g_\tau^0\left[x(t),u(t)\right]\,dt}_{\mathcal{B}_\tau^0(z)}+\underbrace{\int_0^{+\infty}\Delta g_\tau\left[t,x(t),u(t)\right]\,dt}_{\Delta\mathcal{B}_\tau(z)}, \tag{5.19}$$

where $g_\tau^0(\cdot):=g_0^0(\cdot)+\tau_1g_1^0(\cdot)+\cdots+\tau_kg_k^0(\cdot)$ and $\Delta g_\tau(\cdot):=\Delta g_0(\cdot)+\tau_1\Delta g_1(\cdot)+\cdots+\tau_k\Delta g_k(\cdot)$. Assumptions (3.i)–(3.iv) apparently remain valid for $i:=\tau$.

Let $\mathcal{B}_\tau(h) < 0$ for some $h \in \mathfrak{M}$. Since $\Delta\mathcal{B}_\tau(h) \geq 0$ by (3.ii), we have $\mathcal{B}^0_\tau(h) < 0$. Thus, the quadratic form $\mathcal{B}^0_\tau(h)$ is not nonnegative on the subspace $\mathfrak{M}$ of all processes $[x(\cdot), u(\cdot)]$ satisfying (5.2) and (5.3) with $a = 0$. By Lemma 5.3,

$$g^0_\tau(y, v) < 0, \qquad \imath\omega = Ay + Bv \tag{5.20}$$

for some $y \in \mathbb{C}^l,\ v \in \mathbb{C}^m$, and $\omega \in \mathbb{R}$. Here $g^0_\tau(x, u)$ is extended on $\mathbb{C}^l \times \mathbb{C}^m$ as Hermitian form. Using the analogous extension of the form $\Delta g_\tau(t, x, u)$, we make the definition (5.19) of $\mathcal{B}_\tau(z)$ valid for processes $[x(\cdot), u(\cdot)]$ with complex-valued components $x(t) \in \mathbb{C}^l,\ u(t) \in \mathbb{C}^m$.

Since the pair (A, B) is stabilizable, there exists a real $l \times m$ matrix C such that the equation $\dot{x} = (A + BC^*)\, x$ is stable. Its solution $x(t),\ t \geq 0$ with $x(0) = a \in \mathbb{C}^l$ belongs to L_2. Denote $x(t|a) := x(t)$ if $t \geq 0$ and $x(t|a) := 0$ otherwise and put $u(\cdot|a) := C^* x(\cdot|a)$. Then we have

$$\dot{x}(t|a) = Ax(t|a) + Bu(t|a) \quad (\text{if } t \neq 0), \quad x(0 + 0|a) = a,$$

$$|x(\cdot|a)|_2 + |u(\cdot|a)|_2 \leq c_2|a| \qquad \forall a \in \mathbb{C}^l\ . \tag{5.21}$$

Now we are ready to construct a sequence of processes to ensure assumption (C.1) of Lemma 2.3. Namely, choose a sequence $\{t_n\} \subset (0, \infty),\ t_n \to \infty$, denote $\chi_n(t) := 1$ if $0 \leq t \leq t_n$ and $\chi_n(t) := 0$ otherwise, and put $z_n := [x_n(\cdot), u_n(\cdot)]$, where

$$\begin{aligned} x_n(t) :=\ & \underbrace{t_n^{-1/2} y\ e^{\imath\omega t}\chi_n(t)}_{x^0_n(t)} + \underbrace{t_n^{-1/2}\left\{x\,[|t| - y] + x\left[t - t_n|e^{\imath\omega t_n} y\right]\right\}}_{\Delta x_n(t)}, \\ u_n(t) :=\ & \underbrace{t_n^{-1/2} v\ e^{\imath\omega t}\chi_n(t)}_{u^0_n(t)} + \underbrace{t_n^{-1/2}\left\{u\,[|t| - y] + u\left[t - t_n|e^{\imath\omega t_n} y\right]\right\}}_{\Delta u_n(t)}. \end{aligned} \tag{5.22}$$

It is easy to see that $|x_n(\cdot)| + |u_n(\cdot)| \in L_2,\ \dot{x}_n = Ax_n + Bu_n$ for $t \geq 0, t \neq t_n$ and also that $x_n(0 + 0) = 0,\ x_n(t_n - 0) = x_n(t_n + 0)$. The last relation implies that the differential equation is valid for all $t \geq 0$. Therefore, separating the real $z^{(1)}_n$ and the imaginary $z^{(2)}_n$ parts of the process $z_n = z^{(1)}_n + \imath z^{(2)}_n$, we evidently have $z^{(1)}_n,\ z^{(2)}_n \in \mathfrak{M}$. (We recall that the subspace $\mathfrak{M}$ is described by relations (5.2) and (5.3) with $a = 0$.) By (5.19), $\mathcal{B}_\tau(z_n) = \mathcal{B}_\tau(z^{(1)}_n) + \mathcal{B}_\tau(z^{(2)}_n)$ and, consequently,

$$\liminf_{n\to\infty} \mathcal{B}_\tau(z_n) \geq \liminf_{n\to\infty} \mathcal{B}_\tau(z^{(1)}_n) + \liminf_{n\to\infty} \mathcal{B}_\tau(z^{(2)}_n)\,. \tag{5.23}$$

It is easy to see that $\left|x^0_n(\cdot)\right|_2 = |y|,\ \left|u^0_n(\cdot)\right|_2 = |v|$ and, by (5.21), $|\Delta x_n(\cdot)|_2 \leq 2t_n^{-1/2}c_2|y| \to 0,\ |\Delta u_n(\cdot)|_2 \leq 2t_n^{-1/2}c_2|y| \to 0$ as $n \to \infty$. Denoting $z^0_n := \left[x^0_n(\cdot), u^0_n(\cdot)\right]$ and $\Delta z_n := [\Delta x_n(\cdot), \Delta u_n(\cdot)]$, we have $\left|z^0_n\right|^2_H = |y|^2 + |v|^2,\ |\Delta z_n|_H \to 0$ as $n \to \infty$ and so

$$\left|\mathcal{B}_\tau(z_n) - \mathcal{B}_\tau\left(z^0_n\right)\right| \leq 2\left|B_\tau\left(\Delta z_n, z^0_n\right)\right| + \left|\mathcal{B}_\tau(\Delta z_n)\right|$$

$$\leq 2\,\|B_\tau\|\left|z^0_n\right|_H |\Delta z_n|_H + \|B_\tau\|\,|\Delta z_n|^2_H \to 0 \quad \text{as } n \to \infty.$$

This and the definition of $x^0_n(\cdot), u^0_n(\cdot)$ yield that

$$f := \lim_{n\to\infty} \mathcal{B}_\tau(z_n) = \lim_{n\to\infty} \mathcal{B}_\tau\left(z^0_n\right) = \lim_{n\to\infty} \left\{ t_n^{-1} \int_0^{t_n} g^0_\tau\left[e^{\imath\omega t}\ y,\ e^{\imath\omega t}\ v\right]\ dt \right.$$

$$\left. + t_n^{-1} \int_0^{t_n} \Delta g_\tau\left[e^{\imath\omega t}\ y,\ e^{\imath\omega t}\ v\right]\ dt \right\} = g^0_\tau(y, v) + \lim_{n\to\infty} t_n^{-1} \int_0^{t_n} \Delta g_\tau(t, y, v)\ dt.$$

Then, by (5.18) and (5.20), $f = g^0_\tau(y, v) < 0$. In light of (5.23), we see that $\liminf$ from (2.4) is negative at least for one of the two sequences $\{h_n\} := \{z^{(1)}_n\}$, $\{z^{(2)}_n\} \subset \mathfrak{M}$.

To complete the proof of assumption (C.1), it remains to show that $z^{(\nu)}_n \longrightarrow 0$ with respect to the weak topology of $L_2 \times L_2$ or, in other words, that $\langle z^{(\nu)}_n, \widetilde{z}\rangle \to 0$ as $n \to \infty$ for all $\widetilde{z} = [\widetilde{x}(\cdot),\ \widetilde{u}(\cdot)] \in L_2 \times L_2$. It is obvious that

$$|\langle z^{(\nu)}_n, \widetilde{z}\rangle| \le \underbrace{\int_0^\infty |x_n(t)|\,|\widetilde{x}(t)|\ dt}_{\alpha_n} + \underbrace{\int_0^\infty |u_n(t)|\,|\widetilde{u}(t)|\ dt}_{\beta_n},$$

$$\alpha_n \le \int_0^\infty |\Delta x_n(t)|\,|\widetilde{x}(t)|\ dt + \int_0^\infty \left|x^0_n(t)\right|\,|\widetilde{x}(t)|\ dt$$

$$\le |\Delta x_n(\cdot)|_2\,|\widetilde{x}(\cdot)|_2 + \int_0^T \left|x^0_n(t)\right|\,|\widetilde{x}(t)|\ dt + \int_T^\infty \left|x^0_n(t)\right|\,|\widetilde{x}(t)|\ dt$$

for any $T > 0$. Here $\left|x^0_n(t)\right| \le t_n^{-1/2}|y|$ for all $t \ge 0$ and $\left|x^0_n(\cdot)\right|_2 = |y|$. So we have

$$\alpha_n \le |\Delta x_n(\cdot)|_2\,|\widetilde{x}(\cdot)|_2 + t_n^{-1/2}|y| \int_0^T |\widetilde{x}(t)|\ dt + \left|x^0_n(\cdot)\right|_2 \left(\int_T^\infty |\widetilde{x}(t)|^2\ dt\right)^{1/2}$$

$$\le |\Delta x_n(\cdot)|_2\,|\widetilde{x}(\cdot)|_2 + \left(\frac{T}{t_n}\right)^{1/2} |y|\,|\widetilde{x}(\cdot)|_2 + |y| \left(\int_T^\infty |\widetilde{x}(t)|^2\ dt\right)^{1/2}.$$

Successively letting $n \to \infty$ and then $T \to \infty$, we see that $\alpha_n \to 0$ as $n \to \infty$. By analogy, $\beta_n \to 0$ as $n \to \infty$ that completes the proof of assumption (C.1).

Thus, all the assumptions of Theorem 2.2 are fulfilled and, by this theorem, Lemma 5.4 is valid. □

Remarks. 1. Lemma 5.4 remains valid if assumption (3.ii) is replaced by the following more general one.

(3.ii′) *For each* $\tau^* = \tau = \|\tau_i\| \in \mathbb{R}^k$, *denote* $g^0_\tau(\cdot) := g^0_0(\cdot) + \tau_1 g^0_1(\cdot) + \cdots + \tau_k g^0_k(\cdot)$, $\Delta g_\tau(\cdot) := \Delta g_0(\cdot) + \tau_1 \Delta g_1(\cdot) + \cdots + \tau_k \Delta g_k(\cdot)$. *Consider the collection* $\Theta = \{\tau\}$ *of all multipliers* $\tau \ge 0$ *for which the frequency criterion* (5.17) *is fulfilled with* $g := g^0_\tau$.

Given $\tau \in \Theta$, *we have* $\Delta g_\tau(t, x, u) \ge 0$ *for all* x, u *and almost all* $t \ge 0$.

Indeed, assumption (3.ii) was used only to demonstrate assumption (C.1) of Lemma 2.3; i.e., it was used to show that $\boxed{\tau \ge 0 \text{ and } \mathcal{B}_\tau(h) < 0} \Rightarrow f := \underline{\lim}_{n\to\infty} \mathcal{B}_\tau(h_n) < 0$ for the sequence $\{h_n\} \subset \mathfrak{M}$ constructed above. It has been actually demonstrated that $f < 0$ for any $\tau \overline{\in} \Theta$. So the role of (3.ii) was only to ensure the implication $\boxed{\tau \ge 0 \text{ and } \mathcal{B}_\tau(h) < 0} \Rightarrow \tau \overline{\in} \Theta$, which is clearly equivalent to the implication $\tau \in \Theta \Rightarrow \mathcal{B}_\tau(h) \ge 0$ $(\forall h \in \mathfrak{M})$ and now is still true. Indeed, if $\tau \in \Theta$ and $h \in \mathfrak{M}$, then $\mathcal{B}^0_\tau(h) \ge 0$ by Lemma 5.3 and $\Delta\mathcal{B}_\tau(h) \ge 0$ by (5.19) and (3.ii′). Therefore, $\mathcal{B}_\tau(h) \ge 0$.

2. The above considerations show that assumption (3.ii′) can be replaced by the still more general one: *given* $\tau \in \Theta$, *we have* $\mathcal{B}_\tau(h) \ge 0$ *for all* $h \in \mathfrak{M}$.

3. Shifting the functions (5.22) to the right by $s_n \ge 0$ and analyzing the above considerations, we see that assumption (3.iv) can be replaced by the following one.

(3.iv′) *Given* $y \in \mathbb{C}^l$ *and* $v \in \mathbb{C}^m$, *there exist infinite sequences of reals* $\{s_n\} \subset [0,\infty)$ and $\{t_n\} \subset (0,\infty)$ *such that* $t_n \to \infty$ *as* $n \to \infty$ *and*

$$t_n^{-1} \int_{s_n}^{s_n+t_n} \Delta g_i(t,y,v)\, dt \to 0 \quad \text{as} \quad n \to \infty \quad \text{for all} \quad i = 0,\dots,k.$$

5.4. Nonstationary linear-quadratic problem with coefficients, which become constant since some time instant. Consider the problem (5.1)–(5.4) and assume that

(4.i) *the matrix-valued functions* $A(\cdot)$, $B(\cdot)$, $G_i(\cdot)$, $Q_i(\cdot)$, *and* $\Gamma_i(\cdot)$ *are piecewise continuous. There exists a time instant* $t^0 \geq 0$ *such that they are constant in the domain* $t \geq t^0$, *i.e.,* $A(t) = A$, $B(t) = B$, $G_i(t) = G_i^0$, $Q_i(t) = Q_i^0$, *and* $\Gamma_i(t) = \Gamma_i^0$ *for all* $t \geq t^0$;

(4.ii) $\phi_i(t,x,u) = 0$ $(i = 0,\dots,k)$, *the pair* (A,B) *is stabilizable, and the system* $\dot{x} = A(t)x + B(t)u$, $0 \leq t \leq t^0$ *is controllable on any nontrivial subinterval* $[t_1,t_2] \subset \left[0,t^0\right]$; *i.e., given* $t_1,t_2 \in \left[0,t^0\right]$, $t_1 < t_2$, *and* $x_1,x_2 \in \mathbb{R}^l$, *there is a control* $u(\cdot)$, *which brings the system starting at time* t_1 *at* x_1 *to* x_2 *at time* t_2.[5]

To proceed further, we need some notation. Given a Lagrange multiplier $\tau = \|\tau_i\| \in \mathbb{R}^k$, the function $g_\tau(t,x,u) := g_0(t,x,u) + \sum_{i=1}^k \tau_i g_i(t,x,u) = x^*G_\tau(t)x + 2x^*Q_\tau(t)u + u^*\Gamma_\tau(t)u$ is independent of t in the domain $t \geq t^0$; i.e., $g_\tau(t,x,u) = g_\tau^0(x,u)$ for $t \geq t^0$.

(4.D) *Denote by* Θ *the collection of all Lagrange multipliers* $\tau = \|\tau_i\| \in \mathbb{R}^k$ *such that* $\Gamma_\tau(t) \geq 0$ *for* $t \geq 0$, *the frequency criterion* (5.17) *is fulfilled for* $g := g_\tau^0$, *and* $\tau_i \geq 0$ $(i = 1,\dots,k)$. *Introduce also the set* $\Theta^0 := \{\tau \in \Theta : \Gamma_\tau(t \pm 0) > 0$ *for all* $t \geq 0$ *and* (5.17) *is true for* $g(y,v) := g_\tau^0(y,v) - \delta\left(|y|^2 + |v|^2\right)$ *with some* $\delta > 0\}$.

Both the set Θ and the set Θ^0 can be empty. It is only to simplify the further formulations that we assume that

(4.iii) *Either* $\Theta = \emptyset$ *or* $\Theta^0 \neq \emptyset$ (*i.e.,* $\Theta \neq \emptyset \Rightarrow \Theta^0 \neq \emptyset$).

Let $\Theta^0 \neq \emptyset$, and consider $\tau \in \Theta^0$. By the Kalman–Yakubovich–Popov lemma [15, 16, 17, 18], there exist and are unique the real $l \times l$-matrix $P_\tau = P_\tau^*$ and the real $l \times m$ matrix r_τ such that

$$2x^*P_\tau\,(Ax + Bu) + g_\tau^0(x,u) = (u - r_\tau^*x)^*\,\Gamma_\tau^0\,(u - r_\tau^*x) \quad \forall x \in \mathbb{R}^l,\ u \in \mathbb{R}^m \tag{5.24}$$

and the system $\dot{x} = (A + Br_\tau^*)\,x$ is stable. There is known a number of efficient methods to calculate P_τ and r_τ [15, 16, 17, 18]. Introduce also the matrices

$$\begin{gathered}\mathcal{A}_\tau(t) := A(t) - B(t)\Gamma_\tau(t)^{-1}Q_\tau(t)^*, \qquad \mathcal{D}_\tau(t) := B(t)\Gamma_\tau(t)^{-1}B(t)^*\,, \\ \mathcal{C}_\tau(t) := G_\tau(t) - Q_\tau(t)\Gamma_\tau(t)^{-1}Q_\tau(t)^*\end{gathered} \tag{5.25}$$

and define $l \times l$ matrices $X(t)$ and $\Psi(t)$ as the solution of the Cauchy problem

$$\begin{array}{rcl} \dot{X}_\tau(t) &=& \mathcal{A}_\tau(t)X_\tau(t) + \mathcal{D}_\tau(t)\Psi_\tau(t), \\ \dot{\Psi}_\tau(t) &=& \mathcal{C}_\tau(t)X_\tau(t) - \mathcal{A}_\tau(t)^*\Psi_\tau(t), \end{array} \quad 0 \leq t \leq t^0, \tag{5.26}$$

$$X_\tau\left(t^0\right) = I, \qquad \Psi_\tau\left(t^0\right) = -P_\tau.$$

[5] See, for example, [20, pp. 92–96] about criteria for controllability.

The last assumption is the following.

(4.iv) *Given* $\tau \in \Theta^0$, *we have*

$$\det X_\tau(t) \neq 0 \qquad \left(\forall t \in \left(0, t^0\right]\right). \tag{5.27}$$

If $\Theta^0 = \emptyset$, this assumption is meant to be omitted.

In (5.27), the time instant t^0 can be chosen in various ways. However, the validity of (4.iv) is independent of this choice (see Lemma 5.6 below).

LEMMA 5.5. *Let the above assumptions* (4.i)–(4.iv) *hold. Then assertions* $(\mathfrak{A})$ *and* $(\mathfrak{B})$ *are true provided that the problem* (5.1)–(5.4) *is regular.*

We preface the proof of this lemma by two preliminary facts. To state them, we recall that $\mathfrak{M}$ is the collection of all processes $[x(\cdot), u(\cdot)]$ satisfying (5.2) and (5.3) with $a = 0$.

LEMMA 5.6. *Let assumptions* (4.i), (4.ii) *be fulfilled and a multiplier* $\tau \in \Theta^0$ *be given. Define the mapping* $\mathcal{B}_\tau : H \to \mathbb{R}$ *by* (5.14). *Then the relation*

$$\mathcal{B}_\tau(h) \geq 0 \qquad \forall h \in \mathfrak{M} \tag{5.28}$$

is true if and only if condition (5.27) *is fulfilled.*

Note that relation (5.28) does not use the time instant t^0. So the choice of this instant does not affect the validity of condition (5.27).

Proof of Lemma 5.6. (5.27) $\Rightarrow$ (5.28). Let $h = [x(\cdot), u(\cdot)] \in \mathfrak{M}$; i.e., let relations (5.2) and (5.3) be true with $a = 0$. By (5.5), we have

$$\mathcal{B}_\tau(h) = \int_0^{t^0} g_\tau\left[t, x(t), u(t)\right]\, dt + \int_{t^0}^{\infty} g_\tau^0\left[x(t), u(t)\right]\, dt. \tag{5.29}$$

Put $x := x(t)$ and $u := u(t)$ into (5.24) and recall that $\Gamma_\tau^0 \geq 0$ by the definition of the set $\Theta \supset \Theta^0 \ni \tau$. Then, for $t \geq t^0$, we get

$$\begin{aligned} &\frac{d}{dt}\left[x(t)^* P_\tau x(t)\right] + g_\tau^0\left[x(t), u(t)\right] \\ &= \left[u(t) - r_\tau^*(t)x(t)\right]^* \Gamma_\tau^0\left[u(t) - r_\tau^*(t)x(t)\right] \geq 0. \end{aligned} \tag{5.30}$$

Integrating over the interval $\left[t^0, \infty\right)$ results in the inequality

$$\int_{t^0}^{\infty} g_\tau^0\left[x(t), u(t)\right]\, dt \geq x\left(t^0\right)^* P_\tau x\left(t^0\right). \tag{5.31}$$

This permits us to establish a below bound of the quantity (5.29)

$$\mathcal{B}_\tau(h) \geq \mathcal{I} := \int_0^{t^0} g_\tau\left[t, x(t), u(t)\right]\, dt + x\left(t^0\right)^* P_\tau x\left(t^0\right). \tag{5.32}$$

Here

$$\dot{x}(t) = A(t)x(t) + B(t)u(t), \quad 0 \leq t \leq t^0, \quad x(0) = 0. \tag{5.33}$$

It remains to note that relation (5.27) is sufficient and necessary for nonnegativity of the functional $\mathcal{I} = \mathcal{I}\left[x(\cdot), u(\cdot)\right]$ from (5.32) over processes $\left[x(\cdot), u(\cdot)\right]$, $0 \leq t \leq t^0$ satisfying (5.33) [16, 32, 33].

(5.28) $\Rightarrow$ (5.27). By the above citation, it suffices to show that $\mathcal{I} \geq 0$ for any process $[x(\cdot), u(\cdot)]$, $0 \leq t \leq t^0$ satisfying (5.33). Consider such a process. For $t \geq t^0$, define $x(t)$ to be the solution of the Cauchy problem $\dot{x} = (A + Br_\tau^*)\, x,\ x\left(t^0 + 0\right) = x\left(t^0 - 0\right)$ and put $u(t) := r_\tau^* x(t)$ where r_τ^* is the matrix from (5.24). Since the system $\dot{x} = (A + Br_\tau^*)\, x$ is stable, relations (5.2) and (5.3) are true with $a = 0$, i.e., $h := [x(\cdot), u(\cdot)] \in \mathfrak{M}$. Put $x := x(t)$ and $u := u(t)$ into (5.24) and recall that $u(t) = r_\tau^* x(t)$ for $t \geq t^0$. Then we see that, in (5.30)–(5.32), the inequality sign can be replaced by the equality one. In particular, $\mathcal{I} = \mathcal{B}_\tau(h)$ by (5.32) where $\mathcal{B}_\tau(h) \geq 0$ due to (5.28). Thus, we have the inequality desired: $\mathcal{I} \geq 0$. □

In what follows, the symbol $\rightharpoonup$ will denote the convergence with respect to the weak topology of the space $H := L_2\left([0, +\infty) \to \mathbb{R}^l\right) \times L_2\left([0, +\infty) \to \mathbb{R}^m\right)$.

LEMMA 5.7. *Let assumptions* (4.i)–(4.iv) *be fulfilled and a multiplier* $\tau \in \mathbb{R}^k$ *be given. Define the mapping* $\mathcal{B}_\tau : H \to \mathbb{R}$ *by* (5.14) *and consider the set* Θ *from* (4.D).

If $\tau \overline{\in} \Theta$ *and* $\tau_i \geq 0$ *for all* $i = 1, \ldots, k$, *then*

$$f := \liminf_{h \rightharpoonup 0,\ h \in \mathfrak{m}} \mathcal{B}_\tau(h) < 0, \tag{5.34}$$

where lim inf *is with respect to the weak topology.*

Proof. By (4.D), either (1) $v^* \Gamma_\tau(\bar{t} + 0) v < 0$ for some $\bar{t} \geq 0$ and $v \in \mathbb{R}^m$ or (2) relations (5.20) are true for some $\omega \in \mathbb{R},\ y \in \mathbb{C}^l$, and $v \in \mathbb{C}^m$. Consider first the case (1). By (4.ii), there exists an $l \times m$ matrix C such that the equation $\dot{x} = (A + BC^*)\, x$ is stable. Denote by $x(\cdot|a)$ its solution with $x(0|a) = a$ and put $u(\cdot|a) := C^* x(\cdot|a)$. Then relations (5.21) are valid. Choose an instant t_* such that $t_* \geq t^0$ and $t_* \geq \bar{t} + 2$. Given $\epsilon \in (0, 1]$, we put $u_\epsilon(t) := \epsilon^{-1/2} v$ if $\bar{t} \leq t < \bar{t} + \epsilon$, $u_\epsilon(t) := -\epsilon^{-1/2} v$ if $\bar{t} + \epsilon \leq t < \bar{t} + 2\epsilon$, and $u_\epsilon(t) := 0$ if $0 \leq t < \bar{t}$ or $\bar{t} + 2\epsilon \leq t < t_*$. Consider the solution $x_\epsilon(\cdot)$ of the Cauchy problem $\dot{x}_\epsilon(t) = A(t) x_\epsilon(t) + B(t) u_\epsilon(t),\ 0 \leq t \leq t_*,\ x_\epsilon(0) = 0$. As is well known, $|x_\epsilon(\cdot)|_\infty \leq \mathcal{K} |u_\epsilon(\cdot)|_1$ and so

$$\max_{t \in [0, t_*]} |x_\epsilon(t)| \leq \mathcal{K} |u_\epsilon(\cdot)|_1 = \mathcal{K} \epsilon^{-1/2} |v| 2\epsilon \to +0 \ \text{ as } \ \epsilon \to +0. \tag{5.35}$$

Denote $a_\epsilon := x_\epsilon(t_*)$ and define $x_\epsilon(t)$, $u_\epsilon(t)$ for $t \geq t_*$ as follows: $x_\epsilon(t) := x(t - t_* | a_\epsilon)$, $u_\epsilon(t) := u(t - t_* | a_\epsilon)$. It is easy to see that $h_\epsilon := [x_\epsilon(\cdot), u_\epsilon(\cdot)] \in \mathfrak{M}$. By (5.35), $a_\epsilon \to 0$ as $\epsilon \to +0$ and so, due to (5.21) and (5.35), we have

$$|x_\epsilon(\cdot)|_2 \to 0,\ \int_{t_*}^{\infty} |u_\epsilon(t)|^2\ dt \to 0 \ \text{ as } \ \epsilon \to +0,\ \int_0^{t_*} |u_\epsilon(t)|^2\ dt = 2|v|^2. \tag{5.36}$$

Taking into account (5.14), we get

$$\begin{aligned} \mathcal{B}_\tau(h_\epsilon) \ = \ & \underbrace{\int_0^\infty x_\epsilon(t)^* G_\tau(t) x_\epsilon(t)\ dt}_{\delta_1} \ + \ 2 \underbrace{\int_0^\infty x_\epsilon(t)^* Q_\tau(t) u_\epsilon(t)\ dt}_{\delta_2} \\ & + \ \underbrace{\int_{t_*}^\infty u_\epsilon(t)^* \Gamma_\tau(t) u_\epsilon(t)\ dt}_{\delta_3} \ + \ \underbrace{\int_0^{t_*} u_\epsilon(t)^* \Gamma_\tau(t) u_\epsilon(t)\ dt}_{\delta_4}. \end{aligned}$$

Here $|\delta_1| \leq \sup_{t \geq 0} |G_\tau(t)|\, |x_\epsilon(\cdot)|_2^2 \to 0$ as $\epsilon \to +0$ by (5.36) and also $\delta_2 \to 0,\ \delta_3 \to 0$ as $\epsilon \to +0$ due to analogous reasons. Furthermore,

$$\delta_4 = \epsilon^{-1} \left[\int_{\bar{t}}^{\bar{t} + \epsilon} v^* \Gamma_\tau(t) v\ dt + \int_{\bar{t} + \epsilon}^{\bar{t} + 2\epsilon} (-v)^* \Gamma_\tau(t) (-v)\ dt \right] \to 2 v^* \Gamma_\tau(\bar{t} + 0) v \ \text{ as } \ \epsilon \to +0.$$

Thus,

$$\mathcal{B}_\tau\,(h_\epsilon) \to 2v^*\Gamma_\tau(\bar t+0)v<0 \ \text{ as } \ \epsilon\to+0. \tag{5.37}$$

To complete the proof, it remains to show that $h_\epsilon \rightharpoonup 0$ as $\epsilon\to+0$. For $\widetilde h = [\widetilde x(\cdot),\widetilde u(\cdot)]\in H$, we have

$$\Big\langle h_\epsilon,\widetilde h\Big\rangle = \underbrace{\int_0^\infty \widetilde x(t)^* x_\epsilon(t)\,dt}_{\Delta_1} + \underbrace{\int_{t_*}^\infty \widetilde u(t)^* u_\epsilon(t)\,dt}_{\Delta_2} + \underbrace{\int_0^{t_*} \widetilde u(t)^* u_\epsilon(t)\,dt}_{\Delta_3}.$$

Here $|\Delta_1|\le|\widetilde x(\cdot)|_2\,|x_\epsilon(\cdot)|_2\to 0$ as $\epsilon\to+0$ due to (5.36). Likewise, $\Delta_2\to 0$ as $\epsilon\to+0$ and

$$|\Delta_3|\le\int_{\bar t}^{\bar t+2\epsilon}|u_\epsilon(t)|\,|\widetilde u(t)|\;dt=\epsilon^{-1/2}|v|\int_{\bar t}^{\bar t+2\epsilon}|\widetilde u(t)|\;dt$$

$$\le\epsilon^{-1/2}|v|\left(\int_{\bar t}^{\bar t+2\epsilon}\;dt\right)^{1/2}\left(\int_{\bar t}^{\bar t+2\epsilon}|\widetilde u(t)|^2\;dt\right)^{1/2}=\sqrt{2}|v|\left(\int_{\bar t}^{\bar t+2\epsilon}|\widetilde u(t)|^2\;dt\right)^{1/2}\to 0$$

as $\epsilon\to+0$. Thus, $h_\epsilon\rightharpoonup 0$ as $\epsilon\to+0$ where $h_\epsilon\in\mathfrak{M}$ as it was shown above. Therefore, the quantity f in (5.34) does not exceed the limit $\lim_{\epsilon\to+0}\mathcal{B}_\tau\,(h_\epsilon)$, which is negative by (5.37). So relation (5.34) is true in case (1).

Consider case (2). It was shown in subsection 5.3 that there exists a sequence $h_n=[x_n(\cdot),u_n(\cdot)]\in\mathfrak{M},\ n=1,2,\dots$ such that $h_n\rightharpoonup 0$ as $n\to\infty,\ \dot x_n(t)=Ax_n(t)+Bu_n(t),\ 0\le t<\infty,\ x_n(0)=0$, and

$$\liminf_{n\to\infty}\underbrace{\int_0^\infty g_\tau^0\,[x_n(t),u_n(t)]\;dt}_{q_n}<0.$$

Shift the functions $x_n(\cdot)$ and $u_n(\cdot)$ to the right $\widetilde x_n(t):=x_n\big(t-t^0\big),\ \widetilde u_n(t):=u_n\big(t-t^0\big)$ if $t\ge t^0$ and $\widetilde x_n(t):=0,\ \widetilde u_n(t):=0$ otherwise. It is obvious that $\widetilde h_n:=[\widetilde x_n(\cdot),\widetilde u_n(\cdot)]\in\mathfrak{M},\ \widetilde h_n\rightharpoonup 0$ as $n\to\infty$ and, by (5.5), $\mathcal{B}_\tau(\widetilde h_n)=q_n$. It remains to note that the quantity f in (5.34) does not exceed $\liminf_{n\to\infty}\mathcal{B}_\tau(h_n)=\liminf_{n\to\infty}q_n<0$. □

Proof of Lemma 5.5. The reduction to the abstract problem (1.1) is performed just as it was done at the beginning of the section. Equip the space H with the weak topology. The verification of assumptions (A) and (B) of Theorem 2.2 is performed just as it was done in subsection 5.1. To prove (C), consider a multiplier $\tau=\|\tau_i\|\in\mathbb{R}^k$ with $\tau_i\ge 0$. We have to demonstrate relation (2.3), which is equivalent to the implication

$$f:=\liminf_{h\rightharpoonup 0,\ h\in\mathfrak{m}}\mathcal{B}_\tau(h)\ge 0\Rightarrow\mathcal{B}_\tau(h)\ge 0\quad\forall h\in\mathfrak{M}. \tag{5.38}$$

By Lemma 5.7, $f\ge 0\Rightarrow\tau\in\Theta$. So it suffices to show that

$$\tau\in\Theta\Rightarrow\mathcal{B}_\tau(h)\ge 0\quad(\forall h\in\mathfrak{M}). \tag{5.39}$$

For $\tau\in\Theta^0$, the conclusion from (5.39) is justified by Lemma 5.6.

Let $\tau\in\Theta$. By (4.iii), $\Theta^0\ne\emptyset$ and we can choose $\tau^0\in\Theta^0$. It easily follows from (4.D) that $\tau+\epsilon\tau^0\in\Theta^0$ for any $\epsilon>0$. So $\mathcal{B}_{\tau+\epsilon\tau^0}(h)\ge 0$ for all $h\in\mathfrak{M}$. By letting $\epsilon\to+0$, we get the conclusion from (5.39). Thus, (5.39) is true and all the assumptions of Theorem 2.2 are fulfilled. By this theorem, Lemma 5.5 is valid. □

5.5. Counterexample. In the conclusion of the section, we show that, in general, the method (I)–(IV) fails to be applicable to the problem (5.1)–(5.4). The following is the counterexample required

$$\mathcal{G}_0 := -\int_0^T \underbrace{2x_1x_2 + x_2^2 + x_1u}_{\mu}\, dt \to \min \quad \text{subject to} \tag{5.40}$$

$$\dot{x}_1 = x_2, \quad \dot{x}_2 = u, \quad 0 \le t < \infty, \quad x_1(0) = x_2(0) = 0, \quad |x_1(\cdot)| + |x_2(\cdot)| + |u(\cdot)| \in L_2, \tag{5.41}$$

$$\mathcal{G}_1 := 2\int_0^T \underbrace{x_2^2 + x_1u}_{\sigma}\, dt - 1 \le 0,\ \mathcal{G}_2 := -\int_0^T \underbrace{2x_2u + x_2^2 + x_1u}_{\eta}\, dt + 1 \le 0, \tag{5.42}$$

where the real $T > 0$ is fixed. We shall show that (1) *the problem* (5.40)–(5.42) *is regular and has a solution*, but (2) *the method* (I)–(IV) *fails to be applicable to this problem.*

To this end, note first that, by (5.41), $\sigma = d/dt\,(x_1x_2)$, $\eta = d/dt\,\big(x_2^2 + x_1x_2\big)$, and $\mu = d/dt\,\big(x_1^2 + x_1x_2\big)$. So the problem can be rewritten in terms of the state $y := [x_1(T), x_2(T)]$

$$f(y) := -y_1^2 - y_1y_2 \to \min \text{ subject to } y = \|y_i\| \in Z := \mathbb{R}^2, \quad g_1(y) := 2y_1y_2 \le 1, \quad g_2(y) := -y_2^2 - y_1y_2 \le -1. \tag{5.43}$$

(Relations (5.41) do not imply any restrictions on y due to controllability of the object $\dot{x}_1 = x_2,\ \dot{x}_2 = u$.) Consequently, it suffices to prove assertions (1) and (2) with respect to the problem (5.43).

(1) Since $g_1(0,2) = 0 < 1$ and $g_2(0,2) = -4 < -1$, the problem (5.43) is regular. It is straightforward to calculate that $f = -\frac{1}{2}g_1g_2/\left(g_2 + \frac{1}{2}g_1\right)$ for all y_1, y_2. To estimate the infimum of f in (5.43), consider the following chain of apparent implications, which starts with the inequalities $g_1 \le 1$ and $g_2 \le -1$ from (5.43)

$$\boxed{\begin{array}{c}\frac{2g_2}{g_2-1} \ge 1 \ge g_1,\\ g_2 - 1 < 0\end{array}} \Leftarrow \boxed{g_1 \le 1,\ g_2 \le -1} \Rightarrow \boxed{g_2 + 1/2g_1 < 0}$$
$$\Rightarrow \boxed{2g_2 \le (g_2 - 1)\, g_1} \Rightarrow \boxed{\begin{array}{c}2\,(g_2 + 1/2g_1)\\ \le g_1g_2\end{array}} \Rightarrow \boxed{\begin{array}{c}\frac{g_1g_2}{2\left(g_2+\frac{1}{2}g_1\right)}\\ = -f \le 1\end{array}}.$$

Thus, $f \ge -1$ for all points $y = (y_1,\ y_2)$ that satisfy the constraints from (5.43). Since $f\left(1/\sqrt{2},\ 1/\sqrt{2}\right) = -1$, $g_1\left(1/\sqrt{2},\ 1/\sqrt{2}\right) = 1$, and $g_2\left(1/\sqrt{2},\ 1/\sqrt{2}\right) = -1$, we see that the point $y_1 = 1/\sqrt{2},\ y_2 = 1/\sqrt{2}$ is a solution of the problem (5.43).

(2) Given a Lagrange multiplier $\tau = \|\tau_i\| \in \mathbb{R}^2$, the Lagrangian function $S(\tau, y_1, y_2) := f + \tau_1g_1 + \tau_2g_2 = -y_1^2 - y_1y_2 + 2\tau_1y_1y_2 - \tau_2\left(y_2^2 + y_1y_2\right) - \tau_1 + \tau_2$ is a quadratic form in y_1 with the negative main coefficient. So $\inf_{y\in\mathbb{R}^2} S\,(\tau, y_1, y_2) = -\infty$ and relation (1.12) fails to be true because its left-hand side is finite as it was shown above. Therefore, the method (I)–(IV) is valid neither for the problem (5.43) nor for (5.40)–(5.42). □

6. An example of application of the method (I)–(IV). Return to consideration of the problem (5.1)–(5.4) under the assumptions of subsection 5.4. The goal of this section is to illustrate in general outline the usefulness of the method (I)–(IV) for solution of linear-quadratic problems with quadratic constraints. To avoid consideration of details that are not directly related to this goal we shall impose additional simplifying assumptions on the problem. In what follows, we shall use the notations from (4.D), (5.24)–(5.26).

LEMMA 6.1. *Let assumptions* (4.i), (4.ii) *be fulfilled and* $\Theta^0 \neq \emptyset$. *Suppose that the problem* (5.1)–(5.4) *is regular and* $\det X_\tau(t) \neq 0$ *for all* $\tau \in \Theta^0$ *and* $t \in \left[0, t^0\right]$. *Denote* $R_\tau(t) := -\Psi_\tau(t) X_\tau(t)^{-1}$ $(t \in [0, t^0])$, *and assume that the problem*

$$b(\tau) := a^* R_\tau(0) a - \sum_{i=1}^{k} \tau_i \gamma_i \ \to \sup \quad \textit{subject to} \ \ \tau \in \Theta^0 \tag{6.1}$$

has a solution $\tau^0 \in \Theta^0$. (*Here* a *is the initial state from* (5.3).)

Then there exists and is unique the optimal process in the problem (5.1)–(5.4). *This process is generated by the closed-loop controller*

$$\begin{gathered} u = q_{\tau^0}(t)^* x, \\ \textit{where} \ \ q_\tau(t) := \overline{r}_\tau(t) \ \ \textit{for} \ \ 0 \leq t \leq t^0 \ \ \textit{and} \ \ q_\tau(t) := r_\tau \ \ \textit{for} \ \ t > t^0. \end{gathered} \tag{6.2}$$

Here r_τ *is the matrix from* (5.24) *and*

$$\overline{r}_\tau(t) := -\left[Q_\tau(t) + R_\tau(t) B(t)\right] \Gamma_\tau(t)^{-1}. \tag{6.3}$$

Remarks. 1. It readily follows from (5.26) and is also well known that the function $R_\tau(\cdot)$ is the solution of the matrix Riccati equation

$$\dot{R}_\tau(t) + R_\tau(t) A(t) + A(t)^* R_\tau(t) + G_\tau(t) = \overline{r}_\tau(t) \Gamma_\tau(t) \overline{r}_\tau(t)^*, \quad R_\tau\left(t^0\right) = P_\tau. \tag{6.4}$$

Here $\overline{r}_\tau(t)$ is supposed to be replaced by the right-hand side of (6.3) and P_τ is the matrix from (5.24).

2. In (6.1), the domain $\Theta^0 \subset \mathbb{R}^k$ is convex that easily follows from (4.D). It will be shown below that the function $b(\cdot)$ is concave. So the problem (6.1) can be solved by means of the methods of convex programming.

3. Consider the matrices P_τ and r_τ from (5.24) and put $R_\tau := P_\tau$, $\overline{r}_\tau(t) := r_\tau$ for $t \geq t^0$. Then relations (6.3) and (6.4) remain valid for $t \geq t^0$ [16, 17]. This implies that the choice of the time instant t^0 does not affect the coefficient (6.2).

4. It will be shown below that, in the case under consideration, (6.1) is in fact a concretized form of the dual problem (1.7). On the whole, Lemma 6.1 offers the following concretization of the method (I)–(IV).

(1) *Form the set* Θ^0 *in accordance with* (4.D).

(2) *Find a solution* $\tau^0 \in \Theta^0$ *of the problem* (6.1) *where the quantity* $a^* R_\tau(0) a$ *can be calculated as follows. Given* $\tau \in \Theta^0$, *determine the solution* $P_\tau = P_\tau^*$ *and* r_τ *of equations* (5.24) *such that the system* $\dot{x} = (A + B r_\tau^*)\, x$ *is stable,*[6] *and then find the solution* $R_\tau(\cdot)$ *of the Cauchy problem* (6.4).

(3) *Determine* $\overline{r}_\tau(\cdot)$ *by* (6.3). *The optimal process is generated by* (6.2).

Proof of Lemma 6.1. Since $\Theta^0 \neq \emptyset$ and $\det X_\tau(t) \neq 0$ for all $\tau \in \Theta^0$, $t \in \left[0, t^0\right]$, assumptions (4.iii) and (4.iv) are fulfilled. So, by Lemma 5.5, the method (I)–(IV)

[6] See, for example, [15, 16, 17, 18] for methods to calculate P_τ and r_τ.

is applicable and we can use it. In (1.4), now $S(\tau,z)=\mathcal{B}_\tau(z)-\sum_{i=1}^k\tau_i\gamma_i$ due to (1.3), (1.13), (4.ii), and (5.6). Given $\tau\overline{\in}\Theta,\tau\geq 0$, we have $\mathcal{B}_\tau(h)<0$ for some $h\in\mathfrak{M}$ by Lemma 5.7 and (2.2). So, choosing $z'\in Z$, we get $\inf_{z\in Z}S(\tau,z)+\sum_{i=1}^k\tau_i\gamma_i\leq\inf_{\rho\in\mathbb{R}}\mathcal{B}_\tau(z'+\rho h)=\inf_{\rho\in\mathbb{R}}\left[\rho^2\mathcal{B}_\tau(h)+2\rho B_\tau(h,z')+\mathcal{B}_\tau(z')\right]=-\infty$. This means that now the dual problem (1.7) takes the form

$$S_0(\tau)\to\max\ \text{ subject to }\ \tau\in\Theta. \tag{6.5}$$

Here the function $S_0(\tau)$ is concave as the infimum (1.6) of the functions (1.4), which are linear in τ. It easily follows from (4.D) that (1) the sets Θ and Θ^0 are convex; (2) $\mathrm{ri}\Theta^0\subset\Theta^0\subset\Theta=\overline{\Theta}$; and (3) $(1-\epsilon)\tau+\epsilon\tau^0\in\Theta^0$ for any $\epsilon\in(0,1)$, $\tau\in\Theta$, $\tau^0\in\Theta^0$, and so $\overline{\Theta^0}=\Theta$. From this it follows that [8, pp. 46, 55]

$$\max_{\tau\in\Theta}S_0(\tau)=\sup_{\tau\in\Theta^0}S_0(\tau). \tag{6.6}$$

We recall that now the subspace $Z=\{z=[x(\cdot),u(\cdot)]\}$ is described by (5.2) and (5.3). Let $\tau\in\Theta^0$ and $z\in Z$. Taking into account (5.2) and (6.3), (6.4), it is straightforward to compute that

$$\frac{d}{dt}\left[x(t)^*R_\tau(t)x(t)\right]+g_\tau\left[t,x,u\right]=\left[u-\overline{r}_\tau(t)^*x\right]^*\Gamma_\tau(t)\left[u-\overline{r}_\tau(t)^*x\right]$$

for $t\leq t^0$, where $x=x(t)$ and $u=u(t)$. Integrating both this equality and (5.30), we get by (5.29)

$$\begin{aligned}S(\tau,z)&=a^*R_\tau(0)a-\sum_{i=1}^k\tau_i\gamma_i\\&+\int_0^\infty\left[u(t)-q_\tau(t)^*x(t)\right]^*\Gamma_\tau(t)\left[u(t)-q_\tau(t)^*x(t)\right]\,dt,\end{aligned} \tag{6.7}$$

where $q_\tau(\cdot)$ is defined from (6.2). Since $\Gamma_\tau(t\pm 0)>0$ by the definition of the set $\Theta^0\ni\tau$, we have $S_0(\tau):=\inf_{z\in Z}S(\tau,z)=b(\tau)$ where $b(\tau)$ was defined in (6.1). This and (6.6) prove that the multiplier τ^0 from the statement of Lemma 6.1 is a solution for the dual problem (6.5). So we can use τ^0 in the items (III) and (IV) of the method (I)–(IV).

Show that the problem (5.1)–(5.4) has a solution. To this end, consider a minimizing sequence of processes $\{z_n\}_{n=1}^\infty$; i.e., $z_n=[x_n(\cdot),u_n(\cdot)]$ satisfies (5.2), (5.3) (with $x(\cdot):=x_n(\cdot)$, $u(\cdot):=u_n(\cdot)$) and

$$\mathcal{G}_i\left[z_n\right]\leq 0\ \ \forall\, i=1,\dots,k,\ n=1,2,\dots,\ \ \mathcal{G}_0\left[z_n\right]\to\inf_{z\in D}\mathcal{G}_0(z)\ \text{ as }\ n\to\infty, \tag{6.8}$$

where $\mathcal{G}_i$ is defined by (5.4). By the item (IV), $S(\tau^0,z_n)\to S_0(\tau^0)=a^*R_\tau(0)a-\sum_{i=1}^k\tau_i\gamma_i$ as $n\to\infty$. Then, due to (6.7), $\Delta u_n(\cdot):=u_n(\cdot)-q_{\tau^0}(t)^*x_n(\cdot)\to 0$ as $n\to\infty$ with respect to the L_2-norm. Put $x(\cdot):=x_n(\cdot)$ and $u(\cdot):=u_n(\cdot)=q_{\tau^0}(\cdot)^*x_n(\cdot)+\Delta u_n(\cdot)$ into (5.2)

$$\dot{x}_n(t)=\underbrace{\left[A(t)+B(t)q_{\tau^0}(t)^*\right]}_{S(t)}x_n(t)+B(t)\Delta u_n(t),\quad 0\leq t<\infty,\quad x_n(0)=a.$$

For $t\geq t_0$, the matrices $S(t)=S^0$ and $B(t)=B$ are constant due to (4.i) and (6.2). We recall also that the equation $\dot{x}=S^0x=(A+Br_{\tau^0}^*)x$ is stable by the

definition of the matrix r_{τ^0}. So $|x_n(\cdot)-x(\cdot)|_2 \to 0$ as $n \to \infty$ where $x(\cdot)$ is the solution of the Cauchy problem $\dot{x} = S(t)x(t)$, $0 \le t < \infty$, $x(0) = a$. Denoting $u(t) := q_{\tau^0}(t)^* x(t)$, we see that the pair $z_\infty := [x(\cdot), u(\cdot)]$ satisfies (5.2), (5.3) and $|u_n(\cdot) - u(\cdot)|_2 = |\Delta u_n(\cdot) + q_{\tau^0}(\cdot)^* [x_n(\cdot) - x(\cdot)]|_2 \to 0$ as $n \to \infty$. By (5.4), the functionals $\mathcal{G}_i$ are continuous with respect to the $L_2 \times L_2$ norm. So passing to the limit in (6.8), we get $\mathcal{G}_i[z_\infty] \le 0$ for $i = 1, \dots, k$ and $\mathcal{G}_0[z_\infty] = \inf_{z \in D} \mathcal{G}_0(z)$ where $z_\infty \in Z$. Thus, z_∞ is a solution of the problem (5.1)–(5.4).

Consider an optimal process $z^0 = \left[x^0(\cdot), u^0(\cdot)\right]$ in the problem (5.1)–(5.4). In correspondence with the item (III), $S(\tau^0, z^0) = \inf_{z \in Z} S(\tau^0, z) = a^* R_{\tau^0}(0) a - \sum_{i=1}^k \tau_i \gamma_i$ and so, by (6.7), $u^0(t) = q_{\tau^0}(t)^* x^0(t)$ for all $t \ge 0$. This implies that this process is generated by the controller (6.2) and is thereby unique. □

Example. Consider the following problem:

$$\text{minimize } \mathcal{G}_0 := \int_0^{+\infty} \left[u(t)^2 + \sigma x(t)^2\right] \, dt \text{ subject to} \tag{6.9}$$

$$\dot{x}(t) = u(t), \quad 0 \le t < \infty, \quad x(0) = a, \quad |x(\cdot)| + |u(\cdot)| \in L_2, \tag{6.10}$$

$$\mathcal{G}_1 := \int_T^{+\infty} u(t)^2 \, dt - \nu \int_T^{+\infty} x(t)^2 \, dt \le 0, \quad \mathcal{G}_2 := \int_0^T x(t)^2 \, dt \le \alpha, \tag{6.11}$$

where the reals $\sigma > 0, \nu > 0, \alpha > 0, T > 0$, and $a \in \mathbb{R}$ are given. This problem is to minimize the convex functional $\mathcal{G}_0$ on the domain $D := \{z = [x(\cdot), u(\cdot)] \in H := W_2^1(0, +\infty) \times L_2(0, +\infty) : (6.10) \text{ and } (6.11) \text{ are true}\}$, which is not convex. To prove this note first that the domain D is evidently closed with respect to the norm of H. So, were the domain D convex, it would be weakly closed [3, p. 4]. But it is not weakly closed. Indeed, consider a process $z = [x(\cdot), u(\cdot)] \in D' := \{z \in H : (6.10) \text{ is true and } \mathcal{G}_2 \le \alpha\}$. Given $n > 0$ and $\theta > T$, we put

$$\chi_n(t) := \begin{cases} n^{-1} & \text{for } 0 \le t < n, \\ -n^{-1} & \text{for } n \le t < 2n, \\ 0 & \text{otherwise,} \end{cases}$$

$$\Delta u_{n,\theta}(t) := \chi_n(t-\theta),$$
$$\Delta x_{n,\theta}(t) := \begin{cases} (t-\theta)\chi_n(t-\theta) & \text{for } t \le \theta + n, \\ (t-\theta-2n)\chi_n(t-\theta) & \text{for } \theta + n < t. \end{cases}$$

Denote by $\rightharpoonup$ the convergence with respect to the weak topology of H. By analogy with (5.8), we have $\Delta z_{n,\theta} := [\Delta x_{n,\theta}(\cdot), \Delta u_{n,\theta}(\cdot)] \rightharpoonup 0$ as $\theta \to \infty$ provided that the real n is fixed. It is clear that $z_{n,\theta} := z + \Delta z_{n,\theta} \in D'$ and

$$\begin{aligned} \mathcal{G}_1(z_{n,\theta}) &= \underbrace{\mathcal{G}_1(z) + 2n^{-1} - \nu\frac{2n}{3}}_{\Delta(n)} \\ &+ \underbrace{2\int_T^{+\infty} \left[u(t)\Delta u_{n,\theta}(t) - \nu x(t)\Delta x_{n,\theta}(t)\right] \, dt}_{\delta(n,\theta)}, \end{aligned} \tag{6.12}$$

where $\delta(n, \theta) \to 0$ as $\theta \to \infty$ because $\Delta z_{n,\theta} \rightharpoonup 0$ as $\theta \to \infty$. Choose $n^- > 0$ and $n^+ > 0$ such that $\Delta(n^-) < 0$ and $\Delta(n^+) > 0$. Then $\mathcal{G}_1\left(z_{n^-,\theta}\right) < 0$ and $\mathcal{G}_1\left(z_{n^+,\theta}\right) > 0$

provided that the real θ is sufficiently large. In other words, $z_{n^-,\theta} \in D$ and $z_{n^+,\theta} \in D' \setminus D$ where $z_{n,\theta} \rightharpoonup z \in D'$ as $\theta \to \infty$. Thus the set D and its complement $D' \setminus D$ are weakly dense in D'. This proves that the domain D is not weakly closed.

LEMMA 6.2. *There exists and is unique the optimal process in the problem* (6.9)–(6.11). *This process is generated by the closed-loop controller* $u = q(t)x$, *where the function* $q(t)$ *is determined as follows. Put*

$$p_0 := \begin{cases} \sqrt{\sigma} & \text{if } \nu > \sigma, \\ \frac{1}{2}\left(\sqrt{\nu} + \frac{\sigma}{\sqrt{\nu}}\right) & \text{otherwise,} \end{cases} \qquad r_0 := \begin{cases} -\sqrt{\sigma} & \text{if } \nu > \sigma, \\ -\sqrt{\nu} & \text{otherwise,} \end{cases} \tag{6.13}$$

$$\begin{aligned} \varphi(\lambda) &:= \lambda\left\{a^2\left[2p_0 + T\left(p_0^2 - \lambda^2\right)\right] - 2\alpha p_0^2\right\}\tanh^2(\lambda T) \\ &+ \left\{a^2\left(\lambda^2 + p_0^2\right) - 4\alpha\lambda^2 p_0\right\}\tanh(\lambda T) + a^2\lambda T\left(\lambda^2 - p_0^2\right) - 2\alpha\lambda^3. \end{aligned} \tag{6.14}$$

If $\varphi\left(\sqrt{\sigma}\right) \le 0$, *we put* $\lambda_0 := \sqrt{\sigma}$. *If* $\varphi\left(\sqrt{\sigma}\right) > 0$, *then the equation* $\varphi(\lambda) = 0$ *has a single root* λ_* *in the domain* $\lambda \ge \sqrt{\sigma}$ *and we put* $\lambda_0 := \lambda_*$. *The above coefficient* $q(\cdot)$ *is given by the formula*

$$r(t) := -\lambda_0 \frac{\lambda_0 \tanh\left[\lambda_0(T-t)\right] + p_0}{p_0 \tanh\left[\lambda_0(T-t)\right] + \lambda_0} \text{ if } 0 \le t \le T \text{ and } r(t) := r_0 \text{ otherwise.} \tag{6.15}$$

Proof. The problem (6.9)–(6.11) is a particular case of the problem (5.1)–(5.4): $l = m = 1,\ A(t) = 0,\ B(t) = 1,\ k = 2,\ \varphi_i(\cdot) = 0,\ g_0(t,x,u) = u^2 + \sigma x^2,\ g_1(t,x,u) = 0$ for $t \le T$ and $g_1(t,x,u) = u^2 - \nu x^2$ otherwise, $g_2(t,x,u) = x^2$ for $t \le T$ and $g_2(t,x,u) = 0$ otherwise, $\gamma_0 = \gamma_1 = 0,\ \gamma_2 = \alpha$. Furthermore, the problem (6.9)–(6.11) is regular, i.e., there exists a process $\overline{z} = [\overline{x}(\cdot), \overline{u}(\cdot)]$ such that (6.10) is true and $\mathcal{G}_1(\overline{z}) < 0,\ \mathcal{G}_2(\overline{z}) < \alpha$. Indeed, choose $\delta < \min\left\{T, 3\alpha/\left(a^2+1\right)\right\}, \theta > T$, and $n > \sqrt{3/\nu}$. Put $x_\delta(t) := a\left(1 - \delta^{-1}t\right),\ u_\delta(t) := -a\delta^{-1}$ for $0 \le t \le \delta$ and $x_\delta(t) := u_\delta(t) := 0$ for $t > \delta$ and pick $\overline{z} := [x_\delta(\cdot) + \Delta x_{n,\theta}(\cdot), u_\delta(\cdot) + \Delta u_{n,\theta}(\cdot)]$ where the process $[\Delta x_{n,\theta}(\cdot), \Delta u_{n,\theta}(\cdot)]$ was defined above. Then $\overline{z}$ apparently satisfies (6.10) and $\mathcal{G}_2(\overline{z}) = \int_0^T x_\delta^2\ dt = a^2\delta/3 < \alpha$. Putting $x(\cdot) := u(\cdot) := 0$ into (6.12) we get $\mathcal{G}_1(\overline{z}) = \mathcal{G}_1\left[\Delta x_{n,\theta}(\cdot), \Delta u_{n,\theta}(\cdot)\right] = 2n^{-1} - 2n\nu/3 < 0$.

The next step will be to apply Lemma 6.1. Its assumptions (4.i), (4.ii) are fulfilled with $t^0 := T$. Now $\tau = \|\tau_i\| \in \mathbb{R}^2$ and $g_\tau(t,x,u) := g_0(t,x,u) + \sum_{i=1}^2 \tau_i g_i(t,x,u) = u^2 + (\sigma + \tau_2)\,x^2$ for $t \le T$ and $g_\tau(t,x,u) = g_\tau^0(x,u) = (1+\tau_1)\,u^2 + (\sigma - \tau_1\nu)\,x^2$ for $t > T$. For $g(x,u) := g_\tau^0(x,u) - \delta\left(|x|^2 + |u|^2\right)$, the frequency criterion (5.17) takes the form $(1 + \tau_1 - \delta)\,|u|^2 + (\sigma - \tau_1\nu - \delta)\,|x|^2 \ge 0$ for all $\omega \in \mathbb{R}, x, u \in \mathbb{C}$ such that $\imath\omega x = u$ or, equivalently, $(1 + \tau_1 - \delta) + (\sigma - \tau_1\nu - \delta)\,\omega^{-2} \ge 0$ for all $\omega \ne 0$. So, by (4.D), we have $\Theta^0 = \left\{\tau = (\tau_1, \tau_2) : \tau_i \ge 0 \text{ and } \tau_1 < \sigma\nu^{-1}\right\}$. Thus the assumption $\Theta^0 \ne \emptyset$ is valid. Consider the problem (6.1) where the objective function is determined by (5.26) and (5.24). Let $\tau \in \Theta^0$. It is easy to verify that relations (5.24) $2P_\tau xu + (\tau_1 + 1)\,u^2 + (\sigma - \tau_1\nu)\,x^2 = (\tau_1 + 1)\,(u - r_\tau x)^2\ \ (\forall x, u)$ have the solution

$$P_\tau = \sqrt{(1+\tau_1)\,(\sigma - \nu\tau_1)} =: P_{\tau_1}, \qquad r_\tau = -\sqrt{(\sigma - \nu\tau_1)\,(1+\tau_1)^{-1}} \tag{6.16}$$

for which the equation $\dot{x} = (A + Br_\tau)\,x = r_\tau x$ is stable. The solution of the system (5.26) $\dot{X}_\tau = \Psi_\tau,\ \dot{\Psi}_\tau = (\sigma + \tau_2)\,X_\tau,\ 0 \le t \le T,\ X_\tau(T) = 1,\ \Psi_\tau = -P_\tau$ is given by

$$\begin{array}{lll} \Psi_\tau(t) = & -\lambda\sinh\left[\lambda(T-t)\right] & -P_\tau\cosh\left[\lambda(T-t)\right], \\ X_\tau(t) = & \cosh\left[\lambda(T-t)\right] & +\frac{P_\tau}{\lambda}\sinh\left[\lambda(T-t)\right], \end{array} \qquad \lambda := \sqrt{\sigma + \tau_2}, \tag{6.17}$$

and evidently satisfies the assumption $\det X_\tau(t) = X_\tau(t) \neq 0\ (\forall t \in [0,T])$ from Lemma 6.1. By (6.17), the problem (6.1) takes the form

$$b(\tau) = a^2\omega\left[P_{\tau_1}, \sqrt{\sigma+\tau_2}\,\right] - \alpha\tau_2 \to \sup \text{ subject to } \tau_2 \geq 0,\ 0 \leq \tau_1 < \sigma\nu^{-1}, \tag{6.18}$$

where $P_\tau = P_{\tau_1}$ is given by (6.16) and

$$\omega(p,\lambda) := \lambda\frac{p+\lambda\tanh(\lambda T)}{\lambda+p\tanh(\lambda T)}. \tag{6.19}$$

The function $\omega(p,\lambda)$ is strictly monotone with respect to p

$$\frac{\partial\omega}{\partial p}(p,\lambda) = \lambda^2\frac{1-\tanh^2(\lambda T)}{\left[\lambda+p\tanh(\lambda T)\right]^2} > 0.$$

So the problem (6.18) can be rewritten as follows:

$$w\left(\tau_2\right) := a^2\omega\left[p_0, \sqrt{\sigma+\tau_2}\,\right] - \alpha\tau_2 \to \sup \ \text{ subject to } \ \tau_2 \geq 0, \tag{6.20}$$

where $p_0 := \max\{P_{\tau_1} : 0 \leq \tau_1 < \sigma/\nu\}$. The maximum p_0 is achieved by $\tau_1 = \tau_1^0$ where $\tau_1^0 := 0$ if $\nu > \sigma$ and $\tau_1^0 := (2\nu)^{-1}(\sigma-\nu)$ otherwise.

By Remark 2 to Lemma 6.1, the function $b(\tau)$ is concave on Θ^0, and so evidently is the function $w\left(\tau_2\right)$ over $\tau_2 \geq 0$. It follows from (6.19) and (6.20) that $w\left(\tau_2\right) \to -\infty$ as $\tau_2 \to \infty$. So the problem (6.20) has a solution τ_2^0 and

$$\frac{dw}{d\tau_2}(0) \leq 0 \Rightarrow \tau_2^0 = 0, \qquad \frac{dw}{d\tau_2}(0) > 0 \Rightarrow \frac{dw}{d\tau_2}\left(\tau_2^0\right) = 0, \tag{6.21}$$

where the equation $\frac{dw}{d\tau_2}(\tau) = 0$ has no more than one solution. (Indeed, otherwise the set of all its roots is an interval and, therefore, the analytical function $\frac{dw}{d\tau_2}(\cdot)$ vanishes, i.e., $w(\cdot) =$ const that does not take place.) This implies that the problem (6.18) also has the solution $\tau^0 = \left(\tau_1^0, \tau_2^0\right) \in \Theta^0$. Thus all the assumptions of Lemma 6.1 are fulfilled.

By this lemma, the optimal process exists, is unique, and is generated by the closed-loop controller (6.2). Putting $\tau_1 := \tau_1^0$ into (6.16), we get (6.13). So, for $t \geq T$, the coefficient $q(\cdot) = q_{\tau^0}(\cdot)$ from (6.2) takes the form (6.15). Denote $\lambda := \sqrt{\sigma+\tau_2}$. The direct calculation shows that

$$\frac{dw}{d\tau_2}\left(\tau_2\right) = \frac{1}{2\lambda}\frac{\varphi(\lambda)}{\left[\lambda+p\tanh(\lambda T)\right]^2},$$

where $\varphi(\lambda)$ is defined by (6.14). This and (6.21) imply that the quantity $\lambda_0 := \sqrt{\sigma+\tau_2^0}$ is determined as it was indicated in the statement of the lemma. Putting $\tau := \tau^0$ into (6.3), we get $\overline{r}_{\tau^0}(t) = \Psi_{\tau^0}(t)X_{\tau^0}(t)^{-1}$ if $t \leq T$. So, taking into account (6.17), we get (6.15). □

REFERENCES

[1] K. J. Arrow, L. Hurwicz, and H. Uzawa, *Studies in Linear and Non-Linear Programming*, Stanford University Press, Stanford, CA, 1964.

[2] E. G. Gol'shtein, *Dualitätstheorie in der Nichtlinearen Optimierung und ihre Anwendung*, Deutscher Spache, Akademie-Verlag, Berlin, 1975.

[3] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, Studies in Mathematics and Its Applications, Vol. 1, American Elsevier Publishing Company, New York, 1976.
[4] O. L. Mangasarian, *Pseudo-convex functions*, SIAM J. Control Optim., Ser. A, 3 (1965), pp. 281–290.
[5] J. Rissanen, *On duality without convexity*, J. Math. Anal. Appl., 18 (1967), pp. 269–275.
[6] J. V. Outrata and J. Jarušek, *Duality theory in mathematical programming and control*, Kybernetika (Prague), Supplement 20/21, 1984/85.
[7] R. T. Rockafellar, *Convex Functions and Duality in Optimization Problems and Dynamics*, Lecture Notes in Oper. Res. and Math. Ec. II, Springer-Verlag, Berlin, 1969.
[8] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[9] J. F. Toland, *Duality in nonconvex optimization*, J. Math. Anal. Appl., 66 (1978), pp. 399–415.
[10] P. O. Lindberg, *A generalization of Fenchel conjugation giving generalized Lagrangians and symmetric nonconvex duality*, in Survey of Mathematical Programming, Vol. 1, North-Holland, Amsterdam, 1979, pp. 249–267.
[11] V. A. Yakubovich, *Nonconvex optimization problems: The infinite-horizon linear-quadratic problems with quadratic constraints*, Systems Control Lett., 19 (1992), pp. 13–22.
[12] V. A. Yakubovich, *Linear-quadratic optimization problems with quadratic constraints*, in Proc. of the Second European Control Conf., The Netherlands, 1 (1993), pp. 346–359.
[13] A. S. Matveev and V. A. Yakubovich, *Nonconvex problems of global optimization*, Algebra i Analiz, 4 (1992), pp. 189–219 (in Russian); St. Peterburg Math. J., 4 (1993), pp. 1217–1243 (in English).
[14] V. A. Yakubovich, *On a method to solve special problems of global optimization*, Vestnik S.-Petersburgs. Gos. Universiteta, Ser. 1 (1992), pp. 58–68 (in Russian).
[15] R. E. Kalman, *Contributions to the theory of optimal control*, Bol. Soc. Mat. Mexicana (5), (1960), pp. 102–119.
[16] D. H. Jacobson, D. H. Marton, M. Pachter, and T. Geveci, *Extensions of Linear-Quadratic Optimal Control*, Lecture Notes in Control and Inform. Sci. 27, Springer-Verlag, Berlin, Heidelberg, New York, 1980.
[17] V. A. Yakubovich, *A frequency theorem in control theory*, Sibirsk. Mat. Zh., 14 (1973), pp. 384–420 (in Russian); Siberian Math. J., 14 (1973), pp. 265–289 (in English).
[18] J. C. Willems, *Least squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automat. Control, 16 (1971), pp. 621–634.
[19] A. V. Megretskiĭ and V. A. Yakubovich, *Singular stationary nonhomogeneous linear-quadratic optimal control*, Amer. Math. Soc. Transl., 155 (1993), pp. 129–167.
[20] H. Hermes and J. P. LaSalle, *Functional Analysis and Time Optimal Control*, Academic Press, New York, London, 1969.
[21] P. R. Halmos, *A Hilbert Space Problem Book,* D. Van Nostrand, Princeton, NJ, Toronto, London, 1982.
[22] V. A. Yakubovič, *Minimization of quadratic functionals under quadratic constraints and the necessity of a frequency condition in the quadratic criterion for absolute stability of nonlinear control systems*, Dokl. Acad. Nauk SSSR, 209 (1973), pp. 1039–1042 (in Russian); Soviet Math. Dokl., 14 (1973), pp. 593–597 (in English).
[23] A. L. Fradkov, *Duality theorems for certain nonconvex extremal problems*, Sibirsk. Mat. Zh., 14 (1973), pp. 357–383 (in Russian); Siberian Math. J., 14 (1973), pp. 247–264 (in English).
[24] A. Megretsky and S. Treil, *S-Procedure and Power Distribution Inequalities: A New Method in Optimization and Robustness of Uncertain Systems*, Mittag-Leffler Institute, Stockholm, Vol. 1, 1990/1991, preprint.
[25] A. A. Agrachev and R. V. Gamkrelidze, *The index of extremality and quasiextremal controls*, Soviet Math. Dokl., 32 (1985), pp. 478–481.
[26] A. A. Agrachev and A. V. Sarychev, *Abnormal sub-Riemannian geodesics: Morse index and rigidity*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 13 (1996), pp. 635–690.
[27] A. A. Milutin, *On quadratic conditions of extremum in smooth problems with finite-dimensional image*, in Methods of Optimization Theory in Economics, Nauka, Moscow, 1981, pp. 138–177 (in Russian).
[28] S. Boyd, L. E. Ghaoui, E. Feron, and A. V. Balakrishnan, *Linear Matrix Inequalities in Systems and Control Theory*, SIAM, Philadelphia, PA, 1994.
[29] N. Danford and J. T. Schwartz, *Linear Operators. Part* II. *Spectral Theory. Self Adjoint Operators in Hilbert Space*, Interscience, New York, London, 1963.

[30] H. H. Schaefer, *Topological Vector Spaces*, Macmillan, New York, 1966.
[31] L. Amerio and G. Prouse, *Almost-Periodic Functions and Functional Equations*, Van Nostrand Reinhold, New York, Cincinnati, Toronto, London, Melbourne, 1971.
[32] P. L. Zezza, *The Jacobi condition in optimal control theory, stochastic analysis and applications*, World Scientific, River Edge, NJ, 1991, pp. 137–149.
[33] V. A. Yakubovich, *Frequency theorem for periodic systems and the theory of analytical design of regulators*, in Method of Lyapunov Functions in Analysis of Systems Dynamics, Nauka, 1987, pp. 143–151 (in Russian).

SIAM J. CONTROL OPTIM.
Vol. 36, No. 1, pp. 379–390, January 1998

019

GENERIC POLE ASSIGNMENT VIA DYNAMIC FEEDBACK*

SUSUMU ARIKI†

Abstract. An elementary proof of a sufficient condition for the generic pole placement problem based on a new elementary framework is given. This condition is the best at present and was proved by using algebraic geometry. The necessary condition is proved by considering a group action, which also suggests new treatment of the problem.

Key words. linear systems, output feedback pole placement

AMS subject classifications. 93B55, 93B27

PII. S0363012995274684

1. Introduction.

Let $S_{m,p,n}$ be the set of all strictly proper transfer functions of degree n with m inputs, p outputs. For $G(s) \in S_{m,p,n}$, we let $\chi(s)$ be the product of denominators of its Smith–MacMillan form. $\chi(s)$ is the characteristic polynomial of $G(s)$, and the degree of $\chi(s)$ is n. For a proper transfer function $K(s)$ of degree at most d with p inputs, m outputs, $G_{cl}(s) = (I_p - G(s)K(s))^{-1}G(s)$ is the transfer function of the closed-loop system. Its characteristic polynomial is denoted by $\chi_{cl}(s)$. Since $\chi_{cl}(s)$ determines the behavior of the closed-loop system, a basic design problem is to ask if we can control $\chi_{cl}(s)$. Design parameters are K, L, M, N such that $K(s) = M(sI_d - K)^{-1}L + N$. For almost all K, L, M, N, $\chi_{cl}(s)$ is a degree $n+d$ monic polynomial such that coefficients are polynomials with respect to design parameters. But for certain parameter values, cancellation may occur, and $\chi_{cl}(s)$ has less degree. Hence to keep everything continuous, it is better to control this polynomial, say $\chi'_{cl}(s)$, rather than $\chi_{cl}(s)$ itself. Ordinarily, this control problem is called the pole placement problem. For a precise description of $\chi'_{cl}(s)$, see Proposition 2.5 in the body of this paper.

We say that $G(s)$ is pole assignable by degree d compensators if for any polynomial of degree $n + d$ one can choose a $K(s)$ of degree at most d such that $\chi'_{cl}(s)$ coincides with the polynomial. (Both are assumed to be monic without loss of generality.) Since $G(s)$ is pole assignable by degree d compensators if and only if $G(s)^T$ is so, we can assume $p \leq m$.

The generic pole placement problem is to find the minimum d for almost all $G(s)$. (For a special $G(s)$, d may be smaller.)

To reach the goal, several authors study sufficient conditions. The best result among them, recently obtained by X. Wang and J. Rosenthal [8], is that $n < mp + d\max(m, p)$ is sufficient for generic pole placement. The method depends on a compactification of the moduli space of systems.

The purpose of this paper is to show that this result is a consequence of the following elementary fact.

Let X be a subset of a vector space V such that

(1) a neighborhood of the origin is contained in X;

(2) X is stable under multiplication by positive real numbers.

*Received by the editors August 28, 1995; accepted for publication (in revised form) December 9, 1996.

http://www.siam.org/journals/sicon/36-1/27468.html

†Division of Mathematics, Tokyo University of Mercantile Marine, Etchujima 2-1-6, Koto-ku, Tokyo 135, Japan (ariki@ipc.tosho-u.ac.jp).

Then X coincides with V.

Hence we do not need algebraic geometry to reach the best result.

The necessary condition is $n \leq mp + d(m + p - 1)$, which is proved in [6] also by using algebraic geometry. But one can understand this fact by counting parameters as noticed by preceders. Another purpose of this paper is to prove this necessary condition along this idea. To verify the idea, we consider a group action here. It is because this new treatment suggests another approach to the pole placement problem.

(After submitting this paper, The author learned the work of X. Wang [7], [10], [11] and J. Leventides and N. Karcanias [12]. These treat the static feedback case. The latter is of special interest to the author, since the basic idea in the paper is very similar to his. But there are differences, and the results developed in this paper in order to work in dynamic feedback case are new.)

2. Preliminaries. We first review the theory of minimal bases. For the proof, we refer to [2], [4], [5]. They are interesting and not difficult to prove.

DEFINITION 2.1. *Let V be an r-dimensional subspace of $\mathbf{R}(s)^N$.*

A basis $\{g_1(s), \ldots, g_r(s)\}$ of V is called minimal if

(1) *$g_i(s)$'s have polynomial entries;*

(2) *the matrix $B(s) = (g_1(s), \ldots, g_r(s))$ is left invertible; i.e., there exists an $U(s) \in M(r, N, \mathbf{R}[s])$ such that $U(s)B(s) = I_r$;*

(3) *$\hat{B}(s) = B(s)\mathrm{diag}(s^{-\nu_1}, \ldots, s^{-\nu_r})$ is full rank at $s = \infty$, where ν_i is the highest degree of the entries in $g_i(s)$.*

ν_i is called the minimal index of $g_i(s)$. The following theorems are well known.

THEOREM 2.2. (1) *Minimal bases exist.*

(2) *The set of minimal indices is independent of the choice of a minimal basis.*

A minimal basis $B(s)$ is said to be in echelon form if $\nu_1 \geq \cdots \geq \nu_r$ and there exist different r row numbers γ_j $(1 \leq j \leq r)$ such that the following hold.

(1) γ_j increases on intervals $\{i | \nu_i = k\}$ for all k.

(2) The ith entry of the jth column has degree equal to or less than ν_j if $i < \gamma_j$, less than ν_j if $i > \gamma_j$. The γ_jth entry is a monic polynomial of degree ν_j.

(3) The γ_kth entry $(\nu_k \leq \nu_j)$ of the jth column has degree less than ν_k.

Note that if $\nu_k < \nu_j$, then it forces $k > j$, but if $\nu_k = \nu_j$, then $k > j$ and $k < j$ are both possible. It is obvious that there is a unique minimal basis in echelon form.

THEOREM 2.3. *Let (A, B) $(A \in M(n, n, \mathbf{R}), B \in M(n, m, \mathbf{R}))$ be controllable, i.e., $\mathbf{R}[A]\mathrm{Im}B = \mathbf{R}^n$. Then the set of minimal indices of* $\mathrm{Ker}\left(\begin{bmatrix} sI_n - A, & -B \end{bmatrix}\right)$ *coincides with the set of controlability indices as a multiset.*

THEOREM 2.4. (1) *Let $G(s)$ be a strictly proper transfer function of degree n with m inputs, p outputs. Then there exist $A \in M(n, n, \mathbf{R})$, $B \in M(n, m, \mathbf{R})$, $C \in M(p, n, \mathbf{R})$, such that*

(a) *$G(s) = C(sI_n - A)^{-1}B$,*

(b) *the characteristic polynomial $\chi(s)$ of $G(s)$ equals* $\det(sI_n - A)$,

(c) *(A, B) is controllable,*

(d) *(A, C) is observable.*

(2) *If $G(s)$ is of the form $G(s) = C(sI_n - A)^{-1}B$, then there exist $N_0(s) \in M(p, m, \mathbf{R}[s])$, $D_0(s) \in M(m, m, \mathbf{R}[s])$ such that*

(e) *the sum of highest degrees of columns in $D_0(s)$ is n,*

(f) *$\hat{D}_0(s)$ has full rank at $s = \infty$,*

(g) $\det(D_0(s)) = \det(sI_n - A)$,

(h) *$G(s) = N_0(s)D_0(s)^{-1}$.*

If (h) *holds and $G(s)$ is strictly proper, then*

(i) *all m minors of $B_0(s) = \begin{bmatrix} N_0(s) \\ D_0(s) \end{bmatrix}$ have degree equal to or less than n and the only m minor of degree n is* $\det(D_0(s))$.

(3) *If $B_0(s) \in M(m+p, m, \mathbf{R}[s])$ is such that* $\det(D_0(s))$ *is of degree n, other m minors with 1 row from $N_0(s)$, other $(m-1)$ rows from $D_0(s)$ have degrees less than n, then $G(s) = N_0(s)D_0(s)^{-1}$ is strictly proper.*

PROPOSITION 2.5. *Let $G(s)$ be a strictly proper transfer function of degree n with m inputs, p outputs. We take (A, B, C), $N_0(s)$, $D_0(s)$ as in Theorem* 2.4(1) *and* (2).

(1) *For $K(s) = M(sI_d - K)^{-1}L + N$, where $K \in M(d, d, \mathbf{R})$, $L \in M(d, p, \mathbf{R})$, $M \in M(m, d, \mathbf{R})$, $N \in M(m, p, \mathbf{R})$, we set*

$$\chi'_{cl}(s) = \det \left(\begin{bmatrix} I_p & 0 & N_0(s) & 0 \\ 0 & I_d & 0 & I_d \\ N & M & D_0(s) & 0 \\ L & K & 0 & sI_d \end{bmatrix} \right).$$

Then, $\chi_{cl}(s) = \chi'_{cl}(s)$ if and only if
$\left(\begin{bmatrix} A+BNC & BM \\ LC & K \end{bmatrix}, \begin{bmatrix} B \\ 0 \end{bmatrix} \right)$ *is controllable, and*
$\left(\begin{bmatrix} A+BNC & BM \\ LC & K \end{bmatrix}, \begin{bmatrix} C & 0 \end{bmatrix} \right)$ *is observable.*

(2) *All proper transfer functions of degree at most d with p inputs, m outputs may be described as $K(s) = M(sI_d - K)^{-1}L + N$.*

Proof. This can be proved in an elementary way, but here we appeal to the theory of elementary divisors.

We first note that the elementary divisors of the $\mathbf{C}[s]$-module

$$(G(s)\mathbf{C}[s]^m + \mathbf{C}[s]^p)/\mathbf{C}[s]^p$$

are nothing but denominators of the Smith–MacMillan form of $G(s)$. For a $\mathbf{C}[s]$-module A, we denote by $\chi(A)$ the product of the elementary divisors of A. Then if A is a submodule or a quotient module of B, $\chi(A)$ divides $\chi(B)$.

We also note that if we set

$$X(s) = \begin{bmatrix} sI_d - K^T & -M^T & 0 \\ L^T & N^T & I_p \\ 0 & D_0^T(s) & N_0^T(s) \end{bmatrix},$$

then $y = G_{cl}^T(s)r$ can be described as

$$X(s) \begin{bmatrix} z \\ y \\ -u \end{bmatrix} = \begin{bmatrix} 0 \\ -r \\ 0 \end{bmatrix}.$$

Hence we have

$$G_{cl}^T(s) = \begin{bmatrix} 0 & I_m & 0 \end{bmatrix} X(s)^{-1} \begin{bmatrix} 0 \\ -I_p \\ 0 \end{bmatrix}.$$

Therefore, $(G_{cl}^T(s)\mathbf{C}[s]^p + \mathbf{C}[s]^m)/\mathbf{C}[s]^m$ is a subquotient module of

$$(X(s)^{-1}\mathbf{C}[s]^{d+m+p} + \mathbf{C}[s]^{d+m+p})/\mathbf{C}[s]^{d+m+p},$$

and thus $\chi_{cl}(s)$ divides $\det(X(s))$. Since

$$\det\left(\begin{bmatrix} sI_d - K^T & -M^T & 0 \\ L^T & N^T & I_p \\ 0 & D_0^T(s) & N_0^T(s) \end{bmatrix}\right) = \pm\det\left(\begin{bmatrix} K^T & M^T & I_d & 0 \\ L^T & N^T & 0 & I_p \\ 0 & D_0^T(s) & 0 & N_0^T(s) \\ sI_d & 0 & I_d & 0 \end{bmatrix}\right),$$

we know that $\det(X(s))$ coincides with $\chi'_{cl}(s)$ up to sign. In particular, $\chi_{cl}(s)$ and $\chi'_{cl}(s)$ coincide if and only if the degree of $\chi_{cl}(s)$ is $n+d$.

To see when they coincide, we use the following description of $G_{cl}(s)$:

$$G_{cl}(s) = \begin{bmatrix} C & 0 \end{bmatrix} \begin{bmatrix} sI_n - A - BNC & -BM \\ -LC & sI_d - K \end{bmatrix}^{-1} \begin{bmatrix} B \\ 0 \end{bmatrix}.$$

Suppose that this triple for $G_{cl}(s)$ is either uncontrollable or unobservable; then $\chi_{cl}(s)$ has lower degree and never coincides with $\chi'_{cl}(s)$. On the other hand, they coincide if they are both controllable and observable. It has the same proof as Theorem 2.4(1). Hence we have (1). Equation (2) is a direct consequence of Theorem 2.4. □

By Proposition 2.5(1), we have that $\chi_{cl}(s) = \chi'_{cl}(s)$ for almost all K, L, M, N, and the coefficients of $\chi'_{cl}(s)$ are polynomials with respect to these design parameters if (A, B, C) is generic.

We say that $G(s)$ is pole assignable by compensators of degree d if all monic polynomials of degree $n+d$ can be expressed in the form $\chi'_{cl}(s)$ by compensators of degree at most d.

3. Pole placement map.

DEFINITION 3.1. *We denote by $S_{m,p,n}$ the set of strictly proper transfer functions of degree n with m inputs, p outputs. $S^{\le n}_{m,p} = \cup_{1\le k\le n} S_{m,p,k}$.*

For a partition ν of n into at most m parts, and a permutation γ of $\{1,\dots,m\}$ which increases on intervals $\{i|\nu_i = k\}$, we denote by $\Sigma^{\gamma,\nu}_{m,p}$ the set $\left\{B_0(s) = \left[\begin{smallmatrix} N_0(s) \\ D_0(s) \end{smallmatrix}\right]\right\}$ of elements satisfying

1. *$N_0(s) \in M(p, m, \mathbf{R}[s])$, and entries of the jth column have degree less than ν_j;*
2. *$D_0(s) \in M(m, m, \mathbf{R}[s])$, and the ith entry of the jth column have degrees equal to or less than ν_j if $i < \gamma_j$, less than ν_j if $i > \gamma_j$. The γ_jth entry is a monic polynomial of degree ν_j. Further, the γ_kth entry ($\nu_k \le \nu_j$) of the jth column must have degree less than ν_k.*

It is simply denoted $\Sigma^{\nu}_{m,p}$ if γ is the identity. We set $\Sigma^{\le n}_{m,p} = \cup_{|\nu|\le n} \Sigma^{\gamma,\nu}_{m,p}$.

Note that we do not assume that $B_0(s) \in \Sigma^{\gamma,\nu}_{m,p}$ is minimal. But the subset consisting of minimal ones is a dense open subset of $\Sigma^{\gamma,\nu}_{m,p}$.

For $B_0(s) \in \Sigma^{\le n}_{m,p}$, we set

$$N(s) = \begin{bmatrix} N_0(s) & 0 \\ 0 & I_d \end{bmatrix}, \quad D(s) = \begin{bmatrix} D_0(s) & 0 \\ 0 & sI_d \end{bmatrix}, \quad B(s) = \begin{bmatrix} N(s) \\ D(s) \end{bmatrix}.$$

We also set $G = GL(m+p+2d, \mathbf{R})$, and for $g \in G$, the product of g and $B(s)$ is partitioned as

$$gB(s) = \begin{bmatrix} N^{(g)}(s) \\ D^{(g)}(s) \end{bmatrix}.$$

LEMMA 3.2. *$p : \Sigma^{\leq n}_{m,p} \to S^{\leq n}_{m,p}$ defined by $p(B_0(s)) = N_0(s)D_0(s)^{-1}$ is well defined and onto. The subset consisting of minimal bases bijectively corresponds to $S_{m,p,n}$.*

Proof. By Theorem 2.4(3), $p(B_0(s))$ is strictly proper.

Since the characteristic polynomial of $G(s)$ divides $\det(D_0(s))$, its degree is equal to or less than n. Thus we have the well-definedness. For any $G(s)$, take $B_0(s)$ as in Theorem 2.4. Since $\hat{B}_0(s)$ has the form $\begin{bmatrix} G(s) \\ I_m \end{bmatrix} \hat{D}_0(s)$, we have that the highest degree of all entries in the jth column of $N_0(s)$ is smaller than that of $D_0(s)$. Then by multiplying $\hat{D}_0(\infty)^{-1}$ from the right, and after successive elementary column transformations if necessary, we have an element in $\Sigma^{\leq n}_{m,p}$ which maps to $G(s)$. The rest is an obvious consequence of the uniqueness of echelon form. □

Now we give the definition of a new type of a pole placement map, which allows us an elementary proof in an elementary framework.

DEFINITION 3.3. *For $B_0(s) \in \Sigma^{\leq n}_{m,p}$, we define the pole placement map*

$$\rho_{B_0} : G \times M(m+d, p+d, \mathbf{R}) \to \mathbf{R}[s]_{n+d}$$

by the following:

$$\rho_{B_0}(g, X) = \det(g)^{-1} \det \left(\begin{bmatrix} I_{p+d} & N^{(g)}(s) \\ X & D^{(g)}(s) \end{bmatrix} \right).$$

The following proposition is very useful. For the static case, this observation is due to R. W. Brocket and C. I. Byrnes [1]. But we note that their generalization of this technique does not allow us to use the differential map of the pole placement map. Our generalization is a key to the proof we will give in Theorem 4.3.

PROPOSITION 3.4. *Let $G(s) \in S_{m,p,n}$, and we take $B_0(s) \in \Sigma^{\leq n}_{m,p}$ as in Theorem 2.4. Then $G(s)$ is pole assignable by degree d compensators if and only if the pole placement map covers $\mathbf{R}[s]_{n+d} \setminus \mathbf{R}[s]_{n+d-1}$.*

Proof. Assume that $G(s)$ is pole assignable by degree d compensators. Then, by Proposition 2.5, the set

$$\left\{ \det \left(\begin{bmatrix} I_{p+d} & N(s) \\ X & D(s) \end{bmatrix} \right) | X \in M(m+d, p+d, \mathbf{R}) \right\}$$

coincides with the set of all monic polynomials of degree $n+d$. Since the image of the pole placement map is nothing but

$$\left\{ \det \left(\begin{bmatrix} K_1 & N(s) \\ K_2 & D(s) \end{bmatrix} \right) \middle| \begin{bmatrix} K_1 \\ K_2 \end{bmatrix} \text{is full rank} \right\},$$

we can conclude that ρ_{B_0} covers $\mathbf{R}[s]_{n+d} \setminus \mathbf{R}[s]_{n+d-1}$.

Conversely, we assume that ρ_{B_0} covers $\mathbf{R}[s]_{n+d} \setminus \mathbf{R}[s]_{n+d-1}$. Then any monic polynomial of degree $n+d$ is expressed as

$$\det \left(\begin{bmatrix} K_1 & N(s) \\ K_2 & D(s) \end{bmatrix} \right).$$

Laplace expansion tells us that it equals $\det(K_1)\det(D_0(s))+$ (a linear combination of other $m+d$ minors). Since $N(s)D(s)^{-1}$ is strictly proper, we have $\det(K_1) \neq 0$ by Theorem 2.4(2)(i). By putting $X = K_2 K_1^{-1}$, we know that the set of all monic

polynomials of degree $n+d$ coincides with

$$\left\{ \det\left(\begin{bmatrix} I_{p+d} & N(s) \\ X & D(s) \end{bmatrix} \right) \middle| X \in M(m+d, p+d, \mathbf{R}) \right\}.$$

Hence $G(s)$ is pole assignable by degree d compensators. □

4. Generic pole placement. We define a_n, b_n by $n = ma_n + b_n$ $(0 \leq b_n < m)$, and $\nu(n)$ by $\nu(n) = (a_n+1, \ldots, a_n+1, a_n, \ldots, a_n)$ (a_n+1 repeats b_n times).

Among $\Sigma_{m,p}^{\gamma,\nu}$'s, $\Sigma_{m,p}^{\nu(n)}$ is the unique set of the maximum dimension $n(m+p)$. (We know it by counting the number of coefficients.) Hence, to consider the generic pole placement problem, we can restrict ourselves to $p : \Sigma_{m,p}^{\nu(n)} \to S_{m,p}^{\leq n}$ by Lemma 3.2.

REMARK 4.1. *In fact, it is known that we can introduce a topology on $S_{m,p}^{\leq n}$ to make the image of the set of minimal bases in $\Sigma_{m,p}^{\nu(n)}$ the unique open cell* (see [6]).

The following lemma is the key idea to prove Theorem 4.3. Its proof depends on a quite elementary fact which we have explained in the introduction.

LEMMA 4.2. *Let $B_0(s)$ be an element in $\Sigma_{m,p}^{\leq n}$.*

(1) *If $\rho_{B_0}(g,-) : M(m+d, p+d, \mathbf{R}) \to \mathbf{R}[s]_{n+d}$ covers a neighborhood of the origin for some $g \in G$, then $p(B_0(s)) = G(s)$ is pole assignable by degree d compensators.*

(2) *If $\rho_{B_0}(g,-)$ maps 0 to 0, and its differential at $0 \in M(m+d, p+d, \mathbf{R})$ is surjective, then $\rho_{B_0}(g,-)$ covers a neighborhood of the origin.*

Proof. (1) Since the image of ρ_{B_0} is stable under multiplication by positive real numbers, the assumption ensures that ρ_{B_0} is surjective. Thus we have the pole assignability by Proposition 3.4. (2) is obvious. □

We now state the main theorem about generic pole assignability.

THEOREM 4.3. *If $d > \frac{n-mp}{max(m,p)}$, then $S_{m,p,n}$ is generic pole assignable by degree d compensators. More precisely, all elements in $p(U)$ are pole assignable by degree d compensators.*

Before we proceed to the proof, we briefly sketch the logical structure of the proof. We first define an open subset U of $\Sigma_{m,p}^{\nu(n)}$ and prove that U is dense. Next, we prove that all systems in U are pole assignable by degree d compensators. To define U, we first prepare notation.

DEFINITION 4.4. *For $h\begin{bmatrix} N(s) \\ D(s) \end{bmatrix} U(s)$ such that $h \in G$ and an invertible polynomial matrix $U(s)$, we write the first a_n+1 entries of the first column as*

$$\begin{bmatrix} B_1 \\ B_2 \end{bmatrix} \begin{bmatrix} 1 \\ s \\ \cdot \\ \cdot \\ s^{a_n} \end{bmatrix} \qquad \left(\begin{array}{l} B_1 \in M(a_n+1, a_n+1, \mathbf{R}), \\ B_2 \in M(m+p+2d-a_n-1, a_n+1, \mathbf{R}) \end{array} \right).$$

If $\det(B_1) \neq 0$, we define $g \in G$ by

$$g = \begin{bmatrix} B_1^{-1} & 0 \\ -B_2 B_1^{-1} & I_{m+p+2d-a_n-1} \end{bmatrix}.$$

If we compute $gh\begin{bmatrix}N(s)\\D(s)\end{bmatrix}U(s)$, *then we find that the first column has the entries in such a way that* 1 *to* s^{a_n} *in the ascending order in the first* a_n+1 *entries, and* 0*'s in the rest of entries.*

We also note that g *is determined by* $h\in G$ *and a polynomial matrix* $U(s)$.

Let U *be the open subset of* $\Sigma_{m,p}^{\nu(n)}$ *consisting of elements satisfying*

(1) $\det(B_1)\neq 0$ *for some* $h\begin{bmatrix}N(s)\\D(s)\end{bmatrix}U(s)$;

(2) *the set of* $m+d$ *minors which have one row from* $N^{(gh)}(s)$ *and other rows from* $D^{(gh)}(s)$ *span* $\mathbf{R}[s]_{n+d}$.

REMARK 4.5. *We form* $B(s)=\begin{bmatrix}N(s)\\D(s)\end{bmatrix}$ *from* $N_0(s)$ *and* $D_0(s)$ *as in the previous section.*

Note that U is dense if it is nonempty, since for each h, $U(s)$, the above set is defined as the complement of the set of zeros of a polynomial, and U is their union.

It is easy to see that U is nonempty. To find an element in U, we set, for example,

$$N_0(s)=\begin{bmatrix}0&\cdots&0&1\\0&\cdots&1&s^{\nu_m-p+1}\\\cdot&\cdots&\cdot&\cdot\\0&\cdots&1&s^{\nu_m-1}\end{bmatrix},D_0(s)=\begin{bmatrix}s^{\nu_1}&0&\cdots&0&0&0\\1&s^{\nu_2}&\cdots&0&0&0\\\cdot&\cdot&\cdots&\cdot&\cdot&\cdot\\0&0&\cdots&1&s^{\nu_{m-1}}&0\\0&0&\cdots&0&1&s^{\nu_m}\end{bmatrix}$$

for the most tight case $\nu_m=p+d-1$.

It turns out to be an element in U. To see it, it is enough to see the matrix transformation given below. We can treat the other case in the same way. (The constant feedback case is not the exception.)

We first move the mth column to the first column, the $(p+d+m)$th row to the $(p+d+1)$th row, and then the $(p+1)$th to the $(p+d)$th rows to the first d rows. $B(s)=\begin{bmatrix}N(s)\\D(s)\end{bmatrix}$ then becomes

$$\Rightarrow\begin{bmatrix}0&0&\cdots&0&1&0&\cdots&0\\0&0&\cdots&0&0&1&\cdots&0\\\cdot&\cdot&\cdots&\cdot&\cdot&\cdot&\cdots&\\0&0&\cdots&0&0&0&\cdots&1\\1&0&\cdots&0&0&0&\cdots&0\\s^{\nu_m-p+1}&0&\cdots&1&0&0&\cdots&0\\\cdot&\cdot&\cdots&\cdot&\cdot&\cdot&\cdots&\\s^{\nu_m-1}&0&\cdots&1&0&0&\cdots&0\\s^{\nu_m}&0&\cdots&1&0&0&\cdots&0\\0&s^{\nu_1}&\cdots&0&0&0&\cdots&0\\0&1&\cdots&0&0&0&\cdots&0\\\cdot&\cdot&\cdots&\cdot&\cdot&\cdot&\cdots&\\0&0&\cdots&s^{\nu_{m-1}}&0&0&\cdots&0\\0&0&\cdots&0&s&0&\cdots&0\\0&0&\cdots&0&0&s&\cdots&0\\\cdot&\cdot&\cdots&\cdot&\cdot&\cdot&\cdots&\\0&0&\cdots&0&0&0&\cdots&s\end{bmatrix}.$$

Next, we multiply each of the last d columns by a power of s and add in the first column. Then it is further transformed into

$$\Rightarrow \begin{bmatrix}
s^{d-1} & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\
s^{d-2} & 0 & \cdots & 0 & 0 & 1 & \cdots & 0 \\
\cdot & \cdot & \cdots & \cdot & \cdot & \cdot & \cdots & \\
1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 1 \\
1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\
s^{\nu_m-p+1} & 0 & \cdots & 1 & 0 & 0 & \cdots & 0 \\
\cdot & \cdot & \cdots & \cdot & \cdot & \cdot & \cdots & \\
s^{\nu_m-1} & 0 & \cdots & 1 & 0 & 0 & \cdots & 0 \\
s^{\nu_m} & 0 & \cdots & 1 & 0 & 0 & \cdots & 0 \\
0 & s^{\nu_1} & \cdots & 0 & 0 & 0 & \cdots & 0 \\
0 & 1 & \cdots & 0 & 0 & 0 & \cdots & 0 \\
\cdot & \cdot & \cdots & \cdot & \cdot & \cdot & \cdots & \\
0 & 0 & \cdots & s^{\nu_{m-1}} & 0 & 0 & \cdots & 0 \\
s^{d} & 0 & \cdots & 0 & s & 0 & \cdots & 0 \\
s^{d-1} & 0 & \cdots & 0 & 0 & s & \cdots & 0 \\
\cdot & \cdot & \cdots & \cdot & \cdot & \cdot & \cdots & \\
s & 0 & \cdots & 0 & 0 & 0 & \cdots & s
\end{bmatrix}.$$

Now we eliminate the latter part of the first column entries by using elementary row transformations:

$$\Rightarrow \begin{bmatrix}
s^{d-1} & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\
s^{d-2} & 0 & \cdots & 0 & 0 & 1 & \cdots & 0 \\
\cdot & \cdot & \cdots & \cdot & \cdot & \cdot & \cdots & \\
1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 1 \\
1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\
s^{\nu_m-p+1} & 0 & \cdots & 1 & 0 & 0 & \cdots & 0 \\
\cdot & \cdot & \cdots & \cdot & \cdot & \cdot & \cdots & \\
s^{\nu_m-1} & 0 & \cdots & 1 & 0 & 0 & \cdots & 0 \\
s^{\nu_m} & 0 & \cdots & 1 & 0 & 0 & \cdots & 0 \\
0 & s^{\nu_1} & \cdots & 0 & 0 & 0 & \cdots & 0 \\
0 & 1 & \cdots & 0 & 0 & 0 & \cdots & 0 \\
\cdot & \cdot & \cdots & \cdot & \cdot & \cdot & \cdots & \\
0 & 0 & \cdots & s^{\nu_{m-1}} & 0 & 0 & \cdots & 0 \\
0 & 0 & \cdots & -1 & s & 0 & \cdots & 0 \\
0 & 0 & \cdots & 0 & -1 & s & \cdots & 0 \\
\cdot & \cdot & \cdots & \cdot & \cdot & \cdot & \cdots & \\
0 & 0 & \cdots & 0 & 0 & 0 & \cdots -1 & s
\end{bmatrix}.$$

The next step is only to delete 1 in the first column by using the row just below it:

$$\Rightarrow \begin{bmatrix} s^{d-1} & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ s^{d-2} & 0 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot & \cdot & \cdot & \cdots & \\ s & 0 & \cdots & 0 & 0 & 0 & \cdots 1 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 1 \\ 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ s^{\nu_m-p+1} & 0 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot & \cdot & \cdot & \cdots & \\ s^{\nu_m-1} & 0 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ s^{\nu_m} & 0 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ 0 & s^{\nu_1} & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot & \cdot & \cdot & \cdots & \\ 0 & 0 & \cdots & s^{\nu_{m-1}} & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & -1 & s & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & -1 & s & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot & \cdot & \cdot & \cdots & \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots -1 & s \end{bmatrix}.$$

It is now easy to check the condition (1), (2), since by rearranging rows of the above matrix we obtain the final form given below, and we have $B_1 = I$, $B_2 = 0$. We also remark here that $\nu_m = a_n$.

$$h \begin{bmatrix} N(s) \\ D(s) \end{bmatrix} U(s) = \begin{bmatrix} 1 & 0 & \cdots & \cdot & \cdot & \cdots & \cdot & \cdot \\ s & 0 & \cdots & \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdots & \cdot & \cdot & \cdots & \cdot & \cdot \\ s^{\nu_m} & \cdot & \cdots & \cdot & \cdot & \cdots & \cdot & \cdot \\ 0 & s^{\nu_1} & \cdots & \cdot & \cdot & \cdots & \cdot & \cdot \\ 0 & 1 & \cdots & \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdots & \cdot & \cdot & \cdots & \cdot & \cdot \\ 0 & \cdot & \cdots & s^{\nu_{m-1}} & 0 & \cdots & \cdot & \cdot \\ 0 & \cdot & \cdots & -1 & s & \cdots & \cdot & \cdot \\ 0 & \cdot & \cdots & 0 & -1 & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdots & \cdot & \cdot & \cdots & \cdot & \cdot \\ 0 & \cdot & \cdots & \cdot & \cdot & \cdots & -1 & s \\ 0 & \cdot & \cdots & \cdot & \cdot & \cdots & \cdot & 1 \end{bmatrix}.$$

Proof of the main theorem. As was explained in the introduction, we can assume $m \geq p$ without loss of generality. We consider the map ρ_{B_0} in Lemma 4.2. Take any element in U; we will show in the following that the corresponding transfer function has the pole assignability. Note that $\rho_{B_0}(gh, X)$ is a nonzero scalar multiple of

$$\det\left(\begin{bmatrix} I_{p+d} & N^{(gh)}(s) \\ X & D^{(gh)}(s) \end{bmatrix}\begin{bmatrix} I & 0 \\ 0 & U(s) \end{bmatrix}\right),$$

where g is as in the above definition. It is well defined by the condition (1).

Because of $d > \frac{n-mp}{m}$, we have $a_n < p + d$. Hence the first column of $D^{(gh)}(s)$ is zero vector and $\det(D^{(gh)}(s)) = 0$. In other words, the map $\rho_{B_0}(gh, -)$ maps $0 \in M(m + d, p + d, \mathbf{R})$ to the origin.

Since the image of its differential at $0 \in M(m + d, p + d, \mathbf{R})$ is spanned by $m + d$ minors which have one row from $N^{(gh)}(s)$, other rows from $D^{(gh)}(s)$, we know that the differential is surjective by the condition (2).

Therefore, $\rho_{B_0}(gh, -)$ covers a neighborhood of the origin of $\mathbf{R}[s]_{n+d}$, which means that any element in $p(U)$ is pole assignable by degree d compensators. Since we can find an element in U, U is dense, and we have the result.

5. Necessary condition for generic pole placement. The second application of Proposition 3.4 is the following result, which was known, and was "proved" naively by counting parameters [9]. We give a proof to this result along this idea. To verify the idea, we consider a group action. We think that this new treatment, i.e., the Lie group theoretic approach to the pole placement problem, fits well to the further study, since, in Lie group theory, orbit structures are considered over $\mathbf{R}$, compared with algebraic geometry which is mainly considered over $\mathbf{C}$. J. Rosenthal gave another proof using a different method [6]. The following proof is new, but we also remark that for the proof of this necessary condition we replace the general theorem from Lie group theory by an elementary ring theoretic argument, since the difference between $\mathbf{R}$ and $\mathbf{C}$ does not matter at this stage. The real advantage of this Lie group method remains for future research.

THEOREM 5.1. *If $G(s) \in S_{m,p,n}$ is pole assignable by degree d compensators, then $d \geq \frac{n-mp}{m+p-1}$.*

Proof. Let $\rho'_{B_0} : M(m+p+2d, p+d, \mathbf{R}) \to \mathbf{R}[s]_{n+d}$ be an extended pole assignment map given by

$$\rho'_{B_0}\left(\begin{bmatrix} K_1 \\ K_2 \end{bmatrix}\right) = \det\left(\begin{bmatrix} K_1 & N(s) \\ K_2 & D(s) \end{bmatrix}\right).$$

The image of the set of full rank matrices is the same as the image of ρ_{B_0}. ρ'_{B_0} naturally induces the algebra homomorphism

$$\mathbf{R}[a_0, \ldots, a_{n+d}] \to \mathbf{R}[x_{ij}],$$

where a_i's are the coefficients of polynomials in $\mathbf{R}[s]_{n+d}$, x_{ij}'s are matrix coordinates of $M(m+p+2d, p+d, \mathbf{R})$. Since $G(s)$ is pole assignable by degree d compensators, Proposition 3.4 tells us that it is injective.

Let H be the group consisting of the elements

$$h = \begin{bmatrix} h_1 & 0 & 0 \\ 0 & I_m & 0 \\ 0 & 0 & h_2 \end{bmatrix}$$

satisfying

$$h_1 \in GL(p+d, \mathbf{R}), \quad h_2 \in GL(d, \mathbf{R}), \quad \det(h_1) = \det(h_2).$$

We denote $\begin{bmatrix} I_p & 0 \\ 0 & h_2 \end{bmatrix}$, $\begin{bmatrix} I_m & 0 \\ 0 & h_2 \end{bmatrix}$ by $\phi_1(h_2)$, $\phi_2(h_2)$, respectively. H acts on $M(m+p+2d, p+d, \mathbf{R})$ by

$$h \cdot \begin{bmatrix} K_1 \\ K_2 \end{bmatrix} = \begin{bmatrix} \phi_1(h_2) K_1 h_1^{-1} \\ \phi_2(h_2) K_2 h_1^{-1} \end{bmatrix}.$$

We then have

$$\rho'_{B_0}\left(h\cdot\begin{bmatrix} K_1 \\ K_2 \end{bmatrix}\right) = \det\left(\begin{bmatrix} \phi_1(h_2)K_1h_1^{-1} & N(s) \\ \phi_2(h_2)K_2h_1^{-1} & D(s) \end{bmatrix}\right)$$

$$= \det\left(\begin{bmatrix} I_{p+d} & 0 \\ 0 & \phi_2(h_2) \end{bmatrix}\right)\det\left(\begin{bmatrix} \phi_1(h_2)K_1 & N(s) \\ K_2 & \phi_2(h_2)^{-1}D(s) \end{bmatrix}\right)$$

$$\times \det\left(\begin{bmatrix} h_1^{-1} & 0 \\ 0 & I_{m+d} \end{bmatrix}\right).$$

Since $D(s)$ commutes with $\phi_2(h_2)$, it equals

$$= \det(h_2)\det(h_1)^{-1}\det\left(\begin{bmatrix} \phi_1(h_2)K_1 & N(s) \\ K_2 & D(s)\phi_2(h_2)^{-1} \end{bmatrix}\right)$$

$$= \det\left(\begin{bmatrix} \phi_1(h_2)K_1 & N(s)\phi_2((h_2) \\ K_2 & D(s) \end{bmatrix}\right)\det\left(\begin{bmatrix} I_{p+d} & 0 \\ 0 & \phi_2(h_2)^{-1} \end{bmatrix}\right).$$

Similarly, we use $\phi_1(h_2)N(s) = N(s)\phi_2(h_2)$. Then,

$$= \det(h_1)^{-1}\det\left(\begin{bmatrix} \phi_1(h_2)K_1 & \phi_1(h_2)N(s) \\ K_2 & D(s) \end{bmatrix}\right) = \rho'_{B_0}\left(\begin{bmatrix} K_1 \\ K_2 \end{bmatrix}\right).$$

Hence we have that $\mathbf{R}[a_0,\ldots,a_{n+d}]$ is embedded in $\mathbf{R}[x_{ij}]^H$.

Take a Zariski open set U of $M(m+p+2d, p+d, \mathbf{R})$ consisting of elements $\begin{bmatrix} K_1 \\ K_2 \end{bmatrix}$ satisfying

(1) $\det(K_1) \neq 0$,

(2) $K_2 = \begin{bmatrix} N & M \\ L & K \end{bmatrix}$ such that K has d distinct eigenvalues, (K, L) controllable.

Then H acts on U freely. Now we consider an embedding $H \times U_0 \to U : (h, u) \mapsto h \cdot \begin{bmatrix} I_{p+d} \\ u \end{bmatrix}$, where U_0 is the set of $\begin{bmatrix} N & M \\ L & K \end{bmatrix}$ such that

(1) the last column of L is $\begin{bmatrix} 0 & \cdots & 1 \end{bmatrix}^T$,

(2) K is of the form

$$\begin{bmatrix} 0 & 1 & \cdots \\ \cdots & \cdots & \cdots \\ \cdots & \cdots & 1 \\ -k_1 & \ldots & -k_d \end{bmatrix},$$

which has distinct eigenvalues.

H acts on $H \times U_0$ by left multiplication, and the ring of H invariants is the polynomial ring in coordinates for U_0. By comparing the transcendental degrees of polynomial rings $\mathbf{C}[a_0,\ldots,a_{n+d}]$ and the coordinate ring of U_0, we can verify the dimension counting, which is

$$n + d + 1 \leq (m+p+2d)(p+d) - ((p+d)^2 + d^2 - 1) = mp + d(m+p) + 1,$$

which completes the proof. □

REFERENCES

[1] R. W. Brocket and C. I. Byrnes, *Multivariable Nyquist criteria, root loci and pole placement: A geometric viewpoint*, IEEE Trans. Automat. Control, 26 (1981), pp. 271–287.

[2] G. D. Forney, *Minimal bases of rational vector spaces, with applications to multivariable linear systems*, SIAM J. Control Optim., 13 (1975), pp. 493–520.

[3] T. Kailath, *Linear Systems*, Prentice–Hall, Englewood Cliffs, NJ, 1980.

[4] M. Kuijper, *First Order Representations of Linear Systems*, Birkhauser-Verlag, Basel, Switzerland, 1994.

[5] H. H. Rosenbrook, *State Space and Multivariable Theory*, Nelson, London, 1970.

[6] J. Rosenthal, *On dynamic feedback compensation and compactification of systems*, SIAM J. Control Optim., 32 (1994), pp. 279–296.

[7] X. Wang, *Pole placement by static output feedback*, J. Math. Systems Estim. Control, 2 (1993), pp. 205–218.

[8] X. Wang and J. Rosenthal, *Pole placement with small order dynamic compensators*, in 31st IEEE Conference on Decision and Control, Tucson, AZ, 1992, pp. 3098–3099.

[9] J. C. Willems and W. H. Hesselink, *Generic properties of the pole placement problem*, in Proc. IFAC, Helsinki, Finland, 1978, pp. 1725–1729.

[10] J. Rosenthal and X. Wang, *Output feedback pole placement with dynamic compensators*, IEEE Trans. Automat. Control, 41 (1996), pp. 830–843.

[11] X. Wang, *Grassmannian, central projection, and output feedback pole placement of linear systems*, IEEE Trans. Automat. Control, 41 (1996), pp. 786–794.

[12] J. Leventides and N. Karcanias, *Global asymptotic linearization of the pole placement map: A closed form for the constant output feedback*, Automatica J. IFAC, 31 (1995), pp. 1303–1309.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 2, pp. 391–421, March 1998

LOCAL EXACT BOUNDARY CONTROLLABILITY OF THE BOUSSINESQ EQUATION*

A. V. FURSIKOV† AND O. YU. IMANUVILOV‡

Abstract. We study the local exact boundary controllability problem for the Boussinesq equations that describe an incompressible fluid flow coupled to thermal dynamics. The result that we get in this paper is as follows: suppose that $\hat{y}(t,x)$ is a given solution of the Boussinesq equation where $t \in (0,T)$, $x \in \Omega$, Ω is a bounded domain with C^∞-boundary $\partial\Omega$. Let $y_0(x)$ be a given initial condition and $\|\hat{y}(0,\cdot) - y_0\| < \epsilon$ where $\epsilon = \epsilon(\hat{y})$ is small enough. Then there exists boundary control u such that the solution $y(t,x)$ of the Boussinesq equations satisfying

$$y|_{(0,T)\times\partial\Omega} = u, \qquad y|_{t=0} = y_0$$

coincides with $\hat{y}(t,x)$ at the instant T : $y(T,x) \equiv \hat{y}(T,x)$.

Key words. Boussinesq equation, local exact boundary controllability

AMS subject classifications. 76D05, 49J20, 93B05, 93C20

PII. S0363012996296796

Introduction. We study the local exact controllability problem for the Boussinesq equations that describe the incompressible fluid flow coupled to thermal dynamics. The control function is the Dirichlet boundary condition of the velocity and temperature vector field of fluid flow. More precisely, the investigated local exact controllability problem is as follows: suppose that

$$\partial_t y(t,x) + A(y) = f(t,x), \qquad t \in (0,T), \qquad x \in \Omega, \tag{0.1}$$

is a symbolic writing of the Boussinesq equations defined in a bounded domain $\Omega \subset \mathbf{R}^n$, $n = 2, 3$, where $y(t,x)$ is a velocity and temperature vector field, $f(t,x)$ is an external forces vector field, and $t \in (0,T)$ is a time. Assume that a solution $\hat{y}(t,x)$ of (0.1),

$$\partial_t \hat{y}(t,x) + A(\hat{y}) = f(t,x),$$

as well as an initial condition $y_0(x)$, are given and they satisfy the proximity condition

$$\|\hat{y}(0,\cdot) - y_0(\cdot)\| \le \epsilon, \tag{0.2}$$

where $\|\cdot\|$ is the norm of the corresponding initial conditions space and $\epsilon > 0$ is a sufficiently small magnitude. One has to find control u defined on the lateral surface $\Sigma = (0,T) \times \partial\Omega$ of the cylinder $(0,T) \times \Omega$:

$$y|_\Sigma = u \tag{0.3}$$

*Received by the editors January 3, 1996; accepted for publication (in revised form) December 11, 1996.

http://www.siam.org/journals/sicon/36-2/29679.html

†Department of Mechanics and Mathematics, Moscow State University, Moscow 119899, Russia (fursikov@dial01.msu.ru). The research of this author was supported by the International Science Foundation and the government of Russia under grant M 76300 and Russian Foundation of Fundamental Investigations grant 96-01-00947.

‡Korea Institute for Advanced Study, 207-43 Chungryangri-dong, Dongdaemoon-ku, Seoul, Korea 130-012. The research of this author was supported by the International Science Foundation and the government of Russia under grants M 76300 and GARC-KOSEF.

such that the solution $y(t,x)$ of (0.1), (0.3) supplied by the initial condition

$$y|_{t=0} = y_0 \tag{0.4}$$

coincides with the given solution $\hat{y}(t,x)$ at instant $t = T$:

$$y|_{t=T} = \hat{y}|_{t=T}. \tag{0.5}$$

One useful application of the local exact controllability property is as follows. Let $f(t,x) \equiv f(x)$ be independent on t and $\hat{y}(x)$ be a steady-state solution of (0.1) with zero boundary condition which, by definition, is a singular point in the phase space of the dynamical system generated by equation (0.1) supplied by zero boundary conditions. Suppose that this point $\hat{y}(x)$ is an unstable one. Then solvability of problem (0.1), (0.3), (0.5) implies that one can transfer an arbitrary point y_0 belonging to a small neighborhood of $\hat{y}$ to $\hat{y}$ via a solution of (0.1) by appropriate choice of boundary control. Hence, it is possible to suppress the rise of turbulence with the help of boundary control.

The interest in controllability problems for equations simulating a fluid flow was initiated by J.-L. Lions in [29], [30]. Different kinds of approximate controllability results for the Stokes system were obtained by J.-L. Lions in [29], [30], in A. Fursikov and O. Imanuvilov [18], [20], in J. I. Diaz and A. V. Fursikov [6], in J.-L. Lions and E. Zuazua [33]. A close problem was considered in C. Fabre and G. Lebeau [8]. Approximate controllability of the semilinear heat equation with nonlinearity satisfying the global Lipschitz condition was studied by C. Fabre, J.-P. Puel, and E. Zuazua [9], [10]. Approximate controllability for a system associated with Navier–Stokes equations but possessing the sublinear growth of its nonlinearity was proved by C. Fabre [7].

The solvability of (0.1), (0.3)–(0.5) was first proved in A. Fursikov and O. Imanuvilov [15] for the case when (0.1) is the Burgers equation. This problem was solved in the case of Navier–Stokes equations and $\hat{y} \equiv 0$ in A. Fursikov and O. Imanuvilov [16] when the dimension of the system $n = 2$ and in A. Fursikov [12] when $n = 3$. The case of Navier–Stokes equations and $\hat{y} \neq 0$ has been studied in A. Fursikov and O. Imanuvilov [17]. The results proved in this work as well as analogous results for the Navier–Stokes equations were announced in [19].

J.-M. Coron established exact controllability of the two-dimensional Euler equation in [3], [4] and approximate controllability of the two-dimensional Navier–Stokes equations in [5].

At last, exact controllability of a semilinear parabolic equation with coefficients depending on t and x and with nonlinearity satisfying the global Lipschitz condition was established by O. Imanuvilov [23]–[26]. A similar result was obtained by G. Lebeau and L. Robbiano [28] for the linear heat equation on manifolds. We are interested in the Boussinesq equations because the investigation of a fluid flow stability in the free convection problem is important in the theory of hydrodynamical stability (see D. Joseph [27]). Besides, as we think, the exact controllability result for Boussinesq equations should be useful for solution of certain reversibility problems in the theory of climate (see J.-L. Lions [29], [30]).

The first step of the controllability property proof is a reduction of the nonlinear problem (0.1), (0.3)–(0.5) to the solvability of the analogous problem for the linearization of (0.1). We do it with the help of one variant of the implicit function theorem. To establish solvability of the controllability problem we prove that the set of the data

for which the linear controllability problem has a solution is dense (section 3) and is closed (section 5) in the corresponding function space.

The main difficulties of the proof are connected with the pressure term in the Boussinesq equations. To overcome this difficulty we introduce some nonstandard functional spaces and construct in these spaces a decomposition of a vector field on solenoidal and potential components (section 4). This decomposition is based on the Carleman estimate for the Laplace operator (L. Hörmander [21], [22]) and for the heat equation (section 6). Note that in papers [12], [15], [16] mentioned above we also used Carleman estimates, but the method of proof of Carleman estimates which was applied in section 6 is more close to O. Imanuvilov [26] and A. Fursikov and O. Imanuvilov [17].

1. Statement of the problem and formulation of the main result. In a bounded domain $\Omega \subset \mathbf{R}^n$ ($n = 2$ or 3) with C^∞-boundary $\partial\Omega$ we consider the Boussinesq system

$$\partial_t v(t,x) - \Delta v + (v,\nabla)v + \theta(t,x)e_0 + \nabla p(t,x) = f(t,x), \tag{1.1}$$

$$\mathbf{div}\, v \equiv \sum_{j=1}^{n} \partial_{x_j} v_j = 0, \tag{1.2}$$

$$\partial_t \theta(t,x) - \Delta\theta + (v,\nabla\theta) + (v,e_0) = h(t,x), \tag{1.3}$$

$$v(t,x)|_{t=0} = v_0(x), \qquad \theta(t,x)|_{t=0} = \theta_0(x), \tag{1.4}$$

$$v|_\Sigma = u_v, \qquad \theta|_\Sigma = u_\theta. \tag{1.5}$$

Here $(t,x) \in Q = (0,T)\times\Omega$, $v(t,x) = (v_1(t,x),\dots,v_n(t,x))$ is the velocity of a fluid at point x and at instant t, $\theta(t,x)$ is the temperature of a fluid, ∇p is a pressure gradient, $f(t,x)$ is the density of external forces, $h(t,x)$ is the density of external heat sources, $e_0 \in \mathbf{R}^n$ is the vector of the gravity force direction, u_v, u_θ are Dirichlet boundary conditions (in our case, they are control functions), and v_0, θ_0 are initial conditions. Besides, $\Sigma = (0,T)\times\partial\Omega$, $\partial_t = \partial/\partial t$, $\partial_{x_j} = \partial/\partial_{x_j}$, Δ is the Laplace operator, $(v,\nabla)v = \sum v_j \partial_{x_j} v$, and $(v,\nabla\theta) = \sum_{j=1}^n v_j \partial_{x_j}\theta$.

We investigate the local exact controllability problem for Boussinesq equations. Its formulation is as follows. Suppose that $\hat v(t,x)$, $\hat p(t,x)$, $\hat\theta(t,x)$ is a given sufficiently smooth[1] solution of Boussinesq equations (1.1)–(1.3):

$$\partial_t \hat v(t,x) - \Delta\hat v + (\hat v,\nabla)\hat v + \hat\theta(t,x)e_0 + \nabla\hat p(t,x) = f(t,x),$$
$$\mathbf{div}\,\hat v = 0,$$
$$\partial_t\hat\theta(t,x) - \Delta\hat\theta(t,x) + (\hat v,\nabla\hat\theta) + (\hat v,e_0) = h(t,x),$$

and initial conditions $v_0(x)$, $\theta_0(x)$ are sufficiently close to $\hat v(0,x)$, $\hat\theta(0,x)$ with respect to an appropriate norm. One has to find boundary control (u_v,u_θ) defined on Σ such that the solution $(v(t,x),p(t,x),\theta(t,x))$ of boundary value problem (1.1)–(1.5) satisfies at instant $t = T$ the relations

$$v(T,x) \equiv \hat v(T,x), \qquad \theta(T,x) \equiv \hat\theta(T,x). \tag{1.6}$$

Let us introduce the functional spaces to set precisely the controllability problem and to formulate the main result. We use the Sobolev spaces $W_p^k(\Omega)$, $1 \le p < \infty$, k integer

[1] The precise smoothness conditions are formulated below.

not negative, possessing the norm

$$\|u\|_{W_p^k(\Omega)} = \left(\int_\Omega \sum_{|\alpha|\le k} \left| \frac{\partial^{|\alpha|} u(x)}{\partial x_1^{\alpha_1} \dots \partial x_n^{\alpha_n}} \right|^p \, dx \right)^{1/p},$$

where $\alpha = (\alpha_1, \dots, \alpha_n)$, $\mid \alpha \mid = \alpha_1 + \ldots + \alpha_n$. We also use the Sobolev spaces $W_2^\alpha(\Omega) = H^\alpha(\Omega)$ with an arbitrary real α. See their definition in J.-L. Lions and E. Magenes [32].

Define the functional space $V^k(\Omega)$ of solenoidal vector fields

$$V^k(\Omega) = \{v(x) \in (H^k(\Omega))^n : \mathbf{div}\, v(x) = 0\}. \tag{1.7}$$

We need the following spaces of functions defined in the cylinder Q:

$$W^{1,2(k)}(Q) = \{\theta(t,x) \in L_2(0,T; H^{k+2}(\Omega)) : \partial_t \theta \in L_2(0,T; H^k(\Omega))\}, \tag{1.8}$$

$$V^{1,2(k)}(Q) = \{v(t,x) \in (W^{1,2(k)}(Q))^n : \mathbf{div}\, v = 0\}. \tag{1.9}$$

The main result of this paper is as follows.

THEOREM 1.1. *Suppose that $f(t,x) \in (W^{1,2(2)}(Q))^n$, $h(t,x) \in W^{1,2(2)}(Q)$ are given data and $(\hat v(t,x), \hat p(t,x), \hat\theta(t,x)) \in V^{1,2(2)}(Q) \times L_2(0,T; H^3(\Omega)) \times W^{1,2(2)}(Q)$ is a solution of equations* (1.1)–(1.3)*, satisfying the property*

$$\int_{\Gamma_j} (\hat v(t,x), \nu(x))\, d\sigma = 0, \qquad j = 1, \dots, r, \qquad t \in [0,T), \tag{1.10}$$

where Γ_j are components of $\partial\Omega$: $\partial\Omega = \cup_{j=0}^r \Gamma_j$, $\Gamma_j \cap \Gamma_k = \{\emptyset\}$, if $j \ne k$, $\nu(x)$ is the vector field of outside normals to $\partial\Omega$. Suppose that $(v_0(x), \theta_0(x)) \in V^1(\Omega) \times H^1(\Omega)$ is a given initial datum satisfying conditions

$$\int_{\Gamma_j} (v_0(x), \nu(x))\, d\sigma = 0, \qquad j = 1, \dots, r, \tag{1.11}$$

which is close to $(\hat v(0,x), \hat\theta(0,x))$:

$$\|v_0 - \hat v(0,\cdot)\|^2_{V^1(\Omega)} + \|\theta_0 - \hat\theta(0,\cdot)\|^2_{H^1(\Omega)} < \epsilon, \tag{1.12}$$

where $0 < \epsilon \le \epsilon_0$ and ϵ_0 is of sufficiently small magnitude depending on $(\hat v, \hat\theta)$. Then there exists boundary control $(u_v, u_\theta) \in (L_2(\Sigma))^n \times L_2(\Sigma)$ such that there exists the solution $(v, p, \theta) \in V^{1,2(0)}(Q) \times L_2(0,T; H^1(\Omega)) \times W^{1,2(0)}(Q)$ of problem (1.1)–(1.5) *and this solution satisfies condition* (1.6)*. Moreover, there exist constants $\kappa > 0$, $c_1 > 0$ such that*

$$\|v(t,\cdot) - \hat v(t,\cdot)\|^2_{V^1(\Omega)} + \|\theta(t,\cdot) - \hat\theta(t,\cdot)\|^2_{H^1(\Omega)} \le c_1 e^{-\frac{\kappa}{(T-t)}} \qquad \text{as } t \to T. \tag{1.13}$$

In the remaining part of the paper we prove this theorem.

Remark 1.1. The condition $(\hat v, \hat p, \hat\theta) \in V^{1,2(2)}(Q) \times L_2(0,T; H^3(\Omega)) \times W^{1,2(2)}(Q)$ of Theorem 1.1 can be weakened. Namely, the assertion of Theorem 1.1 remains true if, instead of the assumptions mentioned above, we suppose that

$$\hat v(t,x) \in V^{1,2(1/2)}(Q) \cap (L_\infty(Q))^n, \qquad \hat\theta \in W^{1,2(1/2)}(Q). \tag{1.14}$$

In the case of assumption (1.14) we have to add to the proof of Theorem 1.1 with some more or less complicated applications of the Sobolev embedding theorem and also with one technical method mentioned below in Remark 5.1.

2. Reduction to a linear controllability problem.

2.1. Let us begin with a simple but useful remark. We will not specially construct the boundary control (v_u, θ_u). Instead, we will study the solvability of problem (1.1)–(1.4), (1.6), which does not contain boundary conditions (1.5). We will find a boundary control $(v_u,\, \theta_u)$ at the very end of the proof with the help of a restriction of the constructed solution (v, θ) at the boundary Σ.

We show now that one can reduce the controllability problem mentioned above to the case of bounded domain Ω with a connected boundary. Indeed, let Γ_0 be the external component of the boundary $\partial\Omega$. We denote by G the bounded domain with the boundary Γ_0. Evidently

$$G = \Omega \cup \cup_{j=1}^{r} (\Omega_j \cup \Gamma_j),$$

where Ω_j is the bounded domain with the boundary Γ_j. To reduce the proof of Theorem 1.1 to the case of domain G with connected boundary we have to continuously extend functions $(\hat{u}, \hat{p}, \hat{\theta}) \in V^{1,2(2)}(Q) \times L_2(0,T; W_2^3(\Omega)) \times W^{1,2(2)}(Q)$ up to $(\tilde{u}, \tilde{p}, \tilde{\theta}) \in V^{1,2(2)}(\hat{Q}) \times L_2(0,T; W_2^3(G)) \times W^{1,2(2)}(\hat{Q})$ where $\hat{Q} = (0,T) \times G$ and initial conditions $(v_0, \theta_0) \in V^1(\Omega) \times H^1(\Omega)$ up to $(\tilde{v}, \tilde{\theta}) \in V^1(G) \times H^1(G)$. After this extension we substitute $(\tilde{v}, \tilde{p}, \tilde{\theta})$ into (1.1), (1.3) and calculate the right side $(\tilde{f}, \tilde{h})$ of these equations. Naturally, $(\tilde{f}, \tilde{h})$ will be an extension of (f, h). When we prove Theorem 1.1 in the case of domain G, we will restrict the solution of the controllability problem at $\partial\Omega = \cup_{j=0}^{r} \Gamma_j$. Then the constructed function (v_u, θ_u) will be the control which solves the controllability problem in the case of domain Ω with disconnected boundary.

PROPOSITION 2.1. *For an arbitrary natural number l there exists the extension operator $L : L\theta(x)|_\Omega \equiv \theta(x)$ such that the maps*

$$L : H^k(\Omega) \to H^k(G)$$

are bounded for $k = 0, \dots, l$.

Although the proof of this proposition is well known we briefly revisit the construction of the extension, taking into account our future goals. After application of a partition of unity and rectification of the boundary, we obtain the extension problem for a function $u(x)$ defined in $\mathbf{R}^n_+ = \{x = (x_1, \dots, x_n), x_n > 0\}$ up to a function defined on $\mathbf{R}^n$. The extension operator L is now defined by the formula

$$Lu(x', x_n) = \begin{cases} u(x', x_n) & \text{when } x_n \geq 0, \\ \sum_{k=1}^{l} \lambda_k u(x', -x_n/k) & \text{when } x_n < 0, \end{cases}$$

where $\lambda_1, \dots, \lambda_n$ are the solutions of system

$$\sum_{k=1}^{l} \left(-\frac{1}{k}\right)^j \lambda_k = 1 \qquad (j = 0, 1, \dots, l-1).$$

This construction allows us to prove estimates declared in Proposition 2.1 (see V. Babich [2], L. Slobodetskii [37]). This construction and Proposition 2.1 imply the following.

PROPOSITION 2.2. *For an arbitrary natural k there exists a bounded extension operator*

$$L : W^{1,2(k)}(Q) \to W^{1,2(k)}(\hat{Q}), \qquad \hat{Q} = (0,T) \times G,$$
$$L : L_2(0,T; H^k(\Omega)) \to L_2(0,T; H^k(G)).$$

Let us consider functional spaces of solenoidal vector fields. We define the space

$$\hat{V}^k(\Omega) = \{v_0 \in V^k(\Omega) : v_0 \text{ satisfies (1.11)}\}.$$

Remark 2.1. Below we use the operator **rot**. Its definition in the three-dimensional case is well known. In the case when $\dim \Omega = 2$ we define the operator **rot** by formula

$$\mathbf{rot}\, u = \partial_{x_1} u_2 - \partial_{x_2} u_1.$$

PROPOSITION 2.3. (i) *For an arbitrary natural number k there exists the extension operator $\hat{L}$ such that the maps*

$$\hat{L} : \hat{V}^k(\Omega) \to V^k(G)$$

are bounded for $k = 0, 1, \dots, l$.

(ii) *For an arbitrary natural number k there exist bounded extension operators*

$$\hat{L} : V^{1,2(k)}(Q) \to V^{1,2(k)}(\hat{Q}), \qquad \hat{Q} = (0,T) \times G,$$
$$\hat{L} : L_2(0,T; V^k(\Omega)) \to L_2(0,T; V^k(G)).$$

Proof. Denote $H_\sigma = \{v \in V^0(\Omega) : (v,\nu)|_{\partial\Omega} = 0\}$, where (v,ν) is understood as an element belonging to $W^{-1/2}(\Omega)$ (see details in R. Temam [38]). For $u \in V^k(\Omega)$ we consider the boundary value problem

$$\mathbf{rot}\, v = u, \qquad x \in \Omega,$$
$$\mathbf{div}\, v = 0, \qquad x \in \Omega,$$
$$(v,\nu)|_{\partial\Omega} = 0.$$

There exists a solution $v \in V^k(\Omega) \cap H_\sigma$ of this problem, which satisfies the estimate

$$\|v\|_{V^{k+1}(\Omega)} \le c_1(\|u\|_{V^k(\Omega)} + \|v\|_{(L_2(\Omega))^n});$$

moreover, if we will take v from an orthogonal complement to $\operatorname{Ker} \mathbf{rot}\, V^1(\Omega)$ in the space H_σ, then

$$\|v\|_{(L_2(\Omega))^n} \le c_2 \|u\|_{(L_2(\Omega))^n}$$

(see R. Temam [38]). Hence, for such v we have the estimate

$$\|v\|_{V^{k+1}(\Omega)} \le c_3 \|u\|_{V^k(\Omega)}.$$

Now, for $u \in V^k(\Omega)$ we define the extension operator $\hat{L}$ by formula

$$\hat{L}u = \mathbf{rot}\, Lv,$$

where L is the extension operator from Proposition 2.1 and v is the solenoidal vector field constructed above by u. Evidently, the estimate for v written above and Propositions 2.1, 2.2 imply assertions (i) and (ii) of Proposition 2.3. □

2.2. Now we reduce the proof of Theorem 1.1 to the case of a linear controllability problem. Applying the well-known formula of vector analysis,

$$(v, \nabla)v = -v \times \mathbf{rot}\, v + \nabla(|v|^2/2),$$

where $\times$ is the operation of vector multiplication, we can rewrite equation (1.1) in the form

$$\partial_t v(t,x) - \Delta v - v \times \mathbf{rot}\, v + \theta(t,x) e_0 + \nabla p'(t,x) = f(t,x), \tag{2.1}$$

where $\nabla p' = \nabla(p + |v|^2/2)$. We write the solution (v, θ) in the form

$$v(t,x) = \hat{v}(t,x) + w(t,x), \qquad \theta(t,x) = \hat{\theta}(t,x) + \tau(t,x), \tag{2.2}$$

substitute (2.2) into equations (2.1), (1.2), (1.3), and subtract from them the same equations for $(\hat{v}, \hat{p}, \hat{\theta})$. As a result we get

$$\mathcal{N}(w,q,\tau) = \partial_t w(t,x) - \Delta w - \hat{v} \times \mathbf{rot}\, w - w \times \mathbf{rot}\, \hat{v} - w \times \mathbf{rot}\, w + \nabla q + \tau e_0 = 0, \tag{2.3}$$

$$\mathbf{div}\, w = 0, \tag{2.4}$$

$$\mathcal{H}(w,\tau) = \partial_t \tau(t,x) - \Delta \tau + (\hat{v}, \nabla\tau) + (w, \nabla\hat{\theta}) + (w, \nabla\tau) + (w, e_0) = 0, \tag{2.5}$$

where $\nabla q = \nabla p' - \nabla \hat{p}$. The functions w, τ satisfy the initial conditions:

$$w(0,x) = w_0(x), \qquad \tau(0,x) = \tau_0(x), \tag{2.6}$$

where $w_0(x) = v_0(x) - \hat{v}(0,x)$, $\tau_0(x) = \theta_0(x) - \hat{\theta}(0,x)$. Evidently we have reduced our problem to the construction of a solution $(w(t,x), \tau(t,x))$ of problem (2.3)–(2.6), which satisfies the equalities

$$w(T,x) = 0, \qquad \theta(T,x) = 0. \tag{2.7}$$

Remark 2.2. In the two-dimensional case we will rewrite the nonlinear term $(v, \nabla)v$ in the form

$$(v, \nabla)v = (-v_2\, \mathbf{rot}\, v, v_1\, \mathbf{rot}\, v) + \nabla(|v|^2/2)$$

and derive the analog of (2.3)–(2.7) in the same way as above. The obtained system differs from (2.3)–(2.7), but the proof of Theorem 1.1 will not differ from the proof given below for the three-dimensional case.

We will solve problem (2.3)–(2.7) with the help of the following variant of the implicit function theorem.

THEOREM (on a right inverse operator). *Suppose that X, Z are Banach spaces and*

$$\mathcal{A} : X \to Z \tag{2.8}$$

is a continuously differentiable map. We assume that for $x_0 \in X$, $z_0 \in Z$ the equality

$$\mathcal{A}(x_0) = z_0 \tag{2.9}$$

holds and the derivative $\mathcal{A}'(x_0) : X \to Z$ of the map $\mathcal{A}$ at x_0 is a surjective operator. Then there exists $\epsilon > 0$ such that for any $z \in Z$ which satisfies the condition

$$\|z - z_0\|_Z < \epsilon,$$

there exists a solution $x \in X$ of equation

$$\mathcal{A}(x) = z.$$

This theorem is a simple corollary of the generalization of the implicit function theorem which has been proved in V. Alekseev, V. Tikhomirov, and S. Fomin [1]. In our case X will be a space of triplets $x = (w, q, \tau)$,

$$\mathcal{A}(x) = (\mathcal{N}(w, q, \tau), \mathcal{H}(w, \tau), w|_{t=0}, \tau|_{t=0}), \tag{2.10}$$

and the collection of components in (2.10) defines the space Z. We note that (2.7) will be guaranteed by the insertion of special weights on the norm of X. We take $x_0 = (0, 0, 0)$, $z_0 = (0, 0, 0)$. Then equation (2.9) for operator (2.10), (2.3), (2.5) is fulfilled. Thus, the main condition that should be verified to apply the right inverse operator theorem is the assertion of solvability of the equation $\mathcal{A}'(0)x = z$ for any $z \in Z$. This equation in our case has the following form:

$$\mathcal{N}'(0)(v, p, \theta) = \partial_t v(t, x) - \Delta v - \hat{v} \times \mathbf{rot}\, v - v \times \mathbf{rot}\, \hat{v} + \theta e_0 + \nabla p = f, \tag{2.11}$$

$$\mathbf{div}\, v = 0, \tag{2.12}$$

$$\mathcal{H}'(0)(v, \theta) = \partial_t \theta(t, x) - \Delta\theta + (\hat{v}, \nabla\theta) + (v, \nabla\hat{\theta}) + (v, e_0) = h, \tag{2.13}$$

$$v|_{t=0} = v_0, \qquad \theta|_{t=0} = \theta_0, \tag{2.14}$$

$$v|_{t=T} = 0, \qquad \theta|_{t=T} = 0. \tag{2.15}$$

2.3. We now define the functional spaces X, Z corresponding to the problem (2.3)–(2.7). Let

$$\eta(t, x) \equiv \eta^s(t, x) = s(e^{2\hat{x}_1} - e^{x_1})/(T - t) \tag{2.16}$$

be the weight function where $s > 0$ is a parameter which will be chosen below, $\hat{x}_1 = \max_{x=(x_1 \dots x_n)\in\Omega} |x_1|$.

Denote

$$L_2(Q, \eta) \equiv L_2(Q, \eta^s) = \left\{ y(t, x) : \|y\|^2_{L_2(Q,\eta)} = \int_Q e^{2\eta^s} |y|^2 \, dx\, dt < \infty \right\}. \tag{2.17}$$

Below we will also use the space $L_2(Q, \beta)$ with different weights. We define the space $\mathbf{\Theta}(Q, \eta)$ of components $\theta(t, x)$ in (2.11)–(2.15):

$$\begin{aligned} \mathbf{\Theta}(Q, \eta) \equiv \mathbf{\Theta}(Q, \eta^s) = \Bigg\{ & \theta(t, x), (t, x) \in Q : \ \|\theta\|^2_{\mathbf{\Theta}(Q,\eta^s)} \equiv \|\partial_t\theta - \Delta\theta\|^2_{L_2(Q,\eta^s)} \\ & + \|(T - t)^{-3/2}\theta\|^2_{L_2(Q,\eta^s)} + \|(T - t)^{-1/2}|\nabla\theta|\|^2_{L_2(Q,\eta^s)} \\ & + \|(T - t)^{1/2}\partial_t\theta\|^2_{L_2(Q,\eta^s)} + \sum_{i,j=1}^{n} \|(T - t)^{1/2}\partial^2_{x_i x_j}\theta\|^2_{L_2(Q,\eta^s)} < \infty \Bigg\}. \end{aligned} \tag{2.18}$$

The space of right side components f in (2.11)–(2.15) is as follows:

$$\begin{aligned} F(Q, \eta) \equiv F(Q, \eta^s) = \{ & f \in (L_2(Q))^n : \ \exists \ f_1 \in (L_2(Q, \eta))^n, \\ & \exists \ f_2 \in L_2(0, T; H^1(\Omega)) \text{ such that } f = f_1 + \nabla f_2 \}. \end{aligned} \tag{2.19}$$

The norm of the space $F(Q, \eta)$ is defined by the relation

$$\|f\|_{F(Q,\eta^s)} = \inf_{\substack{f_1, \nabla f_2 \\ f = f_1 + \nabla f_2}} (\|f_1\|^2_{(L_2(Q,\eta^s))^n} + \|\nabla f_2\|^2_{(L_2(Q))^n})^{1/2}. \tag{2.20}$$

Remark 2.3. Note that $F(Q,\eta^s)$ is a Hilbert space. Indeed, since the functional $J(f_1,f_2)=(\|f_1\|^2_{(L_2(Q,\eta^s))^n}+\|\nabla f_2\|^2_{(L_2(Q))^n})^{1/2}$ is strictly convex, for an arbitrary $f\in F(Q,\eta^s)$ in the set of pairs $(f_1,\nabla f_2)\in L_2(0,T;H^1(\Omega))\times(L_2(Q,\eta))^n$ satisfying $f=f_1+\nabla f_2$ there exists the unique pair $(\hat f_1,\nabla\hat f_2)\in(L_2(Q,\eta^s))^n\times(L_2(Q))^n$ such that $\|f\|_{F(Q,\eta^s)}=J(\hat f_1,\hat f_2)$. Hence, the map $f=\hat f_1+\nabla\hat f_2\to(\hat f_1,\nabla\hat f_2)$ establishes isometric isomorphism between $F(Q,\eta^s)$ and $L_2(0,T;H^1(\Omega))\times L_2(Q,\eta^s)$, which implies that $F(Q,\eta^s)$ is a Hilbert space.

We define the space $\Xi(Q,\eta)$ of components v in (2.11)–(2.15) with the help of the equality

$$\Xi(Q,\eta)\equiv\Xi(Q,\eta^s)=\Bigg\{v(t,x):\ \ \mathbf{div}\,v=0,\quad \|v\|^2_{\Xi(Q,\eta^s)}$$

$$\equiv\|\partial_t v-\Delta v\|^2_{F(Q,\eta^s)}+\|(T-t)^{-1}v\|^2_{(L_2(Q,\eta^s))^n}+\|\nabla v\|^2_{(L_2(Q,\eta^s))^n}\tag{2.21}$$

$$+\|(T-t)\partial_t v\|^2_{(L_2(Q,\eta^s))^n}+\sum_{i,j=1}^n\|(T-t)\partial^2_{x_ix_j}v\|^2_{(L_2(Q,\eta^s))^n}<\infty\Bigg\}.$$

Now we can define the spaces X and Z in the case of problems (2.3)–(2.7) or (2.11)–(2.15):

$$X=X^s(Q)=\Xi(Q,\eta^s)\times L_2(0,T;H^1(\Omega))\times\mathbf{\Theta}(Q,\eta^s),\tag{2.22}$$

$$Z=Z^s(Q)=F(Q,\eta^s)\times L_2(Q,\eta^s)\times V^1(\Omega)\times H^1(\Omega).\tag{2.23}$$

Since the weight function $\eta^s(t,x)$ increases exponentially as $t\to T$, the functions $v\in\Xi(Q,\eta^s)$, $\theta\in\mathbf{\Theta}(Q,\eta^s)$ decrease exponentially as $t\to T$, and therefore equalities (2.15) are true.

2.4. Let us show that for an arbitrary parameter $s>0$, operator (2.8) and its derivative

$$\mathcal{A}'(0):X^s(Q)\to Z^s(Q)\tag{2.24}$$

are continuous, where $\mathcal{A}(x)$ is defined as in (2.10), (2.3), (2.5).

LEMMA 2.1. *Suppose that* $\hat v\in V^{1,2(2)}(Q)$, $\hat\theta\in W^{1,2(2)}(Q)$,

$$\mathcal{A}'(0)(v,p,\theta)=(\mathcal{N}'(0)(v,p,\theta),\mathcal{H}'(0)(v,\theta),v|_{t=0},\theta|_{t=0}),\tag{2.25}$$

where $\mathcal{N}'(0)$, $\mathcal{H}'(0)$ *are defined by* (2.11), (2.13). *Then, for* $s>0$, *the operator* (2.24) *is continuous.*

Proof. Since $n=\dim\Omega=2,3$, the embeddings $\Xi(Q,\eta)\subset V^{1,2(0)}(Q)$, $\mathbf{\Theta}(Q,\eta)\subset W^{1,2(0)}(Q)$ are continuous. Since the restriction operator $\gamma_0 y=y|_{t=0}$ acts continuously from $W^{1,2(0)}(Q)$ to $H^1(\Omega)$ and from $V^{1,2(0)}(Q)$ to $V^1(\Omega)$ (see [32]), the inequalities

$$\|\gamma_0 v\|_{V^1(\Omega)}\le c_4\|v\|_{\Xi(Q,\eta)},\qquad \|\gamma_0\theta\|_{H^1(\Omega)}\le c_5\|\theta\|_{\Theta(Q,\eta)}\tag{2.26}$$

hold. Let us prove that the operator

$$\mathcal{H}'(0):\Xi(Q,\eta)\times\mathbf{\Theta}(Q,\eta)\ \to L_2(Q,\eta)\tag{2.27}$$

defined in (2.13) is continuous. Since the embeddings

$$V^{1,2(2)}(Q) \subset C(0,T;C^1(\bar\Omega))^n, \qquad W^{1,2(2)}(Q) \subset C(0,T;C^1(\bar\Omega)) \tag{2.28}$$

are continuous for $n \le 3$, we obtain, taking into account (2.13), (2.17), (2.18),

$$\begin{aligned}\|\mathcal{H}'(0)(v,\theta)\|_{L_2(Q,\eta)} &\le \|\partial_t\theta - \Delta\theta\|_{L_2(Q,\eta)} + \|\hat v\|_{(C(\bar Q))^n}\|\nabla\theta\|_{(L_2(Q,\eta))^n}\\ &+ \|\nabla\hat\theta\|_{(C(\bar Q))^n}\|v\|_{(L_2(Q,\eta))^n} \le (1+\|\hat v\|_{V^{1,2(2)}(Q)})\|\theta\|_{\mathbf{\Theta}(Q,\eta)} + \|\nabla\hat\theta\|_{(C(\bar Q))^n}\|v\|_{\Xi(Q,\eta)}.\end{aligned} \tag{2.29}$$

The relations (2.11), (2.17)–(2.21) yield

$$\begin{aligned}\|\mathcal{N}'(0)(v,p,\theta)\|_{F(Q,\eta)} &\le \|\partial_t v - \Delta v - \hat v\times\mathbf{rot}\, v - v\times\mathbf{rot}\,\hat v + \theta e_0\|_{(L_2(Q,\eta))^n}\\ &+ \|\nabla p\|_{(L_2(Q))^n} \le \|\partial_t v + \Delta v\|_{(L_2(Q,\eta))^n} + \|\hat v\|_{C(0,T;(C^1(\bar\Omega))^n)}(\|v\|_{(L_2(Q,\eta))^n}\\ &+\||\nabla v|\|_{(L_2(Q,\eta))^n}) + c_6\|\theta\|_{L_2(Q,\eta)} + \|\nabla p\|_{(L_2(Q))^n}\\ &\le (1+\|\hat v\|_{V^{1,2(2)}(Q)})\|v\|_{\Xi(Q,\eta)} + c_7\|\theta\|_{\mathbf{\Theta}(Q,\eta)} + \|p\|_{L_2(0,T;H^1(\Omega))}.\end{aligned} \tag{2.30}$$

The inequalities (2.26), (2.29), (2.30) imply the desired assertion. □

LEMMA 2.2. *Suppose that $\hat v \in V^{1,2(2)}(Q)$, $\hat\theta \in W^{1,2(2)}(Q)$, and $\mathcal{A}$ is operator (2.10). Then for arbitrary $s > 0$ the operator*

$$\mathcal{A} : X^s(Q) \to Z^s(Q)$$

is continuous.

Proof. To prove this lemma we need only to complete the proof of Lemma 2.1 by the estimate of the terms $w\times\mathbf{rot}\, w$ and $(w,\nabla\tau)$. The Cauchy–Bunyakovskii inequality and the Sobolev embedding theorem yield:

$$\begin{aligned}\|e^\eta(w,\nabla\tau)\|_{L_2(Q)} &\le \int_0^T \|e^{\frac\eta2}w(t,\cdot)\|_{(L_4(\Omega))^n}\|e^{\frac\eta2}|\nabla\tau(t,\cdot)|\|_{L_4(\Omega)}\,dt\\ &\le c_8\int_0^T \|e^{\frac\eta2}w(t,\cdot)\|_{V^1(\Omega)}\|e^{\frac\eta2}\tau(t,\cdot)\|_{W_2^2(\Omega)}\,dt\\ &\le c_9\|e^{\frac\eta2}w\|_{C(0,T;V^1(\Omega))}\|e^{\frac\eta2}\tau\|_{L_2(0,T;W_2^2(\Omega))}\\ &\le c_{10}\|e^{\frac\eta2}w\|_{V^{1,2(0)}(Q)}\|e^{\frac\eta2}\tau\|_{L_2(0,T;W_2^2(\Omega))}.\end{aligned} \tag{2.31}$$

Taking into account (2.16), the evident inequality

$$(T-t)^{-k} \le c(k)e^{\frac\eta2},$$

and the definition of norms in the right side of (2.30), we get the upper bound:

$$\begin{aligned}\|e^{\frac\eta2}w\|_{V^{1,2(0)}(Q)}\|e^{\frac\eta2}\tau\|_{L_2(0,T;W_2^2(\Omega))} &\le c_{11}\Big(\|e^{\frac\eta2}(T-2)^{-2}w\|_{(L_2(Q))^n}\\ &+ \|e^{\frac\eta2}(T-t)^{-1}|\nabla w|\|_{(L_2(Q))^n} + \sum_{i,j=1}^n \|e^{\frac\eta2}\partial^2_{x_ix_j}w\|_{(L_2(Q))^n}\Big)\Bigg(\|e^{\frac\eta2}(T-t)^{-2}\tau\|_{L_2(Q)}\\ &+ \|e^{\frac\eta2}(T-t)^{-1}|\nabla\tau|\|_{L_2(Q)} + \sum_{i,j=1}^n \|e^{\frac\eta2}\partial^2_{x_ix_j}\tau\|_{L_2(Q)}\Bigg) \le c_{12}\|w\|_{\Xi(Q,\eta)}\|v\|_{\mathbf{\Theta}(Q,\eta)}.\end{aligned} \tag{2.32}$$

One can estimate the term $(w,\nabla)w$ analogously. □

Thus, in order to apply the right inverse operator theorem we have to prove surjectivity of the operator $\mathcal{A}'(0) : X^s(Q) \to Z^s(Q)$. To prove this assertion we will show that the image of this operator is dense in $Z^s(Q)$, and besides, it is a closed subset of $Z^s(Q)$ for s sufficiently large. These assertions imply that the image of $\mathcal{A}'(0)$ coincides with the whole space $Z^s(Q)$.

3. The solvability of the linear controllability problem for a dense set of data.

3.1. In order to solve the controllabilty problem for a dense set of data we need the Carleman estimate for elliptic and inverse parabolic equations.

We consider the Cauchy problem for the Laplace operator

$$\Delta z(x) = f(x), \qquad x \in \Omega, \qquad z|_{\partial\Omega} = \left.\frac{\partial z}{\partial \nu}\right|_{\partial\Omega} = 0, \tag{3.1}$$

where $\Omega \subset \mathbf{R}^n$ is a bounded domain with C^∞ boundary and $\partial/\partial\nu$ is the derivative along outside normal ν to $\partial\Omega$.

LEMMA 3.1. *Let $f(x) \in L_2(\Omega)$. There exists $s_0 > 0$ such that for any $s > s_0$ the solution $z(x) \in W_2^2(\Omega)$ of (3.1) satisfies the Carleman estimate:*

$$\int_\Omega \left(\frac{1}{s} \sum_{i,j=1}^n \left| \frac{\partial^2 z(x)}{\partial x_i \partial x_j} \right|^2 + s|\nabla z|^2 + s^3 z^2 \right) \exp(s e^{x_1})\, dx \le c_1 \int_\Omega f^2(x) \exp(s e^{x_1})\, dx, \tag{3.2}$$

where x_1 is the first component of $x = (x_1, \dots, x_n) \in \Omega$ and $c_1 > 0$ does not depend on s.

For the proof of Lemma 3.1 we refer to L. Hörmander [21], [22]. Note that estimate (3.2) can be obtained as a simple corollary of Lemma 3.2, which will be proved in section 6.

Let $\gamma(t)$ be a function satisfying the properties

$$\gamma(t) \in C^\infty(0,T), \qquad 0 < \gamma(t) \le 1 \,\forall\ t \in (0,T), \tag{3.3}$$

$$\gamma(t) = \begin{cases} t & \text{when } t \in (0, T_0), \\ T - t & \text{when } t \in (T - T_0, T), \end{cases} \quad T_0 = \min\left(\frac{T}{3}, \frac{1}{2}\right).$$

We define $\varphi(t,x)$, $\alpha(t,x)$ by relations

$$\varphi(t,x) = \frac{e^{x_1}}{\gamma(t)}, \qquad \alpha(t,x) = (e^{x_1} - e^{2\hat{x}_1})/\gamma(t), \tag{3.4}$$

$$\text{where } \hat{x}_1 = \max_{x=(x_1,\dots,x_n)\in\Omega} |x_1|.$$

COROLLARY 3.1. *Let $f(x) \in L_2(\Omega)$ and s be just the same as in Lemma 3.1. Then for any $t \in (0,T)$ the following estimate is true:*

$$\int_\Omega \left(\frac{\gamma(t)}{s} \sum_{i,j=1}^n \left| \frac{\partial^2 z(x)}{\partial x_i \partial x_j} \right|^2 + \frac{s}{\gamma(t)} |\nabla z|^2 + \frac{s^3}{\gamma(t)^3} |z|^2 \right) e^{s\varphi(t,x)}\, dx \le c_2 \int_\Omega f^2(x) e^{s\varphi(t,x)}\, dx. \tag{3.5}$$

Proof. We substitute $s = s_1(\gamma(t))^{-1}$ into (3.2) and obtain (3.5), where instead of s_1 we write s. In virtue of Lemma 3.1 estimate (3.5) is true when $s > s_0\gamma(t)$. Since $0 < \gamma(t) \le 1$ for $t \in (0,T)$, this inequality is also true when $s > s_0$. □

We consider the inverse heat equation

$$\partial_t z(t,x) + \Delta z = f(t,x), \qquad (t,x) \in Q, \tag{3.6}$$

with the boundary conditions

$$z|_\Sigma = 0, \qquad \left.\frac{\partial z}{\partial \nu}\right|_\Sigma = 0, \tag{3.7}$$

where, recall, $Q = (0,T) \times \Omega$, $\Sigma = (0,T) \times \partial\Omega$.

LEMMA 3.2. *There exists $s_0 > 0$ such that for any $s > s_0$ the solution $z(t,x)$ of* (3.6), (3.7) *satisfies the Carleman estimate:*

$$\int_Q \left((s\varphi)^{-1} \left(|\partial_t z|^2 + \sum_{i,j=1}^n |\partial^2_{x_i x_j} z(t,x)|^2 \right) + s\varphi \sum_{j=1}^n |\partial_{x_j} z|^2 + s^3\varphi^3 z^2 \right) e^{s\alpha(t,x)}\, dx\, dt \le c_3 \int_Q f^2(t,x) e^{s\alpha}\, dx\, dt, \tag{3.8}$$

where the functions $\varphi(t,x)$, $\alpha(t,x)$ are defined in (3.4) *and $c_3 > 0$ does not depend on s.*

We prove this lemma below, in section 6.

3.2. First, instead of problem (2.11)–(2.15), we consider an auxiliary problem. Let $\Omega_0 \subset \mathbf{R}^n$ be a bounded domain with C^∞-boundary $\partial\Omega_0$ which contains the closure $\bar\Omega$ of Ω : $\bar\Omega \subset \Omega_0$ and satisfies the condition $\sup_{x\in\Omega_0} |x_1| < 2 \sup_{x\in\Omega} |x_1|$. Therefore the function η from (2.16) is positive and the function α from (3.4) is negative. We denote

$$Q_0 = (0,T) \times \Omega_0, \quad \Sigma_0 = (0,T) \times \partial\Omega_0, \quad \omega = \Omega_0 \setminus \bar\Omega.$$

In Q_0 we consider the linearized Boussinesq equation with the distributed control concentrated in $(0,T) \times \omega$:

$$\begin{aligned} \acute{\mathcal{N}}'(w,p,\tau,u) &= \partial_t w(t,x) - \Delta w - \hat v \times \mathbf{rot} w + w \times \mathbf{rot}\hat v + \nabla p + \tau(t,x) e_0 + u'(t,x) \\ &= f(t,x), \end{aligned} \tag{3.9}$$

$$\mathbf{div}\, w = 0, \tag{3.10}$$

$$\hat{\mathcal{H}}'(w,\tau,u) = \partial_t \tau(t,x) - \Delta\tau + (w, \nabla\hat\theta) + (\hat v, \nabla\tau) + (w, e_0) + u_{n+1}(t,x) = h(t,x), \tag{3.11}$$

$$w(0,x) = w_0(x), \qquad \tau(0,x) = \tau_0(x), \tag{3.12}$$

$$w(T,x) = 0, \qquad \tau(T,x) = 0, \tag{3.13}$$

where $u(t,x) = (u'(t,x), u_{n+1}(t,x)) = (u_1, \dots, u_n, u_{n+1})$ is the distributed control concentrated in $Q_\omega = (0,T) \times \omega$: $\operatorname{supp} u \subset Q_\omega$. The functional space for data (f,h,w_0,τ_0) of problem (3.9)–(3.13) is as follows:

$$(f,h,w_0,\tau_0) \in \Phi^s(Q_0) = (L_2(Q_0,\eta^s))^n \times L_2(Q_0,\eta^s) \times V^1(\Omega_0) \times H^1(\Omega_0), \tag{3.14}$$

where $s > 0$ is an arbitrary fixed number. We define the functional space of solutions of problem (3.9)–(3.13) by the formula

$$(w, \nabla p, \tau, u) \in U^s(Q_0) \equiv \Xi(Q_0, \eta^s) \times L_2(Q_0, \eta^s) \times \mathbf{\Theta}(Q_0, \eta^s) \times (\hat{L}_2(Q_\omega, \eta^s))^{n+1}, \tag{3.15}$$

where $\hat{L}_2(Q_\omega, \eta^s)$ is the set of functions that belong to $L_2(Q_0, \eta^s)$ and equal zero on the set $Q_0 \setminus Q_\omega$; the constant s in (3.15) is just the same as in (3.14). We suppose that the functions $\hat{v}, \hat{\theta}$ in (3.9), (3.11) satisfy the condition

$$\hat{v} \in V^{1,2(1/2)}(Q_0), \qquad \hat{\theta} \in W^{1,2(1/2)}(Q_0). \tag{3.16}$$

As in Lemma 2.1 one can easily prove that the operator

$$\hat{\mathcal{A}}' : U^s(Q_0) \to \Phi^s(Q_0) \tag{3.17}$$

is continuous, where $\Phi^s(Q)$, $U^s(Q)$ are defined in (3.14), (3.15), and

$$\hat{\mathcal{A}}'(w, \nabla p, \tau, u) = (\hat{\mathcal{N}}'(w, \nabla p, \tau, u), \hat{\mathcal{H}}'(w, \tau, u), \gamma_0 w, \gamma_0 \tau) \tag{3.18}$$

with $\hat{\mathcal{N}}', \hat{\mathcal{H}}'$ defined in (3.9), (3.11).

LEMMA 3.3. *The image of operator* (3.17), (3.18) *is dense in the space* $\Phi^s(Q_0)$.

Proof. Suppose that the assertion of Lemma 3.3 is not true. Then there exists a nonzero collection $\phi \equiv (m(t,x), \zeta(t,x), z_0(x), \psi_0(x)) \in \Phi^s(Q_0)$ such that

$$(\mathcal{A}'(w, \nabla p, \tau, u), \phi)_{\Phi^s(Q_0)} = 0 \qquad \forall\ (w, \nabla p, \tau, u) \in U^s(Q_0). \tag{3.19}$$

We can rewrite equality (3.19) in the form

$$\begin{aligned} \int_{Q_0} &(\partial_t w(t,x) - \Delta w - \hat{v} \times \mathbf{rot}\, w - w \times \mathbf{rot}\, \hat{v} + \nabla p(t,x) + \tau(t,x) e_0 \\ &+ u'(t,x), m(t,x)) e^{2\eta^s(t,x)}\, dx\, dt + \int_{Q_0} (\partial_t \tau(t,x) - \Delta \tau + (w, \nabla \hat{\theta}) + (\hat{v}, \nabla \tau) \\ &+ (w, e_0) + u_{n+1}) \zeta(t,x) e^{2\eta^s}\, dx\, dt + (w(0,\cdot), z_0)_{V^1(\Omega_0)} + (\tau(0,\cdot), \psi_0)_{H^1(\Omega_0)} = 0. \end{aligned} \tag{3.20}$$

We set, in (3.20),

$$z(t,x) = m(t,x) e^{2\eta^s(t,x)}, \qquad \psi(t,x) = \zeta(t,x) e^{2\eta^s(t,x)}, \tag{3.21}$$

$\nabla p(t,x) \equiv 0$, $u(t,x) \equiv 0$, $w \in \Xi(Q_0, \eta) \cap (C_0^\infty(Q_0))^n$, and $\tau \in \mathbf{\Theta}(Q_0, \eta) \cap C_0^\infty(Q_0)$. Then integrating by parts in (3.20) yields the equations

$$\partial_t z + \Delta z = \mathbf{rot}\, (\hat{v} \times z) + z \times \mathbf{rot}\, \hat{v} + \psi(\nabla \hat{\theta} + e_0) + \nabla \tilde{p} \text{ in } Q_0, \tag{3.22}$$

$$\partial_t \psi + \Delta \psi = -\nabla(\psi \hat{v}) + (e_0, z) \qquad \text{in } Q_0. \tag{3.23}$$

If we set, in (3.20), $u \in (\hat{L}_2(Q_\omega, \eta))^{n+1}$, $\nabla p = 0$, $w = 0$, $\tau = 0$, we will obtain the equalities

$$z(t,x) \equiv 0, \qquad \psi(t,x) = 0, \qquad (t,x) \in Q_\omega = Q_0 \setminus Q. \tag{3.24}$$

In particular, (3.24) means that z and ψ equal zero in a neighborhood of $\Sigma_0 = (0,T) \times \partial\Omega_0$. After setting, in (3.20), $\nabla p \in (L_2(Q_0, \eta))^n$, $w = 0$, $\tau = 0$, $u = 0$, and taking into account (3.24), we get

$$\mathbf{div}\, z = 0 \text{ in } Q_0. \tag{3.25}$$

Equalities (3.22), (3.24) yield that

$$\nabla\tilde{p}(t,x)\equiv 0,\qquad (t,x)\in Q_\omega. \tag{3.26}$$

Applying to both parts of (3.22) the operator **div** and taking into account (3.25) and the formula $\mathbf{div}\,\mathbf{rot}\,y=0$, we obtain

$$-\Delta\tilde{p}=\mathbf{div}(z\times\mathbf{rot}\,\hat{v})+\mathbf{div}((\nabla\hat{\theta}+e_0)\psi). \tag{3.27}$$

Our main goal now is to deduce from relations (3.22)–(3.27) that $z\equiv 0$, $\psi\equiv 0$. We will make it with the help of Carleman estimates (3.5), (3.8). We can suppose that s_0 in Lemmas 3.1 and 3.2 are equal. Otherwise, we can replace them in both lemmas with their maximum.

Let

$$\sigma\geq\max(s,s_0), \tag{3.28}$$

where s is the constant from $\Phi^s(Q_0)$ in the formulation of Lemma 3.3. We take magnitude σ instead of s in (3.8) and apply estimate (3.8) to the equations (3.22), (3.23). Note that boundary conditions (3.7) are fulfilled in our case in virtue of (3.24). We have

$$\begin{aligned}\int_{Q_0}(\sigma\varphi|\nabla z|^2+(\sigma\varphi)^3|z|^2)e^{\sigma\alpha(t,x)}\,dx\,dt&\\ +\int_{Q_0}(\sigma\varphi|\nabla\psi|^2+(\sigma\varphi)^3|\psi|^2)e^{\sigma\alpha}\,dx\,dt&\\ \leq c_1\int_{Q_0}e^{\sigma\alpha}(|\hat{v}|^2|\nabla z|^2+|\nabla\hat{v}|^2|z|^2+|\psi|^2(1+|\nabla\hat{\theta}|^2)+|\nabla\tilde{p}|^2&\\ +|\psi|^2|\nabla\hat{v}|^2+|\nabla\psi|^2|\hat{v}|^2+|z|^2)\,dx\,dt.&\end{aligned} \tag{3.29}$$

We need estimate $\nabla\tilde{p}$ in the right side of (3.29). We do it by means of (3.27), (3.26). Note that $\tilde{p}$ is defined to within an arbitrary constant. We fix it by the condition

$$\tilde{p}(t,x)\equiv 0,\qquad (t,x)\in Q_\omega. \tag{3.30}$$

Taking into account (3.26), (3.30) we apply estimate (3.5) to (3.27). After multiplication (3.5) on $(\gamma(t)/\sigma)\exp(-e^{2\tilde{x}_1}/\gamma(t))$ scalarly in $L_2(\Omega)$ and integration with respect to t, we get

$$\begin{aligned}\int_{Q_0}|\nabla\tilde{p}|^2e^{\sigma\alpha}\,dx\,dt\leq c_5\int_{Q_0}\frac{\gamma(t)}{\sigma}(|\nabla z|^2|\nabla\hat{v}|^2+|z|^2|\nabla\mathbf{rot}\hat{v}|^2&\\ +|\nabla\psi|^2(1+|\hat{\theta}|^2)+|\psi|^2|\Delta\hat{\theta}|^2)e^{\sigma\alpha}\,dx\,dt&\\ \leq c_6(\|\hat{v}\|^2_{C(0,T;(C^1(\bar{\Omega}))^n)}+\|\hat{\theta}\|^2_{C(0,T;C^1(\bar{\Omega}))}+1)\int_{Q_0}\frac{\gamma(t)}{\sigma}(|\nabla z|^2+|\nabla\psi|^2)e^{\sigma\alpha}\,dx\,dt&\\ +c_7\left(\|\mathbf{rot}\,\hat{v}\|^2_{L_\infty(0,T;(W_4^1(\Omega_0))^n)}\int_0^T\left(\int_\Omega\left(e^{2\sigma\alpha}\left(\frac{\gamma(t)}{\sigma}\right)^2|z|^4\right)dx\right)^{1/2}dt\right.&\\ \left.+\|\Delta\hat{\theta}\|^2_{L_\infty(0,T;L_4(\Omega_0))}\int_0^T\left(\int_\Omega\left(\frac{\gamma(t)}{\sigma}\right)^2\psi^4e^{2\sigma\alpha}\,dx\right)^{1/2}dt\right).&\end{aligned} \tag{3.31}$$

Let us estimate the right side of (3.31) using the continuity of embeddings $W^{1,2(2)}(Q_0) \subset C(0,T;C^1(\bar{\Omega}_0))$, $W^{1,2(2)}(Q_0) \subset L_\infty(0,T;W_4^2(\Omega_0))$, and $H^1(\Omega_0) \subset L_4(\Omega_0)$ when $\dim \Omega_0 \le 3$, and taking into account that in virtue of (3.4)

$$|\partial_{x_j}(e^{\frac{\sigma\alpha}{2}} z)|^2 \le c_8(\sigma^2\varphi^2|z|^2 + |\nabla z|^2)e^{\sigma\alpha}.$$

As a result we obtain the inequality

$$\int_{Q_0} |\nabla \tilde{p}|^2 e^{\sigma\alpha}\,dx\,dt \le c_9(\|\hat{v}\|^2_{V^{1,2(2)}(Q_0)} + \|\hat{\theta}\|^2_{W^{1,2(2)}(Q_0)} + 1)\int_{Q_0}\left(\frac{\gamma(t)}{\sigma}(|\nabla z|^2 + |\nabla\psi|^2) + \frac{\sigma e^{2x_1}}{\gamma(t)}(|z|^2 + |\psi|^2)\right)e^{\sigma\alpha}\,dx\,dt. \tag{3.32}$$

The substitution of (3.32) into (3.29) and simple transformations give us the upper bound:

$$\begin{aligned}\int_{Q_0}\left(\frac{\sigma e^{x_1}}{\gamma(t)}(|\nabla z|^2 + |\nabla\psi|^2) + \frac{\sigma^3 e^{3x_1}}{\gamma(t)^3}(|z|^2 + |\psi|^2)\right)e^{\sigma\alpha}\,dx\,dt \\ \le c_{10}(\|\hat{v}\|^2_{V^{1,2(2)}(Q_0)} + \|\hat{\theta}\|^2_{W^{1,2(2)}(Q_0)} + 1)\int_{Q_0}\left(\left(\frac{\gamma(t)}{\sigma} + 1\right)(|\nabla z|^2 + |\nabla\psi|^2) \right. \\ \left. + \left(\frac{\sigma\ e^{2x_1}}{\gamma(t)} + 1\right)(|z|^2 + |\psi|^2)\right)e^{\sigma\alpha}\,dx\,dt.\end{aligned} \tag{3.33}$$

Note that (3.33) is true for arbitrary σ satisfying (3.28). We choose σ so large that estimates

$$\begin{aligned}\frac{\sigma e^{x_1}}{\gamma(t)} &> c_{10}(\|\hat{v}\|^2_{V^{1,2(2)}(Q_0)} + \|\hat{\theta}\|^2_{W^{1,2(2)}(Q_0)} + 1)\left(\frac{\gamma(t)}{\sigma} + 1\right), \\ \frac{\sigma^3 e^{3x_1}}{\gamma(t)^3} &> c_{10}(\|\hat{v}\|^2_{V^{1,2(2)}(Q_0)} + \|\hat{\theta}\|^2_{W^{1,2(2)}(Q_0)} + 1)\left(\frac{\sigma e^{2x_1}}{\gamma(t)} + 1\right)\end{aligned}$$

hold for all $(t,x) \in Q_0$. Then (3.33) yields that

$$z(t,x) \equiv 0, \qquad \psi(t,x) \equiv 0. \tag{3.34}$$

Substituting into (3.20) $\nabla p \equiv 0$, $u \equiv 0$, $w \in \Xi(Q_0,\eta)$, and $\tau \equiv 0$, and $\nabla p \equiv 0$, $u \equiv 0$, $w \equiv 0$, $\tau \in \boldsymbol{\Theta}(Q_0,\eta)$ yield the equalities

$$\begin{aligned}(w(0,\cdot), z_0)_{V^1(\Omega_0)} &= (w(0,\cdot), z(0,\cdot))_{L_2(\Omega_0)} = 0, \\ (\tau(0,\cdot), \psi_0)_{H^1(\Omega_0)} &= (\tau(0,\cdot), \psi(0,\cdot))_{L_2(\Omega_0)} = 0.\end{aligned}$$

Therefore

$$z_0 = 0, \qquad \psi_0 = 0. \tag{3.35}$$

Hence, by (3.21), (3.34), and (3.35), we have that $\phi \equiv (m(t,x), \zeta(t,x), z_0(x), \psi_0(x)) \equiv 0$. □

3.3. Now we can prove the main result of this section.

THEOREM 3.1. *Suppose that $\hat{v} \in V^{1,2(2)}(Q), \hat{\theta} \in W^{1,2(2)}(Q)$, the operator $\mathcal{A}'(0)$ is defined in* (2.25), (2.11), (2.13), *the spaces $X^s(Q), Z^s(Q)$ are defined in* (2.16)–(2.23),

and the parameter s of these spaces is an arbitrary positive number. Then the image of the operator

$$\mathcal{A}'(0) : X^s(Q) \to Z^s(Q)$$

is dense in the space $Z^s(Q)$.

Proof. Let Ω_0, Q_0 be the sets introduced in the beginning of section 3.2. By Propositions 2.2, 2.3 we extend the functions $\hat{v}(t,x), \hat{\theta}(t,x)$ continuously from $V^{1,2(2)}(Q)$ up to $V^{1,2(2)}(Q_0)$ and from $W^{1,2(2)}(Q)$ up to $W^{1,2(2)}(Q_0)$ correspondingly and denote these new functions also by $\hat{v}(t,x), \hat{\theta}(t,x)$. Comparing (3.9)–(3.12) and (2.11)–(2.14), we see that the restriction of operator (3.17), (3.18) on the cylinder Q coincides with the operator

$$\mathcal{A}'(0) : U^s(Q) \to \Phi^s(Q). \tag{3.36}$$

Here $\mathcal{A}'(0)$ is operator (2.25), and in contrast to (3.15),

$$U^s(Q) = \Xi(Q,\eta^s) \times L_2(Q,\eta^s) \times \mathbf{\Theta}(Q,\eta^s) \tag{3.37}$$

because the restriction of an arbitrary function from $\hat{L}_2(Q_\omega,\eta^s)$ to Q is identical to zero. Therefore, in virtue of Lemma 3.3, the image of operator (3.36), (2.25) is dense in $\Phi^s(Q)$. Let $(f,h,v_0,\theta_0) \in Z^s(Q)$ (see (2.23)) be an arbitrary element. Since $f \in F(Q,\eta^s)$ (see (2.19)) then $f = f_1 + \nabla f_2$ where $f_1 \in L_2(Q,\eta^s)$, $f_2 \in L_2(0,T;H^1(\Omega))$, and therefore $(f_1,h,v_0,\theta_0) \in \Phi^s(Q)$. By the density of the image of operator (3.36), for any $\epsilon > 0$ there exists $(f_1^\epsilon, h^\epsilon, v^\epsilon, \theta^\epsilon) \in \Phi^s(Q)$ possessing preimage $(v^\epsilon, p^\epsilon, \theta^\epsilon) \in U^s(Q)$,

$$\mathcal{A}'(0)(v^\epsilon, p^\epsilon, \theta^\epsilon) = (f_1^\epsilon, h^\epsilon, v_0^\epsilon, \theta_0^\epsilon), \tag{3.38}$$

and satisfying the inequality

$$\|(f_1 - f_1^\epsilon, h - h^\epsilon, v_0 - v_0^\epsilon, \theta_0 - \theta_0^\epsilon)\|_{\Phi^s(Q)} \le \epsilon. \tag{3.39}$$

By virtue of (2.11) and (3.38),

$$\mathcal{A}'(0)(v^\epsilon, p^\epsilon + f_2, \theta^\epsilon) = (f_1^\epsilon + \nabla f_2, h^\epsilon, v_0^\epsilon, \theta_0^\epsilon). \tag{3.40}$$

Since $f - (f_1^\epsilon + \nabla f_2) = f_1 - f_1^\epsilon$, then by (2.23), (2.20), (3.14), and (3.36) we have

$$\begin{aligned} \|(f - (f_1^\epsilon + \nabla f_2), h - h^\epsilon, v_0 - v_0^\epsilon, \theta_0 - \theta^\epsilon)\|_{Z^s(Q)} \\ \le \|(f - f_1^\epsilon, h - h^\epsilon, v_0 - v_0^\epsilon, \theta_0 - \theta_0^\epsilon)\|_{\Phi^s(Q)} < \epsilon. \end{aligned} \tag{3.41}$$

By (3.15), (2.22) the inclusion $(v^\epsilon, p^\epsilon, \theta^\epsilon) \in U^s(Q)$ involves the inclusion $(v^\epsilon, p^\epsilon + f_2, \theta^\epsilon) \in X^s(Q)$. Hence, by (3.40), $(v^\epsilon, p^\epsilon + f_2, \theta^\epsilon)$ is the preimage of $(f_1^\epsilon + \nabla f_2, h^\epsilon, v_0^\epsilon, \theta_0^\epsilon)$. This proves the theorem. □

Remark 3.1. The method of Lemma 3.3's proof, based on applying the Hahn–Banach theorem and using some uniqueness theorems, is well known (see J.-L. Lions [31]). The density of right-hand sides for which a solution of the corresponding boundary value problem exists was proved in A. Fursikov [13], [14] for certain situations different from those studied above.

4. On a decomposition of Weyl type. In this section we investigate the decomposition of the Weyl type,

$$y(t,x) = v(t,x) + \nabla q, \qquad (t,x) \in Q_0, \tag{4.1}$$

where $\mathbf{div}\, v = 0$ and $\nabla q = (\partial_{x_1} q, \dots, \partial_{x_n} q)$ is the gradient of a function. We do not impose any boundary conditions on v or ∇q but look for v belonging to the space $\Xi(Q_0, \eta)$ when $y \in (\Theta(Q_0, \eta))^n$. We do not look for natural uniqueness conditions for the decomposition (4.1) but need the following assumption to be fulfilled:

$$\text{if } \mathbf{div}\, y(0,x) \equiv 0, \text{ then } y(0,x) \equiv v(0,x). \tag{4.2}$$

To find decomposition (4.1) we consider the extremal problem

$$J(u) = \int_{Q_0} \frac{|u(t,x)|^2 e^{2\eta}}{(T-t)^4}\, dx\, dt \to \inf, \tag{4.3}$$

$$\Delta u(t,x) = \mathbf{div}\, y(t,x), \qquad (t,x) \in Q_0, \tag{4.4}$$

where $y(t,x) \in (\mathbf{\Theta}(Q_0, \eta))^n$ is a given function. If a solution $m(t,x)$ of problem (4.3), (4.4) exists, we will denote $v = y - \nabla m$. Then, by (4.4), the equality $\mathbf{div}\, v = 0$ will hold, and therefore decomposition (4.1) will be true.

LEMMA 4.1. *There exists s_0 such that for $y(t,x) \in (\mathbf{\Theta}(Q, \eta^s))^n$ where $s \geq s_0$, the problem* (4.3), (4.4) *has the unique solution $m(t,x) \in L_2(Q_0, \eta - 2\ln(T-t))$. This solution satisfies the estimates*

$$\int_{Q_0} \frac{|m(t,x)|^2}{(T-t)^4} e^{2\eta^s}\, dx\, dt \leq c_1 \int_{Q_0} \frac{|\mathbf{div}\, y|^2}{(T-t)} e^{2\eta^s}\, dx\, dt, \tag{4.5}$$

$$\int_{Q_0} |\partial_t m(t,x)|^2 e^{2\eta^s}\, dx\, dt \leq c_2 \|y\|^2_{\mathbf{\Theta}(Q_0, \eta^s)}. \tag{4.6}$$

Proof. Let s_0 be defined as in Lemma 3.1. We denote $Q_\epsilon = (0, T-\epsilon) \times \Omega_0$ and instead of (4.3), (4.4), consider the extremal problem

$$J_\epsilon(u) = \int_{Q_\epsilon} \frac{|u(t,x)|^2}{(T-t)^4} e^{2\eta}\, dx\, dt \to \inf, \tag{4.7}$$

$$\Delta u(t,x) = \mathbf{div}\, y(t,x), \qquad (t,x) \in Q_\epsilon. \tag{4.8}$$

The weight $e^{2\eta}(T-t)^{-4}$ is bounded above and below on Q_ϵ. Hence the space $U_\epsilon = \{u \in L_2(Q_\epsilon) : \Delta u \in L_2(Q_\epsilon)\}$ is natural for the problem (4.7), (4.8) and the set of its admissible elements is as follows:

$$A_\epsilon = \{u \in U_\epsilon : \Delta u = \mathbf{div}\, y\}.$$

As is well known, the limit $m_\epsilon \in A_\epsilon$ of a weakly converging subsequence of the minimizing sequence u_k: $J_\epsilon(u_k) \to \inf_{v \in A_\epsilon} J_\epsilon(v)$ is the solution of problem (4.7), (4.8). The uniqueness of m_ϵ follows from the functional J_ϵ strict-convexity. For $\epsilon_1 > \epsilon_2$, $m_{\epsilon_1}(t,x)$ coincides almost everywhere with the restriction of $m_{\epsilon_2}(t,x)$ on Q_{ϵ_1}. Indeed, if it is not so, then $J_{\epsilon_1}(m_{\epsilon_1}) < J_{\epsilon_1}(m_{\epsilon_2})$. But on this occasion, m_{ϵ_2} is not a solution because the function

$$\hat{m}(t,x) = \begin{cases} m_{\epsilon_1}(t,x), & (t,x) \in Q_{\epsilon_1}, \\ m_{\epsilon_2}(t,x), & (t,x) \in Q_{\epsilon_2} \backslash Q_{\epsilon_1} \end{cases}$$

satisfies (4.8) and inequality $J_{\epsilon_2}(\hat{m}) < J_{\epsilon_2}(m_{\epsilon_2})$ holds. That is why below we use the notation $m_\epsilon = m$. Since operator $\Delta : U_\epsilon \to L_2(Q_\epsilon)$ is surjective, we can apply to problem (4.7), (4.8) the Lagrange principle (see [1]). This principle asserts that there exists $p_\epsilon \in (L_2(Q_\epsilon))^n$ such that the Lagrange function

$$\mathcal{L}(u, p_\epsilon) \equiv \int_{Q_\epsilon} \left(\frac{1}{2} \frac{|u(t,x)|^2}{(T-t)^4} e^{2\eta} + (\Delta u - \mathbf{div}\, y) p_\epsilon(t,x) \right) dx\, dt$$

satisfies the equality $\partial_u \mathcal{L}(u, p_\epsilon)|_{u=m} = 0$; i.e., for any $h \in U_\epsilon$,

$$\int_{Q_\epsilon} \left(\frac{m(t,x)h(t,x)}{(T-t)^4} e^{2\eta} + \Delta h p_\epsilon(t,x) \right) dx\, dt = 0. \tag{4.9}$$

It follows from (4.9) that

$$\Delta p_\epsilon(t,x) + \frac{m(t,x)}{(T-t)^4} e^{2\eta} = 0 \text{ in } \Omega_0, \qquad p_\epsilon|_{\partial\Omega_0} = \left. \frac{\partial p_\epsilon}{\partial \nu} \right|_{\partial\Omega_0} = 0. \tag{4.10}$$

Relations (4.10) imply that p_ϵ does not depend on ϵ and therefore, below, we use the notation $p_\epsilon = p$. We apply to (4.10) Carleman estimate (3.2), substitute in this estimate $s = 2s_1(T-t)^{-1}$, multiply it on $(T-t)^4 \exp\{-2s_1 e^{2x_1}/(T-t)\}$, and integrate with respect to t. As a result we have the estimate

$$\int_{Q_\epsilon} (T-t) p^2 e^{-2\eta}\, dx\, dt \le c_3 \int_{Q_\epsilon} \frac{m^2}{(T-t)^4} e^{2\eta}\, dx\, dt, \tag{4.11}$$

where $c_3 > 0$ does not depend on ϵ. We substitute $u = m$ into (4.4), scale the obtained equation by p in $L_2(Q_\epsilon)$, integrate by parts, and apply (4.10). As a result we get

$$\begin{aligned} 0 = \int_{Q_\epsilon} (\Delta m - \mathbf{div}\, y) p\, dx\, dt &= \int_{Q_\epsilon} (m\Delta p - p\, \mathbf{div}\, y)\, dx\, dt \\ &= -\int_{Q_\epsilon} \left(\frac{m^2}{(T-t)^4} e^{2\eta} + p\, \mathbf{div}\, y \right) dx\, dt. \end{aligned}$$

This equality and (4.11) yield

$$\begin{aligned} \int_{Q_\epsilon} \frac{m^2}{(T-t)^4} e^{2\eta}\, dx\, dt &\le c_4 \left(\int_{Q_\epsilon} \frac{|\mathbf{div}\, y|^2}{(T-t)} e^{2\eta} dx dt \right)^{1/2} \left(\int_Q (T-t)|p|^2 e^{-2\eta}\, dx\, dt \right)^{1/2} \\ &\le c_5 \int_{Q_\epsilon} \frac{|\mathbf{div}\, y|^2}{(T-t)} e^{2\eta}\, dx\, dt + \frac{1}{2} \int_{Q_\epsilon} \frac{m^2}{(T-t)^4} e^{2\eta}\, dx\, dt, \end{aligned}$$

which gives us the upper bound

$$\int_{Q_\epsilon} \frac{m^2}{(T-t)^4} e^{2\eta}\, dx\, dt \le c_6 \int_{Q_\epsilon} \frac{|\mathbf{div}\, y|^2}{(T-t)} e^{2\eta}\, dx\, dt, \tag{4.12}$$

where c_6 does not depend on ϵ. Hence, we can pass to the limit in (4.12) as $\epsilon \to 0$ and obtain (4.5). Let $\hat{m}$ be the solution of problem (4.3), (4.4). Since m is the solution of (4.7), (4.8) we have

$$\int_{Q_\epsilon} \frac{m^2}{(T-t)^4} e^{2\eta}\, dx\, dt \le \int_{Q_\epsilon} \frac{\hat{m}^2}{(T-t)^4} e^{2\eta}\, dx\, dt \qquad \forall \epsilon > 0,$$

and therefore

$$\int_{Q_0} \frac{m^2}{(T-t)^4} e^{2\eta}\,dx\,dt = \int_{Q_0} \frac{\hat{m}^2}{(T-t)^4} e^{2\eta}\,dx\,dt.$$

This equation implies the equality $m = \hat{m}$ because the solution of problem (4.3), (4.4) is unique. Differentiation of the equations in (4.4), (4.10) with respect to t yields

$$\Delta \partial_t m = \mathbf{div}\,\partial_t y, \tag{4.13}$$

$$\Delta \partial_t p + (\partial_t m)\frac{e^{2\eta}}{(T-t)^4} + m\partial_t\left(\frac{e^{2\eta}}{(T-t)^4}\right) = 0. \tag{4.14}$$

Applying to (4.14) the Carleman estimate (3.2) in the same way as in (4.11) we obtain

$$\int_{Q_0} |\nabla \partial_t p|^2 (T-t)^7 e^{-2\eta}\,dx\,dt \le c_7 \int_{Q_0} (|\partial_t m|^2 + (T-t)^{-4}|m|^2) e^{2\eta}\,dx\,dt. \tag{4.15}$$

Scaling equation (4.13) by $\partial_t p$ in $L_2(Q_0)$, integration by parts, and application of (4.14) yield

$$\begin{aligned} 0 = \int_{Q_0} (T-t)^4 (\Delta \partial_t m - \mathbf{div}\partial_t y)\partial_t p\,dx\,dt = \int_{Q_0} (T-t)^4 (\partial_t m \Delta \partial_t p \\ -(\partial_t y, \nabla \partial_t p))\,dx\,dt = \int_{Q_0} \left(-|\partial_t m|^2 e^{2\eta} - \left((\partial_t m) m \partial_t \frac{e^{2\eta}}{(T-t)^4}\right)(T-t)^4 \right. \\ \left. - (T-t)^4 (\partial_t y, \nabla \partial_t p)\right) dx\,dt. \end{aligned}$$

This equality and (4.15) imply

$$\begin{aligned} \int_{Q_0} |\partial_t m|^2 e^{2\eta}\,dx\,dt \le c_8 \int_{Q_0} \left(|\partial_t m||m| \frac{e^{2\eta}}{(T-t)^2} \right. \\ \left. + e^{\eta}(T-t)^{1/2}|\partial_t y| e^{-\eta}(T-t)^{\frac{7}{2}} |\nabla \partial_t p|\right) dx\,dt \le \frac{1}{4}\int_{Q_0} |\partial_t m|^2 e^{2\eta}\,dx\,dt \\ + c_9 \int_{Q_0} \left(\frac{|m|^2}{(T-t)^4} e^{2\eta} + (T-t)|\partial_t y|^2 e^{2\eta}\right) dx\,dt. \end{aligned}$$

This inequality and (4.5) give (4.6). □

Let

$$\rho(x) \in C^\infty(\bar{\Omega}_0), \qquad \rho|_{\partial\Omega_0} = 0, \qquad \rho(x) > 0 \ \forall\ x \in \Omega_0.$$

Below, we use the following space:

$$\begin{aligned} M(Q_0,\eta) = \Bigg\{ f = (f_1,\dots,f_n) : \|f\|^2_{M(Q_0,\eta)} = \|(T-t)^{-1} f\|^2_{(L_2(Q_0,\eta^s))^n} \\ + \||\nabla f|\|^2_{(L_2(Q_0,\eta^s))^n} + \|(T-t)\partial_t f\|^2_{(L_2(Q_0,\eta^s))^n} \\ + \sum_{i,j=1}^n \|(T-t)\partial^2_{x_i x_j} f\|^2_{(L_2(Q_0,\eta^s))^n} < \infty \Bigg\}. \end{aligned} \tag{4.16}$$

LEMMA 4.2. *Let $m(t,x)$ be the solution of problem* (4.3), (4.4) *constructed in Lemma* 4.1. *Then*

$$\|\rho^3\nabla m\|^2_{M(Q_0,\eta)} \le c_{10}\|y\|^2_{(\mathbf{\Theta}(Q_0,\eta))^n}. \tag{4.17}$$

Proof. Set $\tilde m = m\rho$. Then (4.4), where $u = m$, implies the following equation for $\tilde m$:

$$\Delta\tilde m = m\Delta\rho + 2(\nabla\rho,\nabla m) + \rho\mathbf{div}\, y. \tag{4.18}$$

We multiply this equation by $-e^{2\eta}\tilde m(T-t)^{-2}$ scalarly in $L_2(Q_0)$, integrate by parts, and have as a result

$$\begin{aligned}\int_{Q_0}|\nabla\tilde m|^2e^{2\eta}(T-t)^{-2}\,dx\,dt &= \int_{Q_0}(T-t)^{-2}\left(\frac{1}{2}|\tilde m|^2\Delta e^{2\eta} - m^2\rho\Delta\rho e^{2\eta}\right.\\ &\left.+\frac{1}{2}m^2(\Delta\rho^2e^{2\eta} + (\nabla\rho^2,\nabla e^{2\eta})) - \rho^2me^{2\eta}\mathbf{div}\, y\right)dx\,dt\\ &\le c_{11}\int_{Q_0}\left(\frac{m^2}{(T-t)^4} + |\mathbf{div}\, y|^2\right)e^{2\eta}\,dx\,dt.\end{aligned}$$

This inequality, (4.5) and the definition (2.18) of the space $\mathbf{\Theta}(Q_0,\eta)$ imply:

$$\begin{aligned}\int_{Q_0}\frac{|\rho\nabla m|^2}{(T-t)^2}e^{2\eta}\,dx\,dt &\le c_{12}\int_{Q_0}\frac{|\nabla(\rho m)|^2}{(T-t)^2}e^{2\eta}\,dx\,dt\\ &+\int_{Q_0}\frac{m^2|\nabla\rho|^2}{(T-t)^2}e^{2\eta}\,dx\,dt \le c_{13}\|y\|^2_{(\mathbf{\Theta}(Q_0,\eta))^n}.\end{aligned} \tag{4.19}$$

Denote $m_0 = m\rho^2e^\eta$. Then we have, analogously to (4.18),

$$\Delta m_0 = g, \qquad m_0|_{\partial\Omega_0} = 0, \tag{4.20}$$

where $g = m\Delta(\rho^2e^\eta) + 2(\nabla(\rho^2e^\eta),\nabla m) + \rho^2e^\eta\mathbf{div}\, y$.

By (4.5), (4.19) we get

$$\|g\|_{L_2(Q)} \le c_{14}\|y\|_{(\mathbf{\Theta}(Q_0,\eta))^n}. \tag{4.21}$$

Applying to elliptic boundary value problem (4.20) the well-known estimate of its solution and taking into account (4.21) we obtain

$$\begin{aligned}\|m_0\|^2_{L_2(0,T;W_2^2(\Omega_0))} &= \|m\rho^2e^\eta\|^2_{L_2(0,T;W_2^2(\Omega_0))}\\ &\le c_{15}\|g\|^2_{L_2(Q_0)} \le c_{16}\|y\|^2_{\mathbf{\Theta}(Q_0,\eta)^n}.\end{aligned} \tag{4.22}$$

Since

$$\begin{aligned}|\partial^2_{x_ix_j}(\rho^2me^\eta)|^2 &\ge \frac{1}{2}|\partial_{x_i}(\rho^2\partial_{x_j}m)e^\eta|^2\\ &- c_{17}(|\rho^2(\partial_{x_j}m)\partial_{x_i}e^\eta|^2 + |(\partial_{x_i}m)\partial_{x_j}(\rho^2e^\eta)|^2 + |m\partial^2_{x_ix_j}(\rho^2e^\eta)|^2),\end{aligned}$$

then inequalities (4.22), (4.19), and (4.5) imply the estimate

$$\begin{aligned}\int_{Q_0}e^{2\eta}\sum_{j=1}^n|\partial_{x_j}(\rho^2\nabla m)|^2\,dx\,dt &\le c_{18}\int_{Q_0}\sum_{i,j=1}^n(|\partial^2_{x_ix_j}(\rho^2me^\eta)|^2\\ &+ |\rho^2(\partial_{x_j}m)\partial_{x_i}e^\eta|^2 + |(\partial_{x_i}m)\partial_{x_j}(\rho^2e^\eta)|^2 + |m\partial^2_{x_ix_j}(\rho^2e^\eta)|^2)\,dx\,dt\\ &\le c_{19}\|y\|^2_{(\mathbf{\Theta}(Q_0,\eta))^n}.\end{aligned} \tag{4.23}$$

Denote $m_i = \rho^3(\partial_{x_i} m)e^{\eta}(T-t)$. Then, by virtue of (4.4) with $u = m$,

$$\Delta m_i = g_i, \qquad m_i|_{\partial\Omega_0} = 0, \tag{4.24}$$

where

$$g_i = (\partial_{x_i} m)\Delta(\rho^3 e^{\eta}(T-t)) + 2(\nabla(\rho^3 e^{\eta}(T-t)), \partial_{x_i}\nabla m) + \rho^3 e^{\eta}(T-t)\partial_{x_i}\mathbf{div}\, y.$$

Apply an estimate of the solution of the Laplace equation to the solution m_i of problem (4.24). Then, as in (4.22), we get, with the help of inequalities (4.5), (4.19), (4.23),

$$\begin{aligned}\|\rho^3(\partial_{x_i} m)e^{\eta}(T-t)\|^2_{L_2(0,T;W_2^2(\Omega_0))} &\le c_{20}\Bigg(\|(\partial_{x_i} m)\Delta(\rho^3 e^{\eta}(T-t))\|^2_{L_2(Q_0)} \\ &+ \left\|\left(\frac{1}{\rho^2}\nabla(\rho^3 e^{\eta}(T-t)), (\partial_{x_i}(\rho^2\nabla m) - 2(\partial_{x_i}\rho)\rho\nabla m)\right)\right\|^2_{L_2(Q_0)} \\ &+ \|\rho^3 e^{\eta}(T-t)\partial_{x_i}\mathbf{div}\, y\|^2_{L_2(Q_0)}\Bigg) \le c_{21}\|y\|^2_{(\mathbf{\Theta}(Q_0,\eta))^n}.\end{aligned} \tag{4.25}$$

As in (4.22), inequalities (4.25) with $i = 1, \dots, n$, and estimates (4.23), (4.19), (4.5) imply:

$$\int_{Q_0} e^{2\eta}\sum_{k,l=1}^{n}|\partial^2_{x_k x_l}(\rho^3\nabla m)|^2(T-t)^2\,dx\,dt \le c_{22}\|y\|^2_{(\mathbf{\Theta}(Q_0,\eta))^n}. \tag{4.26}$$

By virtue of (4.13),

$$\Delta(\rho\partial_t m) = \partial_t m\Delta\rho + 2(\nabla\rho, \nabla\partial_t m) + \rho\,\mathbf{div}\partial_t y. \tag{4.27}$$

Scaling (4.27) by $-(\rho\partial_t m)e^{2\eta}(T-t)^2$ in $L_2(Q_0)$ and integrating by parts we have

$$\begin{aligned}\int_{Q_0}|\nabla(\rho\partial_t m)|^2(T-t)^2\,dx\,dt &= \int_{Q_0}\Bigg(\frac{1}{2}(\rho\partial_t m)^2\Delta e^{2\eta}(T-t)^2 - \rho(\partial_t m)^2\Delta\rho e^{2\eta}(T-t)^2 \\ &+\frac{1}{2}(T-t)^2(\partial_t m)^2\mathbf{div}(e^{\eta}\nabla\rho^2) - \rho^2\partial_t m(\mathbf{div}\,\partial_t y)e^{2\eta}(T-t)^2\Bigg)\,dx\,dt.\end{aligned}$$

This equality implies

$$\begin{aligned}\int_{Q_0}|\rho\nabla\partial_t m|^2(T-t)^2e^{2\eta}\,dx\,dt &\le c_{23}\int_{Q_0}|\partial_t m|^2(c_{24}|\nabla\rho|^2(T-t)^2 + \rho^2 \\ &+ |\rho\Delta\rho|(T-t)^2 + c_{25}(T-t)(|\nabla\rho^2| + (T-t)|\Delta\rho^2|))e^{2\eta}\,dx\,dt \\ &+ \int_{Q_0}[(\rho\nabla\partial_t m, \partial_t y)\rho e^{2\eta}(T-t)^2 + \partial_t m(\nabla(\rho^2 e^{2\eta}), \partial_t y)(T-t)^2]\,dx\,dt \\ &\le c_{26}\int_{Q_0}|\partial_t m|^2e^{2\eta}\,dx\,dt + \frac{1}{2}\int_{Q_0}|\rho\nabla\partial_t m|^2(T-t)^2e^{2\eta}\,dx\,dt \\ &+ c_{27}\int_{Q_0}e^{2\eta}|\partial_t y|^2(T-t)^2\,dx\,dt.\end{aligned} \tag{4.28}$$

After transferring the term with $\rho\nabla\partial_t m$ from the right side of (4.28) to the left side, we get, with the help of (4.6) and (2.18),

$$\int_{Q_0}|\rho\nabla\partial_t m|^2(T-t)^2e^{2\eta}\,dx\,dt \le c_{28}\|y\|^2_{(\mathbf{\Theta}(Q_0,\eta))^n}.$$

This equality and upper bounds (4.19), (4.23), (4.26) imply (4.17). □

We prove now the main result of this section.

THEOREM 4.1. *Let s satisfy the conditions of Lemma* 4.1. *Then an arbitrary vector field $y \in (\mathbf{\Theta}(Q_0, \eta^s))^n$ admits decomposition* (4.1), *where* $\mathbf{div}\, v(t,x) \equiv 0$ *and* $\rho^3 \nabla q \in M(Q_0, \eta^s)$, *and if $y(t,x)$ satisfies equality* $\mathbf{div}\, y(0,x) \equiv 0$, *then* $y(0,x) \equiv z(0,x)$.

Proof. We define the function $\varphi(t) \in C^\infty(0,T)$, such that $\varphi(t) \equiv 0$ when $t \in (0, T/4)$ and $\varphi(t) \equiv 1$ when $t \in [\frac{3}{4}T, T]$. Let $m(t,x)$ be the solution of problem (4.3), (4.4) constructed in Lemma 4.1. Since $y \in (\mathbf{\Theta}(Q_0, \eta))^n$, then for almost all $t \in (0,T)$ the function $\Delta m(t,\cdot) \in L_2(\Omega_0)$, and by virtue of (4.5), $m(t,\cdot) \in L_2(\Omega_0)$. Hence (see J.-L. Lions and E. Magenes [32]) the restriction $m(t,\cdot)|_{\partial\Omega}$ is well defined and belongs to $H^{-1/2}(\partial\Omega_0)$. We introduce the function $\zeta(t,x)$, defined on $(0,T) \times \partial\Omega_0$ by formula

$$\zeta(t,x) = \varphi(t) m(t,x), \qquad t \in (0,T),\ x \in \partial\Omega_0,$$

and consider the following Dirichlet problem:

$$\Delta q(t,x) = \mathbf{div}\, y(t,x), \qquad (t,x) \in Q_0, \tag{4.29}$$

$$q|_{(0,T)\times\partial\Omega_0} = \zeta. \tag{4.30}$$

The unique solution $q(t,x)$ of (4.29), (4.30) exists (see J.-L. Lions and E. Magenes [32]), and by virtue of the properties of $\zeta(t,x)$,

$$q(x,t) \equiv m(t,x) \qquad \forall\ (t,x) \in [3T/4, T] \times \Omega_0 \tag{4.31}$$

and

$$\forall t \in [0, T/4] \qquad \mathbf{div}\, y(t,x) \equiv 0 \qquad \text{implies} \qquad q(t,x) \equiv 0. \tag{4.32}$$

By virtue of (4.31) and (4.17) we have $\rho^3 \nabla q \in M(Q_0, \eta)$. Besides, (4.2) follows from (4.32). □

5. The proof of the main results.

5.1. First, we want to solve the exact controllability problem for linearized Boussinesq equations (2.11)–(2.15). To do it we apply the analogous controllability result for the parabolic equation which is formulated below. We consider the controllability problem for the heat equation

$$\partial_t \theta(t,x) - \Delta\theta(t,x) = h(t,x), \qquad (t,x) \in Q_0, \tag{5.1}$$

$$\theta|_{t=0} = \theta_0(x), \qquad \theta|_{t=T} = 0, \qquad x \in \Omega_0, \tag{5.2}$$

where the functions $h \in L_2(Q_0, \eta)$, $\theta_0 \in H^1(\Omega)$ are given . Many authors studied this controllability problem (see D. Russell [35], H. Fattorini [11], and T. Seidman [36]). But, taking into account the functional spaces where we look for a solution, the following result which can be extracted from A. Fursikov and O. Imanuvilov [15], [16], [26] is convenient for us.

THEOREM 5.1. *There exists a number s_1 such that for any $s > s_1$ and for arbitrary given $\theta_0 \in H^1(\Omega_0)$, $h \in L_2(Q_0, \eta^s)$ there exists the solution $\theta \in \mathbf{\Theta}(Q_0, \eta^s)$ of problem* (5.1), (5.2).

We consider also the controllability problem for the following parabolic system:

$$\partial_t y(t,x) - \Delta y - \hat{v} \times \mathbf{rot}\, y = f, \qquad (t,x) \in Q_0, \tag{5.3}$$

$$y|_{t=0} = y_0, \qquad y|_{t=T} = 0, \qquad x \in \Omega_0. \tag{5.4}$$

THEOREM 5.2. *Let $\hat{v}(t,x) \in V^{1,2(2)}(Q_0)$ be given. Then there exists a number s_2 such that for any $s > s_2$ and for arbitrary given data $y_0 \in (H^1(\Omega_0))^n$, $f \in (L_2(Q_0,\eta^s))^n$, a solution $y \in (\mathbf{\Theta}(Q_0,\eta^s))^n$ of problem* (5.3), (5.4) *exists.*

One can prove Theorem 5.2 absolutely by the same way as in A. Fursikov and O. Imanuvilov [15], [16], [26], [12]. Let us prove one abstract lemma.

LEMMA 5.1. *Suppose that X, Y are separable Hilbert spaces, a bounded linear operator $B : X \to Y$ is surjective, and $K : X \to Y$ is a linear compact operator. Then the image of the operator $B + K$ is closed in Y.*

Proof. For an arbitrary $\epsilon > 0$ there exists the operator K_ϵ that has finite-dimensional image and

$$\|K - K_\epsilon\| < \epsilon \tag{5.5}$$

(see [34]). The equality

$$B + K = B_\epsilon + K_\epsilon \qquad \text{where } B_\epsilon = B + (K - K_\epsilon)$$

is true. If in (5.5) ϵ is small enough, then the image of operator B_ϵ coincides with the whole Y. Thus, we reduce Lemma 5.1 to the case when operator $K : X \to Y$ has a finite-dimensional image.

We can suppose also that $\operatorname{Ker} B \cap \operatorname{Ker} K = 0$. Indeed, if it is not so we introduce the factor space $X_1 = X/(\operatorname{Ker} B \cap \operatorname{Ker} K)$, define operators B_1 and K_1 by formulas

$$B_1\tilde{x} = Bx, \qquad K_1\tilde{x} = Kx, \qquad \text{where} \qquad \tilde{x} = x + \operatorname{Ker} B \cap \operatorname{Ker} K,$$

and consider the problem on the closure of operators $B_1 + K_1 : X_1 \to Y$ image. Since operator K has a finite-dimensional image then there exists a finite linear independent system of vectors $e_1, \dots, e_k \in Y$ and a linear independent system of functionals $f_1, \dots, f_n$ defined and bounded on X such that

$$Kx = \sum_{j=1}^{k} f_j(x) e_j.$$

The linear independentness of $f_1, \dots, f_n$ implies that there exist such linear independent vectors $g_1, \dots, g_k \in X$ that $f_j(g_i) = \delta_{i,j}$, where $\delta_{i,j}$ is Kronecker symbol. Hence, the space X admits the decomposition

$$X = [g_1, \dots, g_k] + \operatorname{Ker} K$$

where $[g_1, \dots, g_k]$ is a linear span of $g_1, \dots, g_k$. Since $\operatorname{Ker} B \cap \operatorname{Ker} K = 0$, then $\dim \operatorname{Ker} B \le k$ and X admits the decomposition

$$X = S + \operatorname{Ker} B + \operatorname{Ker} K,$$

where S is a certain finite-dimensional space. Let B_2, K_2 be the restrictions at the space $S + \operatorname{Ker} K$ of the operators B and K, respectively. Since the operator

$$B : S + \operatorname{Ker} K \to Y$$

is an isomorphism, then by Fredholm theorem the image $B_2 + K_2$ is closed and has a finite codimension in Y. The coincidence $(B_2+K_2)(S+\operatorname{Ker} K) = (B+K)(S+\operatorname{Ker} K)$ implies the embedding

$$(B + K)(S + \operatorname{Ker} K) \subset (B + K)X.$$

Hence $(B+K)X = (B+K)(S+\mathrm{Ker}\ K)+S_1$, where S_1 is a certain finite-dimensional subspace of Y. Being a finite-dimensional space, the subspace S_1 is closed. Hence, $(B+K)$ is closed also. □

5.2. Now we prove the assertion on the closure of a set of data for which the controllability problem for the Boussinesq equations has a solution.

THEOREM 5.3. *Let $\hat{v}(t,x) \in V^{1,2(2)}(Q), \hat{\theta}(t,x) \in W^{1,2(2)}(Q)$. Then the set of data (f,h,v_0,θ_0) for which there exists a solution $(v,p,\theta) \in X^s(Q)$ of problem (2.11)–(2.15) is closed in the space $Z^s(Q)$ when the magnitude of parameter s is sufficiently large (spaces $X^s(Q), Z^s(Q)$ are defined in (2.22), (2.23)).*

Proof. To prove this theorem we intend to apply Lemma 5.1. We decompose the operator generated by problem (2.11)–(2.15) into the sum $B+K$, where B is the operator generated by the problem

$$\partial_t v(t,x) - \Delta v - \hat{v} \times \mathbf{rot}\, v + \nabla p = f(t,x), \qquad \mathbf{div}\, v = 0, \qquad v(0,x) = v_0(x), \tag{5.6}$$

$$\partial_t \theta(t,x) - \Delta\theta = h(t,x), \qquad \theta(0,x) = \theta_0(x), \tag{5.7}$$

$$v(T,x) \equiv 0, \qquad \theta(T,x) \equiv 0. \tag{5.8}$$

The operator K is defined by the formula

$$K(v,p,\theta) = (-v \times \mathbf{rot}\, \hat{v} + \theta e_0, (\hat{v}, \nabla\theta) + (v, \nabla\hat{\theta}) + (v, e_0), 0, 0). \tag{5.9}$$

The boundedness of the operator

$$B : X^s(Q) \to Z^s(Q) \tag{5.10}$$

is proved in Lemma 2.1. To prove that operator (5.10) is surjective we first, instead of (5.6), use more simple equations:

$$\partial_t y(t,x) - \Delta y - \hat{v} \times \mathbf{rot}\, y = f_1(t,x), \qquad y(0,x) = y_0(x). \tag{5.11}$$

Let Q_0, Ω_0 be the set introduced in the beginning of section 3.2. We continuously extend $\hat{v}(t,x)$ from $V^{1,2(2)}(Q)$ to $V^{1,2(2)}(Q_0)$ as well as $\hat{\theta}(t,x)$ from $W^{1,2(2)}(Q)$ to $W^{1,2(2)}(Q_0)$ using Propositions 2.2 and 2.3 and consider the problem (5.11), (5.7) on Q_0. Note that $y_0(x) \in V^1(\Omega_0)$ is an extension of $v_0 \in V^1(\Omega)$.

We choose a parameter s satisfying conditions of Theorems 4.1, 5.1, and 5.2 simultaneously. Then by virtue of these theorems for an arbitrary $(f_1,h,y_0,\theta_0) \in (L_2(Q_0,\eta^s))^n \times L_2(Q_0,\eta^s) \times V^1(\Omega_0) \times H^1(\Omega_0)$, there exists a solution $(y,\theta) \in (\mathbf{\Theta}(Q_0,\eta^s))^n \times \mathbf{\Theta}(Q_0,\eta^s)$ of problem (5.11), (5.7) defined on Q_0. With the help of Theorem 4.1 we decompose the component y of this solution as follows:

$$y(t,x) = v(t,x) + \nabla q, \tag{5.12}$$

where $\mathbf{div}\, v = 0$, $\rho^3 \nabla q \in M(Q_0,\eta^s)$. Here $M(Q_0,\eta)$ is space (4.16) and $y(0,x) = v(0,x) = y_0(x)$. We substitute (5.12) into (5.11) and verify that $v(t,x)$ satisfies the equation

$$\partial_t v(t,x) - \Delta v - \hat{v} \times \mathbf{rot}\, v + \nabla m = f_1(t,x),\ \mathbf{div}\, v = 0,\ v(0,x) = y_0(x), \tag{5.13}$$

$$m = (\partial_t q - \Delta q). \tag{5.14}$$

Now we can prove that operator (5.10) is surjective. Indeed, let $(f,h,v_0,\theta_0) \in Z^s(Q) = F(Q,\eta^s) \times L_2(Q,\eta^s) \times V^1(\Omega) \times H^1(\Omega)$. By the definition of the space $F(Q,\eta)$, the decomposition

$$f = f_1 + \nabla f_2, \qquad f_1 \in (L_2(Q,\eta^s))^n, \qquad f_2 \in L_2(0,T;H^1(\Omega))$$

holds. After extension of f_1, f_2, h from Q onto Q_0 and v_0, θ_0 from Ω onto Ω_0 we get, as was shown above, the function (v, m, θ), which satisfies (5.13), (5.7), (5.8). Evidently, if we define

$$p = m + f_2, \tag{5.15}$$

then (v, p, θ) will satisfy (5.6)–(5.8). After the restriction of (v, p, θ) at Q this triplet satisfies boundary value problem (5.6)–(5.8), which we consider on Q. We made an extension from Q to Q_0, and after that the restriction from Q_0 to Q, to have the equality (5.12) defined on Q with $\nabla q \in M(Q, \eta^s)$ (the restriction onto Q allows us to take off the multiplier ρ^3 including $\rho^3 \nabla q \in M(Q_0, \eta^s)$). Since $\nabla q \in M(Q, \eta^s)$, then by virtue of (5.14), (5.15), $p \in L_2(0, T; H^1(\Omega))$.

Equality (5.12) and inclusions $\nabla q \in M(Q, \eta), y \in (\boldsymbol{\Theta}(Q, \eta))^n$ imply that all terms in definition (2.21) of $\|\cdot\|^2_{\Xi(Q,\eta)}$ for v are finite, except perhaps the term $\|\partial_t v - \Delta v\|_{F(Q,\eta)}$. Let us show that this term is also finite. By virtue of (5.6), (5.15), (5.14)

$$\begin{aligned}\|\partial_t v - \Delta v\|_{F(Q,\eta)} &= \|f_1 + \hat{v} \times \mathbf{rot}\, v + \nabla f_2 - \nabla p\|_{F(Q,\eta)} \le \|f_1 + \hat{v} \times \mathbf{rot}\, v\|_{(L_2(Q,\eta))^n} \\ &+ \|\nabla f_2 - \nabla p\|_{(L_2(Q,\eta))^n} \le c_1(\|f_1\|_{(L_2(Q,\eta))^n} \\ &+ \|\hat{v}\|_{(C(\bar{Q}))^n} \||\nabla v|\|_{(L_2(Q,\eta))^n} + \||\nabla(\partial_t q - \Delta q)|\|_{(L_2(Q,\eta))^n}) < \infty.\end{aligned}$$

Hence, $v \in \Xi(Q, \eta)$, and therefore we have proved that operator (5.10) is surjective. We prove now that the operator

$$K : X^s(Q) \to Z^s(Q) \tag{5.16}$$

is compact, where K is as defined in (5.9). To prove this assertion one has to establish compactness of the operator

$$K_1 : X^s(Q) \to (L_2(Q, \eta))^n \times L_2(Q, \eta), \tag{5.17}$$

where

$$K_1(v, p, \theta) = (-v \times \mathbf{rot}\, \hat{v} + \theta e_0, (\hat{v}, \nabla\theta) + (v, \nabla\hat{\theta}) + (v, e_0)). \tag{5.18}$$

We have

$$\begin{aligned}&\int_{T-\delta}^{T} \int_\Omega e^{2\eta}(|v \times \mathbf{rot}\, \hat{v}|^2 + |(\hat{v}, \nabla\theta) + (v, \nabla\hat{\theta}) + (v, e_0)|^2)\, dx\, dt \qquad (5.19)\\ &\le c_2(\|\hat{v}\|^2_{C(0,T;(C^1(\bar{\Omega}))^n)} + \|\hat{\theta}\|^2_{C(0,T;C^1(\bar{\Omega}))} + 1) \int_{T-\delta}^{T} \int_\Omega e^{2\eta}(|v|^2 + |\theta|^2 + |\nabla\theta|^2)\, dx\, dt \\ &\le c_3(\|\hat{v}\|^2_{V^{1,2(2)}(Q)} + \|\hat{\theta}\|^2_{W^{1,2(2)}(Q)} + 1)\delta \int_{T-\delta}^{T} \int_\Omega e^{2\eta}((T-t)^{-2}(|v|^2 + |\theta|^2) \\ &+ (T-t)^{-1}|\nabla\theta|^2)\, dx\, dt \le c_4 c_3 \delta(\|\hat{v}\|^2_{V^{1,2(2)}(Q)} + \|\hat{\theta}\|^2_{W^{1,2(2)}(Q)})\end{aligned}$$

uniformly with respect to

$$(v, \theta) \in \Phi \equiv \{(v, \theta) : \|v\|^2_{\Xi(Q,\eta)} + \|\theta\|^2_{\boldsymbol{\Theta}(Q,\eta)} \le c_4\}.$$

Evidently, at $Q^\delta = (0, T - \delta) \times \Omega$ we have

$$\Xi(Q^\delta, \eta) = V^{1,2(0)}(Q^\delta), \qquad \boldsymbol{\Theta}(Q^\delta, \eta) = W^{1,2(0)}(Q^\delta), \qquad L_2(Q^\delta, \eta) = L_2(Q^\delta),$$

and by the Sobolev embedding theorem the operator

$$K : V^{1,2(0)}(Q^\delta) \times L_2(0,T;W_2^1(\Omega)) \times W^{1,2(0)}(Q^\delta) \to (L_2(Q^\delta))^{n+1} \times V^1(\Omega) \times H^1(\Omega)$$

is compact. This property of operator K and (5.19) prove the compactness of operator (5.17), (5.18). Hence, all assumptions of Lemma 5.1 are true and by this lemma we get assertion of Theorem 5.3. □

Now we can immediately prove Theorem 5.4.

THEOREM 5.4. *Let $\hat v \in V^{1,2(2)}(Q), \hat\theta \in W^{1,2(2)}(Q)$. Then for an arbitrary data $(f,h,v_0,\theta_0) \in Z^s(Q)$ there exists a solution $(v,p,\theta) \in X^s(Q)$ of problem* (2.11)–(2.15) *when the magnitude of parameter s is sufficiently large*[2].

Proof. By Theorem 3.1 the set of data $(f,h,v_0,\theta_0) \in Z^s(Q)$ for which problem (2.11)–(2.13) has a solution $(v,p,\theta) \in X^s(Q)$ is dense in $Z^s(Q)$, and by Theorem 5.3, it is closed. Hence, this set coincides with $Z^s(Q)$. □

Proof of Theorem 1.1. First we apply the right inverse operator theorem to problem (2.3)–(2.7). Let $\mathcal{A}$ be operator (2.10), (2.3), (2.5) and the spaces $X = X^s(Q)$, $Z = Z^s(Q)$ be defined in (2.22), (2.23), (2.17)–(2.21). Taking into account that $\mathcal{A}$ is a sum of linear and quadratic operators we can assert that continuous differentiability of operator (2.8) follows from Lemmas 2.1 and 2.2. Equality (2.9) is evident for $x_0 = 0$, $z_0 = 0$. At last, the assertion that operator

$$\mathcal{A}'(0) : X^s(Q) \to Z^s(Q)$$

is surjective has been proved in Theorem 5.4. So, all assumptions of the right inverse operator theorem are fulfilled and therefore there exists $\epsilon > 0$ such that for any initial data (w_0,τ_0) satisfying inequality

$$\|w_0\|^2_{V^1(\Omega)} + \|\tau_0\|^2_{H^1(\Omega)} \le \epsilon$$

and for zero right sides of equation (2.3), (2.5) the problem (2.3)–(2.7) possesses the solution $(v,q,\theta) \in \Xi(Q,\eta^s) \times L_2(0,T;H^1(\Omega)) \times \mathbf{\Theta}(Q,\eta^s)$. After returning from problem (2.3)–(2.7) to problem (1.1)–(1.4), (1.6) by change of variables (2.2) we get the assertion of Theorem 1.1. □

Remark 5.1. As we pointed out in Remark 1.1 the smoothness condition imposed on the given solution $(\hat v, \hat p, \hat\theta)$ in Theorem 1.1 can be replaced by the more weak condition (1.14). This change of condition would lead to the following complication of Theorem 5.3's proof, which we show below. We approximate functions $\hat v, \hat\theta$ by functions $\hat v_\epsilon \in V^{1,2(2)}(Q)$, $\hat\theta_\epsilon \in W^{1,2(2)}(Q)$:

$$\|\hat v - \hat v_\epsilon\|_{V^{1,2(1/2)}(Q)\cap(L_\infty(Q))^n} \le \epsilon, \qquad \|\hat\theta - \hat\theta_\epsilon\|_{W^{1,2(1/2)}(Q)\cap L_\infty(Q)} < \epsilon, \tag{5.20}$$

where ϵ is sufficiently small. We can write

$$B + K = B + R_\epsilon + K_\epsilon,$$

where

$$K_\epsilon(v,\theta) = (-v \times \mathbf{rot}\, \hat v_\epsilon + \theta e_0, (\hat v_\epsilon, \nabla\theta) + (v, \nabla\hat\theta_\epsilon) + (v, e_0), 0, 0),$$

$$R_\epsilon(v,\theta) = (-v \times \mathbf{rot}\, (\hat v - \hat v_\epsilon), (\hat v - \hat v_\epsilon, \nabla\theta) + (v, \nabla(\hat\theta - \hat\theta_\epsilon)), 0, 0).$$

By virtue of (5.20) the operator $R_\epsilon : X^s(Q) \to Z^s(Q)$ has a small norm, and therefore the operator $B + R_\epsilon : X^s(Q) \to Z^s(Q)$ is surjective. The compactness of operator $K_\epsilon : X^s(Q) \to Z^s(Q)$ has been proved in Theorem 5.3. Hence, by Lemma 5.1 the image of operator $B + R_\epsilon + K_\epsilon$ coincides with $Z^s(Q)$. □

[2]More precisely, s simultaneously satisfies the conditions of Theorems 4.1, 5.1, 5.2.

6. The proof of a Carleman estimate. In this section we prove Lemma 3.2.

Proof of Lemma 3.2. We make the change of variables

$$z(t,x) = e^{-s\alpha} w(t,x) \tag{6.1}$$

in (3.6), (3.7). As a result, by virtue of (6.1) we get

$$L_1 w(t,x) + L_2 w(t,x) = f_s(t,x), \qquad (t,x) \in Q, \tag{6.2}$$

$$w|_\Sigma = \left.\frac{\partial w}{\partial \nu}\right|_\Sigma = 0, \tag{6.3}$$

where

$$L_1 w = \Delta w + s^2\varphi^2 w - s(\partial_t \alpha) w, \tag{6.4}$$

$$L_2 w = \partial_t w - 2s\varphi \partial_{x_1} w, \tag{6.5}$$

$$f_s = e^{s\alpha} f + s\varphi w. \tag{6.6}$$

Besides, by virtue of (6.1) and by the properties of α we have

$$w|_{t=0} = w|_{t=T} = 0. \tag{6.7}$$

Equation (6.2) implies

$$\|L_1 w\|^2_{L_2(Q)} + \|L_2 w\|^2_{L_2(Q)} + 2(L_1 w, L_2 w)_{L_2(Q)} = \|f_s\|^2_{L_2(Q)}. \tag{6.8}$$

By virtue of (6.4), (6.5) we get

$$(L_1 w, L_2 w)_{L_2(Q)} = I_1 + I_2 + I_3, \tag{6.9}$$

where

$$I_1 = \int_Q (\Delta w + s^2\varphi^2 w - s(\partial_t\alpha) w)\partial_t w \, dx\, dt, \tag{6.10}$$

$$I_2 = -\int_Q \Delta w (2s\varphi \partial_{x_1} w)\, dx\, dt, \tag{6.11}$$

$$I_3 = -\int_Q (s^2\varphi^2 - s(\partial_t\alpha))(2s\varphi w \partial_{x_1} w)\, dx\, dt. \tag{6.12}$$

Let us transform I_1, I_2, I_3. Integration by parts in (6.10) with the help of (6.3), (6.7) yields

$$\begin{aligned} I_1 = \int_Q \left(-\frac{1}{2}\partial_t |\nabla w|^2 + \frac{1}{2}(s^2\varphi^2 - s(\partial_t\alpha))\partial_t |w|^2 \right) dx\, dt \\ = -\int_Q \left(s^2\varphi \partial_t \varphi - \frac{s}{2}\partial^2_{tt}\alpha \right) |w|^2 \, dx\, dt. \end{aligned} \tag{6.13}$$

Analogously, integration by parts with respect to x in (6.12) with help of (6.3) yields

$$\begin{aligned} I_3 = -\int_Q (s^2\varphi^2 - s\partial_t\alpha) s\varphi \partial_{x_1} w^2 \, dx\, dt = \int_Q (3s^3\varphi^3 w^2 \\ - s^2(\partial_t\varphi)\varphi w^2 - s^2(\partial_t\alpha)\varphi w^2)\, dx\, dt. \end{aligned} \tag{6.14}$$

Finally, let us estimate term (6.11). Integration by parts and (6.1) imply

$$I_2 = (\nabla w, \nabla(2s\varphi\partial_{x_1} w))_{(L_2(Q))^n} = \int_Q (2s\varphi(\partial_{x_1} w)^2 + s\varphi\partial_{x_1}|\nabla w|^2)\,dx\,dt = \int_Q (2s\varphi(\partial_{x_1} w)^2 - s\varphi|\nabla w|^2)\,dx\,dt. \tag{6.15}$$

We substitute (6.13), (6.14), (6.15) into (6.9), and after that substitute the obtained equality into (6.8). As a result we have

$$\|L_1 w\|^2_{L_2(Q)} + \|L_2 w\|^2_{L_2(Q)} + 2\int_Q (3s^3\varphi^3|w|^2 - s\varphi|\nabla w|^2 + (\partial_{x_1} w)^2 2s\varphi)\,dx\,dt = \|f_s\|^2_{L_2(Q)} + X_1, \tag{6.16}$$

where

$$X_1 = 2\int_Q \left(s^2\varphi\partial_t\varphi - \frac{s}{2}\partial^2_{tt}\alpha + s^2\varphi(\partial_t\varphi) + s^2\varphi(\partial_t\alpha)\right)|w|^2\,dx\,dt. \tag{6.17}$$

We get, with the help of a simple estimation of (6.6),

$$\|f_s\|^2_{L_2(\Omega)} \le 2\int_Q (e^{2s\alpha}|f|^2 + s^2\varphi^2|w|^2)\,dx. \tag{6.18}$$

Definition (3.4) of φ and α imply the inequalities

$$|\partial_t\varphi| \le c_1\varphi^2, \qquad |\partial_t\alpha| \le c_2\varphi^2, \qquad |\partial^2_{tt}\alpha| \le c_3\varphi^3, \tag{6.19}$$

where c_1, c_2, c_3 do not depend on s, t, x. The estimation of (6.17) with the help of (6.19) yields

$$|X_1| \le c_4\int_Q (1+s^2)\varphi^3|w|^2\,dx\,dt. \tag{6.20}$$

Scaling (6.2) by $s\varphi w$ in $L_2(Q)$ and taking into account (6.4) we get after integration by parts

$$\int_Q f_s s\varphi w\,dx\,dt = \int_Q (L_2 w)s\varphi w\,dx\,dt + \int_Q \left(s^3\varphi^3|w|^2 - s\varphi(\partial_t\alpha)|w|^2 - s\varphi|\nabla w|^2 + \frac{1}{2}s\Delta\varphi|w|^2\right)dx\,dt.$$

We can rewrite this equality in the form

$$\int_Q s\varphi|\nabla w|^2\,dx\,dt = \int_Q s^3\varphi^3|w|^2\,dx\,dt - X_2, \tag{6.21}$$

where

$$X_2 = \int_Q \left(f_s s\varphi w - (L_2 w)s\varphi w + s\varphi(\partial_t\alpha)|w|^2 - \frac{1}{2}s\varphi|w|^2\right)dx\,dt. \tag{6.22}$$

We estimate X_2, taking into account (6.18), (6.19):

$$|X_2| \le \frac{1}{4}\|L_2 w\|^2_{L_2(Q)} + c_5 \int_Q (e^{2s\alpha}|f|^2 + (s^2\varphi^2 + s^2\varphi^3 + s\varphi)|w|^2)\,dx\,dt. \tag{6.23}$$

The estimation of (6.16) by means of (6.20), (6.18) yields:

$$\|L_1 w\|^2_{L_2(Q)} + \|L_2 w\|^2_{L_2(Q)} + 2\int_Q (3s^3\varphi^3|w|^2 - s\varphi|\nabla w|^2)\,dx\,dt \le \int_Q e^{2s\alpha}|f|^2 dx dt + c_6 \int_Q ((1+s^2)\varphi^3 + s^2\varphi^2)|w|^2\,dx\,dt. \tag{6.24}$$

We express the terms $\int_Q s\varphi|\nabla w|^2\,dx\,dt$ in (6.24) by means of (6.21) and after that use estimate (6.23). As a result we get the upper bound

$$\|L_1 w\|^2_{L_2(Q)} + \|L_2 w\|^2_{L_2(Q)} + 2\int_Q 2s^3\varphi^3|w|^2\,dx\,dt \le \frac{1}{2}\|L_2 w\|^2_{L_2(Q)} + c_9 \int_Q (e^{2s\alpha}|f|^2 + s^2\varphi^3 w^2)\,dx\,dt. \tag{6.25}$$

By (6.25) there exists a parameter s_0 such that the following inequality holds:

$$\|L_1 w\|^2_{L_2(Q)} + \|L_2 w\|^2_{L_2(Q)} + \int_Q s^3\varphi^3|w|^2\,dx\,dt \le c_{10}\int_Q e^{2s\alpha}|f|^2\,dx\,dt \ \ \forall\ s \ge s_0, \tag{6.26}$$

where c_{10} does not depend on s. After the estimation of the right side of (6.21) with the help of (6.23), (6.26) we get

$$\int_Q s\varphi|\nabla w|^2\,dx\,dt \le c_{11}\int_Q e^{2s\alpha}|f|^2\,dx\,dt \ \ \forall\ s \ge s_0. \tag{6.27}$$

Multiplying (6.4) on $(s\varphi)^{-1/2}$ and estimating with the help of (6.26), (6.19) we get

$$\int_Q (s\varphi)^{-1}|\Delta w|^2\,dx\,dt \le c_{12}\int_Q ((s\varphi)^{-1}|L_1 w|^2 + s^3\varphi^3 w^2 + s(\varphi)^3 w^2)\,dx\,dt \le c_{13}\int_Q e^{2s\alpha}|f|^2\,dx\,dt \quad \forall\ s > s_0. \tag{6.28}$$

Analogously, multiplying (6.5) on $(s\varphi)^{-1/2}$, we obtain the following inequality by means of (6.26), (6.27):

$$\int_Q (s\varphi)^{-1}|\partial_t w|^2\,dx\,dt \le c_{14}\int_Q e^{2s\alpha}|f|^2\,dx\,dt \quad \forall\ s > s_0. \tag{6.29}$$

After substituting $w = e^{s\alpha}z$ into (6.26)–(6.29), we obtain (3.8). □

REFERENCES

[1] V. Alekseev, V. Tikhomirov, and S. Fomin, *Optimal Control*, Consultants Bureau, New York, 1987.

[2] V. Babich, *On the propagation of functions*, Russian Math. Surveys, 8 (1953), pp. 111–113 (in Russian).

[3] J.-M. Coron, *Contrôlabilité exacte frontière de l'équation d'Euler des fluides parfaits incompressibles bidimensionneles*, C. R. Acad. Sci. Paris Sér. I Math., 317 (1993), pp. 271–276.

[4] J.-M. Coron, *On the controllability of* 2*-D incompressible perfect fluids*, J. Math. Pures Appl., 75 (1996), pp. 155–188.

[5] J.-M. Coron, *On the controllability of the* 2*-D incompressible Navier-Stokes equations with the Navier slip boundary conditions*, European Series in Applied and Industrial Mathematics Control, Optimization and Calculus of Variations, 1 (1996), pp. 35–75.

[6] J.I. Diaz and A.V. Fursikov, *Approximate Controllability of the Stokes system on cylinders by external unidirectional forces*, J. Math. Pures Appl., 76 (1997), pp. 353–375.

[7] C. Fabre, *Résultats d'unicité pour les équations de Stokes et applications au contrôle*, C.R. Acad. Sci. Paris Sér. 1 Math., 322 (1996), pp. 1191–1196.

[8] C. Fabre and G. Lebeau, *Prolongement unique des solutions de l'équation de Stokes*, Preprint.

[9] C. Fabre, J.-P. Puel, and E. Zuazua, *Contrôlabilité approchée de l'équation de la chaleur semi-linéaire*, C. R. Acad. Sci. Paris Sér. 1, 315 (1992), pp. 807–812.

[10] C. Fabre, J.-P. Puel, and E. Zuazua, *Approximate controllability of the semilinear heat equation*, in Proc. Roy. Soc. Edinburgh, 125A (1995), pp. 31–61.

[11] H. Fattorini, *Boundary control of temperature distributions in a parallepipedon*, SIAM J. Control, 13 (1975), pp. 1–13.

[12] A. Fursikov, *Exact boundary zero controllability of three dimensional Navier-Stokes equations*, J. Dynam. Control Systems, 1 (1995), pp. 325–350.

[13] A. Fursikov, *Control problems and theorems concerning the unique solvability of a mixed boundary value problem for the three-dimensional Navier-Stokes and Euler equations*, Math. USSR Sbornik, 43 (1982), pp. 251–273.

[14] A. Fursikov, *Lagrange principle for problems of optimal control of ill-posed or singular distributed systems*, J. Math. Pures Appl., 71 (1992), pp. 139–194.

[15] A. Fursikov and O. Imanuvilov, *On controllability of certain systems simulating a fluid flow*, in Flow Control, IMA Vol. Math. Appl. 68, M.D. Gunzburger, ed., Springer-Verlag, New York, 1995, pp. 149–184.

[16] A. Fursikov and O. Imanuvilov, *On exact boundary zero-controllability of two-dimensional Navier–Stokes equations*, Acta Appl. Math., 37 (1994), pp. 67–76.

[17] A. Fursikov and O. Imanuvilov, *Local exact controllability of two dimensional Navier-Stokes system with control on the part of the boundary*, Sbornik. Math., 187 (1996), pp. 1355–1390.

[18] A. Fursikov and O. Imanuvilov, *On approximate controllability of the Stokes system*, Ann. Faculté Sci. Toulouse, 11 (1993), pp. 205–232.

[19] A. Fursikov and O. Imanuvilov, *Local exact controllability of the Navier-Stokes equations*, C.R. Acad. Sci. Paris Ser. 1 Math., 323, (1996), pp. 275–280.

[20] A. Fursikov and O. Imanuvilov, *On* ϵ*-controllability of the Stokes problem with distributed control concentrated in a subdomain*, Russian Math Surveys, 47 (1992), pp. 255–256.

[21] L. Hörmander, *Linear Partial Differential Operators*, Springer-Verlag, Berlin, 1963.

[22] L. Hörmander, *The Analysis of Linear Partial Differential Operators, Fourier Integral Operators*, Springer-Verlag, Berlin, 1985.

[23] O. Imanuvilov, *Some Problems of Optimization and Exact Controllability*, Thesis, Moscow State University, Moscow, 1991 (in Russian).

[24] O. Imanuvilov, *Exact controllability of the semilinear parabolic equation*, Vestnik R.U.D.N. Ser. Math., 1 (1994), pp. 109–116 (in Russian).

[25] O. Imanuvilov, *Exact boundary controllability of the parabolic equation*, Russian Math. Surveys, 48 (1993), pp. 211–212.

[26] O. Imanuvilov, *Boundary controllability of parabolic equations*, Sb. Math., 186 (1995), pp. 879–900.

[27] D. Joseph, *Stability of Fluid Motions*, Vol. II, Springer-Verlag, New York, 1976.

[28] G. Lebeau and L. Robbiano, *Contrôle exact de l'eq́uation de la chaleur*, Comm. Partial Differential Equations, 20 (1995), pp. 336–356.

[29] J.- L. Lions, *Are there connections between turbulence and controllability?*, In Analyse et optimization des systemes, Lecture Notes in Control and Inform. Sciences 144, Springer-Verlag, New York, 1990.

[30] J.- L. Lions, *Remarques sur la contrôlabilité approché*, in Proceedings of "Jornadas Hispano-Francesas sobre Control de Sistemas Distribuidos," University of Malaga, Spain, 1990.

[31] J.- L. Lions, *Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles*, Dunod Gauthier-Villars, Paris, 1968.

[32] J.-L. Lions and E. Magenes, *Problemes aux limites nonhomogènes et applications*, Vol. 1, Dunod, Paris, 1968.

[33] J.-L. Lions and E. Zuazua, *A generic uniqueness result for the Stokes system and its control theoretical consequences*, in Partial Differential Equatons and Applications, P. Marcellini, G. Talenti, and E. Visentini eds., Lecture Notes Pure Appl. Math. 177, Marcel Dekker, New York, 1996, pp. 221–235.

[34] M. Reed and B. Simon, *Methods of Modern Mathematical Physics. Functional Analysis*, Academic Press, New York, London, 1972.

[35] D. Russell, *Controllability and stabilizability theory for linear partial differential equations. Recent progress and open questions*, SIAM Review, 20 (1978), pp. 639–739.

[36] T. Seidman, *Two results on exact boundary controllability of parabolic equations*, Appl. Math. Optim., 11 (1984), pp. 145–152.

[37] L. Slobodetskii, *Generalized Sobolev spaces and their applications to boundary value problems for partial differential equations*, Science Notices of Leningrad State Pedagogic Institute, 197 (1958), pp. 54–112 (in Russian).

[38] R. Temam, *Navier-Stokes Equations*, North-Holland, Amsterdam, 1979.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 2, pp. 422–446, March 1998

002

DISCRETIZED MAXIMUM LIKELIHOOD AND ALMOST OPTIMAL ADAPTIVE CONTROL OF ERGODIC MARKOV MODELS*

T. E. DUNCAN†, B. PASIK-DUNCAN†, AND L. STETTNER‡

Abstract. Three distinct controlled ergodic Markov models are considered here. The models are a discrete time controlled Markov process with complete observations, a controlled diffusion process with complete observations, and a discrete time controlled Markov process with partial observations. The partial observations for the third model have the special form of complete observations in a fixed recurrent set and noisy observations in its complement. For each of the models an almost self-optimizing adaptive control is given. These adaptive controls are constructed from a family of estimates that use a finite discretization of the parameter set and a finite family of almost optimal ergodic controls by a randomized certainty equivalence method. A continuity property of the information of a model for one parameter value with respect to another is used to establish this almost optimality property.

Key words. adaptive control, ergodic control, Markov processes, controlled Markov processes, almost optimal adaptive control

AMS subject classifications. 93E35, 93C40, 60J05, 62M05

PII. S0363012996298369

1. Introduction.

In many control problems the models are not completely described and there are perturbations or unmodeled dynamics that are described by noise so that the models are stochastic. If some distributions or parameters in the models are unknown then these control problems can be considered as problems of stochastic adaptive control. In this paper, three unknown ergodic Markov models are considered. The models are a discrete time controlled Markov process with complete observations, a controlled diffusion process with complete observations, and a discrete time controlled Markov process with partial observations. The discrete time Markov processes evolve in a compact state space, and the transition densities depend on an unknown parameter. The partial observations of the discrete time Markov process in the third model have the special form of complete observations in a fixed recurrent set and noisy observations in its complement. The controlled diffusion is described by a stochastic differential equation where the unknown parameter appears in the drift vector. The solution of the stochastic differential equation is given in the weak sense. Since there are some basic differences among these three models, it is convenient to treat them separately. Typically, the results that are given here are stated for each of the three models.

Since the true value of the parameter is unknown, it is estimated using the maximum likelihood procedure where the time differences between the successive updates of the estimates are sufficiently large so that an ergodic property of the information and the cost can be used. Since only almost self-optimality is desired, the maximum likelihood procedure is restricted to choosing from a finite set of possible values for

*Received by the editors February 9, 1996; accepted for publication (in revised form) December 11, 1996. This research was supported by the Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania.

http://www.siam.org/journals/sicon/36-2/29836.html

†Department of Mathematics, University of Kansas, Lawrence, KS 66045 (duncan@math.ukans.edu, bozenna@kuhub.cc.ukans.edu). This research was supported in part by NSF grants DMS 9305936 and DMS 9623439.

‡Institute of Mathematics, Sniadeckich 8, 00–950 Warsaw, Poland (stettner@impan.gov.pl).

the parameter that is a discretization of the possible parameter values. The adaptive strategy uses a randomized certainty equivalence control that chooses with probability almost 1 the control that is almost optimal for the current value of the estimates, and with small, positive probability each of the other almost optimal controls. This procedure is shown to give an almost self-optimizing adaptive control.

The adaptive control of ergodic Markov models has been considered elsewhere (e.g., [1, 2, 4, 6, 8, 9, 10, 13]). However, only here is the maximization of the likelihood function restricted to a finite, discretized set of the possible parameter values. The work of Agrawal [1] has motivated the use here of information and the randomized certainty equivalence adaptive control. A cost-biased maximum likelihood method introduced in [13] is used in [6, 8] for two of the models considered here. The methods used here relax some of the assumptions in [4, 6, 8]. For example, the global Lipschitz continuity of the drift vector with respect to the unknown parameter for the controlled diffusion model is replaced by only continuity, and the requirement that the law of large numbers for some martingales be uniform in the parameter, which necessitated some assumptions in [6, 8], is not required.

The three models that are considered here can be generalized in various ways. The discrete time Markov process can be modified to include the discrete time recursive model in [17]. The controlled diffusion model can be generalized by analogy to [7] to include processes that satisfy stochastic differential equations with delays. The partial observations structure used here can be modified to noisy observations everywhere if there is a sequence of random times such that the process at these times is a family of independent, identically distributed random variables.

The three models are specifically described as follows.

Model I—Discrete time controlled Markov process. A Markov process $(X_n,\, n \in \mathbb{N})$ evolves in a compact metric space E with the transition operator $P(x_n, \mathrm{d}y; v_n, \alpha^0)$ at time $n \in \mathbb{N}$, where $\alpha^0 \in \mathcal{A}$ is an unknown fixed parameter and $\mathcal{A}$ is a compact metric space, and the control v_n takes values in a compact metric space U and is adapted to $\sigma(X_0, \ldots, X_n)$. A generic parameter value $\alpha \in \mathcal{A}$ has a transition operator that is described by replacing α^0 by α above. The transition operators have continuous densities with respect to a fixed measure $\varphi(\,\cdot\,)$; that is, for each $B \in \mathcal{B}(E)$, the Borel σ-algebra on E, and each $\alpha \in \mathcal{A}$,

$$P(x, B; v, \alpha) = \int_B p(x, y, v, \alpha)\varphi(\mathrm{d}y), \tag{1}$$

where φ is a probability measure on E and $p : E \times E \times U \times \mathcal{A} \to \mathbb{R}_+$ is continuous. It is assumed that $p(x, y, v, \alpha) > 0$ for all $x, y \in E$, $v \in U$, and $\alpha \in \mathcal{A}$, and $\operatorname{supp}\varphi = E$. The control problem is to minimize the following ergodic cost functional:

$$I^1((v_n,\, n \in \mathbb{N})) = \limsup_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} c(X_i, v_i), \tag{2}$$

where $c : E \times U \to \mathbb{R}_+$ is a bounded, Borel measurable function. The family of controls $(v_n,\, n \in \mathbb{N})$ has the form $v_n = u(X_n)$, where $u \in \mathcal{U}$ and $\mathcal{U}$ is the family of Borel measurable functions from E to U.

Model II—Controlled diffusion process. Let $(X(t),\, t \in \mathbb{R}_+)$ be a controlled diffusion process that satisfies the following stochastic differential equation:

$$\begin{aligned} \mathrm{d}X(t) &= f(X(t))\mathrm{d}t + h(X(t), \alpha^0, v(t))\mathrm{d}t + \sigma(X(t))\mathrm{d}W(t), \\ X(0) &= x, \end{aligned} \tag{3}$$

where $X(t) \in \mathbb{R}^n$, $(W(t), t \geq 0)$ is a standard $\mathbb{R}^n$-valued Wiener process, $\alpha^0 \in \mathcal{A}$ is unknown and $\mathcal{A}$ is compact, $(v(t), t \geq 0)$ is adapted to $\sigma(X(s), s \leq t)$, and $v(t) \in U$, a compact set. The functions f and σ satisfy a global Lipschitz condition, $\sigma(x)\sigma^*(x) \geq cI > 0$ for all $x \in \mathbb{R}^n$, and $h : \mathbb{R}^n \times \mathcal{A} \times U \to \mathbb{R}^n$ is a bounded Borel measurable function and either $h(x, \cdot, v)$ is continuous uniformly in $v \in U$ and x from compact subsets of $\mathbb{R}^n$ or $h(x, \cdot, \cdot)$ is continuous for each $x \in \mathbb{R}^n$. The solution of the stochastic differential equation (3) is given in the weak sense by an absolutely continuous transformation of the measure of the strong solution of

$$\begin{aligned} \mathrm{d}Y(t) &= f(Y(t))\mathrm{d}t + \sigma(Y(t))\mathrm{d}W(t), \\ Y(0) &= x. \end{aligned} \tag{4}$$

The family of controls $(v(t), t \geq 0)$ has the form $v(t) = u(X(t))$, where $u \in \mathcal{U}$ and $\mathcal{U}$ is the family of Borel measurable functions from $\mathbb{R}^n$ into U. Let T_A be the first hitting time of $A \in \mathcal{B}(\mathbb{R}^n)$; that is,

$$T_A = \begin{cases} \inf\{s > 0 : X(s) \in A\}, \\ +\infty \qquad\qquad\qquad\qquad \text{if the above set is empty.} \end{cases}$$

Let Γ_1 and Γ_2 be two spheres in $\mathbb{R}^n$ with centers at 0 and radii $0 < r_1 < r_2$, respectively. Let τ be given as

$$\tau = T_{\Gamma_2} + T_{\Gamma_1} \circ \theta_{T_{\Gamma_2}}, \tag{5}$$

where $(\theta_t, t \geq 0)$ is the family of shift operators acting on $C(\mathbb{R}_+, \mathbb{R}^n)$. The random time τ is the first time that the process $(X(t), t \geq 0)$ hits Γ_1 after hitting Γ_2. It is assumed that

$$\sup_{\alpha \in \mathcal{A}} \sup_{u \in \mathcal{U}} \sup_{x \in \Gamma_1} E_x^{\alpha,u}[\tau^2] < \infty \tag{6}$$

and

$$E_x^{\alpha,u}[T_{\Gamma_1}] < \infty \tag{7}$$

for each $(x, \alpha, u) \in \mathbb{R}^n \times \mathcal{A} \times \mathcal{U}$, where $E_x^{\alpha,u}$ is the expectation with respect to a process $(X(t), t \geq 0)$ that satisfies (3) with α^0 replaced by α and $v(t) = u(X(t))$. The dependence of the solution of (3) on $\alpha \in \mathcal{A}$ and the control $(v(t), t \geq 0)$ is suppressed for notational convenience. However, it is shown explicitly when the expectations of functions of the solution are taken. The control problem is to minimize the ergodic cost functional

$$I^2((v(t), t \geq 0)) = \limsup_{t \to \infty} \frac{1}{t} \int_0^t c(X(s), v(s))\mathrm{d}s, \tag{8}$$

where $c : \mathbb{R}^n \times U \to \mathbb{R}_+$ is a bounded, Borel measurable function.

Model III—A partially observed discrete time controlled Markov process. A controlled Markov process $(X_n, n \in \mathbb{N})$ evolves in a compact subset E of R^d with the transition operator $P(x_n, \mathrm{d}y; v_n, \alpha^0)$ at time $n \in \mathbb{N}$, where $\alpha^0 \in \mathcal{A}$ is an unknown parameter, $\mathcal{A}$ is a compact metric space, and the control v_n takes values in a compact metric space U. A generic parameter value $\alpha \in \mathcal{A}$ has a transition operator

that is described by replacing α^0 by α above. The transition operators have densities with respect to Lebesgue measure; that is,

$$P(x, B; v, \alpha) = \int_B p(x, y, v, \alpha)\mathrm{d}y, \tag{9}$$

where $\alpha \in \mathcal{A}$, $B \in \mathcal{B}(E)$, and $p : E \times E \times U \times \mathcal{A} \to \mathbb{R}_+$ is continuous and $p > 0$ on $E \times E \times U \times \mathcal{A}$. The process $(X_n,\, n \in \mathbb{N})$ is completely observed in a nonempty compact subset $\Gamma \subset E$ and is partially observed in $E \setminus \Gamma$. The observation process $(Y_n,\, n \in \mathbb{N})$ is explicitly described as follows:

$$P(Y_i \in B | X_i, \mathcal{Y}_{i-1}) = 1_{B\cap\Gamma}(X_i) + 1_{\Gamma^{\mathrm{c}}}(X_i) \int_{B\cap\Gamma^{\mathrm{c}}} r(X_i, y)\mathrm{d}y, \tag{10}$$

where $B \in \mathcal{B}(E)$, $\mathcal{Y}_i = \sigma(Y_1, \ldots, Y_i)$, $\mathcal{Y}_0 = \{\emptyset, \Omega\}$, and $r : \Gamma^{\mathrm{c}} \times \Gamma^{\mathrm{c}} \to \mathbb{R}$ is Borel measurable such that $\int_{\Gamma^{\mathrm{c}}} r(x, y)\mathrm{d}y = 1$ for each $x \in \Gamma^{\mathrm{c}}$. A control v is a U-valued, $\mathcal{Y}_n$-adapted process. The control problem is to minimize the ergodic cost functional

$$I^3((v_n,\, n \in \mathbb{N})) = \limsup_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} c(X_i, v_i), \tag{11}$$

where $c \in C(E \times U)$. It is assumed that there is a nonempty compact set $\Gamma_1 \subset \Gamma$ such that for each probability law μ on X_0, control $u = (v_n,\, n \in \mathbb{N})$, and $\alpha \in \mathcal{A}$,

$$E_\mu^{\alpha,u}[T_{\Gamma_1}] < \infty \tag{12}$$

and

$$\sup_{x\in\Gamma_1} \sup_{u\in\mathcal{U}} E_x^{\alpha,u}[\tau^2] < \infty, \tag{13}$$

where T_{Γ_1} is the first hitting time of Γ_1, τ is the first hitting time of Γ_1 after hitting Γ^{c} ($\tau = T_{\Gamma^{\mathrm{c}}} + T_{\Gamma_1} \circ \theta_{T_{\Gamma^{\mathrm{c}}}}$), and $E_\mu^{\alpha,u}$ is the expectation for the process $(X_n,\, n \in \mathbb{N})$ with initial law μ, control u, and parameter $\alpha \in \mathcal{A}$. For a probability law μ for X_0 the measure-valued process $(\Pi_n^{\alpha^0},\, n \in \mathbb{N})$ is defined as follows:

$$\Pi_0^{\alpha^0}(B) = \mu(B), \tag{14}$$

$$\Pi_n^{\alpha^0}(B) = P_n(X_n \in B | \mathcal{Y}_n) \tag{15}$$

for each $B \in \mathcal{B}(E)$. This conditional measure process can be represented more explicitly using (10) (e.g., Lemma 1 of [18]) as follows:

$$\Pi_{n+1}^{\alpha^0}(B) = 1_{B\cap\Gamma}(Y_{n+1}) + 1_{\Gamma^{\mathrm{c}}}(Y_{n+1}) M(Y_{n+1}, \Pi_n^{\alpha^0}, v_n, \alpha^0)(B), \tag{16}$$

where

$$M(y, \nu, v, \alpha^0) = \frac{\displaystyle\int_{B\cap\Gamma^{\mathrm{c}}} r(z, y) p(\nu, z, v, \alpha^0)\mathrm{d}z}{\displaystyle\int_{\Gamma^{\mathrm{c}}} r(z, y) p(\nu, z, v, \alpha^0)\mathrm{d}z} \tag{17}$$

and

$$p(\nu, z, v, \alpha^0) = \int_E p(x, z, v, \alpha^0)\nu(\mathrm{d}x). \tag{18}$$

2. A finite family of almost optimal controls. For the adaptive control of Models I, II, and III a finite family of controls is constructed that includes at least one that is almost optimal for each parameter value $\alpha \in \mathcal{A}$.

To determine the almost optimal controls the averaged versions of the ergodic cost functionals (2), (8), (11) are used. These are denoted as follows:

$$J_\mu^{\alpha^0,1}((v_n,\, n\in\mathbb{N})) = \limsup_{n\to\infty} \frac{1}{n} E_\mu^{\alpha^0,v}\left[\sum_{i=0}^{n-1} c(X_i, v_i)\right], \tag{19}$$

$$J_\mu^{\alpha^0,2}((v(t),\, t\geq 0)) = \limsup_{t\to\infty} \frac{1}{t} E_\mu^{\alpha^0,v}\left[\int_0^t c(X(s), v(s))\mathrm{d}s\right], \tag{20}$$

$$J_\mu^{\alpha^0,3}((v_n,\, n\in\mathbb{N})) = \limsup_{n\to\infty} \frac{1}{n} E_\mu^{\alpha^0,v}\left[\sum_{i=0}^{n-1} c(X_i, v_i)\right], \tag{21}$$

where μ is the probability law for X_0 and $\alpha^0 \in \mathcal{A}$ is the true parameter value. The finite families of almost optimal controls for Models I, II, and III are constructed for the cost functionals $J_\mu^{\alpha^0,1}$, $J_\mu^{\alpha^0,2}$, and $J_\mu^{\alpha^0,3}$, respectively.

Model I. It is assumed that it suffices to consider controls of the form $v_n = u_n(X_n)$, where $u_n \in \mathcal{U} = B(E,U)$, the family of Borel measurable functions from E to U. Clearly, this restriction is satisfied if c is a continuous, bounded function because by (1) for $B \in \mathcal{B}(E)$, $x \in E$, $v \in U$, and $\alpha \in \mathcal{A}$,

$$P^{v,\alpha}(x,B) \geq \inf_{x,y\in E} \inf_{v\in U} \inf_{\alpha\in\mathcal{A}} p(x,y,v,\alpha)\varphi(B) \tag{22}$$

and

$$\inf_{x,y\in E} \inf_{v\in U} \inf_{\alpha\in\mathcal{A}} p(x,y,v,\alpha) > 0,$$

and (see Theorem 2.2 and Corollary 3.6 in Chap. 3 of [10]) for each $\alpha \in \mathcal{A}$, there is a $u_\alpha \in B(E,U)$ such that $J_\mu^{\alpha,1}(u_\alpha)$ is optimal.

For Model I there is a uniform ergodicity property and a finite family of almost optimal controls.

PROPOSITION 1. *For Model* I *with $\alpha \in \mathcal{A}$ and $u \in \mathcal{U}$ there is a probability measure π_u^α on $\mathcal{B}(E)$ such that*

$$\sup_{u\in\mathcal{U}} \sup_{\alpha\in\mathcal{A}} \sup_{x\in E} \|(P^{u,\alpha})^n(x,\cdot) - \pi_u^\alpha(\cdot)\|_{\mathrm{var}} \leq 2\gamma_0^{n-1} \tag{23}$$

where $\|\cdot\|_{\mathrm{var}}$ is the variation norm and

$$\gamma_0 = 1 - \inf_{x,y\in E} \inf_{v\in U} \inf_{\alpha\in\mathcal{A}} p(x,y,v,\alpha).$$

There is a constant K_1 such that for $\alpha, \beta \in \mathcal{A}$ and $u \in \mathcal{U}$

$$\|\pi_u^\alpha - \pi_u^\beta\|_{\mathrm{var}} \leq K_1 \sup_{x\in E} \|P^{u,\alpha}(x,\cdot) - P^{u,\beta}(x,\cdot)\|_{\mathrm{var}}. \tag{24}$$

Furthermore, given $\varepsilon > 0$ there is a finite family of controls $\mathcal{U}^1(\varepsilon) = \{u^1, \ldots, u^{r(\varepsilon)}\}$ such that for each $\alpha \in \mathcal{A}$ there is a $k \in \{1, \ldots, r(\varepsilon)\}$ and

$$J_\mu^{\alpha,1}(u^k(X(\cdot))) \leq \lambda_1(\alpha) + \varepsilon, \tag{25}$$

where

$$\lambda_1(\alpha) = \inf_{u\in\mathcal{U}} J_\mu^{\alpha,1}(u(X(\cdot))). \tag{26}$$

Proof. From (22) and (5.6) of [5], (23) is verified. The inequality (24) follows from the proof of Proposition 1 of [17]. The existence of a finite family $\mathcal{U}^1(\varepsilon)$ satisfying (25) follows from (24) and the proof of Lemma 2 of [17]. □

Model II. Let $(\tau_n,\ n\in\mathbb{N})$ be an increasing sequence of random times such that $\tau_1=\tau$ and $\tau_{n+1}=\tau_n\circ\theta_{\tau_n}$ for $n>1$. For a given control $u\in\mathcal{U}$ and parameter $\alpha^0\in\mathcal{A}$ there is a unique invariant measure $\eta_u^{\alpha^0}$ for the embedded Markov chain $(X_{\tau_n},\ n\in\mathbb{N})$ and $X_0\in\Gamma_1$ (e.g., [6]). Furthermore, there is a unique invariant measure $\pi_u^{\alpha^0}$ for the process $(X(t),\ t\ge 0)$ with $v(t)=u(X(t))$, and it has the form

$$\pi_u^{\alpha^0}(B) = \frac{\displaystyle\int_{\Gamma_1} E_x^{\alpha^0,u}\left[\int_0^\tau 1_B(X(s))\mathrm{d}s\right]\eta_u^{\alpha^0}(\mathrm{d}x)}{\displaystyle\int_{\Gamma_1} E_x^{\alpha^0,u}[\tau]\eta_u^{\alpha^0}(\mathrm{d}x)}. \tag{27}$$

For Model II there is an analogue of Proposition 1.

PROPOSITION 2. *For Model* II *there is* $\gamma_0\in(0,1)$ *such that*

$$\sup_{u\in\mathcal{U}}\sup_{\alpha\in\mathcal{A}}\sup_{x\in\Gamma_1}\sup_{B\in\mathcal{B}(\Gamma_1)} |P_x^{\alpha,u}(X(\tau_n)\in B)-\eta_u^\alpha(B)|\le\gamma_0^n, \tag{28}$$

where η_u^α *is the unique invariant measure for the embedded Markov chain. There is a constant* K_1 *such that for* $\alpha,\beta\in\mathcal{A}$ *and* $u\in\mathcal{U}$,

$$\|\eta_u^\alpha-\eta_u^\beta\|_{\mathrm{var}}\le K_1\sup_{x\in E}\sup_{B\in\mathcal{B}(\Gamma_1)}|P_x^{\alpha,u}(X(\tau)\in B)-P_x^{\beta,u}(X(\tau)\in B)|. \tag{29}$$

Furthermore, given $\varepsilon>0$ *there is a* $\delta>0$ *such that if* $\alpha,\beta\in\mathcal{A}$ *and* $\rho_{\mathcal{A}}(\alpha,\beta)<\delta$ *then*

$$\sup_{u\in\mathcal{U}}\sup_{x\in\Gamma_1}\sup_{B\in\mathcal{B}(\Gamma_1)}|P_x^{\alpha,u}(X(\tau)\in B)-P_x^{\beta,u}(X(\tau)\in B)|<\varepsilon \tag{30}$$

and

$$\sup_{u\in\mathcal{U}}\sup_{x\in\Gamma_1}\sup_{B\in\mathcal{B}(\Gamma_1)}\left|E_x^{\alpha,u}\left[\int_0^\tau 1_B(X(s))\mathrm{d}s\right]-E_x^{\beta,u}\left[\int_0^\tau 1_B(X(s))\mathrm{d}s\right]\right|<\varepsilon, \tag{31}$$

where $\rho_{\mathcal{A}}$ *is a metric on* $\mathcal{A}$ *compatible with its topology. Furthermore, given* $\varepsilon>0$, *there is a finite family of controls* $\mathcal{U}^2(\varepsilon)=\{u^1,\dots,u^{r(\varepsilon)}\}$ *such that for each* $\alpha\in\mathcal{A}$ *there is a* $k\in\{1,\dots,r(\varepsilon)\}$ *and*

$$J_\mu^{\alpha,2}(u^k(X(\cdot)))\le\lambda_2(\alpha)+\varepsilon, \tag{32}$$

where

$$\lambda_2(\alpha)=\inf_{u\in\mathcal{U}} J_\mu^{\alpha,2}(u(X(\cdot))). \tag{33}$$

Proof. By Proposition 2.2 of [6] and Theorem 4.1 of [3] the inequality (28) follows. Using the proof of Proposition 1 of [6], as in our Proposition 1, the inequality (29)

follows. The uniform continuity properties (30), (31) can be verified as for (10) and (19) of [6]. Since h is not assumed to be Lipschitz continuous with respect to $\alpha \in \mathcal{A}$, it is necessary to verify that the map

$$H : \Gamma_1 \times \mathcal{A} \to \mathbb{R} \tag{34}$$

given by

$$H(x,\alpha) = E_x\left[\int_0^t |\sigma^{-1}(Y(s))h(Y(s),\alpha,u(Y(s)))|^2 \mathrm{d}s\right] \tag{35}$$

is continuous uniformly in $u \in \mathcal{U}$, where E_x is the expectation for P_x that is the measure for the solution of (4), and

$$E_x\left[\int_0^t |\sigma^{-1}(Y(s))(h(Y(s),\alpha,u(Y(s))) - h(Y(s),\beta,u(Y(s))))|^2 \mathrm{d}s\right] \to 0$$

as $\rho_{\mathcal{A}}(\alpha,\beta) \to 0$ uniformly in $u \in U$. The proof of this last continuity is similar to the verification of the continuity of H, so only the verification of H is given. Since h is bounded, it is sufficient to verify the continuity of $\bar{H} : \Gamma_1 \times \mathcal{A} \to \mathbb{R}$ given by

$$\bar{H}(x,\alpha) = \int_{t_1}^t E_x|\sigma^{-1}(Y(s))h(Y(s),\alpha,u(Y(s)))|^2 \mathrm{d}s \tag{36}$$

for each $t_1 < t$ uniformly in $u \in U$. To verify this continuity note that the map $(s,x) \in (0,\infty) \times \mathbb{R}^d \mapsto P_x(Y(s) \in \cdot)$ is continuous in the variation norm topology. In fact, by Lemma 9.22 of [19], for $s_n \to s > 0$ and $x_n \to x$ the family of measures $(P_{x_n}(Y(s_n) \in \cdot),\, n \in \mathbb{N})$ is tight, so for any $\varepsilon > 0$ there is a compact set $K \subset \mathbb{R}^d$ such that

$$P_y(Y(s_n) \in K^{\mathrm{c}}) < \varepsilon$$

for all $y \in \{x, x_1, x_2, \ldots\}$. By Theorem 3.2.1 of [19] the measures $(P_\cdot(Y(s) \in \cdot),\, s > 0)$ have continuous densities. As $n \to \infty$ the following inequality is easily verified from the previous inequality:

$$\sup_{B \in \mathcal{B}(\mathbb{R}^d)} |P_{x_n}(Y(s_n) \in B) - P_x(Y(s) \in B)| \le 2\varepsilon + \int_K |p(s_n,x_n,y) - p(s,x,y)| \mathrm{d}y.$$

Thus the continuity in the variation norm of $P_\cdot(Y(s) \in \cdot)$ is verified. By this continuity and the continuity of $\alpha \mapsto h(y,\alpha,v)$ that is uniform in $v \in U$, the continuity of (36) and therefore (34) follows. Now only the verification of (32) remains. By (27), (29), (30), (31), given $\varepsilon > 0$, there is a $\delta > 0$ such that if $\alpha, \beta \in \mathcal{A}$ and $\rho_{\mathcal{A}}(\alpha,\beta) < \delta$, then

$$\sup_{u \in \mathcal{U}} \|\pi_u^\alpha - \pi_u^\beta\|_{\mathrm{var}} < \varepsilon.$$

So by Propositions 2.3 and 2.4 of [6] there is a finite family of controls $\mathcal{U}^2(\varepsilon)$ such that the inequality (32) is satisfied. □

Model III. Let $\tilde{\mathcal{U}}$ be a fixed compact subset of $C(\mathcal{P}(E),U)$ where $P(E)$ is the family of probability measures on E with the vague topology, and let $u(\alpha)$ be the control sequence such that $v_n = u(\Pi_n^\alpha)$. Define λ_3 as follows:

$$\lambda_3(\alpha) = \inf_{u \in \tilde{\mathcal{U}}} J_\mu^{\alpha,3}(u(\Pi_n^\alpha)) \tag{37}$$

and an increasing sequence of random times $(\tau_n,\, n \in \mathbb{N})$ as $\tau_1 = \tau$, $\tau_{n+1} = \tau_n + \tau \circ \theta_{\tau_n}$, where $(\theta_t,\, t \geq 0)$ is the family of shift operators acting on $C(\mathbb{R}_+, \mathbb{R}^d)$.

Some additional assumptions are made on Model III.

(A1) The function r given in (10) is continuous on $\Gamma^{\mathrm{c}} \times \Gamma^{\mathrm{c}}$ and bounded on $\Gamma^{\mathrm{c}}_\delta \times \Gamma^{\mathrm{c}}_\delta$ for some $\delta > 0$ where $\Gamma^{\mathrm{c}}_\delta = \{(y,z) \in \Gamma^{\mathrm{c}} \times \Gamma^{\mathrm{c}} : \rho_E(y,\Gamma) \geq \delta\} \cup \{(y,z) \in \Gamma^{\mathrm{c}} \times \Gamma^{\mathrm{c}} : \rho_E(y,z) \geq \delta\}$ and ρ_E is a metric on E that is compatible with its topology. If $(y_n,\, n \in \mathbb{N})$ is a sequence in Γ^{c} such that $y_n \to y \in \Gamma$ as $n \to \infty$, and for $\delta > 0$, $B(y,\delta) = \{z \in \Gamma^{\mathrm{c}} : \rho_E(z,y) \leq \delta\}$, then

$$\lim_{n\to\infty} \inf_{\alpha\in\mathcal{A}} \inf_{v\in\tilde{\mathcal{U}}} \inf_{x\in K} \int_{B(y,\delta)} r(z,y_n) P^{\alpha,v}(x,\mathrm{d}z) = \infty \tag{38}$$

for any compact subset $K \subset E$.

(A2) If $(z_n,\, n \in \mathbb{N})$ is a sequence in Γ^{c} that converges to z, then

$$\lim_{n\to\infty} R(z_n,\cdot) = R(z,\cdot),$$

where the topology is the vague convergence of measures and

$$R(z,A) = \begin{cases} \displaystyle\int_{A\cap\Gamma^{\mathrm{c}}} r(z,y)\mathrm{d}y & \text{for } z \in \Gamma^{\mathrm{c}}, \\ 1_A(z) & \text{for } z \in \Gamma \end{cases} \tag{39}$$

for $A \in \mathcal{B}(E)$.

Using (A1) and (A2) an analogue of Propositions 1 and 2 is given for Model III.

PROPOSITION 3. *For Model* III, *if* (A1) *and* (A2) *are satisfied, then there is a* $\gamma_0 \in (0,1)$ *and a measure* $\eta^\alpha_{u(\beta)}$ *on* Γ *such that*

$$\sup_{u\in\tilde{\mathcal{U}}} \sup_{\alpha,\beta\in\mathcal{A}} \sup_{x\in\Gamma} \sup_{B\in\mathcal{B}(\Gamma_1)} |P^{\alpha,u(\beta)}_x(X_{\tau_n} \in B) - \eta^\alpha_{u(\beta)}(B)| < \gamma_0^n. \tag{40}$$

Given $\varepsilon > 0$ *there is a finite family of controls* $\tilde{\mathcal{U}}(\varepsilon) = \{u^1,\dots,u^{r(\varepsilon)}\} \subset \tilde{\mathcal{U}}$ *and* $\delta_0 > 0$ *such that if* $\rho_{\mathcal{A}}(\alpha,\beta) < \delta_0$, *then*

$$\lambda_3(\beta) - \varepsilon \leq J^{\beta,3}_\mu(u^k(\Pi^\alpha_n)) \leq \lambda_3(\beta) + \varepsilon \tag{41}$$

for some $k \in \{1,\dots,r(\varepsilon)\}$.

Proof. The continuity and the positivity of the transition density p and (13) imply (40). Using the proof of Lemma 3 of [17], it follows that

$$J^{\beta,3}_\mu(u^k(\Pi^\alpha_n)) = \frac{\displaystyle\int_\Gamma E^{\beta,u(\alpha)}_x \left[\sum_{i=0}^{\tau-1} \int_E c(z,u(\Pi^\alpha_i))\Pi^\beta_i(\mathrm{d}z)\eta^\beta_{u(\alpha)}(\mathrm{d}x)\right]}{\displaystyle\int_\Gamma E^{\beta,u(\alpha)}_x[\tau]\eta^\beta_{u(\alpha)}(\mathrm{d}x)}.$$

By the proofs of Lemma 2 of [17] and Proposition 2.4 of [6], for the verification of (41) it suffices to show that given $\varepsilon > 0$, there is a $\delta > 0$ such that for $\alpha,\beta,\alpha_1,\beta_1 \in \mathcal{A}$, if $\rho_{\mathcal{A}}(\alpha,\alpha_1) < \delta$ and $\rho_{\mathcal{A}}(\beta,\beta_1) < \delta$ then

$$\sup_{u\in\tilde{\mathcal{U}}} |J^{\beta,3}_\mu(u(\Pi^\alpha_n)) - J^{\beta_1,3}_\mu(u(\Pi^{\alpha_1}_n))| < \varepsilon. \tag{42}$$

If the inequality (42) is not satisfied then there are sequences $(\alpha^m,\ m\in\mathbb{N})$, $(\alpha_1^m,\ m\in\mathbb{N})$ such that $\alpha^m\to\alpha$, $\alpha_1^m\to\alpha_1$, and $u_m(v)\to u(v)$ uniformly in $v\in\mathcal{P}(E)$ as $m\to\infty$ and

$$|J_\mu^{\alpha^m,3}(u_m(\Pi_n^{\alpha^m}))-J_\mu^{\alpha_1^m,3}(u_m(\Pi_n^{\alpha_1^m}))|\ge\varepsilon>0 \tag{43}$$

for all $m\in\mathbb{N}$. Using some continuity arguments in the proofs of Theorems 1 and 6 of [18], where the pair $(\alpha,v)\in\mathcal{A}\times U$ is considered as the control, it follows that

$$\lim_{m\to\infty}J_\mu^{\alpha^m,3}(u_m(\Pi_n^{\alpha^m}))=J_\mu^{\alpha,3}(u(\Pi_n^\alpha))$$

and

$$\lim_{m\to\infty}J_\mu^{\alpha_1^m,3}(u_m(\Pi_n^{\alpha_1^m}))=J_\mu^{\alpha,3}(u(\Pi_n^\alpha)),$$

which contradict the inequality (43). Thus (42) is satisfied and there is a finite family $\tilde{\mathcal{U}}(\varepsilon)$ of controls such that (41) is satisfied. □

Let $\tilde{\mathcal{A}}(\delta_0)=\{\alpha(1),\dots,\alpha(k(\delta_0))\}$ be distinguished points, one from each of a finite δ_0 net in $\mathcal{A}$. By Proposition 3, given $\varepsilon>0$, there is an $\tilde{\mathcal{A}}(\delta_0)$ from a δ_0 net of $\mathcal{A}$ such that the controls $(u^k(\Pi_n^\alpha),\ k\in\{1,\dots,r(\varepsilon)\}$ and $\alpha\in\tilde{\mathcal{A}}(\delta_0))$ form the family of ε optimal controls.

Remark. A finite family of controls for $\mathcal{U}^{(1)}(\varepsilon)$ can be obtained from a discretization of the Bellman equation (cf. [6] and Section 3.5 of [10]). A finite family of controls for $\mathcal{U}^{(2)}(\varepsilon)$ can be obtained using [14], and a finite family of controls for $\tilde{\mathcal{U}}(\varepsilon)$ can be obtained using [15].

3. The information for different parameters. Kullback and Leibler [11] have used a notion of information in statistics. For the adaptive control problems for the three models considered here the information is computed from the probability densities for different values of the unknown parameter. It is described in [16] as the information of one parameter value with respect to another. It is shown in [12] that it is naturally related to the notion of information in information theory. This quantity has a different form for each of the Markov models. It is denoted K^i, $i=1,2,3$, for the three models.

$$K_u^1(\alpha,\beta)=\int_E\int_E\ln\left(\frac{p(x,y,u(x),\beta)}{p(x,y,u(x),\alpha)}\right)p(x,y,u(x),\beta)\varphi(\mathrm{d}y)\pi_u^\beta(\mathrm{d}x), \tag{44}$$

where φ is given in (1), π is the invariant measure given in (23), $\alpha,\beta\in\mathcal{A}$, and $u\in\mathcal{U}$.

$$\begin{aligned}K_u^2(\alpha,\beta)&\\ &=\frac{1}{2}\int_{\Gamma_1}E_x^{\beta,u}\left[\int_0^\tau|\sigma^{-1}(X(s))(h(X(s),\alpha,u(X(s)))-h(X(s),\beta,u(X(s))))|^2\mathrm{d}s\right]\\ &\qquad\cdot\eta_u^\beta(\mathrm{d}x)\left(\int_{\Gamma_1}E_x^{\beta,u}[\tau]\eta_u^\beta(\mathrm{d}x)\right)^{-1},\end{aligned} \tag{45}$$

where η is the invariant measure for the embedded Markov chain given in (27), $\alpha,\beta\in\mathcal{A}$, and $u\in\mathcal{U}$.

$$K_u^3(\alpha,\beta,\gamma)=\int_\Gamma E_x^{\beta,u(\gamma)}\left[\sum_{i=0}^{\tau-1}\ln\left(\frac{F(\Pi_i^\beta,u(\Pi_i^\gamma),\beta)(Y_{i+1})}{F(\Pi_i^\alpha,u(\Pi_i^\gamma),\alpha)(Y_{i+1})}\right)\right]$$

$$\cdot\eta^{\beta}_{u(\gamma)}(\mathrm{d}x)\left(\int_{\Gamma} E_x^{\beta,u(\gamma)}[\tau]\eta^{\beta}_{u(\gamma)}(\mathrm{d}x)\right)^{-1}, \tag{46}$$

where η is given in (40), $\alpha,\beta,\gamma\in\mathcal{A}$, $u\in\tilde{\mathcal{U}}$, $u(\gamma)$ in (46) indicates that the control $u(\Pi_i^{\gamma})$ is used, and

$$\begin{aligned} F(\nu,v,\alpha)(y) &:= 1_{\Gamma}(y)p(\nu,y,v,\alpha) \\ &\quad + 1_{E\setminus\Gamma}(y)\int_{E\setminus\Gamma} r(z,y)p(\nu,z,v,\alpha)\mathrm{d}z. \end{aligned} \tag{47}$$

Now some important properties are verified for K^i, $i=1,2,3$.

PROPOSITION 4. *Consider Model* I *with the assumptions imposed on it. For each* $u\in\mathcal{U}$ *the map* $K^1_u:\mathcal{A}\times\mathcal{A}\to\mathbb{R}$ *is continuous. Furthermore, if* $K^1_u(\alpha,\beta)=0$, *then* $\pi^{\alpha}_u=\pi^{\beta}_u$.

Proof. Let $L:\mathcal{A}\times\mathcal{A}\times E\to\mathbb{R}$ be given by

$$L(\alpha,\beta,x)=\int_E \ln\left(\frac{p(x,y,u(x),\beta)}{p(x,y,u(x),\alpha)}\right)p(x,y,u(x),\beta)\varphi(\mathrm{d}y).$$

$L(\cdot,\cdot,x)$ is continuous and bounded uniformly in $x\in E$, so the continuity of K^1_u follows by (24).

For $x\in E$, Jensen's inequality implies that

$$\int_E \ln\left(\frac{p(x,y,u(x),\beta)}{p(x,y,u(x),\alpha)}\right)p(x,y,u(x),\beta)\varphi(\mathrm{d}y)\geq 0.$$

For each $B\in\mathcal{B}(E)$ it follows by the definition of invariant measures that

$$\pi^{\beta}_u(B)\geq \inf_{x,y\in E}\inf_{v\in U,\ \beta\in\mathcal{A}} p(x,y,v,\beta)\varphi(B).$$

If $K^1_u(\alpha,\beta)=0$, then

$$\int_E \ln\left(\frac{p(x,y,u(x),\beta)}{p(x,y,u(x),\alpha)}\right)p(x,y,u(x),\beta)\varphi(\mathrm{d}y)=0 \tag{48}$$

for (φ) almost all $x\in E$. Since $\ln(\cdot)$ is a strongly convex function it follows by Jensen's inequality that

$$p(x,y,u(x),\alpha)=p(x,y,u(x),\beta) \tag{49}$$

for (φ) almost all $x\in E$ and (φ) almost all $y\in E$. Thus for $B\in\mathcal{B}(E)$,

$$\begin{aligned} \pi^{\beta}_u(B) &= \int_E\int_B p(x,y,u(x),\beta)\varphi(\mathrm{d}y)\pi^{\beta}_u(\mathrm{d}x) \\ &= \int_E\int_B p(x,y,u(x),\alpha)\varphi(\mathrm{d}y)\pi^{\beta}_u(\mathrm{d}x) \\ &= \int_E P^{u,\alpha}(x,B)\pi^{\beta}_u(\mathrm{d}x). \end{aligned}$$

The last equality implies that π^{β}_u is an invariant measure for the transition operator $P^{\alpha,u}$. The uniqueness of the invariant measure for $P^{\alpha,u}$, which follows from (23), implies that $\pi^{\beta}_u=\pi^{\alpha}_u$. □

Now a result analogous to the above proposition is obtained for Model II.

PROPOSITION 5. *Consider Model* II *with the assumptions imposed on it. For each* $u \in \mathcal{U}$*, the map* $K_u^2 : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$ *is continuous. Furthermore, if* $K_u^2(\alpha, \beta) = 0$*, then* $\pi_u^\alpha = \pi_u^\beta$.

Proof. The continuity of K_u^2 follows from (29), (30), (31) and the continuity of $h(x, \cdot, v)$. If $K_u^2(\alpha, \beta) = 0$, then by Lemma 3.4 of [6] it follows that

$$h(x, \alpha, u(x)) = h(x, \beta, u(x))$$

for all $x \in \mathbb{R}^n \setminus D$ where $\lambda(D) = 0$ and λ is an n-dimensional Lebesgue measure. Thus for (λ) almost all $x \in \mathbb{R}^n$, $t > 0$, and $B \in \mathcal{B}(\mathbb{R}^n)$,

$$P_x^{\alpha,u}(X(t) \in B) = P_x^{\beta,u}(X(t) \in B).$$

The uniqueness of the invariant measures, as in the proof of Proposition 4, implies that $\pi_u^\alpha = \pi_u^\beta$. □

For Model III an additional assumption is introduced:

$$\int_{E \setminus \Gamma} r(x, y)(p(\nu_1, x, v, \alpha) - p(\nu_2, x, v, \beta))\mathrm{d}x = 0 \tag{A5}$$

for almost all $y \in E \setminus \Gamma$ if and only if

$$p(\nu_1, x, v, \alpha) = p(\nu_2, x, v, \beta)$$

for almost all $x \in E \setminus \Gamma$.

Now an analogue of the previous two propositions is verified for Model III.

PROPOSITION 6. *If* (A1)–(A5) *are satisfied for Model* III*, then for each* $u \in \tilde{\mathcal{U}}$ *the map* $K_u^3 : \mathcal{A} \times \mathcal{A} \times \mathcal{A} \to \mathbb{R}$ *is continuous. Furthermore, if* $K_u^3(\alpha, \beta, \gamma) = 0$ *then the measures* $\Psi_{u(\gamma)}^\alpha$ *and* $\Psi_{u(\gamma)}^\beta$ *on the Borel* σ*-algebra of* $\mathcal{P}(E) \times \mathcal{P}(E)$ *coincide, where*

$$\Psi_{u(\gamma)}^\delta(B) = \frac{\displaystyle\int_\Gamma E_x^{\delta,u(\gamma)} \left[\sum_{i=0}^{\tau-1} 1_B(\Pi_i^\alpha, \Pi_i^\beta)\right] \eta_{u(\gamma)}^\delta(\mathrm{d}x)}{\displaystyle\int_\Gamma E_x^{\delta,u(\gamma)}[\tau]\eta_{u(\gamma)}^\delta(\mathrm{d}x)} \tag{50}$$

and $\delta = \alpha, \beta$ *and* $B \in \sigma(\mathcal{P}(E) \times \mathcal{P}(E))$.

Proof. The verification of the continuity of K_u^3 follows from the boundedness and continuity of $F(\cdot, \cdot, \cdot)(\cdot) : \mathcal{P}(E) \times U \times \mathcal{A} \times E \setminus \partial\Gamma \to \mathbb{R}$ (cf. Theorems 1 and 6 of [18]).

If $K_u^3(\alpha, \beta, \gamma) = 0$, then by the strict positivity of p it follows that

$$E_x^{\beta,u(\gamma)} \left[\sum_{i=0}^{\tau-1} \ln\left(\frac{F(\Pi_i^\beta, u(\Pi_i^\gamma), \beta)(Y_{i+1})}{F(\Pi_i^\alpha, u(\Pi_i^\gamma), \alpha)(Y_{i+1})}\right)\right] = 0$$

for almost all $x \in \Gamma$. The map $L : \Gamma \to \mathbb{R}$, where

$$L(x) = E_x^{\beta,u(\gamma)} \left[\sum_{i=0}^{\tau-1} \ln\left(\frac{F(\Pi_i^\beta, u(\Pi_i^\gamma), \beta)(Y_{i+1})}{F(\Pi_i^\alpha, u(\Pi_i^\gamma), \alpha)(Y_{i+1})}\right)\right]$$

is continuous for each $\beta \in \mathcal{A}$, and $u \in U$ (cf. Lemma 8 of [18]), so for all $x \in \Gamma$, $L(x) = 0$. Using (A5) and Lemma 4 of [8] it follows that

$$P_x^{\beta,u(\gamma)}(\Pi_i^\alpha = \Pi_i^\beta \text{ for } i \in \{0,\dots,\tau-1\}) = 1$$

for each $x \in \Gamma$. It follows from Corollary 2 of [8] that

$$E_x^{\beta,u(\gamma)}\left[\sum_{i=0}^{\tau-1} 1_B(\Pi_i^\beta, \Pi_i^\gamma)\right] = E_x^{\alpha,u(\gamma)}\left[\sum_{i=0}^{\tau-1} 1_B(\Pi_i^\beta, \Pi_i^\gamma)\right]$$

for each B in the Borel σ-algebra of $\mathcal{P}(E) \times \mathcal{P}(E)$, $x \in \Gamma_1$, and $\eta_{u(\gamma)}^\alpha = \eta_{u(\gamma)}^\beta$. Thus $\Psi_{u(\gamma)}^\alpha = \Psi_{u(\gamma)}^\beta$. □

4. Almost self-optimal adaptive strategies. For $\varepsilon > 0$ fixed, the controls are restricted to the finite families $\mathcal{U}^1(\varepsilon)$, $\mathcal{U}^2(\varepsilon)$, and $\tilde{\mathcal{U}}(\varepsilon)$ of ε optimal controls for Models I, II, and III, respectively. For $\varepsilon > 0$ there is a $\delta_0 > 0$ given in Proposition 3 and a δ_0 net of $\mathcal{A}$ with a distinguished point from each element of the net $\tilde{\mathcal{A}}(\delta_0) = \{\alpha(1),\dots,\alpha(k(\delta_0))\}$.

For a randomization of an adaptive control the following subsets of $\mathbb{R}$ are used. For $\varepsilon > 0$ let

$$\begin{aligned} S(\varepsilon) = \{&\beta^i(j) : i \in \mathbb{N},\ j \in \{1,2,\dots,r(\varepsilon)\} \text{ and for each } i \in \mathbb{N} \\ &\text{there is a } j_i \in \{1,\dots,r(\varepsilon)\} \text{ such that } \beta^i(j_i) = 1 - \varepsilon/\|c\| \\ &\text{and for } j \neq j_i,\ \beta^i(j) = \varepsilon/[(r(\varepsilon)-1)\|c\|]\} \end{aligned} \tag{51}$$

and

$$\begin{aligned} \tilde{S}(\varepsilon,\delta_0) = \{&\beta^i(j,k) : i \in \mathbb{N},\ j \in \{1,\dots,r(\varepsilon)\},\ k \in \{1,\dots,k(\delta_0)\}, \\ &\text{and for each } i \in \mathbb{N} \text{ there are } j_i \text{ and } k_i \text{ such that } \beta^i(j_i,k_i) = 1 - \varepsilon/\|c\| \\ &\text{and for } j \neq j_i \text{ or } k \neq k_i,\ \beta^i(j,k) = \varepsilon/[(r(\varepsilon)k(\delta_0)-1)\|c\|]\}, \end{aligned} \tag{52}$$

where $\|\cdot\|$ is the supremum norm.

The following result is a continuity property of the invariant measures for the three models and is naturally associated with Propositions 4, 5, and 6.

PROPOSITION 7. i) *Consider Model* I *with the assumptions imposed on it. For* $\varepsilon' > 0$ *there is a* $\delta > 0$ *such that if* $\{\tilde{\beta}^i(j)\} \subset S(\varepsilon)$, $\alpha, \beta \in \mathcal{A}$, $u^j \in \mathcal{U}^1(\varepsilon)$, *and*

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \left| \sum_{j=1}^{r(\varepsilon)} \tilde{\beta}^i(j) K_{u^j}^1(\alpha,\beta) \right| < \delta,$$

then

$$\sup_{u \in \mathcal{U}^1(\varepsilon)} \|\pi_u^\alpha - \pi_u^\beta\|_{\text{var}} < \varepsilon'.$$

ii) *Consider Model* II *with the assumptions imposed on it. For* $\varepsilon' > 0$ *there is a* $\delta > 0$ *such that if* $\{\tilde{\beta}^i(j)\} \subset S(\varepsilon)$, $\alpha, \beta \in \mathcal{A}$, $u^j \in \mathcal{U}^2(\varepsilon)$, *and*

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \left| \sum_{j=1}^{r(\varepsilon)} \tilde{\beta}^i(j) K_{u^j}^2(\alpha,\beta) \right| < \delta,$$

then

$$\sup_{u\in\mathcal{U}^2(\varepsilon)} \|\pi_u^\alpha - \pi_u^\beta\|_{\text{var}} < \varepsilon'.$$

iii) *Consider Model* III *with the assumptions* (A1)–(A5). *For* $\varepsilon' > 0$ *there is a* $\delta > 0$ *such that if* $\{\tilde{\beta}^i(j,k)\} \subset \tilde{S}(\varepsilon,\delta_0)$, $\alpha,\beta \in \mathcal{A}$, $u^j \in \tilde{\mathcal{U}}(\varepsilon)$, *and*

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \left| \sum_{j=1}^{r(\varepsilon)} \sum_{k=1}^{k(\delta_0)} \tilde{\beta}^i(j,k) K_{u^j}^3(\alpha,\beta,\alpha(k)) \right| < \delta,$$

then

$$\sup_{u\in\tilde{\mathcal{U}}(\varepsilon)} \sup_{\gamma\in\tilde{\mathcal{A}}(\delta_0)} \left| \int_{\mathcal{P}(E)\times\mathcal{P}(E)} \int_E c(z,u(\nu_2))\nu_1(\mathrm{d}z)\Psi_{u(\gamma)}^\alpha(\mathrm{d}\nu_1,\mathrm{d}\nu_2) \right.$$

$$\left. - \int_{\mathcal{P}(E)\times\mathcal{P}(E)} \int_E c(z,u(\nu_2))\nu_1(\mathrm{d}z)\Psi_{u(\gamma)}^\beta(\mathrm{d}\nu_1,\mathrm{d}\nu_2) \right| < \varepsilon'.$$

Proof. Only the verifications of i) and iii) are given because the verification of ii) is similar to that of i).

Verifying by contradiction, assume that i) is not true. Then there are sequences $(\alpha_m,\, m\in\mathbb{N})$ and $(\beta_m,\, m\in\mathbb{N})$ and $\{\tilde{\beta}^{i_m}(j)\} \subset S(\varepsilon)$ such that $\alpha_m \to \alpha$, $\beta_m \to \beta$ as $m\to\infty$,

$$\limsup_{m\to\infty} \liminf_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \left| \sum_{j=1}^{r(\varepsilon)} \tilde{\beta}^{i_m}(j) K_{u^j}^1(\alpha_m,\beta_m) \right| = 0, \tag{53}$$

and

$$\sup_{u\in\mathcal{U}^1(\varepsilon)} \|\pi_u^{\alpha_m} - \pi_u^{\beta_m}\|_{\text{var}} > \varepsilon'. \tag{54}$$

By (53) and the definition of $S(\varepsilon)$, it follows that

$$\lim_{m\to\infty} K_{u^j}^1(\alpha_m,\beta_m) = 0$$

for each $j\in\{1,\dots,r(\varepsilon)\}$. Thus, by Proposition 4, $K_{u^j}^1(\alpha,\beta) = 0$ for $j\in\{1,\dots,r(\varepsilon)\}$ and $\pi_{u^j}^\alpha = \pi_{u^j}^\beta$ for $j\in\{1,\dots,r(\varepsilon)\}$. By (24) it follows that

$$\lim_{m\to\infty} \sup_{u\in\mathcal{U}^1(\varepsilon)} \|\pi_u^{\alpha_m} - \pi_u^\alpha\|_{\text{var}} = 0$$

and

$$\lim_{m\to\infty} \sup_{u\in\mathcal{U}^1(\varepsilon)} \|\pi_u^{\beta_m} - \pi_u^\beta\|_{\text{var}} = 0.$$

These last two equalities contradict (54). This contradiction verifies i).

Now assume that iii) is not satisfied. Then there are sequences $(\alpha_m,\, m \in \mathbb{N})$ and $(\beta_m,\, m \in \mathbb{N})$ and $\{\tilde{\beta}^{i_m}(j,k)\} \subset \tilde{S}(\varepsilon, \delta_0)$ such that $\alpha_m \to \alpha$, $\beta_m \to \beta$ as $m \to \infty$,

$$\limsup_{m\to\infty} \liminf_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \left| \sum_{j=1}^{r(\varepsilon)} \sum_{k=1}^{k(\delta_0)} \tilde{\beta}^{i_m}(j,k) K^3_{u^j}(\alpha_m, \beta_m, \alpha(k)) \right| = 0 \tag{55}$$

and

$$\sup_{u\in\tilde{\mathcal{U}}(\varepsilon)} \left| \int_{\mathcal{P}(E)\times\mathcal{P}(E)} \int_E c(z, u(\nu_2)) \nu_1(\mathrm{d}z) \Psi^{\alpha_m}_{u(\gamma)}(\mathrm{d}\nu_1, \mathrm{d}\nu_2) \right.$$

$$\left. - \int_{\mathcal{P}(E)\times\mathcal{P}(E)} \int_E c(z, u(\nu_2)) \nu_1(\mathrm{d}z) \Psi^{\alpha}_{u(\gamma)}(\mathrm{d}\nu_1, \mathrm{d}\nu_2) \right| \geq \varepsilon'. \tag{56}$$

By (55) and the definition of $\tilde{S}(\varepsilon, \delta_0)$, it follows that

$$\lim_{m\to\infty} K^3_{u^j}(\alpha_m, \beta_m, \alpha(k)) = 0$$

for $j \in \{1, \ldots, r(\varepsilon)\}$ and $k \in \{1, \ldots, k(\delta_0)\}$. Thus, by Proposition 6, $K^3_{u^j}(\alpha, \beta, \alpha(k)) = 0$ for $j \in \{1, \ldots, r(\varepsilon)\}$, $k \in \{1, \ldots, k(\delta_0)\}$, and

$$\Psi^{\alpha}_{u(\gamma)} = \Psi^{\beta}_{u(\gamma)}$$

for $u \in \tilde{\mathcal{U}}(\varepsilon)$ and $\gamma \in \tilde{\mathcal{A}}(\delta_0)$. In the proof of Theorem 6 of [18] it is shown that

$$\lim_{m\to\infty} \Psi^{\alpha_m}_{u(\gamma)} = \Psi^{\alpha}_{u(\gamma)}$$

and

$$\lim_{m\to\infty} \Psi^{\beta_m}_{u(\gamma)} = \Psi^{\beta}_{u(\gamma)}$$

in the weak* topology of $\mathcal{P}(E) \times \mathcal{P}(E)$. By the continuity of c in the cost functional (11) there is a contradiction to (56). This contradiction verifies iii). □

Fix $\varepsilon > 0$. For Models I and II let $\varepsilon' = \varepsilon/\|c\|$, and for Model III let $\varepsilon' = \varepsilon$. Using this ε' by Proposition 7, there is a $\delta > 0$ such that i), ii), and iii) are satisfied for Models I, II, and III, respectively. There is a $\bar{\delta} > 0$ such that the following are satisfied.

i) For Model I and $\alpha, \beta \in \mathcal{A}$, if $\rho_{\mathcal{A}}(\alpha, \beta) < \bar{\delta}$, then for each $u \in \mathcal{U}^1(\varepsilon)$

$$|K^1_u(\alpha, \beta)| < \delta/3 \tag{57}$$

and

$$\|\pi^{\alpha}_u - \pi^{\beta}_u\|_{\mathrm{var}} \leq \frac{\varepsilon}{\|c\|}. \tag{58}$$

ii) For Model II and $\alpha, \beta \in \mathcal{A}$, if $\rho_{\mathcal{A}}(\alpha, \beta) < \bar{\delta}$, then for each $u \in \mathcal{U}^2(\varepsilon)$

$$|K^2_u(\alpha, \beta)| < \delta/3 \tag{59}$$

and

$$\|\pi_u^\alpha - \pi_u^\beta\|_{\text{var}} \le \frac{\varepsilon}{\|c\|}. \tag{60}$$

iii) For Model III and $\alpha, \beta \in \mathcal{A}$, if $\rho_{\mathcal{A}}(\alpha,\beta) < \bar{\delta}$, then for each $u \in \tilde{\mathcal{U}}(\varepsilon)$ and $\gamma \in \tilde{\mathcal{A}}(\delta_0)$

$$|K_u^3(\alpha,\beta,\gamma)| < \delta/3 \tag{61}$$

and

$$\left| \int_{\mathcal{P}(E)\times\mathcal{P}(E)} \int_E c(z,u(\nu_2))\nu_1(\mathrm{d}z)\Psi_{u(\gamma)}^{\alpha_m}(\mathrm{d}\nu_1,\mathrm{d}\nu_2) \right.$$

$$\left. - \int_{\mathcal{P}(E)\times\mathcal{P}(E)} \int_E c(z,u(\nu_2))\nu_1(\mathrm{d}z)\Psi_{u(\gamma)}^{\alpha}(\mathrm{d}\nu_1,\mathrm{d}\nu_2) \right| \le \varepsilon. \tag{62}$$

The existence of $\bar{\delta} > 0$ follows in i) from Proposition 4 and (24), in ii) from Proposition 5 and (27), (29), (30), (31), and in iii) from Proposition 6 and the continuity of the map $\Psi : \mathcal{A}\times\mathcal{A} \to \mathcal{P}(E)\times\mathcal{P}(E)$, which follows from the proof of Theorem 6 of [18].

For $\bar{\delta} > 0$ there is a finite covering of $\mathcal{A}$ by balls of radius $\bar{\delta}$ with centers at distinguished points that is denoted $\mathcal{A}(\bar{\delta})$. For $\varepsilon > 0$ and $\delta > 0$ given above, there is a positive integer N whose existence is justified subsequently such that

i) for Model I

$$\sup_{x\in E}\sup_{u\in\mathcal{U}^1(\varepsilon)}\sup_{\alpha\in\mathcal{A}(\bar{\delta})}\sup_{\beta\in\mathcal{A}} \left| \frac{1}{N}\sum_{i=1}^{N-1} E_x^{\beta,u}\left[\ln\left(\frac{p(X_i,X_{i+1},u(X_i),\beta)}{p(X_i,x_{i+1},u(X_i),\alpha)}\right)\right] - K_u^1(\alpha,\beta)\right| < \delta/3 \tag{63}$$

and

$$\sup_{x\in E}\sup_{u\in\mathcal{U}^1(\varepsilon)}\sup_{\beta\in\mathcal{A}} \left| \frac{1}{N}\sum_{i=1}^{N-1} E_x^{\beta,u}[c(X_i,u(X_i))] - \int_E c(z,u(z))\pi_u^\beta(\mathrm{d}z)\right| < \varepsilon; \tag{64}$$

ii) for Model II

$$\sup_{x\in\Gamma}\sup_{u\in\mathcal{U}^2(\varepsilon)}\sup_{\alpha\in\mathcal{A}(\bar{\delta})}\sup_{\beta\in\mathcal{A}} \left| E_x^{\beta,u}\left[\frac{1}{2}\int_0^{\tau_N} |\sigma^{-1}(X(s))(h(X(s),\alpha,u(X(s)))\right.\right.$$

$$\left.\left. - h(X(s),\beta,u(X(s))))|^2\mathrm{d}s\right](E_x^{\beta,u}[\tau_N])^{-1} - K_u^2(\alpha,\beta)\right| < \delta/3 \tag{65}$$

and

$$\sup_{x\in\Gamma_1}\sup_{u\in\mathcal{U}^2(\varepsilon)}\sup_{\beta\in A} \left| E_x^{\beta,u}\left[\int_0^{\tau_N} c(X(s),u(X(s)))]\mathrm{d}s\right.\right.$$

$$\left.\left. \cdot(E_x^{\beta,u}[\tau_N])^{-1} - \int_E c(z,u(z))\pi_u^\beta(\mathrm{d}z)\right| < \varepsilon; \tag{66}$$

iii) for Model III

$$\sup_{x\in\Gamma}\sup_{u\in\tilde{\mathcal{U}}(\varepsilon)}\sup_{\alpha\in\mathcal{A}(\bar{\delta})}\sup_{\gamma\in\tilde{\mathcal{A}}(\delta_0)}\sup_{\beta\in\mathcal{A}}\left|E_x^{\beta,u(\gamma)}\left[\sum_{i=0}^{\tau_N-1}\ln\left(\frac{F(\Pi_i^{\beta},u(\Pi_i^{\gamma}),\beta)(Y_{i+1})}{F(\Pi_i^{\alpha},u(\Pi_i^{\gamma}),\alpha)(Y_{i+1})}\right)\right]\right.$$

$$\left.\cdot(E_x^{\beta,u}[\tau_N])^{-1}-K_u^3(\alpha,\beta,\gamma)\right|<\delta/3 \tag{67}$$

and

$$\sup_{x\in\Gamma}\sup_{u\in\tilde{\mathcal{U}}(\varepsilon)}\sup_{\gamma\in\tilde{\mathcal{A}}(\delta_0)}\sup_{\beta\in\mathcal{A}}\left|E_x^{\beta,u(\gamma)}\left[\sum_{i=0}^{\tau_N-1}\int_E c(z,u(\Pi_i^{\alpha}))\Pi_i^{\beta}(\mathrm{d}z)\right](E_x^{\beta,U(\gamma)}[\tau_N])^{-1}\right.$$

$$\left.-\int_{\mathcal{P}(E)\times\mathcal{P}(E)}\int_E c(z,u(\nu_2))\nu_1(\mathrm{d}z)\Psi_{u(\gamma)}^{\beta}(\mathrm{d}\nu_1,\mathrm{d}\nu_2)\right|<\varepsilon, \tag{68}$$

where Ψ is given in (50).

The following three lemmas justify the existence of N in (63)–(68).

LEMMA 1. *For Model* I *let*

$$L=\sup_{x,y\in E}\sup_{v\in U}\sup_{\alpha,\beta\in\mathcal{A}}\left|\ln\frac{p(x,y,v,\beta)}{p(x,y,v,\alpha)}\right|.$$

Then for $N\geq 3L/(\delta(1-\gamma_0))$ *the inequality* (63) *is satisfied, and for* $N\geq\|c\|/(\varepsilon(1-\gamma_0))$ *the inequality* (64) *is satisfied where* γ_0 *is as given in Proposition* 1.

Proof. By (23) it follows that

$$\left|E_x^{\beta,u}\left[\ln\frac{p(X_i,X_{i+1},u(X_i),\beta)}{p(X_i,X_{i+1},u(X_i),\alpha)}\right]-K_u^1(\alpha,\beta)\right|$$

$$=\left|E_x^{\beta,u}\left[\int_E\ln\left(\frac{p(X_i,y,u(X_i),\beta)}{p(X_i,y,u(X_i),\alpha)}\right)p(X_i,y,u(X_i),\beta)\varphi(\mathrm{d}y)\right]-K_u^1(\alpha,\beta)\right|\leq L\gamma_0^{i-1}$$

for $x\in E$, $\alpha,\beta\in\mathcal{A}$, $u\in\mathcal{U}^1(\varepsilon)$, and $i\in\mathbb{N}$. Thus for $N\geq 3L/\delta(1-\gamma_0)$, the inequality (63) is satisfied. In a similar way by (23) it follows that the inequality (64) is satisfied for $N\geq\|c\|/\varepsilon(1-\gamma_0)$. □

LEMMA 2. *For Model* II *let*

$$L=\sup_{x\in\mathbb{R}^n}\sup_{\alpha\in\mathcal{A}}\sup_{v\in U}\|h(x,\alpha,v)^*\sigma^{-1}(x)\|,$$

$$M_1=\sup_{x\in\Gamma_1}\sup_{\alpha\in\mathcal{A}}\sup_{u\in\mathcal{U}}E_x^{\alpha,u}[\tau],$$

and

$$M_2=\inf_{x\in\Gamma_1}\inf_{\alpha\in\mathcal{A}}\inf_{u\in\mathcal{U}}E_x^{\alpha,u}[\tau].$$

For

$$N\geq\frac{12}{\delta}L^2M_1\frac{1}{M_2(1-\gamma_0)}\left(1+\frac{M_1}{M_2}\right),$$

the inequality (65) *is satisfied, and for*

$$N \geq \frac{1}{\varepsilon}\|c\| M_1 \frac{1}{M_2(1-\gamma_0)}\left(1+\frac{M_1}{M_2}\right),$$

where γ_0 *is given in Proposition* 2, *the inequality* (66) *is satisfied.*

Proof. By (28) it follows that

$$\left|\frac{1}{i}E_x^{\beta,u}\left[\int_0^{\tau_{i+1}} |\sigma^{-1}(X(s))(h(X(s),\alpha,u(X(s))) - h(X(s),\beta,u(X(s))))|^2 \mathrm{d}s\right]\right.$$

$$\left. - \int_{\Gamma_1} E_x^{\beta,u}\left[\int_0^{\tau} |\sigma^{-1}(X(s))(h(X(s),\alpha,u(X(s)) - h(X(s),\beta,u(X(s)))|^2 \mathrm{d}s\right]\eta_u^{\beta}(\mathrm{d}z)\right|$$

$$\leq \frac{1}{i}\sum_{j=0}^{i}\gamma_0^i 4L^2 M_1 = \frac{1}{i}\frac{1-\gamma_0^{i+1}}{1-\gamma_0}4L^2M_1$$

and

$$\left|\frac{1}{i}E_x^{\beta,u}[\tau_{i+1}] - \int_{\Gamma_1} E_z^{\beta,u}[\tau]\eta_u^{\beta}(\mathrm{d}z)\right| \leq \frac{1}{i}\frac{1-\gamma_0^{i+1}}{1-\gamma_0}M_1$$

for $x \in \Gamma_1$, $\beta \in \mathcal{A}$, and $u \in \mathcal{U}$.

Combining the above two inequalities, (65) is satisfied for N stated in the lemma.

In a similar way, (66) is verified. □

LEMMA 3. *For Model* III *let*

$$L_1 = \sup_{x,y\in E}\sup_{v\in U}\sup_{\alpha\in\mathcal{A}} p(x,y,v,\alpha),$$

$$L_2 = \inf_{x,y\in E}\inf_{v\in U}\inf_{\alpha\in\mathcal{A}} p(x,y,v,\alpha),$$

$$M_1 = \sup_{x\in\Gamma}\sup_{u\in\tilde{\mathcal{U}}(\varepsilon)}\sup_{\gamma\in\tilde{\mathcal{A}}(\delta_0)}\sup_{\beta\in\mathcal{A}} E_x^{\beta,u(\gamma)}[\tau],$$

and

$$M_2 = \inf_{x\in\Gamma}\inf_{u\in\tilde{\mathcal{U}}(\varepsilon)}\inf_{\gamma\in\tilde{\mathcal{A}}(\delta_0)}\inf_{\beta\in\mathcal{A}} E_x^{\beta,u(\gamma)}[\tau].$$

For

$$N \geq \frac{3}{\delta}\ln\left(\frac{L_1}{L_2}\right) M_1 \frac{1}{M_2(1-\gamma_0)}\left(1+\frac{M_1}{M_2}\right),$$

the inequality (67) *is satisfied, and for*

$$N \geq \frac{\|c\|}{\varepsilon} M_1 \frac{1}{(1-\gamma_0)M_2}\left(1+\frac{M_1}{M_2}\right),$$

the inequality (68) *is satisfied where* γ_0 *is given in Proposition* 3.

Proof. As in the verification of Lemma 2 it follows by (40) that

$$\left|\frac{1}{i}E_x^{\beta,u(\gamma)}\left[\sum_{j=0}^{\tau_{i+1}-1}\ln\frac{F(\Pi_j^\beta,u(\Pi_j^\gamma),\beta)(Y_{j+1})}{F(\Pi_j^\alpha,u(\Pi_j^\gamma),\alpha)(Y_{j+1})}\right]\right.$$

$$\left.-\int_\Gamma E_x^{\beta,u(\gamma)}\left[\sum_{j=0}^{\tau-1}\ln\frac{F(\Pi_j^\beta,u(\Pi_j^\gamma),\beta)(Y_{j+1})}{F(\Pi_j^\alpha,u(\Pi_j^\gamma),\alpha)(Y_{j+1})}\right]\cdot\eta_{u(\gamma)}^\beta(\mathrm{d}x)\right|\le\frac{1}{i}\frac{1-\gamma_0^{i+1}}{1-\gamma_0}\ln\left(\frac{L_1}{L_2}\right)M_1$$

and

$$\left|\frac{1}{i}E_x^{\beta,u(\gamma)}[\tau_{i+1}]-\int_\Gamma E_z^{\beta,u(\gamma)}[\tau]\eta_{u(\gamma)}^\beta(\mathrm{d}z)\right|\le\frac{1}{i}\frac{1-\gamma_0^{i+1}}{1-\gamma_0}M_1$$

for $x\in\Gamma$, $\beta\in\mathcal{A}$, $\gamma\in\tilde{\mathcal{A}}(\delta_0)$, $u\in\tilde{\mathcal{U}}(\varepsilon)$. These inequalities imply the inequalities for N for which (67) and (68) are satisfied. □

Now the construction of the almost self-optimal controls can be completed. Again, it is subdivided into the three models.

i) For Model I let $\hat{\alpha}_{jN}$ be a maximizer of

$$L_{jN}^1(\alpha)=\sum_{i=0}^{jN-1}\ln p(X_i,X_{i+1},v_i,\alpha) \tag{69}$$

over $\alpha\in\mathcal{A}(\bar{\delta})$, where v_i is the control at time i. The control v_i is a randomized certainty equivalence control. For $i\in\{jN:j\in\mathbb{N}\}$, choose the control $u_{jN}\in\mathcal{U}^1(\varepsilon)$ randomly among $(u^k,\,k=1,\dots,r(\varepsilon))$ as

$$P(u_{jN}=u^{k_0}|X(0),\dots,X(jN))=1-\frac{\varepsilon}{\|c\|}, \tag{70}$$

where u^{k_0} is the almost optimal control for $\alpha=\hat{\alpha}_{jN}$ in $\mathcal{U}^1(\varepsilon)$ and

$$P(u_{jN}=u^k|X(0),\dots,X(jN))=\frac{\varepsilon}{(r(\varepsilon)-1)\|c\|} \tag{71}$$

for $k=\{1,\dots,r(\varepsilon)\}\setminus\{k_0\}$. The control u_{jN} is also used at the times $jN+1,\dots,(j+1)N-1$; that is,

$$v_i=u_{jN}(X_i) \tag{72}$$

for $i=\{jN,\dots,(j+1)N-1\}$.

ii) For Model II, let $\hat{\alpha}(\tau_{jN})$ be a minimizer of

$$L^2(\tau_{jN})=\int_0^{\tau_{jN}}|\sigma^{-1}(X(s))(h(X(s),\alpha,v(s))-h(X(s),\alpha^0,v(s)))|^2\mathrm{d}s \tag{73}$$

over $\alpha\in\mathcal{A}(\delta)$. The control in $[\tau_{jN},\,\tau_{(j+1)N})$ is $u(\tau_{jN})\in U^2(\varepsilon)$, that is, a randomized certainty equivalence control such that

$$P(u(\tau_{jN})=u^{k_0}|X(s),\,0\le s\le\tau_{jN})=1-\frac{\varepsilon}{\|c\|} \tag{74}$$

and

$$P(u(\tau_{jN}) = u^k | X(s),\, 0 \le s \le \tau_{jN}) = \frac{\varepsilon}{(r(\varepsilon)-1)\|c\|} \tag{75}$$

for $k = \{1, \dots, r(\varepsilon)\} \setminus \{k_0\}$ and u^{k_0} is almost optimal for $\hat{\alpha}(\tau_{jN})$.

iii) For Model III let $\hat{\alpha}(\tau_{jN})$ be a maximizer of

$$L^3_{\tau_{jN}}(\alpha) = \sum_{i=0}^{\tau_{jN}-1} \ln F(\Pi^{\alpha}_i, v_i, \alpha)(Y_{i+1}) \tag{76}$$

over $\alpha \in \mathcal{A}(\bar{\delta})$. The controls $\hat{v}(\tau_{jN}), \hat{v}(\tau_{jN+1}), \dots, \hat{v}(\tau_{(j+1)N-1})$ are selected by a randomized certainty equivalence rule such that

$$P(\hat{v}(\tau_{jN}) = u^{k_0}(\Pi^{\alpha(l_0)}_{\tau_{jN}}), \dots, \hat{v}(\tau_{(j+1)N-1}) = u^{k_0}(\Pi^{\alpha(l_0)}_{\tau_{(j+1)N-1}}) | \mathcal{Y}(\tau_{jN})) = 1 - \frac{\varepsilon}{\|c\|} \tag{77}$$

and

$$\begin{aligned} P(\hat{v}(\tau_{jN}) &= u^k(\Pi^{\alpha(l)}_{\tau_{jN}}), \dots, \hat{v}(\tau_{(j+1)N-1}) = u^k(\Pi^{\alpha(l)}_{\tau_{(j+1)N-1}}) | \mathcal{Y}(\tau_{jN})) \\ &= \frac{\varepsilon}{(r(\varepsilon)k(\varepsilon)-1)\|c\|}, \end{aligned} \tag{78}$$

where $k \in \{1, \dots, r(\varepsilon)\} \setminus \{k_0\}$, $j \in \{1, \dots, \mathrm{card}(\tilde{\mathcal{A}}(\delta_{k_0}))\}$, and $u^{k_0}(\Pi^{\alpha(l_0)}_{\tau_{jN}})$ is almost optimal for $\hat{\alpha}(\tau_{jN})$.

Let $(\hat{v}_i,\, i \in \mathbb{N})$ or $(\hat{v}(s),\, s \ge 0)$ be the discrete or the continuous time randomized certainty equivalence control defined in i), ii), or iii) above. Let $(\bar{\beta}_i,\, i \in \mathbb{N})$, $(\bar{\beta}(s),\, s \ge 0)$, and $(\tilde{\beta}_i,\, i \in \mathbb{N})$ be processes with values in $\{1, \dots, r(\varepsilon)\}$ for the first two processes and in $\{(j,k) : j = 1, \dots, r(\varepsilon) \text{ and } k = 1, \dots, k(\delta_0)\}$ for the third process such that the first two processes correspond to the index of the control in $\mathcal{U}^i(\varepsilon)$, $i = 1, 2$, at each time and the third process (for Model III) is the index of the control function and the index of the element of $\tilde{\mathcal{A}}(\delta_0)$.

The following result is the almost self-optimality of the randomized certainty equivalence control for the three Models I, II, and III.

THEOREM 1. *Let $\varepsilon > 0$ be fixed. Let I^i be the pathwise cost functional for $i = 1, 2, 3$ given by* (2), (8), *and* (11), *respectively, for Models* I, II, *and* III, *respectively, and let $\lambda_i(\alpha^0)$ be the optimal cost for $i = 1, 2, 3$ for parameter α^0. Let $(\hat{v}_i,\, i \in \mathbb{N})$ and $(\hat{v}(s),\, s \ge 0)$ be the randomized certainty equivalence controls given above. For the Models* I, II, *and* III *the following inequalities are satisfied:*

$$\text{(i) } I^1((\hat{v}_n,\, n \in \mathbb{N})) \le \lambda_1(\alpha^0) + 6\varepsilon \quad a.s., \tag{79}$$

$$\text{(ii) } I^2((\hat{v}(s),\, s \ge 0)) \le \lambda_2(\alpha^0) + 6\varepsilon \quad a.s., \tag{80}$$

$$\text{(iii) } I^3((\hat{v}_n,\, n \in \mathbb{N})) \le \lambda_3(\alpha^0) + 6\varepsilon \quad a.s. \tag{81}$$

Proof. Initially consider Model I. By the definition of $\hat{\alpha}_{jN}$, it follows that

$$\sum_{i=0}^{jN-1} \ln \frac{p(X_i, X_{i+1}, \hat{v}_i, \hat{\alpha}_{jN})}{p(X_i, X_{i+1}, \hat{v}_i, \alpha^0)} \ge \sum_{i=0}^{jN-1} \ln \frac{p(X_i, X_{i+1}, \hat{v}_i, \bar{\alpha}^0)}{p(X_i, X_{i+1}, \hat{v}_i, \alpha^0)}, \tag{82}$$

where $\bar{\alpha}^0$ is an element in $\mathcal{A}(\bar{\delta})$ of minimum distance to α^0.

By the law of large numbers for martingales for $\alpha \in \mathcal{A}(\bar{\delta})$,

$$\lim_{n\to\infty} \frac{1}{n}\Bigg(\sum_{i=0}^{nN-1} \ln \frac{p(X_i, X_{i+1}, \hat{v}_i, \alpha)}{p(X_i, X_{i+1}, \hat{v}_i, \alpha^0)}$$

(83)
$$- \sum_{i=0}^{n-1} E^{\alpha^0}\Bigg[\sum_{j=iN}^{(i+1)N-1} \ln \frac{p(X_i, X_{i+1}, \hat{v}_i, \alpha)}{p(X_i, X_{i+1}, \hat{v}_i, \alpha^0)} \Bigg| X_0, \ldots, X_{iN}, \bar{\beta}_0, \ldots, \bar{\beta}_{iN} \Bigg] \Bigg) = 0 \quad \text{a.s.},$$

$$\lim_{n\to\infty} \frac{1}{n}\Bigg(\sum_{i=0}^{nN-1} c(X_i, \hat{v}_i)$$

$$(84) \qquad - \sum_{i=0}^{n-1} E^{\alpha^0}\Bigg[\sum_{j=iN}^{(i+1)N-1} c(X_j, \hat{v}_j) \Bigg| X_0, \ldots, X_{iN}, \bar{\beta}_0, \ldots, \bar{\beta}_{iN} \Bigg] \Bigg) = 0 \quad \text{a.s.},$$

$$(85) \qquad \limsup_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \left(K^1_{u_{iN}}(\alpha, \alpha^0) - \sum_{j=1}^{r(\varepsilon)} \beta^i(j) K^1_{u^j}(\alpha, \alpha^0) \right) = 0 \quad \text{a.s.},$$

and

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} J^{\hat{\alpha}_{iN},1}((u_{iN}(X_l),\, l \in \mathbb{N}))$$

$$(86) \qquad - \sum_{j=1}^{r(\varepsilon)} \beta^i(j) J^{\hat{\alpha}_{iN},1}((u^j(X_l),\, l \in \mathbb{N})) = 0 \quad \text{a.s.},$$

where $\beta^i(j) = P(\bar{\beta}_i = j | X_0, \ldots, X_{iN})$, and in the last equality the control functions u_{iN} and u^j are used and their costs are evaluated. Let $\mathcal{N}$ be a null set such that the above four equalities are satisfied on $\mathcal{N}^{\mathrm{c}}$. Let $F(\bar{\delta}) = \{\alpha \in \mathcal{A}(\bar{\delta}) : \text{there is an } \omega \in \Omega \backslash \mathcal{N}$ such that α is a frequent point of $(\hat{\alpha}_{jN}(\omega))\}$. In many subsequent expressions, the random variables are evaluated at some $\omega \in \Omega \setminus \mathcal{N}$ but this evaluation is suppressed for notational convenience. If $\alpha \in F(\bar{\delta})$ then for a corresponding $\omega \in \Omega \setminus \mathcal{N}$ it follows from (82), (83) that

$$(87) \qquad \limsup_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} E^{\alpha^0, u_{iN}}_{X_{iN}} \left[\sum_{j=iN}^{(i+1)N-1} \ln \frac{p(X_j, X_{j+1}, u_{iN}(X_j), \alpha^0)}{p(X_j, X_{j+1}, u_{iN}(X_j), \alpha)} \right]$$

$$\geq \limsup_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} E^{\alpha^0, u_{iN}}_{X_{iN}} \left[\sum_{j=iN}^{(i+1)N-1} \ln \frac{p(X_j, X_{j+1}, u_{iN}(X_j), \alpha^0)}{p(X_j, X_{j+1}, u_{iN}(X_j), \bar{\alpha})} \right]$$

for each $\bar{\alpha} \in \mathcal{A}(\bar{\delta})$. By (63) it follows that

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} K^1_{u_{iN}}(\alpha, \alpha^0) + \frac{\delta}{3} \geq \limsup_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} K^1_{u_{iN}}(\bar{\alpha}, \alpha^0) - \frac{\delta}{3}. \tag{88}$$

Thus, by (57),

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} K^1_{u_{iN}}(\alpha, \alpha^0) \geq -\delta, \tag{89}$$

and by (85),

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \sum_{j=1}^{r(\varepsilon)} \beta^i(j) K^1_{u^j}(\alpha, \alpha^0) \geq -\delta. \tag{90}$$

Therefore, by Proposition 7 with $\varepsilon^1 = \varepsilon/\|c\|$,

$$\sup_{u \in \mathcal{U}^1(\varepsilon)} \|\pi^{\alpha}_u - \pi^{\alpha^0}_u\|_{\text{var}} < \frac{\varepsilon}{\|c\|}. \tag{91}$$

By (84)

$$I^1((\hat{v}_n,\, n \in \mathbb{N})) = \limsup_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} E^{\alpha^0, u_{iN}}_{X_{iN}} \left[\sum_{j=iN}^{(i+1)N-1} c(X_j, u_{iN}(X_j)) \right] \quad \text{a.s.}, \tag{92}$$

so by (64),

$$I^1((\hat{v}_n,\, n \in \mathbb{N})) \leq \limsup_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \int_E c(z, u_{iN}(z)) \pi^{\alpha^0}_{u_{iN}}(\mathrm{d}z) + \varepsilon \quad \text{a.s.} \tag{93}$$

For $\omega \in \Omega \setminus \mathcal{N}$ it follows from (93) that

$$I^1((\hat{v}_n,\, n \in \mathbb{N})) \leq \limsup_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} 1_{F(\bar{\delta})}(\hat{\alpha}_{iN}) \int_E c(z, u_{iN}(z)) \pi^{\alpha^0}_{u_{iN}}(\mathrm{d}z) + \varepsilon. \tag{94}$$

For $\alpha \in F(\bar{\delta})$, (91) is satisfied, so for $\omega \in \Omega \setminus \mathcal{N}$

$$I^1((\hat{v}_n,\, n \in \mathbb{N})) \leq \limsup_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} 1_{F(\bar{\delta})}(\hat{\alpha}_{iN}) \int_E c(z, u_{iN}(z)) \pi^{\hat{\alpha}_{iN}}_{u_{iN}}(\mathrm{d}z) + 2\varepsilon$$

$$= \limsup_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} J^{\hat{\alpha}_{iN},1}((u_{iN}(X_l),\, l \in \mathbb{N})) + 2\varepsilon. \tag{95}$$

For $\omega \in \Omega \setminus \mathcal{N}$ it follows by (86) that

$$I^1((\hat{v}_n,\, n \in \mathbb{N})) \leq \limsup_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \sum_{j=1}^{r(\varepsilon)} \beta_i(j) J^{\hat{\alpha}_{iN},1}((u^j(X_l),\, l \in \mathbb{N})) + 2\varepsilon. \tag{96}$$

Using the definition of $\beta^i(j)$ and (25) it follows that for $\omega \in \Omega \setminus \mathcal{N}$

$$I^1((\hat{v}_n,\, n \in \mathbb{N})) \le \limsup_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \left(1 - \frac{\varepsilon}{\|c\|}\right) \lambda_1(\hat{\alpha}_{iN}) + 3\varepsilon$$

$$\le \limsup_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \lambda_1(\hat{\alpha}_{iN}) + 3\varepsilon \le \limsup_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} 1_{F(\bar{\delta})}(\hat{\alpha}_{iN}) \lambda_1(\hat{\alpha}_{iN}) + 3\varepsilon. \tag{97}$$

By (25) and (91) it follows that for $\alpha \in F(\bar{\delta})$

$$|\lambda_1(\alpha) - \lambda_1(\alpha^0)| \le 2\varepsilon + \sup_{u \in \mathcal{U}^1(\varepsilon)} \left| \int_E c(x, u(x)) (\pi_u^{\alpha}(\mathrm{d}x) - \pi_u^{\alpha^0}(\mathrm{d}x)) \right| \le 3\varepsilon. \tag{98}$$

Thus for $\omega \in \Omega \setminus \mathcal{N}$

$$I^1((\hat{v}_n,\, n \in \mathbb{N})) \le \limsup_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} 1_{F(\bar{\delta})}(\hat{\alpha}_{iN}) \lambda_1(\alpha^0) + 6\varepsilon = \lambda_1(\alpha^0) + 6\varepsilon.$$

The inequality (79) is verified.

Now consider Model III. By the construction of $\hat{\alpha}(\tau_{jN})$ it follows that

$$\sum_{i=0}^{\tau_{jN}-1} \ln \frac{F(\Pi_i^{\hat{\alpha}_{\tau_{jN}}}, \hat{v}_i, \hat{\alpha}_{\tau_{jN}})(Y_{i+1})}{F(\Pi_i^{\alpha^0}, \hat{v}_i, \alpha^0)(Y_{i+1})} \ge \sum_{i=0}^{\tau_{jN}-1} \ln \frac{F(\Pi_i^{\bar{\alpha}_i}, \hat{v}_i, \bar{\alpha}^0)(Y_{i+1})}{F(\Pi_i^{\alpha^0}, \hat{v}_i, \alpha^0)(Y_{i+1})}. \tag{99}$$

In analogy to (83)–(86), by the law of large numbers for martingales it follows that for $\alpha \in \mathcal{A}(\bar{\delta})$,

$$\lim_{n\to\infty} \frac{1}{n} \Bigg(\sum_{i=0}^{\tau_{nN}-1} \ln \frac{F(\Pi_i^{\bar{\alpha}_i}, \hat{v}_i, \alpha^0)(Y_{i+1})}{F(\Pi_i^{\alpha^0}, \hat{v}_i, \alpha)(Y_{i+1})}$$

$$- \sum_{i=0}^{\tau_{nN}-1} E^{\alpha^0} \Bigg[\sum_{j=\tau_{iN}}^{\tau_{(i+1)N}-1} \ln \frac{F(\Pi_i^{\bar{\alpha}_i}, \hat{v}_i, \alpha^0)(Y_{i+1})}{F(\Pi_i^{\alpha^0}, \hat{v}_i, \alpha)(Y_{i+1})} \Big| \mathcal{Y}(\tau_{iN}), \bar{\beta}_0, \ldots, \bar{\beta}_{\tau_{iN}} \Bigg] \Bigg) = 0 \qquad \text{a.s.}, \tag{100}$$

$$\lim_{n\to\infty} \frac{1}{n} \left(\tau_{nN} - \sum_{i=0}^{n-1} E^{\alpha^0} \left[\tau_{(i+1)N} - \tau_{iN} | \mathcal{Y}(\tau_{iN}), \bar{\beta}_0, \ldots, \bar{\beta}_{\tau_{iN}} \right] \right) = 0 \qquad \text{a.s.}, \tag{101}$$

$$\lim_{n\to\infty} \frac{1}{n} \Bigg(\sum_{i=0}^{\tau_{nN}} c(X_i, \hat{v}_i)$$

$$- \sum_{i=0}^{n-1} E^{\alpha^0} \Bigg[\sum_{j=\tau_{iN}}^{\tau_{(i+1)N}-1} \int_E c(z, \hat{v}_i) \Pi_i^{\alpha^0}(\mathrm{d}z) | \mathcal{Y}(\tau_{iN}), \bar{\beta}_0, \ldots, \bar{\beta}_{\tau_{iN}} \Bigg] \Bigg) = 0 \qquad \text{a.s.}, \tag{102}$$

(103)
$$\limsup_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1}\left(K^3_{u_{\tau_{iN}}}(\alpha,\alpha^0,\tilde{\alpha}_{\tau_{iN}})-\sum_{j=1}^{r(\varepsilon)}\sum_{k=1}^{k(\delta_0)}\beta^i(j,k)K^3_{u^j}(\alpha,\alpha^0,\alpha(k))\right)=0 \qquad \text{a.s.},$$

$$\limsup_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1}(J^{\hat{\alpha}_{\tau_{iN}},3}(u_{\tau_{iN}}(\Pi_l^{\tilde{\alpha}_{\tau_{iN}}}),\, l\in\mathbb{N}))$$

(104)
$$-\sum_{j=1}^{r(\varepsilon)}\sum_{k=1}^{k(\delta_0)}\beta^i(j,k)(J^{\hat{\alpha}_{\tau_{iN}},3}(u^j(\Pi_l^{\tilde{\alpha}_{\tau_{iN}}}),\, l\in\mathbb{N}))=0 \qquad \text{a.s.},$$

where $\tilde{\alpha}_{\tau_{iN}}$ is the element of $\tilde{\mathcal{A}}(\delta_0)$ chosen at time τ_{iN} in the construction of the control $\hat{v}_{\tau_{iN}}$,

$$\beta_i(j,k)=P(\tilde{\beta}_{\tau_{iN}}=(j,k)|\mathcal{Y}(\tau_{iN})),$$

and $J^{\alpha,3}$ is the evaluation of the average cost for $\alpha\in\mathcal{A}$. If $\alpha\in\mathcal{A}(\bar{\delta})$ for some $\omega\in\Omega\setminus\mathcal{N}$ is a frequent point of the estimation, then similar to (87)–(90), by (62), (67), (99) it follows that

$$\liminf_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1}\sum_{j=1}^{r(\varepsilon)}\sum_{k=1}^{k(\delta_0)}\beta^i(j,k)K^3_{u^j}(\alpha,\alpha^0,\alpha(k))\geq-\delta,$$

so by Proposition 7

$$\sup_{u\in\tilde{\mathcal{U}}(\varepsilon)}\sup_{\beta\in\tilde{\mathcal{A}}(\delta_0)}\left|\int_{\mathcal{P}(E)\times\mathcal{P}(E)}\int_E c(z,u(\nu_2))\nu_1(\mathrm{d}z)\Psi^{\alpha}_{u(\beta)}(\mathrm{d}\nu_1,\mathrm{d}\nu_2)\right.$$

(105)
$$\left.-\int_{\mathcal{P}(E)\times\mathcal{P}(E)}\int_E c(z,u(\nu_2))\nu_1(\mathrm{d}z)\Psi^{\alpha^0}_{u(\beta)}(\mathrm{d}\nu_1,\mathrm{d}\nu_2)\right|<\varepsilon.$$

By (68), (101), (102) it follows that

$$J^{\alpha^0,3}((\hat{v}_n,\, n\in\mathbb{N}))$$

$$\leq\limsup_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1}\int_{\mathcal{P}(E)\times\mathcal{P}(E)}\int_E c(z,u(\nu_2))\nu_1(\mathrm{d}z)\Psi^{\alpha^0}_{u_{\tau_{iN}}(\tilde{\alpha}_{\tau_{iN}})}(\mathrm{d}\nu_1,\mathrm{d}\nu_2)+\varepsilon.$$

Let $F(\bar{\delta})$ be the set given by $F(\bar{\delta})=\{\alpha\in\mathcal{A}(\bar{\delta}):$ there is an $\omega\in\Omega\setminus\mathcal{N}$ such that α is a frequent point of $\hat{\alpha}_{\tau_{iN}}(\omega)\}$. By (104), (105) it follows that

$$I^3((\hat{v}_n,\, n\in\mathbb{N}))\leq\limsup_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1}1_{F(\delta)}(\hat{\alpha}_{\tau_{iN}})$$

$$\cdot\int_{\mathcal{P}(E)\times\mathcal{P}(E)}\int_E c(z,u(\nu_2))\nu_1(\mathrm{d}z)\Psi^{\hat{\alpha}_{\tau_{iN}}}_{u_{\tau_{iN}}(\tilde{\alpha}_{\tau_{iN}})}(\mathrm{d}\nu_1,\mathrm{d}\nu_2)+2\varepsilon,$$

so by (104)

$$I^3((\hat{v}_n,\, n \in \mathbb{N})) \le \limsup_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \sum_{j=1}^{r(\varepsilon)} \sum_{k=1}^{k(\delta_0)} \beta^i(j,k)$$

$$\cdot J^{\hat{\alpha}_{\tau_{iN}},3}(u^j(\pi_l^{\alpha(k)}),\, l \in \mathbb{N}) + 2\varepsilon.$$

Similar to (97), it follows by (104) that

$$I^3((\hat{v}_n,\, n \in \mathbb{N})) \le \limsup_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} 1_{F(\bar{\delta})}(\hat{\alpha}_{\tau_{iN}}) \lambda_3(\hat{\alpha}_{\tau_{iN}}) + 3\varepsilon.$$

For $\alpha \in F(\bar{\delta})$ it follows from (41), (104) that

$$|\lambda_3(\alpha) - \lambda_3(\alpha^0)| \le 3\varepsilon$$

and

$$I^3((\hat{v}_n,\, n \in \mathbb{N})) \le \lambda_3(\alpha^0) + 6\varepsilon \quad \text{a.s.}$$

This verifies the inequality (81).

The verification of the inequality (80) for Model II is similar to the verification of (79) and (81), and is thereby omitted. □

5. Some other adaptive algorithms. The existence of a finite family of almost optimal controls that is shown in section 2 can be used in the construction of some other algorithms. Three such algorithms are i) maximum likelihood estimation with forcing, ii) cost watching with forcing, and iii) cost watching with randomization. The forcing algorithms (cf., e.g., [2]) are based on the forced use of all of the almost optimal controls successively at times $(T_n,\, n \in \mathbb{N})$ such that

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} 1_{\{T_j,\, j\in\mathbb{N}\}}(i) = 0,$$

so that the forcing does not affect the value of the cost functional. The notion of cost watching is to compare the average costs incurred for each of the almost optimal controls. For cost watching during the nonforcing times, a control that has minimal average cost is used. For cost watching with randomization the control for which the current average cost is minimal is chosen with probability almost 1, and the other controls are chosen with small probability. It seems that the algorithms with forcing should converge slowly, but they have a simple construction. The algorithm given in section 4 is more complicated and requires some continuity properties of the invariant measures with respect to the information. A comparative analysis of the above algorithms requires further study.

REFERENCES

[1] R. Agrawal, *Minimizing the learning loss in adaptive control of Markov chains under the weak accessibility condition*, J. Appl. Prob., 28 (1991), pp. 779–790.

[2] R. Agrawal and D. Teneketzis, *Certainty equivalence control with forcing: Revisited*, Systems Control Lett., 13 (1989), pp. 405–412.

[3] A. Bensoussan, *Perturbation Methods in Optimal Control*, J. Wiley, New York, 1988.
[4] V. S. Borkar, *Self-tuning control of diffusions without the identifiability condition*, J. Optim. Theory Appl., 68 (1991), pp. 117–138.
[5] J. L. Doob, *Stochastic Processes*, J. Wiley, New York, 1953.
[6] T. E. Duncan, B. Pasik-Duncan, and L. Stettner, *Almost self-optimizing strategies for the adaptive control of diffusion processes*, J. Optim. Theory Appl., 81 (1994), pp. 479–507.
[7] T. E. Duncan, B. Pasik-Duncan, and L. Stettner, *On the ergodic and the adaptive control of stochastic delay systems*, J. Optim. Theory Appl., 81 (1994), pp. 509–531.
[8] T. E. Duncan, B. Pasik-Duncan, and L. Stettner, *Adaptive control of a partially observed discrete time Markov process*, J. Appl. Math. Optim., 1998, to appear.
[9] E. Fernandez-Gaucherand, A. Arapostathis, and S. I. Marcus, *Analysis of an adaptive control scheme for a partially observed Markov chain*, IEEE Trans. Automat. Control, 38 (1993), pp. 987–993.
[10] O. Hernandez-Lerma, *Adaptive Markov Control Processes*, Springer-Verlag, Berlin, 1989.
[11] S. Kullback and R. A. Leibler, *On information and sufficiency*, Ann. Math. Stat., 22 (1951), pp. 79–86.
[12] S. Kullback, *Information Theory and Statistics*, J. Wiley, New York, 1959.
[13] P. R. Kumar and A. Becker, *A new family of adaptive controllers for Markov chains*, IEEE Trans. Automat. Control, 27 (1982), pp. 137–146.
[14] H. J. Kushner, *Approximation methods for minimum average cost per unit time problems with a diffusion model*, in Approximate Solutions of Random Equations, A. T. Bharucha-Reid, ed., North-Holland, Amsterdam, 1979, pp. 109–126.
[15] W. J. Runggaldier and L. Stettner, *Approximations of Discrete Time Partially Observed Control Problems*, Appl. Math. Monographs 6, C.N.R., Pisa, 1994.
[16] L. J. Savage, *The Foundations of Statistics*, J. Wiley, New York, 1954.
[17] L. Stettner, *On nearly self-optimizing strategies for a discrete-time uniformly ergodic adaptive model*, J. Appl. Math. Optim., 27 (1993), pp. 161–177.
[18] L. Stettner, *Ergodic control of Markov process with mixed observation structure*, Dissertationes Math., 341 (1995), pp. 1–36.
[19] D. W. Stroock and S. R. S. Varadhan, *Multidimensional Diffusion Processes*, Springer-Verlag, New York, 1979.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 2, pp. 447–487, March 1998

003

DYNAMICS OF TIME-VARYING DISCRETE-TIME LINEAR SYSTEMS: SPECTRAL THEORY AND THE PROJECTED SYSTEM*

FABIAN WIRTH†

Abstract. We study structural properties of linear time-varying discrete-time systems. At first an associated system on projective space is introduced as a basic tool to understand the linear dynamics. We study controllability properties of this system and characterize in particular the control sets and their cores. Sufficient conditions for an upper bound on the number of control sets with nonempty interior are given. Furthermore, exponential growth rates of the linear system are studied. Using finite-time controllability properties in the cores of control sets the Floquet spectrum of the linear system may be described. In particular, the closure of the Floquet spectrum is contained in the Lyapunov spectrum.

Key words. time-varying, discrete-time, control sets, universal controls, projected system, Bohl exponents, Floquet exponents, Lyapunov exponents

AMS subject classifications. 93B05, 93C10, 93C50, 93C55, 58F25, 34D08

PII. S0363012996299600

1. Introduction.

In recent years spectral theory for time-varying linear systems has attracted renewed interest. While the foundations of the theory were laid by Floquet [25], Lyapunov [40], and Bohl [16] the introduction of the problems and considerations of control posed new questions to which different approaches have been proposed.

Here we present an approach to the spectral theory of families of discrete-time time-varying linear systems of the form

$$x(t+1) = A(u(t))x(t), \quad t \in \mathbb{N},$$

where the entries of A depend analytically on the time-varying parameter u, which takes values in a prescribed set. In order to gain insight into the dynamics of this system the system that is obtained by projecting on projective space is analyzed. This approach leads to two generalizations of objects well understood for time-invariant systems. The concept of eigenspace is extended to what is called a control set on projective space that is a set characterized by certain controllability properties. Eigenvalues find natural and well-understood generalizations in Floquet, Lyapunov, and Bohl exponents. We examine these different exponential growth rates and how control sets may be employed to characterize them.

Exponential stability is characterized by the Bohl exponent of a time-varying linear system [16]; see also [24]. Przyłuski and Rolewicz studied Bohl exponents (or generalized spectral radii, in their terminology) for discrete-time systems in [46], with further work appearing in [43]–[45]. On the other hand, Lyapunov exponents characterize exponential growth along trajectories. Properties of Lyapunov exponents were studied by Barabanov in [8]–[11], where sufficient conditions so that Lyapunov exponents characterize exponential stability for families of time-varying systems are shown. Berger and Wang [15], Lagarias and Wang [38], and Gurvits [27] study the

*Received by the editors February 26, 1996; accepted for publication (in revised form) December 11, 1996.

http://www.siam.org/journals/sicon/36-2/29960.html

†Institut für Dynamische Systeme, Universität Bremen, D-28344 Bremen, Germany (fabian@math.uni-bremen.de)

joint and the generalized spectral radii given by a discrete inclusion (not to be confused with the notion of generalized spectral radius due to Przyłuski and Rolewicz). The works cited so far are concerned mainly with the largest exponents characterizing stability. In this article we are interested in the complete spectrum of exponential growth rates associated with the system. Also, we will briefly discuss the relation between the different notions appearing in the literature.

The basic idea of our approach is to study a system on projective space that can be constructed from the linear system by Bogolyubov's projection introduced by Has'minskii [28]. The study of this projection in connection with control theory has found numerous applications for continuous-time systems in the analysis of the Lyapunov spectrum. For deterministic systems, the work of Colonius and Kliemann [20], [21], [22] presents a full picture of what is known. In particular, the relation to exponential dichotomies and the dynamical spectrum as studied by Sacker and Sell [47] and Johnson, Palmer, and Sell [35] is analyzed in these references.

Interest in the complete spectrum of the linear system stems from diverse lines of research. One of these is the question of robust stability. Let $A(u_0)$ be a Hurwitz stable matrix (i.e., the spectrum of $A(u_0)$ consists of values with negative real part) and interpret U as a set determining the structure of possible perturbations to the time-invariant system given by $\dot{x} = A(u_0)x$. The problem of robust stability is to determine whether the perturbed system is exponentially stable under all possible perturbations $u : \mathbb{R} \to U$ that are, e.g., piecewise continuous; see Hinrichsen, Ilchmann, and Pritchard [29], [30] and Colonius and Kliemann [19]. The discrete-time problem has been treated by Wirth and Hinrichsen in [57], [55].

Interpreting u as a control term, knowledge about the set of exponential growth rates or Lyapunov exponents can be employed in the stabilization of such systems; see Colonius, Kliemann, and Krull [23] and Grüne [26].

If $u(t)$ is given a stochastic interpretation we are in the realm of stochastic systems. This problem was treated for continuous-time systems by Has'minskii [28], Arnold, Kliemann, and Oeljeklaus [6], Arnold and Kliemann [5], and Arnold and San Martin [7]. The discrete-time case was studied by Homblé in [32], [33] and Baxendale and Has'minskii [14], however, with the restriction that the discrete-time system is invertible.

In this article we wish to lay the foundation for the theory and treat some of the difficulties inherent in the discrete-time case. It is explained how the problem of noninvertibility can be partially overcome while retaining the possibility of obtaining a reasonable system on projective space. We study asymptotic properties of the projected system, show the existence of controls with universal properties, and examine the controllability structure of the projected system. This supplies the tools we need for an analysis of the different spectra.

We proceed as follows. Section 2 contains the problem statement along with the assumptions we make. In section 3 we study accessibility, transitivity, and regularity of discrete-time systems. Orbits and regular orbits are introduced and it is explained why forward accessibility can be characterized by the rank of a Jacobian. This has been noted by several other authors [41], [32]. What is particularly useful in the case of the projected system is that by Proposition 3.6 it is not necessary to check this in local coordinates on the projective space $\mathbb{P}^{n-1}_{\mathbb{K}}$.

In section 4 we exhibit some asymptotic properties of the system on projective space. The study of ω-limit sets follows the approach of Colonius and Kliemann [20] and is standard if the projections of linear systems on projective space are studied.

Using the regularity arguments from section 3 we obtain sufficient conditions for the generalized eigenspace of a transition matrix to project to a region of exact controllability.

In section 5 we state a result on universally regular controls and a controllability property that can be proved using the existence of universally regular controls. In spite of the activity in the study of accessibility of discrete-time systems, the existence of universal controls has only recently been investigated [54], [50]. In [49] Sontag shows the existence of universally regular (universal nonsingular, in his terminology) controls for analytic, strongly accessible continuous-time systems. Related, and at first glance more interesting, is the existence of universally distinguishing controls which has been studied by Sussmann [51] and Sontag and Wang [48]. It cannot be overemphasized, however, that without the existence of universally regular controls, the following results would lose a considerable amount of strength. The main result of this section is that forward accessibility on projective space implies that a whole linear subspace may be steered so as to simultaneously avoid a complementary linear subspace. An analogue of this statement (Proposition 5.3) has to our knowledge not been studied in continuous time.

A starting point in the study of nonlinear control systems are questions of controllability of a system. Unlike the linear case where controllability is a global property in the state space, nonlinear systems may possess several regions of controllability. An important conceptual tool is to study sets, where it is possible to steer arbitrarily close from any one point to any other. These are the so-called *control sets*, which are introduced in section 6.

Kliemann [37] studied properties of control sets of locally accessible systems on smooth manifolds in continuous time. For the projected system obtained in the continuous-time case, an upper bound on the number of control sets with nonvoid interior has been obtained in [20]. An improved version of this result has been given by Braga and San Martin [12], where smaller upper bounds than the dimension of the state space have been given depending on the group that is acting on projective space. In the discrete-time case control sets have been studied by Albertini and Sontag [3], [4], [2], who also introduced the concept of the core of a control set, which is a strictly discrete-time concept. Introducing a further assumption, we define regular cores which can be shown to enjoy the same properties one would expect for cores; in fact, for the class of systems studied in [3] the definitions of core and regular core coincide. We give an example of a system where the interior of a control set and its regular core do not coincide.

What is surprising is that in neither the continuous- nor the discrete-time case has an effort been undertaken to study control sets for *complex* systems, although it has been known for some time that even for real systems it is useful to study complex perturbations by the results of Hinrichsen and Pritchard [31].

A first observation for our system on projective space is that the generalized eigenspaces corresponding to universally regular controls project to the cores of appropriate control sets. Using this property we show in section 7 that under weak assumptions there exist a unique invariant control set and a unique open control set on projective space. These are maximal (resp., minimal) in the control order on the control sets. Here is the first time where the importance of the universally regular controls becomes clear, as their existence yields an easy proof for the existence of the maximal and minimal control sets. This is also the point where we have to depart from lines of proof available in the literature that are based on properties of Lie groups, if we do not want to restrict ourselves to the invertible case.

In section 8 further results on control sets with nonempty interior are presented. For these it is important what the minimal possible rank drop on a path connecting two admissible invertible matrices is. Depending on this singularity index, we show that the eigenspaces of universally regular controls corresponding to an eigenvalue whose modulus has index greater than the singularity index project to a control set uniquely determined by the index of the modulus. Control sets with this property are called main control sets. This leads to a sufficient condition in terms of the singularity index guaranteeing that there exist at most n control sets with nonempty interior, which are all main control sets. It is briefly explained in what sense control sets may be viewed as a generalization of generalized eigenspaces.

In section 9 we begin our discussion of spectral theory by introducing the different exponents we want to study. Our definition of Floquet and Lyapunov spectra follows Colonius and Kliemann [20], [22], with the exception that in these references the collection of the ith Floquet exponents is not introduced.

In section 10 the Floquet spectrum of the discrete-time system is analyzed. We study Floquet spectra corresponding to control sets with nonempty core. To each such control set an associated set of Floquet exponents is defined. The idea of the proof that the closure of such a set is an interval follows the continuous-time case. The key here is a finite-time controllability property in the cores of control sets. In section 11 we study Lyapunov and Bohl spectra and their relation to the Floquet spectrum. Using an idea already developed in [18] we show under which conditions it is possible to approximate Lyapunov exponents by periodic controls. Furthermore, it is shown that without any further assumptions the closure of a Floquet spectrum of a control set actually consists of Lyapunov exponents corresponding to trajectories that remain in that control set. This is the statement of Theorem 11.1 (ii). It follows that the closure of the Floquet spectrum is contained in the Lyapunov spectrum. It has been shown by Berger and Wang [15] that the joint and generalized spectral radii of a discrete inclusion given by a bounded set of matrices are equal. For our systems, this implies the equality of the suprema of Bohl, Floquet, and Lyapunov spectra. We show that the infima of Floquet and Lyapunov spectra coincide as well.

To indicate a further line of research let us point out that an extension to the theory of control sets is given by the so-called chain control sets, which have been introduced by Colonius and Kliemann [20], [22]. The idea is to consider not trajectories of the system but (ε, T)-chains to define chain-orbits and to use these to define chain control sets. For discrete-time systems, this has been studied by Albertini and Sontag in [4]. The extension of these concepts to the kind of systems we have studied will be an interesting direction for further research, since with chain control sets, it is possible to describe the Morse spectrum of the discrete-time system, which is an outer approximation of the set of Lyapunov exponents.

2. Problem statement. Let $\mathbb{K} = \mathbb{R}, \mathbb{C}$ and let $\tilde{U} \subset \mathbb{K}^m$ be open and connected. For an analytic map

$$A : \tilde{U} \to \mathbb{K}^{n\times n}\,, \tag{2.1}$$

we consider a family of time-varying linear systems of the form

$$x(t+1) = A(u(t))x(t)\,, \quad t \in \mathbb{N}, \tag{2.2}$$

$$x(0) = x_0 \quad \in \mathbb{K}^n, \tag{2.3}$$

where $u : \mathbb{N} \to U \subset \tilde{U}$. The set-up we have chosen contains in particular systems affine in u and positive systems as subclasses. Also, it naturally extends to periodic systems.

For $t \in \mathbb{N}$, U^t denotes the set of admissible finite control sequences $u = (u(0), \dots, u(t-1))$, while $U^{\mathbb{N}}$ is the set of infinite control sequences $u = (u(0), u(1), \dots)$. It will always be clear from the context whether u denotes an element of U, U^t, or $U^{\mathbb{N}}$.

For two finite control sequences $u_1 \in U^{t_1}$, $u_2 \in U^{t_2}$ we define the concatenation (u_1, u_2) to be the sequence in $U^{t_1+t_2}$ given by $(u_1, u_2) = (u_1(0), \dots, u_1(t_1 - 1), u_2(0), \dots, u_2(t_2 - 1))$. The k-times repeated concatenation of $u \in U^t$ is denoted by $(u)^k \in U^{tk}$. For infinite control sequences $u \in U^{\mathbb{N}}$ we consider for $t \in \mathbb{N}$, $u_{[0,t-1]} := (u(0), \dots, u(t-1)) \in U^t$, the "first part" of the control sequence u. The evolution operator generated by a control sequence $u \in U^{\mathbb{N}}$ is defined by

$$\Phi_u(s,s) = I\,, \quad \Phi_u(t+1,s) = A(u(t))\Phi_u(t,s)\,, \quad t \geq s \in \mathbb{N}\,. \tag{2.4}$$

With this notation, $\Phi_u(t,0)x_0$ is the solution of (2.2) corresponding to the initial value x_0 and the control u at time t.

We denote by U_{inv} the set $\{u \in U;\ \det A(u) \neq 0\}$, which is clearly the complement of a set defined by analytic equations in U. Thus U_{inv} is either ω-generic in U or empty, where we call a set ω-generic if its complement is contained in a proper analytic subset of $\tilde{U}$. The term generic will be used for sets whose complements are contained in closed subanalytic sets of dimension strictly less than the manifold considered. For details on the theory of analytic and subanalytic sets we refer the reader to [42], [36], and [52]. Below, we will have to make use of the existence of invertible matrices $A(u)$, so that we have to assume that $U_{inv} \neq \emptyset$.

The following general assumption will be made throughout the remainder of this article. Note, however, that the first one is just for convenience and without loss of generality.

ASSUMPTION 2.1. *Let $\mathbb{K} = \mathbb{R}, \mathbb{C}$ and consider system* (2.2)*. We assume that the map A in* (2.1) *and the sets $U \subset \tilde{U} \subset \mathbb{K}^m$ are such that*

(i) $0 \in U$,
(ii) *the set U_{inv} is ω-generic in U,*
(iii) $\operatorname{int} U$ *is connected,*
(iv) $U \subset \operatorname{cl} \operatorname{int} U \subset \tilde{U}$,
(v) U *is bounded.*

One tool for the study of Lyapunov exponents has been the projection onto the projective space, known as Bogolyubov's projection. It is based on the fact that in continuous time the angular component of the system may be decoupled from the radial and studied independently.

In our discrete-time system we do not exclude the possibility that the origin may be reached from nonzero states. If this is regarded from the point of view of stability or robust stability, it poses no problem, for once system (2.2) is at zero it remains there, as it is totally uncontrollable at zero. However, this means that system (2.2) as such may not be projected onto projective space. First the maximal subsystem that can be projected has to be identified.

To consider the discrete-time analogue of Bogolyubov's projection, we define for $x \in \mathbb{K}^n$

$$U(x) := \{u \in U; A(u)x \neq 0\}\,,$$

and with a slight abuse of notation the analogous sets for finite and infinite control sequences are denoted by $U^t(x)$ and $U^{\mathbb{N}}(x)$.

As $U_{inv} \subset U(x)$ and $U^t_{inv} := (U_{inv})^t \subset U^t(x)$ for all $x \in \mathbb{K}^n \backslash \{0\}$ it follows that for $x \neq 0$ the sets $U(x)$ and $U^t(x)$ are ω-generic in U (resp., U^t). Below, $\mathbb{P}^{n-1}_{\mathbb{K}}$ denotes the $(n-1)$-dimensional projective space, and for $W \subset \mathbb{K}^n$, $\mathbb{P}W$ denotes the natural projection of $W \backslash \{0\}$ onto the projective space $\mathbb{P}^{n-1}_{\mathbb{K}}$.

For $\xi \in \mathbb{P}^{n-1}_{\mathbb{K}}$ we define the admissible control values for ξ by

$$U(\xi) := U(x) \quad \text{iff } \xi = \mathbb{P}x\,,$$

and in an analogous fashion $U^t(\xi), U^{\mathbb{N}}(\xi)$. This is well defined, as $\operatorname{Ker} A(u)$ is a linear subspace. With this notation the projected system corresponding to our linear system (2.2) is given by

$$\xi(t+1) = \mathbb{P}A(u(t))\xi(t)\,, \quad t \in \mathbb{N}, \tag{2.5}$$

$$\xi(0) = \xi_0 \in \mathbb{P}^{n-1}_{\mathbb{K}}, \tag{2.6}$$

$$u \in U^{\mathbb{N}}(\xi_0). \tag{2.7}$$

We denote the solution of (2.5) corresponding to an initial value ξ_0 and a control sequence $u \in U^{\mathbb{N}}(\xi_0)$ by $\xi(\cdot; \xi_0, u)$. For a subset $V \subset \mathbb{P}^{n-1}_{\mathbb{K}}$, $t \in \mathbb{N}$, $u \in U^t$ the notation $\xi(t; V, u) := \{\xi(t; \eta, u); \eta \in V \text{ such that } u \in U^t(\eta)\}$ will be used.

3. Accessibility, transitivity, and regularity. Let us now study the projected system (2.5) from a control point of view. The variable "u" will be treated as if it were available for control of the system. A basic question in control theory is that of accessibility. We begin with the following basic definitions.

DEFINITION 3.1 (orbits). *Let $\mathbb{K} = \mathbb{R}, \mathbb{C}$. Consider system* (2.5). *The forward orbit of ξ at time t is defined as*

$$\mathcal{O}^+_t(\xi) := \{\eta \in \mathbb{P}^{n-1}_{\mathbb{K}}; \exists\, u \in U^t(\xi) \text{ with } \eta = \xi(t; \xi, u)\}\,.$$

The forward orbit of ξ is then defined by $\mathcal{O}^+(\xi) := \cup_{t \in \mathbb{N}} \mathcal{O}^+_t(\xi)$. The backward orbit of ξ at time t is given by

$$\mathcal{O}^-_t(\xi) := \{\eta \in \mathbb{P}^{n-1}_{\mathbb{K}}; \exists\, u \in U^t(\eta) \text{ with } \xi = \xi(t; \eta, u)\},$$

which leads to a definition of $\mathcal{O}^-(\xi)$ analogous to that of the positive forward orbit. Let

$$\mathcal{O}_0(\xi) := \{\xi\} \quad \mathcal{O}_{t+1}(\xi) := \bigcup_{\eta \in \mathcal{O}_t(\xi)} \mathcal{O}^+(\eta) \cup \mathcal{O}^-(\eta)\,, \quad t \in \mathbb{N}.$$

The orbit of ξ is then defined by

$$\mathcal{O}(\xi) = \bigcup_{t \in \mathbb{N}} \mathcal{O}_t(\xi). \tag{3.1}$$

DEFINITION 3.2 (accessibility). *The system* (2.5) *is called* forward accessible from ξ *if* $\operatorname{int} \mathcal{O}^+(\xi) \neq \emptyset$, backward accessible from ξ *if* $\operatorname{int} \mathcal{O}^-(\xi) \neq \emptyset$, and forward (*resp.,* backward) accessible *if it is forward* (*backward*) *accessible from all $\xi \in \mathbb{P}^{n-1}_{\mathbb{K}}$.*
System (2.5) *is called* transitive, *if* $\operatorname{int} \mathcal{O}(\xi) \neq \emptyset$ *for all $\xi \in \mathbb{P}^{n-1}_{\mathbb{K}}$.*

We note the following properties of the forward orbit.

LEMMA 3.3. *Let $\mathbb{K} = \mathbb{R}, \mathbb{C}$. Consider system* (2.5).

(i) *Let $\xi_1, \xi_2 \in \mathbb{P}^{n-1}_{\mathbb{K}}$. If $\xi_2 \in \operatorname{cl} \mathcal{O}^+(\xi_1)$, then $\operatorname{cl} \mathcal{O}^+(\xi_2) \subset \operatorname{cl} \mathcal{O}^+(\xi_1)$.*

(ii) *For all $t \in \mathbb{N}$, $\xi \in \mathbb{P}^{n-1}_{\mathbb{K}}$ it holds that $\operatorname{cl} \mathcal{O}^+_t(\xi)$ is connected.*

Proof. (i) follows from a simple continuity argument. In order to prove (ii) we proceed by induction over $t \in \mathbb{N}$. Let $t = 1$ and $0 \neq x \in \mathbb{K}^n$. For an analytic path $\gamma : [0,1] \to \operatorname{int} U$ with $A(\gamma(\tau))x \not\equiv 0$, we will show that $\operatorname{cl} \mathbb{P}\{A(\gamma(\tau))x; \tau \in [0,1]\}$ is pathwise connected. Assume that $A(\gamma(\tau_0))x = 0$ (there are at most finitely many such τ). Let $k \in \mathbb{N}$ be the smallest integer such that $\frac{d^k}{d\tau^k} A(\gamma(\tau))x|_{\tau=\tau_0} \neq 0$, and without loss of generality, assume that the first component of this vector is nonzero. In standard local coordinates around (1,0,...,0) we obtain a neighborhood of τ_0 where for $\tau \neq \tau_0$ it holds that

$$\mathbb{P}A(\gamma(\tau))x = \left(1, \frac{(A(\gamma(\tau))x)_2}{(A(\gamma(\tau))x)_1}, \cdots, \frac{(A(\gamma(\tau))x)_n}{(A(\gamma(\tau))x)_1}\right). \tag{3.2}$$

Using the rule of de l'Hospital we obtain that $\lim_{\tau \to \tau_0} \mathbb{P}A(\gamma(\tau))x$ exists, which shows our claim. As for every $u_1, u_2 \in \operatorname{int} U$ there exists a piecewise polynomial path connecting them, and using Assumption 2.1 (iv), we see that $\operatorname{cl} \mathcal{O}_1^+(\xi)$ is connected.

Assume now that $\operatorname{cl} \mathcal{O}_t^+(\xi)$ is connected. Then, for $u_0 \in U_{inv}$, it holds that $\mathbb{P}A(u_0) \operatorname{cl} \mathcal{O}_t^+(\xi)$ is connected as the continuous image of a connected set. Thus

$$\operatorname{cl} \mathcal{O}_{t+1}^+(\xi) = \operatorname{cl} \bigcup_{\eta \in \operatorname{cl} \mathcal{O}_t^+(\xi)} \operatorname{cl} \mathcal{O}_1^+(\eta)$$

is connected, as each of the sets in the union is connected and each of the sets intersects the connected set $\mathbb{P}A(u_0) \operatorname{cl} \mathcal{O}_t^+(\xi)$. □

It has been shown that forward accessibility is intimately related to the rank of a certain mapping in the case of smooth invertible systems [4]. To carry this result over to our case, let for every $t \in \mathbb{N}$

$$W_t := \left\{(\xi, u) \in \mathbb{P}_{\mathbb{K}}^{n-1} \times \operatorname{int} U^t; \quad u \in U^t(\xi)\right\} \tag{3.3}$$

and consider the map

$$F_t : W_t \to \mathbb{P}_{\mathbb{K}}^{n-1}, \quad F_t(\xi, u) := \xi(t; \xi, u). \tag{3.4}$$

For fixed $\xi \in \mathbb{P}_{\mathbb{K}}^{n-1}$ and $u_0 \in \operatorname{int} U^t(\xi)$ we consider the rank of the linearization of $F_t(\xi, \cdot) : U^t(\xi) \to \mathbb{P}_{\mathbb{K}}^{n-1}$ at $u_0 \in U^t \subset \mathbb{K}^{mt}$ with respect to $u = (u(0)_1, \ldots, u(0)_m, u(1)_1, \ldots, u(t-1)_1, \ldots, u(t-1)_m)$. We define the following shorthand notation:

$$\frac{\partial F_t}{\partial u}(\xi, u_0) = \left(\frac{\partial F_{t,i}}{\partial u(s)_j}(\xi, u_0)\right)_{i=1,\ldots,n-1; s=0,\ldots,t-1; j=1,\ldots,m},$$

where the $F_{t,i}$ are the ith components of the map $F_t(\xi, \cdot)$ with respect to some coordinate chart around $F_t(\xi, u_0)$. The important detail for us is the rank of this Jacobian which is denoted by

$$r(t; \xi, u_0) := \operatorname{rk} \frac{\partial F_t}{\partial u}(\xi, u_0). \tag{3.5}$$

DEFINITION 3.4 (regularity). *A pair* $(\xi, u) \in \mathbb{P}_{\mathbb{K}}^{n-1} \times \operatorname{int} U^t$ *is called* regular*, if* $u \in \operatorname{int} U^t(\xi)$ *and* $r(t; \xi, u) = n - 1$. *A control* $u \in \operatorname{int} U^t$ *is called* universally regular *if* (ξ, u) *is a regular pair for all* $\xi \in \mathbb{P}_{\mathbb{K}}^{n-1}$.

The following lemma summarizes some easy properties in connection with regularity.

LEMMA 3.5. *Let* $\mathbb{K} = \mathbb{R}, \mathbb{C}$, $u_0 \in \operatorname{int} U^t$, $v_0 \in \operatorname{int} U^s$. *For* $\xi_0 \in \mathbb{P}_{\mathbb{K}}^{n-1}$ *let* $F_{t+s}(\xi_0, (u_0, v_0))$ *be defined. Then*

(i) $r(t+s;\xi_0,(u_0,v_0)) \geq r(s;\xi(t;\xi_0,u_0),v_0)$;
(ii) *if* $v_0 \in \operatorname{int} U^s_{inv}$ *then* $r(t+s;\xi_0,(u_0,v_0)) \geq r(t;\xi_0,u_0)$.

Proof. Both assertions follow from an application of the chain rule. □

PROPOSITION 3.6. *Let* $\mathbb{K} = \mathbb{R}, \mathbb{C}$ *and consider system* (2.5). *For all* $x \in \mathbb{K}^n \backslash \{0\}, t \in \mathbb{N}, u \in \operatorname{int} U^t$, *the following statements are equivalent.*

(i) $(\mathbb{P}x, u)$ *is a regular pair.*
(ii) $\Phi_u(t,0)x \neq 0$ *and the following rank condition holds:*

$$\operatorname{rk} G_t(x,u) := \operatorname{rk} \left[\Phi_u(t,0)x \vdots \frac{\partial}{\partial u}\Phi_u(t,0)x \right] = n\,. \tag{3.6}$$

Proof. It is clear that $\Phi_u(t,0)x \neq 0$ is necessary for regularity. An application of the chain rule and a simple calculation in local coordinates yields the desired result. □

The preceding criterion will be frequently used, as it is easily handled in lower dimensions, where all our examples will be situated. Of course, if the dimension is high or the structure of the map A is complicated, this criterion is much too involved to yield a feasible procedure for checking whether a system is forward accessible.

DEFINITION 3.7 (regular orbit). *Let* $\mathbb{K} = \mathbb{R}, \mathbb{C}$ *and consider system* (2.5). *For* $\xi \in \mathbb{P}^{n-1}_{\mathbb{K}}$ *we define the* regular forward orbit *and* regular backward orbit *by*

$$\hat{\mathcal{O}}^+_t(\xi) := \{\eta;\ \exists u \in \operatorname{int} U^t(\xi)\ \textit{s.t.}\ (\xi,u)\ \textit{is regular and}\ \eta = \xi(t;\xi,u)\}\,, \tag{3.7}$$

$$\hat{\mathcal{O}}^-_t(\xi) := \{\eta;\ \exists u \in \operatorname{int} U^t(\eta)\ \textit{s.t.}\ (\eta,u)\ \textit{is regular and}\ \xi = \xi(t;\eta,u)\}. \tag{3.8}$$

The definitions of $\hat{\mathcal{O}}^+(\xi)$ *and* $\hat{\mathcal{O}}^-(x)$ *are then analogous to Definition* 3.1.

For subsets $V \subset \mathbb{P}^{n-1}_{\mathbb{K}}$ we will use the notations $\hat{\mathcal{O}}^+(V) := \bigcup_{\xi\in V} \hat{\mathcal{O}}^+(\xi)$, etc. The following results exhibit some properties of the regular forward orbits. Items (iii) and (v) are shown in [4] for analytic invertible systems, and similar arguments are applicable here.

LEMMA 3.8. *For* $\mathbb{K} = \mathbb{R}, \mathbb{C}$ *consider system* (2.5). *Let* $\xi \in \mathbb{P}^{n-1}_{\mathbb{K}}$; *then*

(i) $\hat{\mathcal{O}}^+_t(\xi)$ *is open in* $\mathbb{P}^{n-1}_{\mathbb{K}}$;
(ii) $\hat{\mathcal{O}}^-_t(\xi)$ *is open in* $\mathbb{P}^{n-1}_{\mathbb{K}}$;
(iii) $\operatorname{int} \mathcal{O}^+_t(\xi) \neq \emptyset$ *iff* $\hat{\mathcal{O}}^+_t(\xi) \neq \emptyset$;
(iv) *if, for* $t \in \mathbb{N}$, $\hat{\mathcal{O}}^+_t(\xi) \neq \emptyset$, *then* $\hat{\mathcal{O}}^+_s(\xi) \neq \emptyset$ *for all* $s \geq t$;
(v) $\operatorname{int} \mathcal{O}^+_t(\xi) \neq \emptyset \Rightarrow \operatorname{cl} \mathcal{O}^+_t(\xi) = \operatorname{cl} \hat{\mathcal{O}}^+_t(\xi)$.

In the case when ξ is a fixed point under a control u such that (ξ,u) is a regular pair, the following property is immediately obtained.

PROPOSITION 3.9. *Let* $\mathbb{K} = \mathbb{R}, \mathbb{C}$. *For* $\xi \in \mathbb{P}^{n-1}_{\mathbb{K}}$ *there exist* $u_\xi \in \operatorname{int} U^t, t \in \mathbb{N}$ *such that* (ξ, u_ξ) *is a regular pair and*

$$\xi = \xi(t;\xi,u_\xi) \tag{3.9}$$

iff there exists an open neighborhood V *of* ξ *such that* $V \subset \hat{\mathcal{O}}^+_t(\xi) \cap \hat{\mathcal{O}}^-_t(\xi)$.

Proof. "⇒": This follows as $\xi \in \hat{\mathcal{O}}^+_t(\xi) \cap \hat{\mathcal{O}}^-_t(\xi)$ and the fact that both $\hat{\mathcal{O}}^+_t(\xi)$ and $\hat{\mathcal{O}}^-_t(\xi)$ are open by Lemma 3.8. "⇐": This is obvious as $\xi \in \hat{\mathcal{O}}^+_t(\xi)$. □

Let us now extend this property to connected sets.

LEMMA 3.10. *Let $\mathbb{K} = \mathbb{R}, \mathbb{C}$. If $\Gamma \subset \mathbb{P}^{n-1}_{\mathbb{K}}$ is a connected set such that for every $\xi \in \Gamma$ the assumption of Proposition* 3.9 *holds for some $t(\xi) \in \mathbb{N}$, then there exists a connected open set V such that*

$$\Gamma \subset V \subset \bigcap_{\xi \in \Gamma} \hat{\mathcal{O}}^+(\xi) \cap \hat{\mathcal{O}}^-(\xi)\,. \tag{3.10}$$

Proof. Let $\xi \in \Gamma$ and consider the set $\hat{\mathcal{O}}^+(\xi) \cap \Gamma$, which is clearly open in Γ. Let $\eta \in \Gamma \cap \operatorname{cl} \hat{\mathcal{O}}^+(\xi)$. As $\eta \in \hat{\mathcal{O}}^-(\eta)$, which is open, it follows that $\hat{\mathcal{O}}^+(\xi) \cap \hat{\mathcal{O}}^-(\eta) \neq \emptyset$ and hence $\eta \in \hat{\mathcal{O}}^+(\xi)$. Thus $\hat{\mathcal{O}}^+(\xi) \cap \Gamma$ is open and closed in Γ and nonempty. As Γ is connected it follows that $\Gamma \subset \hat{\mathcal{O}}^+(\xi)$, and as $\xi \in \Gamma$ was arbitrary, it holds for all $\xi_1, \xi_2 \in \Gamma$ that $\xi_1 \in \hat{\mathcal{O}}^+(\xi_2)$ and thus $\hat{\mathcal{O}}^+(\xi_1) \subset \hat{\mathcal{O}}^+(\xi_2)$ by Lemma 3.5 (i). By symmetry, we obtain $\hat{\mathcal{O}}^+(\xi_1) = \hat{\mathcal{O}}^+(\xi_2)$. Furthermore, it follows for every $\eta \in \Gamma$ that $\Gamma \subset \hat{\mathcal{O}}^-(\eta)$, and again, for all $\xi_1, \xi_2 \in \Gamma$, it holds that $\hat{\mathcal{O}}^-(\xi_1) = \hat{\mathcal{O}}^-(\xi_2)$. As Γ is connected, we can thus choose V to be the connected component of $\hat{\mathcal{O}}^+(\xi) \cap \hat{\mathcal{O}}^-(\xi)$ containing Γ for some $\xi \in \Gamma$. ∎

4. Asymptotic properties on projective space. A first step in the study of the discrete-time system on projective space is the study of the ω-limit sets defined by constant matrices in Jordan block form, where we follow the argumentation from [20] and extend the arguments used there so that we may treat cases not considered in that reference. The following notation is used from now on.

Let $B \in \mathbb{K}^{n\times n}$. For an eigenvalue $\lambda \in \sigma(B) \cap \mathbb{K}$, $E(\lambda)$ denotes the eigenspace and $GE(\lambda)$ denotes the generalized eigenspace corresponding to λ. If $B \in \mathbb{R}^{n\times n}$ and $\lambda \in \sigma(B)$ is complex then $E(\lambda)$ denotes the real part of the sum of the eigenspaces corresponding to the eigenvalues $\lambda, \overline{\lambda}$. $GE(\lambda)$ denotes the appropriate generalized eigenspaces.

It will also be convenient to consider the set of absolute values of the eigenvalues defined by $|\sigma(B)| := \{|\lambda|\,;\, \lambda \in \sigma(B)\}$. For $1 \leq i \leq n$, let $r_i(B)$ be equal to the ith entry of the ordered sequence $|\lambda_1| \leq \cdots \leq |\lambda_n|$, where each element of the spectrum of B appears according to its algebraic multiplicity. For $r \in |\sigma(B)|$ we denote

$$E(r) = \bigoplus_{\substack{\lambda \in \sigma(B) \\ |\lambda| = r}} E(\lambda)\,, \quad GE(r) = \bigoplus_{\substack{\lambda \in \sigma(B) \\ |\lambda| = r}} GE(\lambda)\,. \tag{4.1}$$

Below, we will be concerned with eigenspaces of $\Phi_u(t,0)$ generated by some finite control sequence $u \in U^t$. To make the dependence on u explicit we write $E(\lambda, u), E(r, u)$, etc. The projection of generalized eigenspaces is particularly important if regularity arguments can be applied.

DEFINITION 4.1. *Let $\mathbb{K} = \mathbb{R}, \mathbb{C}$, $t \in \mathbb{N}$, $u \in U^t$, $r \in |\sigma(\Phi_u(t,0))|$. If $r > 0$, we call* $\mathbb{P}GE(r,u)$ regular *if u can be partitioned as $u = (u_1, u_2)$ with $u_1 \in U^{t_1}$, $u_2 \in \operatorname{int} U^{t_2}$, and $t = t_1 + t_2$, and it holds that*

$$(\xi, u_2) \text{ is a regular pair for every } \xi \in \mathbb{P}\Phi_{u_1}(t_1, 0) GE(r, u)\,. \tag{4.2}$$

DEFINITION 4.2 (limit sets). *Let $\mathbb{K} = \mathbb{R}, \mathbb{C}$, $u \in U^{\mathbb{N}}$, $\xi \in \mathbb{P}^{n-1}_{\mathbb{K}}$. The positive ω-limit set is defined by*

$$\omega^+(\xi, u) := \left\{ \eta \in \mathbb{P}^{n-1}_{\mathbb{K}}; \exists \{t_k\}_{k\in\mathbb{N}} \subset \mathbb{N}, \lim_{k\to\infty} t_k = \infty \text{ such that } \eta = \lim_{k\to\infty} \xi(t_k; \xi, u) \right\}\,. \tag{4.3}$$

The negative ω-limit set is defined by

$$\omega^-(\xi,u) := \Big\{ \eta \in \mathbb{P}^{n-1}_{\mathbb{K}}; \exists \{t_k\}_{k\in\mathbb{N}} \subset \mathbb{N}, \lim_{k\to\infty} t_k = \infty,\ \exists \{\eta_k\} \subset \mathbb{P}^{n-1}_{\mathbb{K}},$$

$$\xi = \xi(t_k;\eta_k,u) \text{ such that } \eta = \lim_{k\to\infty} \eta_k \Big\} \,. \tag{4.4}$$

For $t \in \mathbb{N}$, $u \in U^t$, $\xi \in \mathbb{P}^{n-1}_{\mathbb{K}}$, $\omega^+(\xi,u)$ (resp., $\omega^-(\xi,u)$) denotes the positive (resp., negative) ω-limit set that is obtained by applying the t-periodic continuation of u.

Note that with this definition we do not exclude the possibility that ω-limit sets may be empty, e.g., if $u \notin U^{\mathbb{N}}(\xi)$. For a discussion of the concept of ω-limit sets we refer the reader to [1, Chapter 1]. In the following lemma we collect some simple properties of limit sets pertinent to our problem. The proof is left to the reader.

LEMMA 4.3. *Let $\mathbb{K} = \mathbb{R}, \mathbb{C}$, $t \in \mathbb{N}, u \in U^t$, $\xi \in \mathbb{P}^{n-1}_{\mathbb{K}}$.*

(i) *$\omega^+(\xi,u)$, $\omega^-(\xi,u)$ are closed.*

(ii) *$\Phi_u(t,0)\omega^+(\xi,u) = \omega^+(\xi,u)$.*

(iii) *If $\xi = \mathbb{P}x = \mathbb{P}\sum_{j=1}^{l} x_j$ with $x_j \in GE(r_j,u)$ is the spectral decomposition of ξ and $r_1 < r_2 < \cdots < r_l$, then*

$$\omega^+(\xi,u) \subset \mathbb{P}GE(r_l)\,. \tag{4.5}$$

If $r_1 = 0$, then $\omega^-(\xi,u) = \emptyset$; otherwise

$$\omega^-(\xi,u) \subset \mathbb{P}GE(r_1)\,. \tag{4.6}$$

(iv) *If $r > 0$, then $\xi \in \mathbb{P}E(r,u) \Rightarrow \xi \in \omega^+(\xi,u) = \omega^-(\xi,u) \subset \mathbb{P}E(r,u)$.*

The following lemma states the fundamental asymptotic property of the projected system.

LEMMA 4.4. *Let $\mathbb{K} = \mathbb{R}, \mathbb{C}$.*

(i) *Let $J_n(\lambda)$ denote an $n \times n$ Jordan block to an eigenvalue $\lambda \in \mathbb{K}\backslash\{0\}$. Then for any $x \in \mathbb{K}^n\backslash\{0\}$*

$$\lim_{t\to\pm\infty} \mathbb{P}J_n(\lambda)^t x = \mathbb{P}[1,0,\ldots,0]'\,. \tag{4.7}$$

(ii) *Let $\mathbb{K} = \mathbb{R}$ and let $J_n(\lambda,\bar\lambda)$ denote a $2n \times 2n$ Jordan block to a complex pair of eigenvalues $\lambda, \bar\lambda$. Then, for any Riemannian metric d on $\mathbb{P}^{2n-1}_{\mathbb{R}}$ and any $x \in \mathbb{R}^{2n}\backslash\{0\}$, it holds that*

$$\lim_{t\to\pm\infty} d(\mathbb{P}J_n(\lambda,\bar\lambda)^t x, \mathbb{P}\mathrm{span}\{[1,0,\ldots,0]', [0,1,0,\ldots,0]'\}) = 0\,. \tag{4.8}$$

Proof.

(i) For $\lambda \in \mathbb{K}\backslash\{0\}$, $t > n$ it holds that

$$J_n(\lambda)^t = \begin{bmatrix} \lambda^t & t\lambda^{t-1} & \ldots & \ldots & \binom{t}{t-(n-1)}\lambda^{t-(n-1)} \\ 0 & \lambda^t & t\lambda^{t-1} & \ldots & \binom{t}{t-(n-2)}\lambda^{t-(n-2)} \\ 0 & 0 & \lambda^t & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & & \ldots & 0 & \lambda^t \end{bmatrix}. \tag{4.9}$$

For $i > 1$ it follows immediately that

$$\lim_{t\to\infty} \left| \frac{(J_n(\lambda)^t e_j)_i}{(J_n(\lambda)^t e_j)_1} \right| = 0\,, \tag{4.10}$$

which proves the assertion in the limit $t \to +\infty$. The assertion for $t \to -\infty$ follows upon noting that $J_n(\lambda)^{-t}$ is similar to $J_n(\frac{1}{\lambda})^t$, where the vector e_1 is fixed under the similarity transformation.

(ii) The proof for the complex pair of eigenvalues follows the same pattern and is omitted. ∎

COROLLARY 4.5. *Let $\mathbb{K} = \mathbb{R}, \mathbb{C}$, $t \in \mathbb{N}$, $u \in U^t$. If for $r \in |\sigma(\Phi_u(t,0))|$, $r > 0$, the generalized eigenspace $\mathbb{P}GE(r,u)$ is regular, then there exists an open set V such that*

$$\mathbb{P}E(r,u) \subset V \subset \bigcap_{\xi\in\mathbb{P}E(r,u)} \hat{\mathcal{O}}^+(\xi) \cap \hat{\mathcal{O}}^-(\xi)\,. \tag{4.11}$$

Proof. Let $u = (u_1, u_2)$ be partitioned in accordance with Definition 4.1. If $\xi_0 \in \mathbb{P}E(r,u)$, then there exists a $\xi_2 \in \mathbb{P}\Phi_{u_1}(t_1,0)E(r,u)$ such that $\xi_0 = \xi(t_2;\xi_2,u_2)$ and (ξ_2,u_2) is regular. Furthermore it holds by Lemma 4.3 (iv) that $\xi_2 \in \mathbb{P}\Phi_{u_1}(t_1,0)\omega^+(\xi_0,u)$ and so $\mathcal{O}^+(\xi_0)\cap\hat{\mathcal{O}}^-(\xi_0) \neq \emptyset$ since the regular backward orbit is open by Lemma 3.8 (ii). Using Lemma 3.8 (v) it follows that $\xi_0 \in \hat{\mathcal{O}}^+(\xi_0)$.

As $\mathbb{P}E(r,u)$ is connected, the assertion follows from Lemma 3.10. ∎

COROLLARY 4.6. *Let $\mathbb{K} = \mathbb{R}, \mathbb{C}$, $t \in \mathbb{N}$, $u \in U^t$. If for $r \in |\sigma(\Phi_u(t,0))|$, $r > 0$, the generalized eigenspace $\mathbb{P}GE(r,u)$ is regular, then there exists an open set W such that*

$$\mathbb{P}GE(r,u) \subset W \subset \bigcap_{\xi\in\mathbb{P}GE(r,u)} \hat{\mathcal{O}}^+(\xi) \cap \hat{\mathcal{O}}^-(\xi)\,. \tag{4.12}$$

Proof. Let $u = (u_1, u_2)$ be partitioned in accordance with Definition 4.1 and $\xi \in \mathbb{P}GE(r,u)$. By Lemma 4.4 and Corollary 4.5 there exists $\eta \in \mathcal{O}^+(\xi) \cap \hat{\mathcal{O}}^-(\mathbb{P}E(r,u))$ and it follows that $\mathbb{P}E(r,u) \subset \hat{\mathcal{O}}^+(\xi)$. On the other hand, $\omega^-(\xi,u) \subset \mathbb{P}E(r,u)$ and so by Corollary 4.5 there exists an $\eta \in \hat{\mathcal{O}}^+(\mathbb{P}E(r,u))$ and a $k \in \mathbb{N}$ such that $\xi = \xi(kt;\eta,(u)^k)$. By regularity of the pair $(\xi((k-1)t+t_1;\eta,((u)^{k-1},u_1)),u_2)$ and using the fact that $\hat{\mathcal{O}}^-_{t_2}(\xi)$ is open, we see that $\eta \in \hat{\mathcal{O}}^-_{kt}(\xi)$. Hence $\mathbb{P}E(r,u) \subset \hat{\mathcal{O}}^-(\xi)$. It follows that $\xi \in \hat{\mathcal{O}}^+(\xi)$, and an application of Lemma 3.10 completes the proof. ∎

Now that we have seen that for generalized eigenspaces in projective space certain controllability properties hold if a regularity condition is satisfied, it is reasonable to ask whether we can, for certain controls, guarantee that this condition holds. This is discussed in the next section.

5. Universally regular controls. A crucial point in the development of the theory is the construction of universally regular controls and the proof of their genericity in U^t for t large enough. The following result is largely a restatement of results shown in [54] and [50].

PROPOSITION 5.1. *Let $\mathbb{K} = \mathbb{R}, \mathbb{C}$. For the projected system (2.5) the following statements are equivalent.*

(i) *System (2.5) is forward accessible.*

(ii) *There exist $t \in \mathbb{N}$, $u^* \in \operatorname{int} U^t$ such that u^* is universally regular.*

(iii) *There exists a $t^* \in \mathbb{N}$ such that for all $t > t^*$ the set of universally regular control sequences is generic in $\operatorname{int} U^t$.*

(iv) *There exists a* $t \in \mathbb{N}$, $u \in \operatorname{int} U^t$ *such that for every* $r \in |\sigma(\Phi_u(t,0))|$ *the generalized eigenspace* $\mathbb{P}GE(r,u)$ *is regular.*

Proof. The equivalence of (i), (ii), and (iii) follows from Corollaries 3.2 and 3.3 in [50]. For this, note in particular that by Proposition 3.6 the set of nonregular pairs in $\mathbb{P}^{n-1}_{\mathbb{K}} \times U^t$ is analytic. To complete the proof note that "(ii) $\Rightarrow$ (iv)" is obvious. For the converse direction let u be such that (iv) is satisfied. For any $\xi \in \mathbb{P}^{n-1}_{\mathbb{K}}$ Lemma 4.4 implies that $\omega^+(\xi,u) \subset \mathbb{P}E(r,u)$ for some $r \in |\sigma(\Phi_u(t,0))|$. Corollary 4.6 implies that $\hat{\mathcal{O}}^+(\xi) \neq \emptyset$, so (i) holds. ∎

The set of universally regular $u \in U^t$ will be denoted by U^t_{reg}, while t^* denotes the smallest $t \in \mathbb{N}$ such that $U^t_{reg} \neq \emptyset$. It follows from the results in [50] that if $\operatorname{int} \mathcal{O}^+_t(\xi) \neq \emptyset$ for all $\xi \in \mathbb{P}^{n-1}_{\mathbb{K}}$, then $t^* \leq tn$. Note that U^t_{reg} is open for all $t \in \mathbb{N}$.

Remark 5.2. Let us point out that we use the term *generic* for sets that are the complement of closed subanalytic sets of lower dimension in the real case or proper analytic subsets in the complex case. The reason that we work with analytically defined sets lies in the analytic dependence of A on u. In particular, we use in the proof of Proposition 8.1 that if the complement of a set Z is generic, then from every $x \in Z$ there exists a path that starts in $x \in Z$ and leaves Z immediately. This is due to the fact that subanalytic sets can be represented as a locally finite union of embedded analytic submanifolds; see [52].

PROPOSITION 5.3. *Let* $\mathbb{K} = \mathbb{R}, \mathbb{C}$. *Assume that* (2.5) *is forward accessible; then, for linear subspaces* $X, Y \subset \mathbb{K}^n$ *such that*

$$\dim X + \dim Y \leq n, \tag{5.1}$$

the set $\{u \in U^t_{reg}; \Phi_u(t,0)X \cap Y = \{0\}\}$ *is generic in* $\operatorname{int} U^t$ *for all* $t \geq t^*$.

Proof. For $X = \{0\}$ there is nothing to show, so assume $\dim X \geq 1$. Note that the set $\{(\xi,u) \in \mathbb{P}X \times \operatorname{int} U^t; u \notin U^t(\xi)$ or $[u \in U^t(\xi)$ and $\xi(t;\xi,u) \in \mathbb{P}Y]\}$ is analytic in $\mathbb{P}X \times \operatorname{int} U^t$. From Remmert's proper mapping theorem [36, Theorem 45.17] (resp., the definition of subanalytic sets [52, section 8]) it follows that the projection of this set given by

$$\{u \in \operatorname{int} U^t; \exists \xi \in \mathbb{P}X \text{ such that } [u \notin U^t(\xi) \text{ or } \xi(t;\xi,u) \in \mathbb{P}Y]\} \tag{5.2}$$

is analytic in $\operatorname{int} U^t$ for $\mathbb{K} = \mathbb{C}$ or subanalytic in $\operatorname{int} U^t$ for $\mathbb{K} = \mathbb{R}$. As the set is clearly closed and the intersection of two generic sets is generic, the assertion is thus proved in the real and the complex cases if the following statement is shown:

$$\begin{aligned} &\text{if } t \geq t^* \text{ and } u \in \operatorname{int} U^t, \text{ then in any open neighborhood of } u \\ &\qquad \text{there exists a } v \in U^t_{reg} \text{ such that } \Phi_v(t,0)X \cap Y = \{0\}. \end{aligned} \tag{5.3}$$

We prove (5.3) by induction over $\dim X$. Let $\dim X = 1$. Due to $u \in \operatorname{cl} U^t_{reg}$ it holds that $\xi(t;\xi,u) \in \operatorname{cl} \hat{\mathcal{O}}^+_t(\xi)$ for $\xi = \mathbb{P}X$, so (5.3) follows immediately. Assume that (5.3) is shown for $\dim X = k < n-1$ and let $X = \operatorname{span}\{x_1, \ldots, x_{k+1}\}$ for a linearly independent set of vectors $x_i \in \mathbb{K}^n$, $i = 1, \ldots, k+1$. Without loss of generality, let $Y \subset \operatorname{span}\{e_{k+2}, \ldots, e_n\}$. Denote $X' = \operatorname{span}\{x_1, \ldots, x_k\}$. Fix $u \in \operatorname{int} U^t$ and an open neighborhood $V \subset \operatorname{int} U^t$ of u. Thus there exists $v \in V \cap U^t_{reg}$ such that

$$\Phi_v(t,0)X' \cap \operatorname{span}\{e_{k+2}, \ldots, e_n\} = \{0\}. \tag{5.4}$$

Due to forward accessibility v may be chosen such that

$$\Phi_v(t,0)x_{k+1} \notin \operatorname{span}\{e_{k+2}, \ldots, e_n\}. \tag{5.5}$$

Let $W \subset V \cap U^t_{reg}$ be a neighborhood of v such that (5.4) and (5.5) are satisfied for all $v' \in W$. Let $P \in \mathbb{K}^{k+1\times n}$ be defined by

$$P = \begin{bmatrix} 1 & & 0 & 0 & \cdots & 0 \\ & \ddots & & \vdots & & \\ 0 & & 1 & 0 & \cdots & 0 \end{bmatrix}; \tag{5.6}$$

then

$$\mathrm{rk} P\Phi_v(t,0)[x_1 \vdots \cdots \vdots x_{k+1}] \geq k\,. \tag{5.7}$$

If the rank is equal to $k+1$, then indeed

$$\Phi_v(t,0)X \cap \mathrm{span}\{e_{k+2},\ldots,e_n\} = \{0\}\,. \tag{5.8}$$

Let $u' \in \mathbb{K}^{mt}$ and consider the mappings

$$h_i : (-\varepsilon,\varepsilon) \to \mathbb{K}^{k+1}, \tag{5.9}$$

$$h_i(\tau) = P\Phi_{v+\tau u'}(t,0)x_i \tag{5.10}$$

for $i = 1,\ldots,k+1$, where ε is small enough such that $v + \tau u' \in W$ for $|\tau| < \varepsilon$. We claim that there exist $u' \in \mathbb{K}^{mt}$ such that (5.8) holds for $\Phi_{v+\tau u'}(t,0)$ for some $|\tau| < \varepsilon$. Assume this is not the case. Then $h_{k+1}(\tau) \in \mathrm{span}\{h_i(\tau)\}_{i=1,\ldots,k}$ for all $|\tau| < \varepsilon$. Hence there exist continuously differentiable functions

$$\mu_i : (-\varepsilon,\varepsilon) \to \mathbb{K}, \quad i = 1,\ldots,k \tag{5.11}$$

such that

$$h_{k+1}(\tau) = \sum_{i=1}^{k} \mu_i(\tau)h_i(\tau), \tag{5.12}$$

where the differentiability follows from the differentiability of the h_i and the fact that the $h_i(\tau)$, $i = 1,\ldots,k$, are linearly independent. Hence, if we differentiate with respect to τ at $\tau = 0$,

$$h'_{k+1}(0) = \sum_{i=1}^{k} \mu'_i(0)h_i(0) + \mu_i(0)h'_i(0), \tag{5.13}$$

or equivalently, using the chain rule,

$$P\frac{\partial \Phi_v(t,0)x_{k+1}}{\partial u}\cdot u' = \sum_{i=1}^{k} \mu'_i(0)h_i(0) + \mu_i(0)P\frac{\partial \Phi_v(t,0)x_i}{\partial u}\cdot u'\,. \tag{5.14}$$

Let

$$B_i := P\frac{\partial \Phi_v(t,0)x_i}{\partial u} \in \mathbb{K}^{k+1\times mt} \tag{5.15}$$

be our shorthand notation; then we obtain that if u' is such that $B_i u' \in \mathrm{span}\{h_1(0), \ldots, h_k(0)\}$ for $i = 1,\ldots,k$ and $B_{k+1}u' \notin \mathrm{span}\{h_1(0),\ldots,h_k(0)\}$, then (5.14) cannot be solved and there exist τ arbitrarily small such that

$$\mathrm{rk} P\Phi_{u+\tau u'}(t,0)[x_1 \vdots \cdots \vdots x_{k+1}] = \mathrm{rk}\Phi_{u+\tau u'}(t,0)[x_1 \vdots \cdots \vdots x_{k+1}] = k+1 \tag{5.16}$$

and hence (5.3) holds.

Assume there is no u' satisfying these properties; i.e., for all $u \in \mathbb{K}^{mt}$ it holds that

$$B_i u \in \operatorname{span}\{h_1(0), \dots, h_k(0)\}, \quad i = 1, \dots, k \Rightarrow B_{k+1} u \in \operatorname{span}\{h_1(0), \dots, h_k(0)\}.$$

By (5.4) we obtain $\dim\{u \in \mathbb{K}^{mt};\, B_i u \in \operatorname{span}\{h_1(0), \dots, h_k(0)\}\} \geq mt - 1$ and thus

$$\dim\{u \in \mathbb{K}^{mt};\, B_i u \in \operatorname{span}\{h_1(0), \dots, h_k(0)\} \text{ for } i = 1, \dots, k\} \tag{5.17}$$

$$= \dim\{u \in \mathbb{K}^{mt};\, B_i u \in \operatorname{span}\{h_1(0), \dots, h_k(0)\} \text{ for } i = 1, \dots, k+1\} \tag{5.18}$$

$$\geq mt - k\,. \tag{5.19}$$

Hence, in suitable coordinates, the matrices B_i, $i = 1, \dots, k+1$, are of the form

$$\begin{array}{c} \\ k \\ \\ \\ 1 \end{array} \begin{array}{c} \begin{array}{cccccc} \multicolumn{3}{c}{mt-k} & & \multicolumn{2}{c}{k} \end{array} \\ \left[\begin{array}{ccccccc} * & \dots & * & & * & \dots & * \\ \vdots & & \vdots & & \vdots & & \vdots \\ * & \dots & * & & * & \dots & * \\ 0 & \dots & 0 & & * & \dots & * \end{array}\right] \end{array}, \tag{5.20}$$

and there are scalars $\nu_i \in \mathbb{K}$, $i = 1, \dots, k+1$, not all zero such that

$$\operatorname{Im} \sum_{i=1}^{k+1} \nu_i B_i \subseteq \operatorname{span}\{h_1(0), \dots, h_k(0)\}. \tag{5.21}$$

Now for the vector $\overline{x} = \sum_{i=1}^{k+1} \nu_i x_i \neq 0$ and with G_t as defined in (3.6) it follows that

$$\operatorname{rk} PG_t(\overline{x}, v) = \operatorname{rk} \left[\begin{array}{c:c} \sum_{i=1}^{k+1} \nu_i h_i(0) & \sum_{i=1}^{k+1} \nu_i B_i \end{array}\right] \leq k, \tag{5.22}$$

which contradicts the universal regularity of v by Proposition 3.6. ∎

In [34], [2], [4] accessibility and transitivity properties of analytic, invertible systems have been studied. In particular, Lie algebraic characterizations of these properties were obtained. It was also obtained that on compact manifolds, transitivity, forward, and backward accessibility are all equivalent. Since forward accessibility of system (2.5) implies the existence of a universally regular control, we can state the following proposition.

PROPOSITION 5.4. *Let $\mathbb{K} = \mathbb{R}, \mathbb{C}$. Assume system (2.5) is forward accessible. Then it is backward accessible and transitive. Furthermore, it holds for every $\xi \in \mathbb{P}^{n-1}_{\mathbb{K}}$ that $\hat{\mathcal{O}}^-(\xi) \neq \emptyset$.*

Proof. This is clear from the existence of a universally regular control. ∎

This proposition shows in particular that the criteria for accessibility developed in [34], [4] can be brought to use in our case, even though the full requirements of the theorems stated in these references are not met. This is due to the fact that system (2.5) is forward accessible iff there is an analytic invertible subsystem that is forward accessible. We call a system a subsystem if the map A is the same but the set of control values is restricted. Thus we could choose an open subset U' of

U_{inv} that is relatively compact in U_{inv}. For the system with control values in U' it is clear that its forward accessibility implies forward accessibility of the original system. But the converse is also true, as forward accessibility implies the generic existence of universally regular controls. This implies that there exists a universally regular control in $U'^{t^*}_{reg}$, where t^* is the constant of the original system.

The converse of the statement in Proposition 5.4 does not hold, as shown by the following example.

Example 5.5. Let $\mathbb{K} = \mathbb{R}, \mathbb{C}$ and $U = \mathbb{K}$. Define

$$A(u) := \begin{bmatrix} 1+2u & 0 \\ 0 & 1+u \end{bmatrix}.$$

Then the system

$$\xi(t+1) = \mathbb{P}A(u(t))\xi(t), \quad t \in \mathbb{N},$$

is clearly not forward accessible, as $\mathcal{O}^+(\mathbb{P}[1,0]') = \{\mathbb{P}[1,0]'\}$, $\mathcal{O}^+(\mathbb{P}[0,1]') = \{\mathbb{P}[0,1]'\}$. However, an open set can be steered to $\mathbb{P}[1,0]'$ by applying the constant control given by $\lambda = -1$ (resp., $\mathbb{P}[0,1]'$ and $\lambda = -\frac{1}{2}$). It is then easy to see that $\operatorname{int}\mathcal{O}_1^-(\xi) \neq \emptyset$ for all $\xi \in \mathbb{P}^1_{\mathbb{K}}$. So the system is backward accessible.

6. Control sets. Let us now give a precise meaning to the words "sets where it is possible to steer arbitrarily close from one point to another." Control sets are defined as maximal sets where a controllability property holds. Precontrol sets satisfy the same controllability properties without being maximal. We note that different control sets are disjoint, and that for every precontrol set there exists a unique control set containing it. Furthermore, for every point in a control set there exists a control sequence such that the corresponding trajectory stays in that control set for all times, and the closures of the forward orbits of two points contained in the same control set coincide.

DEFINITION 6.1 (control set). *Let* $\mathbb{K} = \mathbb{R}, \mathbb{C}$. *Consider system* (2.5). *A set* $\emptyset \neq D \subset \mathbb{P}^{n-1}_{\mathbb{K}}$ *is called a* precontrol set *if*

(i) $D \subset \operatorname{cl}\mathcal{O}^+(\xi), \quad \forall \xi \in D$;

(ii) *for every* $\xi \in D$ *there exists a* $u \in U^{\mathbb{N}}(\xi)$ *and an increasing sequence* $(t_k)_{k\in\mathbb{N}} \subset \mathbb{N}$ *such that* $\xi(t_k; \xi, u) \in D$ *for all* $k \in \mathbb{N}$.

A precontrol set D *is called a* control set *if, furthermore,*

(iii) D *is a maximal set with respect to inclusion satisfying* (i).

A control set C *is called an* invariant control set *if*

$$\operatorname{cl} C = \operatorname{cl}\mathcal{O}^+(\xi) \, \forall \xi \in C. \tag{6.1}$$

With this definition we obtain the following basic results.

PROPOSITION 6.2. *Let* $\mathbb{K} = \mathbb{R}, \mathbb{C}$ *and consider system* (2.5).

(i) *For two control sets* D_1, $D_2 \subset \mathbb{P}^{n-1}_{\mathbb{K}}$ *it holds that either* $D_1 = D_2$ *or* $D_1 \cap D_2 = \emptyset$.

(ii) *For every precontrol set* $D' \subset \mathbb{P}^{n-1}_{\mathbb{K}}$ *there exists a unique control set* D *such that* $D' \subset D$.

(iii) *If* $\xi_1, \xi_2 \in D$ *for some control set* $D \subset \mathbb{P}^{n-1}_{\mathbb{K}}$ *and for some* $u \in U^t$ *it holds that*

$$\xi_2 = \xi(t; \xi_1, u), \tag{6.2}$$

then

$$\xi(s;\xi_1,u)\in D \text{ for } s=0,\dots,t\,. \tag{6.3}$$

(iv) *For a control set D it holds that*

$$\operatorname{cl}\mathcal{O}^+(\xi_1)=\operatorname{cl}\mathcal{O}^+(\xi_2)\quad\forall\xi_1,\xi_2\in D\,. \tag{6.4}$$

(v) *Let D be a control set. For every $\xi\in D$ there exists a control $u\in U^{\mathbb{N}}(\xi)$ such that*

$$\xi(t;\xi,u)\in D\quad\forall t\in\mathbb{N}\,. \tag{6.5}$$

(vi) *Let D be a control set. For every $\xi\in D$ and every $T\in\mathbb{N}$ it holds that*

$$\operatorname{cl}\mathcal{O}^+(\xi)=\operatorname{cl}\bigcup_{t=T}^{\infty}\mathcal{O}_t^+(\xi)\,. \tag{6.6}$$

In the forward accessible case, invariant control sets enjoy further useful properties.

PROPOSITION 6.3. *Let $\mathbb{K}=\mathbb{R},\mathbb{C}$. Assume that* (2.5) *is forward accessible. A control set C is invariant iff it is closed and satisfies* $\operatorname{int}C\neq\emptyset$.

Proof. "⇒": If $C=\mathbb{P}^{n-1}_{\mathbb{K}}$ there is nothing to show. Assume that $\xi\in\operatorname{cl}C\setminus C$. This implies that $\xi\in\operatorname{cl}\mathcal{O}^+(\eta)$ for all $\eta\in C$. As $\xi\notin C$ it follows that $\mathcal{O}^+(\xi)\cap C=\emptyset$, for otherwise $C\subset\operatorname{cl}\mathcal{O}^+(\xi)$, and this would imply $\xi\in C$. By assumption, there exist $t\in\mathbb{N}$, $u\in U^t(\xi)$ such that $\xi(t;\xi,u)\in\operatorname{int}\mathcal{O}^+(\xi)$. By continuity, there exists a neighborhood V of ξ that is steered to $\operatorname{int}\mathcal{O}^+(\xi)$, and therefore there exists $\eta\in C$ such that $\mathcal{O}^+(\eta)\cap\operatorname{int}\mathcal{O}^+(\xi)\neq\emptyset$. But $\mathcal{O}^+(\eta)\subset\operatorname{cl}C$, a contradiction. Hence C is closed, and $C=\operatorname{cl}\mathcal{O}^+(\xi)$ for $\xi\in C$. As $\operatorname{int}\mathcal{O}^+(\xi)\neq\emptyset$ it follows that $\operatorname{int}C\neq\emptyset$.

"⇐": Let C be a closed control set with $\operatorname{int}C\neq\emptyset$. If $C=\mathbb{P}^{n-1}_{\mathbb{K}}$ there is nothing to show. Otherwise we have to show for every $\xi\in C$ that $\operatorname{cl}\mathcal{O}^+(\xi)\subset C$, or equivalently, as C is closed, $\mathcal{O}^+(\xi)\subset C$. For every $\eta\in C$ there exists $t\in\mathbb{N}$, $u\in U^t(\eta)$ such that $\xi(t;\eta,u)\in\operatorname{int}C$. By continuous dependence on the initial values there exists an open neighborhood $V(\eta)$ of η such that $\xi(t;V(\eta),u)\subset\operatorname{int}C$. Hence there exists an open set $V\supset C$ such that $\mathcal{O}^+(\xi)\cap\operatorname{int}C\neq\emptyset$ and therefore $C\subset\operatorname{cl}\mathcal{O}^+(\xi)$ for every $\xi\in V$.

Assume now that there exists a $\xi\in C$ and a $u\in U(\xi)$ such that $\xi(1;\xi,u)\notin C$. As $C\subset\operatorname{cl}\mathcal{O}^+(\xi)$ there exists an $\eta\in\mathcal{O}^+(\xi)\cap C$, and Proposition 6.2 (iii) guarantees that there exists a $v\in U(\xi)$ such that $\xi(1;\xi,v)\in C$. Now $\operatorname{cl}\mathcal{O}_1^+(\xi)\cap C\neq\emptyset$, but also $\operatorname{cl}\mathcal{O}_1^+(\xi)\not\subset C$. Since $\operatorname{cl}\mathcal{O}_1^+(\xi)$ is connected, it follows that there exists a $\zeta\in\mathcal{O}_1^+(\xi)\cap(V\setminus C)$. But then $\zeta\in\operatorname{cl}\mathcal{O}^+(\eta)$ for all $\eta\in C$ and $C\subset\operatorname{cl}\mathcal{O}^+(\zeta)$, and thus $\zeta\in C$, which is a contradiction. □

Cores of control sets, a strictly discrete-time concept, have been introduced in [4]. We give a definition of the core that differs slightly from the original definition in that we require a regularity condition to hold. So, to contrast it, it might be called the *regular core* of a control set. It should, however, be noted that for the systems studied in [4], core and regular core of a control set coincide.

DEFINITION 6.4 (regular core). *Let $\mathbb{K}=\mathbb{R},\mathbb{C}$. Let $D\subset\mathbb{P}^{n-1}_{\mathbb{K}}$ be a control set with* $\operatorname{int}D\neq\emptyset$. *The* (regular) core *of D is defined as*

$$\operatorname{core}(D):=\{\xi\in D;\hat{\mathcal{O}}^+(\xi)\cap D\neq\emptyset \text{ and } \hat{\mathcal{O}}^-(\xi)\cap D\neq\emptyset\}\,. \tag{6.7}$$

PROPOSITION 6.5. *Let $\mathbb{K} = \mathbb{R}, \mathbb{C}$ and consider system* (2.5). *It holds that $\xi \in \hat{\mathcal{O}}^+(\xi)$ iff there exists a control set D such that $\xi \in \operatorname{core}(D)$.*

Proof. "⇒": This follows from Proposition 3.9.

"⇐": Let $\eta \in \hat{\mathcal{O}}^-(\xi) \cap D$. By the implicit function theorem there exists a neighborhood V of η with $V \subset \hat{\mathcal{O}}^-(\xi)$. As $\eta \in D$ it follows that $V \cap \mathcal{O}^+(\xi) \neq \emptyset$. Therefore, $\hat{\mathcal{O}}^-(\xi) \cap \mathcal{O}^+(\xi) \neq \emptyset$, and so $\xi \in \hat{\mathcal{O}}^+(\xi)$. □

PROPOSITION 6.6. *Let $\mathbb{K} = \mathbb{R}, \mathbb{C}$ and consider system* (2.5). *Let $D \subset \mathbb{P}^{n-1}_{\mathbb{K}}$ be a control set with $\operatorname{int} D \neq \emptyset$. If system* (2.5) *is forward accessible from every $\xi \in D$, then*

(i) $\operatorname{core}(D)$ *is open in* $\mathbb{P}^{n-1}_{\mathbb{K}}$;

(ii) $\operatorname{cl}\operatorname{core}(D) = \operatorname{cl}\operatorname{int}(D) = \operatorname{cl} D$;

(iii) *if $\xi \in \operatorname{core}(D)$ then $\operatorname{core}(D) \subset \hat{\mathcal{O}}^+(\xi)$ and $D \subset \hat{\mathcal{O}}^-(\xi)$;*

(iv) *if $\xi \in \operatorname{core}(D)$, $t \in \mathbb{N}$, $u \in \operatorname{int} U^t_{inv}$, and $\xi(t;\xi,u) \in D$, then $\xi(s;\xi,u) \in \operatorname{core}(D)$ for $s = 0, \ldots, t$.*

Proof. (i) If $\xi \in \operatorname{core}(D)$, then by Proposition 6.5, $\xi \in \hat{\mathcal{O}}^+(\xi)$. Thus the assertion follows from Proposition 3.9, as there exists an open neighborhood V of ξ satisfying $V \subset \hat{\mathcal{O}}^+(\xi) \cap \hat{\mathcal{O}}^-(\xi)$. V is a precontrol set satisfying the rank condition in (6.7), and thus contained in $\operatorname{core}(D)$.

(ii) Clearly, $\operatorname{cl}\operatorname{core}(D) \subset \operatorname{cl}\operatorname{int} D \subset \operatorname{cl} D$. Let $\xi \in \operatorname{cl} D$ and V be any open neighborhood of ξ. Let $\eta \in \operatorname{int} D$. By Lemma 3.8 (v) and Proposition 6.2 (vi) we have $D \subset \operatorname{cl} \hat{\mathcal{O}}^+(\eta)$. Thus we may choose $\zeta \in D \cap V \cap \hat{\mathcal{O}}^+(\eta)$, and it follows that $\hat{\mathcal{O}}^-(\zeta) \cap \operatorname{int} D \neq \emptyset$. As $\zeta \in D$ we have as before that $\operatorname{int} D \subset D \subset \hat{\mathcal{O}}^+(\zeta)$ and so also $\hat{\mathcal{O}}^+(\zeta) \cap \operatorname{int} D \neq \emptyset$. Thus $\zeta \in \operatorname{core}(D) \cap V$.

(iii) If $\xi \in \operatorname{core}(D)$, then $\xi \in \operatorname{cl} \mathcal{O}^+(\eta)$ for every $\eta \in D$. By Proposition 3.9, $\xi \in \hat{\mathcal{O}}^-(\xi)$ and so $\mathcal{O}^+(\eta) \cap \hat{\mathcal{O}}^-(\xi) \neq \emptyset$, and hence $\eta \in \hat{\mathcal{O}}^-(\xi)$. This shows that $D \subset \hat{\mathcal{O}}^-(\xi)$. As $\xi \in \operatorname{core}(D)$ was arbitrary, this implies also that $\operatorname{core}(D) \subset \hat{\mathcal{O}}^+(\xi)$ for every $\xi \in \operatorname{core}(D)$.

(iv) This is clear as $D \subset \hat{\mathcal{O}}^-(\xi) \subset \hat{\mathcal{O}}^-(\xi(s;\xi,u))$ for $s = 0, \ldots, t$ by Lemma 3.5, and $\operatorname{core}(D) \subset \hat{\mathcal{O}}^+(\xi(t;\xi,u)) \subset \hat{\mathcal{O}}^+(\xi(s;\xi,u))$. □

From now on, control sets of the system on projective space are studied using the underlying linear structure which allows more precise statements. We begin by considering projected generalized eigenspaces that satisfy a regularity condition.

PROPOSITION 6.7. *Let $\mathbb{K} = \mathbb{R}, \mathbb{C}$, $t \in \mathbb{N}$, $u \in \operatorname{int} U^t$. Assume that for $r \in |\sigma(\Phi_u(t,0))|$, $r > 0$, the generalized eigenspace $\mathbb{P}GE(r,u)$ is regular. Then there exists a control set D such that*

$$\mathbb{P}GE(r,u) \subset \operatorname{core}(D)\,. \tag{6.8}$$

Proof. This follows from Corollary 4.6, and the fact that for every precontrol set, there is a control set containing it. □

PROPOSITION 6.8. *Let $\mathbb{K} = \mathbb{R}, \mathbb{C}$, $t \in \mathbb{N}$. Assume that $\gamma : [0,1] \to U^t$ is a continuous path with an associated continuous path $\gamma_2 : [0,1] \to \mathbb{R}$ such that for every $\tau \in [0,1]$*

$$0 \neq \gamma_2(\tau) \in \left|\sigma(\Phi_{\gamma(\tau)}(t,0))\right| \tag{6.9}$$

and $\mathbb{P}GE(\gamma_2(\tau), \gamma(\tau))$ is regular. Then there exists a connected open precontrol set D

contained in the core of a control set with

$$\bigcup_{\tau\in[0,1]} \mathbb{P}GE(\gamma_2(\tau),\gamma(\tau)) \subset D\,. \tag{6.10}$$

Proof. By Proposition 6.7, for every $\tau \in [0,1]$, there exists an open precontrol set $V(\tau) \supset \mathbb{P}GE(\gamma_2(\tau),\gamma(\tau))$ which we may assume without loss of generality to be connected and contained in the core of a control set. By the continuity properties of the eigenprojections (see [13, Chapter II.8]), for every $\tau \in [0,1]$ there exists an $\varepsilon(\tau) > 0$ such that $\mathbb{P}GE(\gamma_2(\tau'),\gamma(\tau')) \subset V(\tau)$ if $|\tau - \tau'| < \varepsilon(\tau)$. This shows that

$$D := \bigcup_{\tau\in[0,1]} V(\tau) \tag{6.11}$$

is connected and, by Lemma 3.10, is a precontrol set with the desired properties. □

For the system (2.5) the core of a control set corresponds to regular pairs (ξ, u), where ξ is an eigenvector of $\Phi_u(t,0)$ by Proposition 6.5. For a forward accessible system, even more is true. For any control set D with nonempty core we may find universally regular controls u that generate an eigenspace whose projection lies in any prescribed open subset of $\operatorname{core}(D)$.

PROPOSITION 6.9. *Let $\mathbb{K} = \mathbb{R}, \mathbb{C}$. Assume that system (2.5) is forward accessible. For every control set $D \subset \mathbb{P}^{n-1}_{\mathbb{K}}$ with $\operatorname{core}(D) \neq \emptyset$ and every open set $\emptyset \neq V \subset \operatorname{core}(D)$, there exist $t \in \mathbb{N}$, $u \in U^t_{reg}$ such that for some $r \in |\sigma(\Phi_u(t,0))|$*

$$\mathbb{P}E(r,u) \cap V \neq \emptyset\,, \text{ and } \mathbb{P}GE(r,u) \subset \operatorname{core}(D)\,. \tag{6.12}$$

Proof. Let $\xi \in V$. By Proposition 6.5, $\xi \in \hat{\mathcal{O}}^+(\xi)$, and we can choose $t \in \mathbb{N}$, $u \in \operatorname{int} U^t$ such that $r(t;\xi,u) = n-1$ and

$$\xi = \xi(t;\xi,u)\,. \tag{6.13}$$

Without loss of generality let $t > t^*$. Since the set of universally regular controls is generic in $\operatorname{int} U^t$, and by Proposition 6.5, we can choose $u_1 \in U^t_{reg}$ such that $\eta_1 := \xi(t;\xi,u_1) \in \hat{\mathcal{O}}^+_t(\xi) \cap \hat{\mathcal{O}}^-_t(\xi) \cap V$. Using the universal regularity of u_1 and applying the implicit function theorem it may be concluded that there exists an open neighborhood $V(\xi) \subset V$ such that for every $\eta \in V(\xi)$ there exists a universally regular $u(\eta) \in U^t_{reg}$ with $\eta_1 = \xi(t;\eta,u(\eta))$. Furthermore, as $\eta_1 \in \hat{\mathcal{O}}^-_t(\xi)$, we may choose $u_2 \in \operatorname{int} U^t_{inv}$ such that $\eta_2 := \xi(t;\eta_1,u_2) \in V(\xi)$. Hence

$$\eta_1 = \xi\left(2t;\eta_1,(u_2,u(\eta_2))\right), \tag{6.14}$$

and as $u_2 \in \operatorname{int} U^t_{inv}$ and $u(\eta_2) \in U^t_{reg}$, it follows by Lemma 3.5 that $(u_2,u(\eta_2))$ is universally regular. Now η_1 is the projection of an eigenvector of $\Phi_{(u_2,u(\eta_2))}(2t,0)$, which proves the first half of (6.12). To complete the proof note that by Proposition 6.7 there exists a control set $D_2 \supset \mathbb{P}GE(r,u)$. But then $D \cap D_2 \neq \emptyset$, and hence $D = D_2$ by Proposition 6.2 (i). □

It should be noted that elements of the cores of control sets need not be eigenvectors corresponding to eigenvalues for *universally regular* controls even though it holds that $\xi \in \hat{\mathcal{O}}^+(\xi) \Leftrightarrow \xi \in \operatorname{core}(D)$ for some control set D. This phenomenon will be exhibited in the following example. The small sidestep necessary in the proof of the previous Proposition 6.9 is thus explained.

Example 6.10. Let $\mathbb{K} = \mathbb{R}, \mathbb{C}$, and consider the map

$$A : \mathbb{K}^2 \to \mathbb{K}^{2\times 2}, \quad A(a,b) = \begin{bmatrix} 1 & a \\ b & 0 \end{bmatrix}. \tag{6.15}$$

Define $U := \{[a\,,\, b]' \in \mathbb{K}^2;\ |a| < 1,\ |b| < \frac{1}{4}\}$. The system (2.5) given with these data is forward accessible, which is most easily seen using the rank criterion of Proposition 3.6. Define

$$V := \left\{ [x_1, x_2]' \in \mathbb{K}^2; x_1 \neq 0,\ \frac{|x_2|}{|x_1|} < \frac{1}{2} \right\}. \tag{6.16}$$

It is easy to show that $\mathbb{P}V$ is an invariant subset of $\mathbb{P}^1_{\mathbb{K}}$. Also, for the point $\xi_0 := \mathbb{P}[1,0]' \in V$ and the control $u_0 = (0,0)$ it may be seen that $\xi_0 = \mathbb{P}A(u_0)\xi_0$ and (ξ_0, u_0) is a regular pair. Thus, by Proposition 6.5 there exists a control set D satisfying $\xi_0 \in \operatorname{core}(D)$, and by invariance of $\mathbb{P}V$ it holds that $D \subset \operatorname{cl}\mathbb{P}V$. (In fact, D is the unique invariant control set, but this will be shown later.) However, ξ_0 does not belong to the projection of a generalized eigenspace of a universally regular control. Note that there is no generalized eigenspace of dimension 2 corresponding to a universally regular control as otherwise $\mathbb{P}^1_{\mathbb{K}}$ would be contained in the core of a control set (by Proposition 6.7), which contradicts the invariance of V. It is easy to see that if $\det A(u) \neq 0$ and $\xi_0 = \mathbb{P}A(u)\eta_0$, then $\eta_0 = \mathbb{P}[0,1]' \notin \operatorname{cl}\mathbb{P}V$. As universal regularity implies invertibility it follows that if $\xi_0 = \xi(t;\xi_0,u)$ for some universally regular control u then $\xi(t-1;\xi_0,u) = \eta_0$, contradicting the invariance of $\mathbb{P}V$.

This difference between the projected eigenspaces of universally regular controls and the regions of complete controllability is unique for discrete systems and does not occur in continuous time. Compare [20, Proposition 3.8]. The reason appears to be the noninvertibility possible in discrete time.

PROPOSITION 6.11. *Let $\mathbb{K} = \mathbb{R}, \mathbb{C}$. Assume that system* (2.5) *is forward accessible. If $U = U_{inv}$, then*

$$\xi \in \operatorname{core}(D) \Leftrightarrow \exists t \in \mathbb{N},\ u \in U^t_{reg} \text{ such that } \xi = \xi(t;\xi,u), \tag{6.17}$$

where D is some control set.

Proof. "$\Leftarrow$" This is clear from Proposition 6.5.

"$\Rightarrow$" As $\xi \in \operatorname{core}(D)$ by Proposition 6.6 (iii) it follows that $\operatorname{core}(D) \subset \hat{\mathcal{O}}^+(\xi)$. Hence there exist $t \in \mathbb{N}$, $u \in U^t_{reg}$ such that $\xi(t;\xi,u) \in \operatorname{core}(D)$. As $\operatorname{core}(D) \subset \hat{\mathcal{O}}^-(\xi)$, there exist $s \in \mathbb{N}$, $v \in \operatorname{int} U^s = \operatorname{int} U^s_{inv}$ such that $\xi = \xi(t+s;\xi,(u,v))$. By Lemma 3.5, $(u,v) \in \operatorname{int} U^{t+s}$ is universally regular. □

A slight modification of Example 6.10 will show that there indeed exist cases where $\operatorname{core}(D) \neq \operatorname{int} D$ for control sets D.

Example 6.12. Let $\mathbb{K} = \mathbb{R}, \mathbb{C}$,

$$A : \mathbb{K}^2 \to \mathbb{K}^{2\times 2}, \quad A(a,b) = \begin{bmatrix} 1 & a^3 \\ b^3 & 0 \end{bmatrix}. \tag{6.18}$$

Let $U := \{[a\,,\, b]' \in \mathbb{K}^2;\ |a| < 1,\ |b| < (\frac{1}{4})^{\frac{1}{3}}\}$. Note that with this definition the system defined by (6.18) behaves no differently from the system in Example 6.10, in the sense that for every point ξ the forward and backward orbits of the two systems coincide, which is clear from the definitions of A and U. Hence there exists the same control set D as in Example 6.10. But still the point ξ_0 that was critical in

the previous example now does not even belong to the core of D. For this we show that $\hat{\mathcal{O}}^-(\mathbb{P}[1,0]') = \hat{\mathcal{O}}^-(\mathbb{P}[0,1]')$. Let $t \in \mathbb{N}$, $u = (u(0), \dots, u(t-1)) \in U^t$ with $u(t-1) = [a,0]'$, and $x \notin \operatorname{Ker} \Phi_u(t,0)$. Then $\xi(t; \mathbb{P}x, u) = \mathbb{P}[1,0]' = \xi_0$ but

$$G_t(x,u) = \left[\begin{array}{c|cc} * & 1 & a^3 \\ \vdots & & \\ 0 & 0 & 0 \end{array}\right] \cdot \frac{\partial \Phi_u(t-1,0)x}{\partial u} \begin{array}{cc} * & 0 \\ & \\ 0 & 0 \end{array}, \tag{6.19}$$

and hence $\operatorname{rk} G_t(x,u) = 1$ and $(\mathbb{P}x, u)$ is not a regular pair. If $b \neq 0$ and $\xi_0 = \mathbb{P}A(a,b)x$, it follows that $A(a,b)$ is invertible. As we have seen in Example 6.10, if $\det A(u) \neq 0$ then any trajectory going to the point ξ_0 must first go through $\eta_0 = \mathbb{P}[0,1]' \notin \operatorname{cl} \mathcal{O}^+(\xi)$ for all $\xi \in \mathbb{P}V$. So $\xi_0 \notin \hat{\mathcal{O}}^+(\xi_0)$ and hence $\xi_0 \in \operatorname{int} D \setminus \operatorname{core}(D)$.

It should also be noted that it cannot be concluded that the projection of an arbitrary eigenspace corresponding to any control is contained in the closure of a control set with nonempty interior. In fact, in the following example we show that any point of the projective space may be a precontrol set, but the control sets with nonempty interior do not cover the whole projective space. Note that the following example is given here, as it fits well in our discussion of control sets and generalized eigenspaces. We do, however, use a fact from the next section, namely, the existence of a unique open and a unique invariant control set.

Example 6.13. Let $\mathbb{K} = \mathbb{R}$,

$$A : \mathbb{R}^2 \to \mathbb{R}^{2\times 2}, \quad A(a,b) = \begin{bmatrix} 1 & ab \\ a & 1 \end{bmatrix}. \tag{6.20}$$

Define $U = \left\{[a,b]' \in \mathbb{R}^2;\, 0 \leq a \leq \frac{1}{2},\, 2 \leq b \leq 4\right\}$. Then, clearly, choosing $a = 0$ leads to a transition matrix for which every $\xi \in \mathbb{P}^1_{\mathbb{R}}$ is a fixed point. Furthermore, b may be chosen such that the rank condition (3.6) is satisfied. However, the controls for which this is possible are not in the interior of U, and hence the statements made until now do not infer that the system (2.5) is completely controllable on $\mathbb{P}^1_{\mathbb{R}}$. In fact, for the set

$$V := \{[x_1, x_2]' \in \mathbb{R}^2;\, 0 < x_2 < x_1\}, \tag{6.21}$$

$\mathbb{P}V$ is an invariant set of system (2.5), and thus the invariant control set satisfies $C \subset \operatorname{cl} \mathbb{P}V$. On the other hand, we have that the open control set satisfies

$$C^- \subset \mathbb{P}\{[x_1, x_2]' \in \mathbb{R}^2;\, x_1 x_2 \leq 0\}, \tag{6.22}$$

as for every $t \in \mathbb{N}$, $u \in \operatorname{int} U^t$ the matrix $\Phi_u(t,0)$ has only strictly positive entries. Thus, by the Perron–Frobenius theory for positive matrices, $\Phi_u(t,0)$ does not have two linearly independent nonnegative eigenvectors, and the eigenvalue corresponding to the nonnegative eigendirection has algebraic multiplicity 1 (see [39, Chapter 15.3, Theorem 1, and Exercise 11]). This implies that for any $u \in U^t_{reg}$ the evolution operator $\Phi_u(t,0)$ has an eigenvector $x = [x_1, x_2]'$ satisfying $x_1 x_2 \leq 0$ corresponding to an eigenvalue of algebraic multiplicity 1, while the eigenvector corresponding to the other eigenvalue of algebraic multiplicity 1 projects to $\mathbb{P}V$. As for every control set D with nonempty interior, there exists a universally regular control u such that $\mathbb{P}GE(r,u) \subset D$ for a suitable value r by Proposition 6.9, it follows that the set $\mathbb{P}\{[x_1, x_2]' \in \mathbb{R}^2;\ 0 < x_1 < x_2\}$ does not intersect a control set with nonempty interior, although every point in this set is a precontrol set.

7. The maximal and minimal control sets. It is now shown that there exist a unique invariant and a unique open control set. These two can be described in a particularly easy fashion: they are the intersection of the closures of forward orbits, respectively, in the interior of the intersection of closures of backward orbits. We call these control sets the maximal, respectively, minimal, control sets. This terminology is justified, as we may introduce a natural order on the set of all control sets on $\mathbb{P}^{n-1}_{\mathbb{K}}$, in which the maximal control set is the invariant one and the minimal is open.

THEOREM 7.1. *Let* $\mathbb{K} = \mathbb{R}, \mathbb{C}$. *Assume that system* (2.5) *is forward accessible. Then*

(i) *there exists a unique invariant control set* $C \subset \mathbb{P}^{n-1}_{\mathbb{K}}$, *given by*

$$C := \bigcap_{\xi \in \mathbb{P}^{n-1}_{\mathbb{K}}} \operatorname{cl} \mathcal{O}^+(\xi); \tag{7.1}$$

(ii) *there exists a unique open control set* $C^- \subset \mathbb{P}^{n-1}_{\mathbb{K}}$, *which satisfies*

$$\operatorname{cl} C^- = C^* := \bigcap_{\xi \in \mathbb{P}^{n-1}_{\mathbb{K}}} \operatorname{cl} \mathcal{O}^-(\xi)\,. \tag{7.2}$$

Moreover, it holds that $\operatorname{core}(C^-) = C^-$.

Proof. (i) To begin with, it has to be shown that C as defined by (7.1) is not empty. Let $u \in U^{t^*}_{reg}$ and $|\sigma(\Phi_u(t,0))| = \{r_1, \dots, r_\nu\}$ with $r_1 < \cdots < r_\nu$. By Proposition 6.7 there exists a control set D such that $\mathbb{P}GE(r_\nu, u) \subset \operatorname{core}(D)$. By Lemma 4.3 it holds for all $\xi \notin \mathbb{P} \bigoplus_{j=1}^{\nu-1} GE(r_j, u)$ that

$$\omega^+(\xi, u) \subset \mathbb{P}GE(r_\nu, u)\,. \tag{7.3}$$

Note that the set of ξ for which (7.3) holds is generic in $\mathbb{P}^{n-1}_{\mathbb{K}}$. By forward accessibility we may steer from any point into that generic set, as the interior of each forward orbit is open, and it follows that $\mathbb{P}GE(r_\nu, u) \cap \operatorname{cl} \mathcal{O}^+(\xi) \neq \emptyset$ for all $\xi \in \mathbb{P}^{n-1}_{\mathbb{K}}$. However, we know that $\mathbb{P}GE(r_\nu, u) \subset \operatorname{core}(D)$ so that $\mathcal{O}^+(\xi) \cap \operatorname{core}(D) \neq \emptyset$, and therefore $\operatorname{core}(D) \subset \mathcal{O}^+(\xi)$. In all, we have obtained that $\operatorname{core}(D) \subset C$. By the definition of C, it follows furthermore that $D = C$, for if $\xi \in C$, then $\operatorname{core}(D) \subset \operatorname{cl} \mathcal{O}^+(\xi)$, and also $\xi \in \operatorname{cl} \mathcal{O}^+(\eta)$ for all $\eta \in D$, so that $\xi \in D$. C is therefore a closed control set with nonempty interior and invariant by Proposition 6.3. As $C \subset \operatorname{cl} \mathcal{O}^+(\eta)$ for every $\eta \in \mathbb{P}^{n-1}_{\mathbb{K}}$, there can be no other invariant control set.

(ii) Let D be the control set with $\mathbb{P}GE(r_1, u) \subset \operatorname{core}(D)$. Recall that by Proposition 5.4, $\hat{\mathcal{O}}^-(\xi) \neq \emptyset$ for all $\xi \in \mathbb{P}^{n-1}_{\mathbb{K}}$. Hence for all $\xi \in \mathbb{P}^{n-1}_{\mathbb{K}}$, we may choose a control $v \in \operatorname{int} U^{t^*}$ and $\xi_2 \in \mathbb{P}^{n-1}_{\mathbb{K}}$ such that (ξ_2, v) is a regular pair, $\xi_2 \notin \mathbb{P} \bigoplus_{j=2}^{\nu} GE(r_j, u)$, and $\xi = \xi(t^*; \xi_2, v)$. By Lemma 4.3 (iii) it follows that $\omega^-(\xi_2, u) \subset \mathbb{P}GE(r_1, u)$. Thus there exists a $\xi_3 \in \operatorname{core}(D)$ such that $\xi_3 \in \hat{\mathcal{O}}^-(\xi)$. Since by Proposition 6.6 $\operatorname{core}(D) \subset \hat{\mathcal{O}}^-(\xi)$ for $\xi \in \operatorname{core}(D)$, it follows that $\operatorname{core}(D) \subset \hat{\mathcal{O}}^-(\xi)$ for all $\xi \in \mathbb{P}^{n-1}_{\mathbb{K}}$ and thus $\operatorname{core}(D) \subset C^*$.

In particular, for $\eta \in D$ it is obtained that $\operatorname{core}(D) \subset \hat{\mathcal{O}}^-(\eta)$ and thus $\eta \in \operatorname{core}(D)$. This implies that D is an open control set by Proposition 6.6 (i).

Finally, it has to be shown that $\operatorname{cl} D = C^*$. Let $\eta \in C^* \setminus \operatorname{cl} D$. As $\eta \in C^*$ it follows that $\eta \in \operatorname{cl} \mathcal{O}^-(\xi)$ for all $\xi \in D$. Hence, in every neighborhood of η there exists a ζ such that $D \subset \mathcal{O}^+(\zeta)$. On the other hand, $D \subset \hat{\mathcal{O}}^-(\zeta)$ and thus $\zeta \in D$, by

maximality. This, however, implies that $\eta \in \operatorname{cl} D$, a contradiction. Thus $\operatorname{cl} D = C^*$ and hence $C^- = D$ is the only open control set contained in C^*.

It remains to show that there is no other open control set in $\mathbb{P}^{n-1}_{\mathbb{K}}$. If D is a control set with $\operatorname{core}(D) \neq \emptyset$, then by Proposition 6.5 there exists $\xi \in \operatorname{core}(D)$, $t \in \mathbb{N}$, $u \in U^t$ such that (ξ, u) is regular and $\xi = \xi(t; \xi, u)$. By Proposition 6.9 we may assume that u is universally regular. Let $|\sigma(\Phi_u(t,0))| = \{r_1, \dots, r_\nu\}$, $r_1 < \dots < r_\nu$. Thus $\xi \in \mathbb{P}GE(r_i, u)$ for some $i > 1$, for otherwise $\xi \in C^-$, which may be seen using the previous arguments. Now for $\eta \in \mathbb{P}(GE(r_i,u) \oplus GE(r_1,u)) \backslash \mathbb{P}GE(r_1,u)$ it holds that $\omega^+(\eta, u) \subset \mathbb{P}GE(r_i, u)$ by Lemma 4.3 (iii). Thus $\partial D \cap \mathbb{P}(GE(r_i,u) \oplus GE(r_1,u)) \subset D$ and D is not open. □

COROLLARY 7.2. *Let $\mathbb{K} = \mathbb{R}, \mathbb{C}$. Assume that system (2.5) is forward accessible. If there exists exactly one control set D in $\mathbb{P}^{n-1}_{\mathbb{K}}$, then $D = \operatorname{core}(D) = \mathbb{P}^{n-1}_{\mathbb{K}}$.*

Proof. By Theorem 7.1 it follows that $D = C = C^- = \operatorname{core}(C^-)$. Thus D is open and closed and not empty, which shows that $D = \operatorname{core}(D) = \mathbb{P}^{n-1}_{\mathbb{K}}$. □

Using the fact that the invariant control set is closed, we may prove the following result on its connectedness.

PROPOSITION 7.3. *Let $\mathbb{K} = \mathbb{R}, \mathbb{C}$. Assume that (2.5) is forward accessible. Then the invariant control set C is connected.*

Proof. For each connected component Y of C and $u \in U_{inv}$ the image $\mathbb{P}A(u)Y$ is connected as the continuous image of a connected set. Since $\mathcal{O}^+_1(\xi)$ is also connected for all $\xi \in Y$ and $\operatorname{cl} \mathcal{O}^+_1(Y) \subset C$, it follows that there exists a connected component Y' of C such that $\operatorname{cl} \mathcal{O}^+_1(Y) \subset Y'$. Let $u \in U^{t^*}_{reg}$. For the connected component of C satisfying $\mathbb{P}GE(r_n, u) \subset Y$, which exists as $\mathbb{P}GE(r_n, u)$ is connected, it follows that $\mathcal{O}^+_{t^*}(Y) \subset Y$, but then $C = \operatorname{cl} \mathcal{O}^+(Y) \subset \bigcup_{s=1}^{t^*} \operatorname{cl} \mathcal{O}^+_s(Y) \subset C$, so that there are $k \leq t^*$ connected components of C. Hence we may assume that the connected components of C are ordered in such a way that

$$\operatorname{cl} \mathcal{O}^+_1(Y_i) \subset Y_{i+1}, \quad i = 1, \dots, k-1\,,$$

and

$$\operatorname{cl} \mathcal{O}^+_1(Y_k) \subset Y_1\,.$$

Let $v \in U^{kt^*+1}_{reg}$. Then, by universal regularity of v,

$$\mathbb{P}GE(r_n, v) \subset \operatorname{core}(C)\,,$$

and for every $i = 1, \dots, k$ it holds that

$$\xi \in Y_i \Rightarrow \xi(kt^* + 1; \xi, v) \in Y_{i \bmod k+1}\,.$$

But if $\xi \in \mathbb{P}GE(r_n, v)$ then clearly $\xi(kt^* + 1; \xi, v) \in \mathbb{P}GE(r_n, v)$ and $\mathbb{P}GE(r_n, v)$ is connected. So $i = i \bmod k + 1$ and thus $k = 1$. □

Remark 7.4. (i) The uniqueness of the invariant control set system (2.5) has been shown in [32] for the case in which all system matrices $A(u)$ are invertible. The proof relies, however, on a theorem in [7], where it has to be assumed that the group generated by $\{A(u); u \in U\}$ is a Lie group. We have shown that in our case these assumptions are not necessary.

(ii) From the proof of Theorem 7.1 it follows that for all $t \geq t^*$, $u \in U^t_{reg}$ we have

$$\mathbb{P}GE(r_1(\Phi_u(t,0)), u) \subset C^-\,, \tag{7.4}$$

$$\mathbb{P}GE(r_n(\Phi_u(t,0)), u) \subset C\,. \tag{7.5}$$

The last argument in the proof of Theorem 7.1 contains the fundamental idea on what order is in a sense natural on the set of control sets.

Let D_1, D_2 be control sets in $\mathbb{P}^{n-1}_{\mathbb{K}}$ for the system (2.5). We define

$$D_1 \leq D_2 :\Leftrightarrow \text{ there exist } \xi \in D_1,\ t \in \mathbb{N},\ u \in U^t \text{ such that } \xi(t;\xi,u) \in D_2\,. \tag{7.6}$$

A priori, this defines only a partial order on the control sets. What is, however, evident at this point is the following.

PROPOSITION 7.5. *Let $\mathbb{K} = \mathbb{R}, \mathbb{C}$. Assume that system* (2.5) *is forward accessible.*

(i) *C is the unique maximal control set with respect to the order "$\leq$" on the control sets.*

(ii) *C^- is the unique minimal control set with respect to the order "$\leq$" on the control sets.*

Proof. (i) is immediate from (7.1), while (ii) follows from (7.2). □

8. Main control sets. In this section we give sufficient conditions for which it is possible to recover exactly those results that are known in the continuous-time case. Namely, the number of control sets with nonvoid interior is bounded by n, the dimension of the state space; the control sets are completely ordered with respect to the order defined in the previous section; and to each control set an index may be assigned as the sum of the algebraic multiplicities of all the eigenvalues corresponding to a universally regular u, whose generalized eigenspace is projected into the core of that control set. Furthermore, in the complex or real invertible case, the control sets are connected.

We begin with the following definition. For every $t \in \mathbb{N}$, $u \in U^t$, we will from now on consider the set $\{r_1, \ldots, r_n\}$, where $r_i \in |\sigma(\Phi_u(t,0))|$, $r_1 \leq \cdots \leq r_n$, and each r_i occurs as often as the sum of the algebraic multiplicities of those $\lambda \in \sigma(\Phi_u(t,0))$ with $r_i = |\lambda|$. We define for $i = 1, \ldots, n$

$$Q_i(t) := \bigcup_{u \in U^t_{reg}} \mathbb{P}GE(r_i, u), \quad Q_i := \bigcup_{t=1}^{\infty} Q_i(t). \tag{8.1}$$

Furthermore, for a map $A : \tilde{U} \to \mathbb{R}^{n\times n}$, we introduce the following index, which is a measure of what sets of rank deficient matrices separate $A(\operatorname{int} U)$. Define the sets

$$U_i := \{u \in U;\, \dim \operatorname{Ker} A(u) \leq i\} \tag{8.2}$$

and the singularity index

$$\bar{i}(A,U) := \min\{i;\, \operatorname{int} U_i \text{ is pathwise connected}\}\,. \tag{8.3}$$

Note that all the sets U_i are generic in U, as $U_i \supset U_{inv} \neq \emptyset$. Moreover, $\mathbb{K} = \mathbb{C}$ implies that $\bar{i}(A,U) = 0$, as proper analytic subsets are nowhere separating in the complex case; see [36, Proposition 7.4]. The significance of the indices $i > \bar{i}(A,U)$ is explained in the following proposition.

PROPOSITION 8.1. *Let $\mathbb{K} = \mathbb{R}, \mathbb{C}$. Assume that* (2.5) *is forward accessible. If $i > \bar{i}(A,U)$ then Q_i is contained in a precontrol set.*

Proof. Let $u, v \in U^t_{reg}$, where we assume without loss of generality that the length of the sequences is the same and that $t \geq t^* + 1$. Denote $u = (u(0), u')$ and $v = (v(0), v')$ where $u(0), v(0) \in \operatorname{int} U_{inv}$ and $u', v' \in \operatorname{int} U^{t-1}_{inv}$. Let $\gamma_1 : [0,1] \to \operatorname{int} U$ be a continuous path connecting $u(0)$ and $v(0)$. γ_1 can be chosen piecewise analytic.

Hence we may assume there is a finite number of points τ_j, $j = 1, \dots, k$, such that $\det(A(\gamma_1(\tau_j))) = 0$. By definition we may assume that $\dim \operatorname{Ker} A(\gamma_1(\tau_j)) \leq \bar{i}(A, U)$ for $j = 1, \dots, k$. By Proposition 5.3, the set

$$Z := \left\{ u \in U_{reg}^{t-1};\, \Phi_u(t-1, 0) \operatorname{Im} A(\gamma_1(\tau_j)) \cap \operatorname{Ker} A(\gamma_1(\tau_j)) = \{0\} \text{ for } j = 1, \dots, k \right\} \tag{8.4}$$

is generic in $\operatorname{int} U^{t-1}$ since it is the finite intersection of generic sets. We may therefore choose a continuous path $\gamma_2 : [0, 1] \to \operatorname{int} U^{t-1}$ such that $\gamma_2(0) = u'$ and $\gamma_2(\tau) \in Z$ for all $\tau \in (0, 1]$. Let $\tilde{u}' := \gamma_2(1)$.

Now consider the path:

$$\gamma_3 : [0, 2] \to \operatorname{int} U^t, \tag{8.5}$$

$$\gamma_3(\tau) = \begin{cases} (u(0), \gamma_2(\tau)), & 0 \leq \tau \leq 1, \\ (\gamma_1(\tau - 1), \tilde{u}'), & 1 \leq \tau \leq 2. \end{cases} \tag{8.6}$$

For $0 \leq \tau \leq 1$, $\gamma_3(\tau)$ is universally regular as $\gamma_3(0) = u$ and $\gamma_2(\tau) \in U_{reg}^{t-1}$ for $\tau \in (0, 1]$. Furthermore, we obtain for $1 \leq \tau \leq 2$ and $i > \bar{i}(A, U)$ that $r_i(\Phi_{\gamma_3(\tau)}(t, 0)) > 0$. This is clear if $\det(A(\gamma_1(\tau - 1))) \neq 0$. For $\tau = 1 + \tau_j$, $j = 1, \dots, k$, we have that

$$\Phi_{\tilde{u}'}(t-1, 0) \operatorname{Im} A(\gamma_1(\tau_j)) \cap \operatorname{Ker} A(\gamma_1(\tau_j)) = \{0\},$$

and hence for the eigenvalue 0 of $\Phi_{\tilde{u}'}(t-1, 0) A(\gamma_1(\tau_j))$, algebraic and geometric multiplicity coincide.

In all we have constructed a continuous path from $u = (u(0), u')$ to $(v(0), \tilde{u}')$ such that $r_i > 0$ along this path if $i > \bar{i}(A, U)$ and furthermore $\tilde{u}' \in Z$ can be chosen arbitrarily close to u'. We wish to continue this procedure in an inductive manner. Assume that for some $0 < j < t - 1$ we have constructed a continuous path from u to $(v'(0), \dots, v'(j-1), v(j), w(j+1), \dots, w(t-1)) \in U_{reg}^t$, where $(v'(0), \dots, v'(j-1))$ is arbitrarily close to $(v(0), \dots, v(j-1))$ and $(w(j+1), \dots, w(t-1))$ is arbitrarily close to $(u(j+1), \dots, u(t-1))$. Furthermore, along this path the ith entry in the ordered spectrum is never 0 if $i > \bar{i}(A, U)$.

Since for all $w_1' \in U^{t-j-2}, w_2' \in U, w_3' \in U_{inv}^j$ the Jordan structures of

$$\Phi_{w_1'}(t-j-2, 0) A(w_2') \Phi_{w_3'}(j+1, 0)$$

and

$$\Phi_{w_3'}(j+1, 0) \Phi_{w_1'}(t-j-2, 0) A(w_2')$$

coincide by similarity, we may work as in the first part to construct a path with the desired properties from $(w(j+1), w(j+2), \dots, w(t-1), v'(0), \dots, v'(j-1), v(j))$ to $(v(j+1), \tilde{w})$, where $\tilde{w}$ may be chosen arbitrarily close to $(w(j+2), \dots, w(t-1), v'(0), \dots, v'(j-1), v(j))$. Note that this rearrangement does not destroy universal regularity by Lemma 3.5. By rearranging the sequence in the original order, we obtain the desired path in the jth step.

Continuing this procedure we obtain a continuous path γ_4 from u to $\tilde{v}$, where $\tilde{v}$ may be chosen arbitrarily close to v. As $v \in U_{reg}^t$, the path may be assumed to go from u to v.

By construction, $r_i(\Phi_{\gamma_4(\tau)}(t, 0)) > 0$ along this path if $i > \bar{i}(A, U)$. Now consider the continuous paths

$$\gamma_5, \gamma_6 : [0, 1] \to \operatorname{int} U^t, \tag{8.7}$$

$$\gamma_5(\tau) = (\gamma_4(\tau), u), \tag{8.8}$$

$$\gamma_6(\tau) = (\gamma_4(1 - \tau), v) \tag{8.9}$$

connecting (u,u) with (v,u) and (u,v) with (v,v), respectively. As u and v are universally regular, we have that for $i > \bar{i}(A,U)$ and all $\tau \in [0,1]$ the sets

$$\mathbb{P}GE(r_i(\tau), \gamma_5(\tau)),\ \mathbb{P}GE(r_i(\tau), \gamma_6(\tau)) \tag{8.10}$$

are regular. Hence each of the sets

$$\bigcup_{\tau\in[0,1]} \mathbb{P}GE(r_i(\tau), \gamma_5(\tau))\,, \tag{8.11}$$

$$\bigcup_{\tau\in[0,1]} \mathbb{P}GE(r_i(\tau), \gamma_6(\tau)) \tag{8.12}$$

is contained in an open precontrol set by Proposition 6.8. Furthermore, it holds that

$$\mathbb{P}GE(r_i, (v,u)) = \mathbb{P}\Phi_v(t,0)GE(r_i, (u,v))\,, \tag{8.13}$$

which is clear by the relation $\Phi_{(u,v)}(2t,0) = \Phi_v(t,0)\Phi_{(v,u)}(2t,0)\Phi_v(t,0)^{-1}$. By symmetry, we obtain furthermore that

$$\mathbb{P}GE(r_i, (u,v)) = \mathbb{P}\Phi_u(t,0)GE(r_i, (v,u)). \tag{8.14}$$

Thus it may be concluded that

$$\bigcup_{\tau\in[0,1]} \mathbb{P}GE(r_i(\tau), \gamma_5(\tau)) \cup \bigcup_{\tau\in[0,1]} \mathbb{P}GE(r_i(\tau), \gamma_6(\tau)) \tag{8.15}$$

is contained in a precontrol set. The proof is completed by fixing one universally regular control and noting that we may apply the procedure of this proof for a path to any other universally regular control. □

Remark 8.2. In the preceding theorem we did not make a statement about connectedness. In Example 6.12, in the case $\mathbb{K} = \mathbb{R}$, we have seen a system where indeed the core of the invariant control set C is not connected. On the other hand, we know by Proposition 6.9 and by the fact that Q_1 is contained in the open control set C^- that for every connected component W of core(C) it holds that $Q_2 \cap W \neq \emptyset$. Note that in this example the index $\bar{i}(A,U) = 1$ as $\mathrm{rk}A(u) \geq 1$ for all $u \in U$, and the controls $(\frac{1}{2}, -\varepsilon)$ $(\frac{1}{2}, \varepsilon)$ can only be connected through a point of the form $(a,0)$ which leads to a rank drop.

The following statement includes in particular the case of real invertible and complex systems.

PROPOSITION 8.3. *Let* $\mathbb{K} = \mathbb{R}, \mathbb{C}$. *Assume that* (2.5) *is forward accessible. If* $\bar{i}(A,U) = 0$, *then for each* $i = 1, \ldots, n$, *the set* Q_i *is contained in a connected component of* core(D) *for some control set* D.

Proof. Fix $u_1, u_2 \in U^t_{reg}$ (where again, without loss of generality, the length of the control sequences is the same) and let $\gamma : [0,1] \to \operatorname{int} U^t_{inv}$ be a continuous connecting path. Such a path exists as $\operatorname{int} U^t_{inv}$ is connected, but there may be $\tau \in [0,1]$ such that $\gamma(\tau)$ is not universally regular. Then the path

$$\gamma_2 : [0,2] \to \operatorname{int} U^{2t}, \tag{8.16}$$

$$\gamma_2(\tau) = \begin{cases} (u_1, \gamma(\tau)), & 0 \leq \tau \leq 1, \\ (\gamma(\tau - 1), u_2), & 1 \leq \tau \leq 2, \end{cases} \tag{8.17}$$

is a continuous path connecting (u_1, u_1) and (u_2, u_2) in $\operatorname{int} U^{2t}$. By Lemma 3.5, the invertibility of $A(\gamma(\tau))$ and the universal regularity of u_1, u_2, it follows furthermore

that $\gamma_2(\tau)$ is universally regular for all $\tau \in [0,2]$. The assertion now follows due to Proposition 6.8. □

THEOREM 8.4. *Let $\mathbb{K} = \mathbb{R}, \mathbb{C}$. Assume that (2.5) is forward accessible and that $\bar{i}(A,U) \leq 1$. Then the following statements hold.*

(i) *The number κ of control sets $D_1, \dots, D_\kappa$ with nonempty interior satisfies*

$$1 \leq \kappa \leq n\,. \tag{8.18}$$

(ii) *For every $t > 0$, $u \in U^t_{reg}$, $r \in |\sigma(\Phi_u(t,0))|$ there exists a control set D_i, $1 \leq i \leq \kappa$, such that*

$$\mathbb{P}GE(r,u) \subset \operatorname{core}(D_i)\,. \tag{8.19}$$

(iii) *The core of the control sets $D_1, \dots, D_\kappa$ consists of exactly those elements $\xi \in \mathbb{P}^{n-1}_{\mathbb{K}}$ which are eigenvectors to a nonzero eigenvalue of some $\Phi_u(t,0)$ where (ξ,u) is a regular pair. If $U = U_{inv}$ the control may be chosen to be universally regular.*

(iv) *For every $t > 0$, $u \in U^t$, $r \in |\sigma(\Phi_u(t,0))|$ there exists a $j \in \{1,\dots,\kappa\}$ with $\mathbb{P}GE(r,u) \cap \operatorname{cl} D_j \neq \emptyset$. Also, for every $t \in \mathbb{N}$, $u \in U^t$, and every $j = 1,\dots,\kappa$, there exists an $r \in |\sigma(\Phi_u(t,0))|$ with $\mathbb{P}GE(r,u) \cap \operatorname{cl} D_j \neq \emptyset$.*

Proof.

(i) Let D be a control set with $\operatorname{core}(D) \neq \emptyset$. By Proposition 6.9 there exists an $i \in \{1,\dots,n\}$ such that $Q_i \cap D \neq \emptyset$. If $i = 1$ then Q_i is contained in a control set by Remark 7.4 (ii). Using Proposition 8.1 it follows that $Q_i \subset D$. Thus the number of control sets with nonempty interior is bounded by n, the number of the sets Q_i.

(ii) This follows from Corollary 4.6 and (i).

(iii) This follows from Propositions 6.5 and 6.11.

(iv) The statement is clear for $u \in U^t_{reg}$. If $t < t^*$ choose l such that $lt \geq t^*$ and consider the control $(u)^l$. If $t \geq t^*$ and $u \notin U^t_{reg}$ by genericity of the universally regular controls, there exists a sequence $(u_k)_{k\in\mathbb{N}} \subset U^t_{reg}$ with $\lim_{k\to\infty} u_k = u$. Using again the continuity properties of the eigenprojections (Chapter II.8 in [13]) it follows that for $r_i \in |\sigma(\Phi_u(t,0))|$ it holds that $\mathbb{P}GE(r_i,u) \cap \operatorname{cl} Q_i \neq \emptyset$. This implies the assertion. □

It has been shown that under the assumption of the previous theorem for every $i \in \{1,\dots,n\}$, there exists a control set D such that $Q_i \subset D$. From now on, the following terminology is used.

DEFINITION 8.5 (main control set). *Let $\mathbb{K} = \mathbb{R}, \mathbb{C}$. Assume that (2.5) is forward accessible. A control set D is called* main control set *if for every index $1 \leq i \leq n$ it holds that*

$$Q_i \cap D \neq \emptyset \;\Rightarrow\; Q_i \subset D\,.$$

The result of the previous theorem may then be paraphrased by saying that in the case where $\bar{i}(A,U) \leq 1$, i.e., in particular, in the complex or real invertible case, the only control sets with nonempty core are main control sets. Let us now examine further properties of main control sets. Recall that $n(\lambda,u)$ denotes the dimension of the generalized eigenspace of the eigenvalue λ of $\Phi_u(t,0)$.

THEOREM 8.6. *Let $\mathbb{K} = \mathbb{R}, \mathbb{C}$. Assume that (2.5) is forward accessible. Then the following holds.*

(i) *If $\bar{i}(A,U) = 0$ then the core of every main control set is connected.*

(ii) *The main control sets are completely ordered with respect to the order "$\leq$."*

(iii) *For each main control set D the number*

$$m(D) = \sum_{\mathbb{P}GE(\lambda,u)\subset \mathrm{core}(D)} n(\lambda,u) \tag{8.20}$$

is independent of $u \in U^t_{reg}$ and $t \in \mathbb{N}$.

Proof.

(i) Let D be a main control set. For any open set $W \subset \mathrm{core}(D)$ there exists an i such that $W \cap Q_i \neq \emptyset$ by Proposition 6.9. As the sets Q_i are contained in connected components of the core by Proposition 8.3, it is sufficient to show the following. If there exists $i, j \in \{1, \dots, n\}$, $i \neq j$, such that $Q_i, Q_j \subset D$, then there exists a $1 \leq k \leq n$ such that $Q_k \subset D$, $Q_i \cap Q_k \neq \emptyset$, and $Q_j \cap Q_k \neq \emptyset$.

Let $\xi \in Q_i$, $\eta \in Q_j$. Hence there exist $t, s \in \mathbb{N}$, $u \in U^t_{reg}$, $v \in U^s_{reg}$ such that

$$\eta = \xi(t; \xi, u)\,, \tag{8.21}$$

$$\xi = \xi(s; \eta, v)\,. \tag{8.22}$$

(Indeed, if $\xi \in \mathbb{P}GE(r_i, u')$ for $u' \in U^{t'}_{reg}$ and $\xi = \xi(t'; \xi', u')$, then by the implicit function theorem there is an open neighborhood of ξ' that can be steered to ξ with universally regular controls. Into this neighborhood we can steer from η using an invertible control. A concatenation yields the desired control.)

Now $(v, u), (u, v) \in U^{t+s}_{reg}$. Furthermore, as $\sigma(\Phi_v(s, 0)\, \Phi_u(t, 0)) = \sigma(\Phi_u(t, 0)\, \Phi_v(s, 0))$ it follows that there exists a $\lambda \in \mathbb{C}^*$ such that

$$\xi \in \mathbb{P}GE(\lambda, (v, u)) \tag{8.23}$$

and

$$\eta \in \mathbb{P}GE(\lambda, (u, v))\,. \tag{8.24}$$

If $|\lambda| = r_k(\Phi_{(u,v)}(s + t, 0)) = r_k(\Phi_{(v,u)}(s + t, 0))$, it follows that $\xi, \eta \in Q_k$. Hence $Q_k \subset D_i$ and $Q_i \cup Q_j \cup Q_k$ is contained in a connected component of the core of D.

(ii) Let D_1, D_2 be two main control sets. Then there exists $Q_i \subset D_1$, $Q_j \subset D_2$. Let us assume that $i \leq j$. Then we claim that $D_1 \leq D_2$. Indeed, let $u \in U^{t^*}_{reg}$ and $\xi \in \mathbb{P}(GE(r_i, u) \oplus GE(r_j, u))$. As $r_i \leq r_j$ it follows that

$$\omega^+(\xi, u) \subset \begin{cases} \mathbb{P}GE(r_i, u) & \text{if } \xi \in \mathbb{P}GE(r_i, u), \\ \mathbb{P}GE(r_j, u) & \text{otherwise.} \end{cases} \tag{8.25}$$

As $\mathbb{P}GE(r_i, u) \subset \mathrm{core}(D_1)$ there exists $\eta \in D_1$ such that $\omega^+(\eta, u) \subset \mathrm{core}(D_2)$. This proves the assertion.

(iii) It is clear that

$$m(D) = \#\{1 \leq i \leq n;\, Q_i \subset D\}\,, \tag{8.26}$$

which is independent of $u \in U^t_{reg}$, $t \in \mathbb{N}$. □

As a result of Theorem 8.6 the following definition is straightforward.

DEFINITION 8.7 (index of a main control set). *Assume that* (2.5) *is forward accessible. For a main control set $D \subset \mathbb{P}^{n-1}_{\mathbb{K}}$ the number $m(D)$ is called the* index of the control set *D.*

It remains to analyze the case where $\bar{i}(A, U) > 1$. By the discussion up to this point it is clear that for $i = 1, n$ and $i > \bar{i}(A, U)$, there exists a main control set D_i

such that $Q_i \subset D_i$. For the remainder of the indices the question of whether there exists a unique control set with this property must for the moment be left unresolved.

To summarize, we have obtained the following picture of the control structure of the system on projective space. For a map A and a set of admissible controls U such that the system on $\mathbb{P}^{n-1}_{\mathbb{K}}$ is forward accessible and $\bar{i}(A,U) \leq 1$, there exists a sequence of indices $i_1, \ldots, i_\kappa$, with $\sum_{j=1}^{\kappa} i_j = n$.

To each index i_j there exists a control set D_j such that $m(D_j) = i_j$. More specifically, it is shown in [56] that if we write

$$\mu_j = \sum_{l=1}^{j} i_l$$

for $j = 1, \ldots, \kappa$, then

$$\bigcup_{i=\mu j-1+1}^{\mu_j} Q_i \subsetneqq \operatorname{core}(D_j),$$

where equality holds if $U = U_{inv}$. So the numbers from 1 to n are partitioned into κ noninterlacing subsequences which represent the indices i such that $Q_i \subset \operatorname{core}(D_j)$:

$$\underbrace{1, \ldots, \mu_1}_{D_1}, \underbrace{\mu_1 + 1, \ldots, \mu_2}_{D_2}, \underbrace{\mu_2 + 1, \ldots}_{\cdots\cdots}, \cdots, \underbrace{\ldots, \mu_{\kappa-1}}_{\cdots\cdots}, \underbrace{\mu_{\kappa-1} + 1, \ldots, n}_{D_\kappa}.$$

The order between the main control sets is simply reflected in the order of the subsequences. In case there are control sets with nonempty core that are not main control sets, this can be extended in a natural way by considering indices that do not correspond to main control sets but to control set clusters; see [53].

With this notation we may formulate the following invariance principle which also motivates the interpretation of control sets and their indices as an extension of eigenspaces and their dimension. For a proof, we refer to [53] or [56].

THEOREM 8.8. *Let $\mathbb{K} = \mathbb{R}, \mathbb{C}$ and assume that (2.5) is forward accessible. For $u \in U^{\mathbb{N}}$ define $d(u) := \max_{t \in \mathbb{N}} \dim \ker \Phi_u(t,0)$. Let $\mu_1, \ldots, \mu_\kappa$ be the indices for the control set structure as described above.*

(i) *For every main control set D_j with $\mu_{j-1} > d(u)$, there exists a linear subspace $X_j(u)$ satisfying*

$$\dim X_j(u) = m(D_j) = \mu_j - \mu_{j-1},$$

$$\text{for all } t \in \mathbb{N} \text{ it holds that } \mathbb{P}\Phi_u(t,0)X_j(u) \subset \operatorname{cl} D_j.$$

(ii) *If $d(u) > 0$ and a main control set D_j exists such that $\mu_{j-1} < d(u) < \mu_j$ then there exists a linear subspace $X_j(u)$ satisfying*

$$\dim X_j(u) = \mu_j - d(u),$$

$$\text{for all } t \in \mathbb{N} \text{ it holds that } \mathbb{P}\Phi_u(t,0)X_j(u) \subset \operatorname{cl} D_j.$$

9. Characteristic exponents. Up to now we have described the control structure of a system on projective space. With the insight that has been gained, let us now discuss properties of the set of characteristic exponents that may be deduced from our knowledge about the control sets.

For systems of the form (2.2) let $\lambda(x_0, u)$ denote the Lyapunov exponent corresponding to the initial value $(0, x_0) \in \mathbb{N}\times\mathbb{K}^n$ and the sequence $A(u(\cdot)) \in \ell^\infty(\mathbb{N}, \mathbb{K}^{n\times n})$ determined by $u \in U^{\mathbb{N}}$, i.e. the exponential growth rate of the corresponding solution:

$$\lambda(x_0, u) = \limsup_{t\to\infty} \frac{1}{t} \log \|\Phi_u(t, 0)x_0\|,$$

while $\beta(u)$ denotes the Bohl exponent determined by $u \in U^{\mathbb{N}}$:

$$\beta(u) = \limsup_{s,t-s\to\infty} \frac{1}{t-s} \log \|\Phi_u(t, s)\| .$$

Note that it is sufficient to study Lyapunov exponents corresponding to the initial time 0, as control sequences may be shifted; i.e., the Lyapunov exponent to the initial value (t, x_t) and the control sequence $u \in U^{\mathbb{N}}$ may be recaptured by studying the initial value $(0, x_t)$ and the control sequence $v \in U^{\mathbb{N}}$ defined by $v(s) = u(s + t)$. It is known that in general $\max_{x_0\neq 0} \lambda(x_0, u) \leq \beta(u)$ where strict inequality is possible; see [24].

Floquet exponents are the Lyapunov exponents corresponding to periodic sequences $u \in U^{\mathbb{N}}$. For $t \in \mathbb{N}$, $u \in U^t$ it is easy to see that the set of Floquet exponents determined by the t-periodic continuation of u is given by

$$\sigma_{Fl}(u) := \left\{ \frac{1}{t} \log r;\ r \in |\sigma(\Phi_u(t, 0))| \right\}, \tag{9.1}$$

where we continue to use the convention $\log 0 = -\infty$. For a system of the form (2.2) determined by the map A and the set of admissible controls U, the Lyapunov spectrum is defined as the union

$$\Sigma_{Ly}(A, U) := \{\lambda(x_0, u)\,;\, x_0 \in \mathbb{K}^n\backslash\{0\}, u \in U^{\mathbb{N}}\} . \tag{9.2}$$

The Floquet spectrum of (2.2) is defined by

$$\Sigma_{Fl}(A, U) := \bigcup_{t\geq 1, u\in U^t} \sigma_{Fl}(u). \tag{9.3}$$

Furthermore, we define

$$\Sigma_{Fl,i}(A, U) := \left\{ \frac{1}{t} \log r_i(\Phi_u(t, 0)); t \geq 1\,,\ u \in U^t \right\}. \tag{9.4}$$

Recall that $\mathbb{P}GE(r, u)$ is called *regular* if $u = (u_1, u_2)$ and (ξ, u_2) is a regular pair for all $\xi \in \mathbb{P}\Phi_{u_1}(t_1, 0)GE(r, u)$. For a control set D with nonempty core we define the Floquet spectrum of D to be

$$\Sigma_{Fl}(D) := \bigcup_{t\geq 1, u\in U^t} \left\{ \frac{1}{t} \log r;\ r \in |\sigma(\Phi_u(t, 0))|,\ \mathbb{P}GE(r, u) \subset \operatorname{core}(D) \right.$$

$$\left. \text{and } \mathbb{P}GE(r, u) \text{ is regular} \right\}. \tag{9.5}$$

Finally, we consider the Bohl spectrum of (2.2) defined as the set of all Bohl exponents the system can generate:

$$\Sigma_{Bo}(A, U) := \{\beta(u); u \in U^{\mathbb{N}}\} . \tag{9.6}$$

Let us begin by explaining how to obtain the Lyapunov exponent $\lambda(x_0, u)$ from the trajectory $\xi(\cdot; \mathbb{P}x_0, u)$ of the projected system. For $\xi \in \mathbb{P}^{n-1}_{\mathbb{K}}$, $u \in U(\xi)$ define

$$q(\xi, u) := \log \frac{\|A(u)x\|}{\|x\|}, \quad \text{where} \quad x \neq 0,\ \mathbb{P}x = \xi. \tag{9.7}$$

This is well defined, as multiplication of x with a nonzero scalar does not alter the value of $q(\xi, u)$. For $\xi \in \mathbb{P}^{n-1}_{\mathbb{K}}$, $t \in \mathbb{N}$, $u \in U^t(\xi)$ define

$$J(t; \xi, u) = \sum_{s=0}^{t-1} q(\xi(s; \xi, u), u(s)). \tag{9.8}$$

Then we obtain the following expression for Lyapunov exponents.

LEMMA 9.1. *Let $\mathbb{K} = \mathbb{R}, \mathbb{C}$. For $x_0 \in \mathbb{K}^n \setminus \{0\}$, $u \in U^{\mathbb{N}}$ it holds that*

$$\lambda(x_0, u) = \begin{cases} \limsup_{t\to\infty} \frac{1}{t} J(t; \mathbb{P}x_0, u) & \text{if} \quad u \in U^{\mathbb{N}}(x_0), \\ -\infty & \text{otherwise.} \end{cases} \tag{9.9}$$

Proof. This can be shown by a straightforward calculation. ∎

The previous lemma shows that we may speak of the Lyapunov exponent corresponding to $(\xi_0, u) \in \mathbb{P}^{n-1}_{\mathbb{K}} \times U^{\mathbb{N}}$, which we denote by $\lambda(\xi_0, u)$.

10. The Floquet spectrum. The Floquet spectrum is closely related to the structure of the control sets examined up to now. In order to explore this relationship we need a controllability property in the cores of control sets. Let $\mathbb{K} = \mathbb{R}, \mathbb{C}$, and consider system (2.5) on $\mathbb{P}^{n-1}_{\mathbb{K}}$. Consider the function

$$h : \mathbb{P}^{n-1}_{\mathbb{K}} \times \mathbb{P}^{n-1}_{\mathbb{K}} \to \mathbb{N} \cup \{\infty\}, \tag{10.1}$$

$$h(\xi, \eta) := \min\{t \in \mathbb{N};\ \text{there is a } u \in U^t \text{ such that } \xi(t; \xi, u) = \eta\},$$

where $\min \emptyset = \infty$.

The previous definition is the discrete-time analogue of the *first-time hitting map*, as defined, for instance, in [17], [18]. Since we treat noninvertible systems as well, it is important for us to obtain information not only on the time that elapses to steer from ξ to η, but also on the "cost" incurred in doing so. For the projected system (2.5) and the function q interpreted as a cost, $|q(\xi, u)|$ may be arbitrarily large if u is chosen such that $A(u)$ is almost singular. In analogy to the first-time hitting map, we define the *minimal absolute cost map* by

$$H : M \times M \to \mathbb{R}_+ \cup \{\infty\}, \tag{10.2}$$

$$H(\xi, \eta) := \inf\{\max_{1\le s\le t} |J(s; \xi, u)|; \quad t \in \mathbb{N};\ u \in U^t \text{ such that } \xi(t; \xi, u) = \eta\},$$

where $\inf \emptyset = \infty$. The essential point is that both these values may be simultaneously bounded if one tries to reach a compact subset of the core of a control set.

LEMMA 10.1. *Let $\mathbb{K} = \mathbb{R}, \mathbb{C}$ and assume that system (2.5) is forward accessible. Let $D \subset \mathbb{P}^{n-1}_{\mathbb{K}}$ be a control set. Assume there are two nonvoid compact sets K_1, K_2 with $K_1 \subset \mathcal{O}^-(D)$ and $K_2 \subset \operatorname{core}(D)$. Then the following statements hold.*

(i) *There are constants $\overline{h} \in \mathbb{N}, \overline{H} \in \mathbb{R}_+$ such that*

$$h(\xi, \eta) \le \overline{h} \ \text{ for all } \xi \in K_1,\ \eta \in K_2, \tag{10.3}$$

$$H(\xi, \eta) \le \overline{H} \ \text{ for all } \xi \in K_1,\ \eta \in K_2. \tag{10.4}$$

(ii) *If* $K_2 = \mathbb{P}GE(r,u)$ *for some* $t \in \mathbb{N}$, $u \in U^t_{reg}$, *and* $r \in |\sigma(\Phi_u(t,0))|$, *then* $\overline{h}$, $\overline{H}$ *may be chosen such that for all* $\xi \in K_1$, $\eta \in K_2$ *there exists* $v \in U^t_{reg}$ *with*

$$\eta = \xi(t;\xi,v), \tag{10.5}$$

$$t \le \overline{h}, \tag{10.6}$$

$$\max_{1\le s\le t} |J(s;\xi,v)| \le \overline{H}. \tag{10.7}$$

Proof. (i) Let $\xi \in K_1$, $\eta \in K_2$. Choose any point $\zeta \in \operatorname{core}(D) \cap \hat{\mathcal{O}}^+(\xi)$, which is possible by Lemma 3.8 (i) and Proposition 6.6 (iii). Thus there exist $u_1 \in \operatorname{int} U^{t_1}(\xi)$ such that $\zeta = \xi(t_1;\xi,u_1)$ and (ξ,u_1) is a regular pair. By the implicit function theorem there exist open neighborhoods V_1 of ξ, W_1 of u_1, and a continuous function $w : V_1 \to W_1$ such that $\zeta = \xi(t_1,\xi',w(\xi'))$ for every $\xi' \in V_1$. This shows that $h(\xi',\zeta) \le t_1$ for all $\xi' \in V_1$. Furthermore, by continuous dependence of $J(s;\xi',w(\xi'))$ on ξ', it may also be obtained that $H(\xi',\zeta) \le H_1$ for some suitable constant $H_1 \in \mathbb{R}$ and all $\xi' \in V_1$, where possibly V_1 has to be chosen to be smaller than the original choice.

On the other hand, there exist $t_2 \in \mathbb{N}$, $u_2 \in \operatorname{int} U^{t_2}(\zeta)$ such that $\eta = \xi(t_2;\zeta,u_2)$ and (ζ,u_2) is a regular pair. By regularity, for any open neighborhood W_2 of u_2, the set $\{\xi(t_2;\zeta,u');\ u' \in W_2\}$ contains an open neighborhood V_2 of η. Choosing W_2 small enough so that $\operatorname{cl} W_2 \subset \operatorname{int} U^{t_2}(\zeta)$ we see that $h(\zeta,\eta') \le t_2$ for all $\eta' \in V_2$ and also $H(\zeta,\eta') \le H_2$ for all $\eta' \in V_2$ and some suitable constant H_2.

In all we have obtained that

$$h(\xi',\eta') \le t_1 + t_2 \quad \text{for all} \quad \xi' \in V_1, \eta' \in V_2$$

and

$$H(\xi',\eta') \le H_1 + H_2 \quad \text{for all} \quad \xi' \in V_1, \eta' \in V_2.$$

The assertion now follows because we may choose a finite subcover of the open cover

$$\{V_1(\xi) \times V_2(\eta); \quad \xi \in K_1, \eta \in K_2\}$$

of the compact set $K_1 \times K_2$.

(ii) Let $\xi \in K_1$, $\eta \in K_2$. Choose ζ' such that $\xi(t;\zeta',u) = \eta$. As u is universally regular, there exists an open neighborhood V of ζ', $V \subset \operatorname{core}(D)$, such that for every $\zeta'' \in V$ there exists $u(\zeta'') \in U^t_{reg}$ with $\eta = \xi(t;\zeta'',u(\zeta''))$. As $\zeta' \in \operatorname{core}(D)$ there exists $t_1 \in \mathbb{N}$, $u_1 \in \operatorname{int} U^{t_1}_{inv}$ such that $\zeta := \xi(t_1;\xi,u_1) \in V$. Let $t_2 = t$, $u_2 = u(\zeta)$; then $\eta = \xi(t_1+t_2;\xi,(u_1,u_2))$, (u_1,u_2) is universally regular and we may proceed as in the proof of part (i) by genericity of $U^{t_1}_{inv}$ and $U^{t_2}_{reg}$. □

With this result in hand we may start to examine the structure of the set of Floquet exponents.

PROPOSITION 10.2. *Let* $\mathbb{K} = \mathbb{R}, \mathbb{C}$. *The set* $\Sigma_{Fl,i}(A,U)$ *is an interval.*

Proof. Consider the function

$$\begin{aligned} \lambda_{i,t} : \quad U^t &\to \mathbb{R} \cup \{-\infty\}, \\ u &\mapsto \frac{1}{t} \log r_i(\Phi_u(t,0)). \end{aligned}$$

By Chapter II.8 in [13], $\lambda_{i,t}$ is continuous and therefore $\lambda_{i,t}(U^t)$ is connected as the continuous image of a connected set, and thus an interval. Now

$$\Sigma_{Fl,i}(A,U) = \bigcup_{t=1}^{\infty} \lambda_{i,t}(U^t),$$

and furthermore, for $u \in U$ and all $t \geq 1$

$$\log |r_i(A(u))| \in \lambda_{i,t}(U^t),$$

as we may simply consider the sequence $(u)^t$. Thus the assertion follows. □

Thus, from the connectedness of the set of admissible controls, it is immediately obtained that the Floquet spectrum is the union of at most n intervals. However, a weak point of this statement is that it totally ignores the dynamics of the system. The interplay between Floquet spectrum and dynamical behavior is studied from now on.

PROPOSITION 10.3. *Let $\mathbb{K} = \mathbb{R}, \mathbb{C}$, and assume that (2.5) is forward accessible. For a control set D with $\operatorname{core}(D) \neq \emptyset$ the set $\operatorname{cl} \Sigma_{Fl}(D)$ is an interval.*

Proof. By Proposition 6.9 there exist $t \geq t^*$, $u_1 \in U^t_{reg}$, and $\lambda_1 \in \sigma(\Phi_{u_1}(t,0))$ such that $\mathbb{P}GE(\lambda_1, u_1) \subset \operatorname{core}(D)$. It is sufficient to show that for any $\lambda \in \Sigma_{Fl}(D)$ the Floquet exponents of D are dense in the interval determined by λ and $\frac{1}{t}\log|\lambda_1|$.

Let $t_2 \in \mathbb{N}$, $u_2 \in \operatorname{int} U^{t_2}$ be such that for some $\lambda_2 \in \sigma(\Phi_{u_2}(t_2,0))$ we have that $\mathbb{P}GE(\lambda_2, u_2) \subset \operatorname{core}(D)$ and the eigenspace is regular. Without loss of generality we may assume that $t = t_2$ and $|\lambda_1| \leq |\lambda_2|$.

By Lemma 10.1 there exist constants $\overline{h}$, $\overline{H}$ such that for any $\xi, \eta \in \mathbb{P}E(\lambda_1, u_1) \cup \mathbb{P}E(\lambda_2, u_2)$ it holds that

$$h(\xi, \eta) \leq \overline{h},$$

$$H(\xi, \eta) \leq \overline{H},$$

where furthermore the corresponding control steering from ξ to η may be chosen to be universally regular if $\eta \in \mathbb{P}E(\lambda_1, u_1)$. Choose $\xi_j \in \mathbb{P}E(\lambda_j, u_j)$, $j = 1, 2$. Clearly, it holds for $j = 1, 2$ that

$$\lambda(\xi_j, u_j) = \frac{1}{t}\log|\lambda_j|.$$

We wish to construct controls such that the corresponding Floquet exponents are dense in the interval $[\frac{1}{t}\log|\lambda_1|, \frac{1}{t}\log|\lambda_2|]$. To this end define the control $u_{k,l,m}$, $k, l, m \in \mathbb{N}$ by

$$u_{k,l,m} := ((u_1)^{mk}, v_{1,k,m}, (u_2)^{ml}, v_{2,m,l}),$$

where $s_{1,k,m}, s_{2,m,l} \leq \overline{h}$ and $v_{1,k,m} \in \operatorname{int} U^{s_{1,k,m}}$ is chosen such that

$$\xi(s_{1,k,m}; \xi(mkt; \xi_1, (u_1)^{mk}), v_{1,k,m}) = \xi_2,$$

and analogously, $\xi(s_{2,l,m}; \xi(mlt; \xi_2, (u_2)^{ml}), v_{2,l,m}) = \xi_1$ for a universally regular control $v_{2,l,m}$, which is possible by Lemma 10.1(ii). We obtain in all that $\xi_1 = \xi(m(k+l)t + s_{1,k,m} + s_{2,k,m}; \xi_1, u_{k,l,m})$. Thus, for some $r \in \mathbb{R}$, it holds that $\xi_1 \in \mathbb{P}GE(r, u_{k,l,m})$. This projected sum of generalized eigenspaces is regular by the universal regularity of $v_{2,l,m}$. The corresponding Floquet exponent is given by

$$\lambda(\xi_1, u_{k,l,m}) = \frac{J(mkt; \xi_1, (u_1)^{mk}) + J(mlt; \xi_2, (u_2)^{ml}) + H(k,l,m)}{m(k+l)t + h(k,l,m)},$$

where $h(k,l,m) \leq 2\overline{h}$ and $|H(k,l,m)| \leq 2\overline{H}$ for all $k, l, m \in \mathbb{N}$. Thus for $k, l \geq 1$ it may be seen that

$$\lim_{m\to\infty} \lambda(\xi_1, u_{k,l,m}) = \lim_{m\to\infty} \frac{1}{m(k+l)t}(J(mkt; \xi_1, (u_1)^{mk}) + J(mlt; \xi_2, (u_2)^{ml}))$$

$$= \frac{k\lambda(\xi_1, u_1) + l\lambda(\xi_2, u_2)}{k+l} \in \operatorname{cl} \Sigma_{Fl}(D).$$

Clearly, the set of points that may be obtained by choosing different $k, l \in \mathbb{N}$ is dense in $[\lambda(\xi_1, u_1), \lambda(\xi_2, u_2)]$. □

COROLLARY 10.4. *Assume that* (2.5) *is forward accessible.*

(i) *If* $\mathbb{K} = \mathbb{R}, \mathbb{C}$, *then for every control set* D *with* $\operatorname{core}(D) \neq \emptyset$ *it holds that*

$$\operatorname{cl} \Sigma_{Fl}(D) = \operatorname{cl} \bigcup_{t \in \mathbb{N}, u \in U^t_{reg}} \left\{ \frac{1}{t} \log |\lambda|;\ \lambda \in \sigma(\Phi_u(t,0)),\ \mathbb{P}GE(\lambda, u) \subset \operatorname{core}(D) \right\}. \tag{10.8}$$

(ii) *If* $\mathbb{K} = \mathbb{R}$, *then for every control set* D *with nonempty core,*

$$\operatorname{cl} \Sigma_{Fl}(D) = \operatorname{cl} \bigcup_{t \in \mathbb{N}, u \in U^t_{reg}} \left\{ \frac{1}{t} \log |\lambda| \in \Sigma_{Fl}(D);\ \lambda \in \sigma(\Phi_u(t,0)) \cap \mathbb{R} \right\}. \tag{10.9}$$

Proof.

(i) If for some $u \in U^t$ and $r \in |\sigma(\Phi_u(t,0))|$ it holds that $\mathbb{P}GE(r, u) \subset \operatorname{core}(D)$, then by the genericity of the universally regular controls and the continuity of the eigenvalues and eigenprojections we may choose universally regular controls whose eigenspaces project to the core of D and whose corresponding Floquet exponents approximate the Floquet exponent $\frac{1}{t} \log r$ arbitrarily close. This shows the assertion.

(ii) Since the intermediate values $\lambda(\xi_1, u_{k,l,m})$ constructed in the previous proof are in fact Floquet exponents corresponding to a real eigenvalue of $\Phi_{u_{k,l,m}}(m(k+l)t + s_{1,k,m} + s_{2,l,m}, 0)$, it follows that it is sufficient to consider real eigenvalues. Now we may argue as in part (i). □

THEOREM 10.5. *Let* $\mathbb{K} = \mathbb{R}, \mathbb{C}$ *and assume that* (2.5) *is forward accessible. Let* κ *be equal to the number of main control sets.*

(i) *For each main control set* D_j, $j = 1, \ldots, \kappa$, *the closed Floquet spectrum is an interval. We define*

$$\operatorname{cl} \Sigma_{FL}(D_j) =: [\alpha_j, \beta_j], \quad \alpha_j \leq \beta_j\,. \tag{10.10}$$

(ii) *If all control sets with nonempty interior are main control sets, then*

$$\operatorname{cl} \Sigma_{FL}(A, U) = \bigcup_{j=1}^{\kappa} [\alpha_j, \beta_j]\,. \tag{10.11}$$

(iii) *If there exist control sets with nonempty interior that are not main control sets, then there exists a constant* $\bar{\beta} \in \mathbb{R}$ *such that*

$$\operatorname{cl} \Sigma_{FL}(A, U) = \bigcup_{j=1}^{\kappa} [\alpha_j, \beta_j] \cup [-\infty, \bar{\beta}]\,. \tag{10.12}$$

(iv) *If for two main control sets,* $D_{j_1} < D_{j_2}$, *then*

$$\alpha_{j_1} \leq \alpha_{j_2}, \tag{10.13}$$

$$\beta_{j_1} \leq \beta_{j_2}. \tag{10.14}$$

(v) *For $j = 1, \dots, \kappa$ it holds that*

$$\#\operatorname{cl}\Sigma_{FL}(D_j) \setminus \Sigma_{FL}(D_j) \leq m(D_j) + 1\,. \tag{10.15}$$

Proof.

(i) This is clear by Proposition 10.3.

(ii) Let $t \in \mathbb{N}$, $u \in U^t$ and consider $\sigma_{Fl}(u)$. As the Floquet spectrum of u does not change if we consider $(u)^l$ for some $l \geq 1$, we may assume that $t \geq t^*$. Hence, we may choose a sequence $\{u_k\}_{k\in\mathbb{N}} \subset U^t_{reg}$ converging to u for k tending to infinity. By the continuity of the spectrum it follows that $\sigma_{Fl}(u) \subset \bigcup_{j=1}^{\kappa}[\alpha_j, \beta_j]$.

(iii) For a control set D with nonempty interior that is not a main control set it holds by Proposition 8.1 that $\inf \Sigma_{Fl}(D) = -\infty$. The assertion thus follows from Proposition 10.3 and the argumentation of (ii).

(iv) If for two main control sets, $D_{j_1} \leq D_{j_2}$, then $Q_i \subset D_{j_1}$ and $Q_j \subset D_{j_2}$ implies that $i < j$. Thus the assertion follows from the obvious inequalities $\inf \Sigma_{Fl,i}(A,U) \leq \inf \Sigma_{Fl,j}(A,U)$ and $\sup \Sigma_{Fl,i}(A,U) \leq \sup \Sigma_{Fl,j}(A,U)$ if $i < j$.

(v) Let $t \in \mathbb{N}$ and $u, v \in U^t_{reg}$. Consider a continuous path $\gamma : [0,1] \to \operatorname{int} U^t$ with $\gamma(0) = u$ and $\gamma(1) = v$. Now consider the control $(\gamma(\tau), u)$. For every $\tau \in [0,1]$ and every $i \in \{1, \dots, n\}$ it holds by the universal regularity of u that $r_i(\tau) := r_i(\Phi_{(\gamma(\tau),u)}(2t, 0)) > 0$ iff $\mathbb{P}GE(r_i(\tau), (\gamma(\tau), u))$ is regular. Thus it follows for every $i \in \{1, \dots, n\}$ that the interval $[r_i(\Phi_{(u,u)}(2t,0)), r_i(\Phi_{(v,u)}(2t,0))]$ is contained in $\Sigma_{Fl}(D)$ for some control set D by Proposition 6.8. Since the sets $\operatorname{int}\Sigma_{Fl,i}(A,U)$ are intervals and by $\operatorname{cl}\Sigma_{Fl}(D_j) = \bigcup_{Q_i \subset D_j} \operatorname{cl}\Sigma_{Fl,i}(A,U)$, it follows that the only points where the Floquet spectrum of a main control set and its closure may differ are the endpoints of the intervals $\Sigma_{Fl,i}(A,U)$. Of these there are at most $m(D_j) + 1$, which shows the assertion. □

It should be noted that the spectral intervals corresponding to different main control sets may overlap, i.e., that the statement $\alpha_i \leq \alpha_j$, $\beta_i \leq \beta_j$ in Theorem 10.5 in no way excludes the possibility that $\beta_i > \alpha_j$. In fact, it is even possible that $\alpha_i = \alpha_j$ and $\beta_i = \beta_j$ for $i \neq j$. To illustrate this phenomenon consider the following example.

Example 10.6. Let $\mathbb{K} = \mathbb{R}$. Define

$$A :\ \mathbb{R}^4 \longrightarrow \mathbb{R}^{2\times 2},$$

$$A(a,b,c,d) := \begin{bmatrix} a & b \\ c & d \end{bmatrix}.$$

Let $\mathbb{R}^n_{\geq 0}$ denote the set of vectors with nonnegative real entries. Define

$$U := \{[a\ b\ c\ d] \in \mathbb{R}^4_{\geq 0};\quad a + c \leq 1;\, b + d \leq 1\}.$$

Then $A(U)$ is exactly the set of nonnegative matrices in $\mathbb{R}^{2\times 2}$ with 1-norm less than or equal to 1. As the set of nonnegative vectors in $\mathbb{R}^2$ is invariant under $A(u)$ for any $u \in U$, i.e., $A(u)\mathbb{R}^2_{\geq 0} \subset \mathbb{R}^2_{\geq 0}$, it follows that the invariant control set $D_2 = C \subset \mathbb{P}\mathbb{R}^2_{\geq 0}$. Hence there also exists a minimal control set $D_1 = C^-$ and no other control set D with $\operatorname{core}(D) \neq \emptyset$.

Clearly, $\alpha_1 = \alpha_2 = -\infty$ as $0 \in A(U)$. Let us show also that $\beta_1 = \beta_2 = 0$. For any $t \geq 1$, $u \in U^t$, it holds that

$$r(\Phi_u(t,0)) \leq \|\Phi_u(t,0)\|_1 \leq \|A(u(t-1))\|_1 \cdot \ldots \cdot \|A(u(0))\|_1 \leq 1.$$

Hence $\beta_1, \beta_2 \leq \log 1 = 0$. On the other hand, $I \in A(U)$, and so $0 \in \operatorname{cl}\Sigma_{FL}(C)$, $0 \in \operatorname{cl}\Sigma_{FL}(C^-)$, and $\beta_1, \beta_2 \geq 0$.

In order to construct a two-dimensional example with identical spectral intervals and $U = U_{inv}$, it is sufficient to replace the set U of the previous example by $U' := \{u \in U;\ \det(A(u)) > 0\}$. If we require that $\operatorname{cl} A(U)$ consist of invertible matrices, then it is still possible to make upper or lower boundaries of spectral intervals equal, e.g., if the map A is replaced by $u \mapsto \exp(A(u))$. Note also that $\exp(A(U))$ consists of nonnegative matrices. Then, as in the preceding discussion, it is possible to obtain that for this modified example, $\beta_1 = \beta_2 = 1$. However, this comes with the price that $\alpha_1 = -1 \neq \alpha_2 = 0$. It is not known whether identical spectral intervals to different main control sets are possible if it is assumed that $\det(A(u)) \neq 0$ for all $u \in \operatorname{cl} U$.

11. The Lyapunov and Bohl spectrums. Let us now discuss how the results on the Floquet spectrum can be related to the other spectra of characteristic exponents. We begin by showing that the Lyapunov exponents corresponding to trajectories that evolve in a specific way in the core of control sets are contained in the closure of the associated Floquet interval. On the other hand, for every element of the closure of the Floquet interval of a control set there exists a control sequence that realizes this number as a Lyapunov exponent.

THEOREM 11.1. *Let $\mathbb{K} = \mathbb{R}, \mathbb{C}$, and assume that (2.5) is forward accessible.*

(i) *Let D be a control set, with $\operatorname{core}(D) \neq \emptyset$. Assume that $(\xi_0, u) \in \mathbb{P}^{n-1}_{\mathbb{K}} \times U^{\mathbb{N}}(\xi_0)$ are given with $\omega^+(\xi_0, u) \subset D$. If there exists a $t_0 \in \mathbb{N}$ with $\xi(t_0; \xi_0, u) \in \operatorname{core}(D)$, then $\lambda(\xi_0, u) \in \operatorname{cl}\Sigma_{Fl}(D)$.*

(ii) *Let D be a control set, with $\operatorname{core}(D) \neq \emptyset$. Then for every $\lambda \in \operatorname{cl}\Sigma_{Fl}(D)$ there exist $\xi_0 \in \operatorname{core}(D)$ and $u \in U^{\mathbb{N}}$ such that $\lambda = \lambda(\xi, u)$. In particular, it holds that*

$$\operatorname{cl}\Sigma_{Fl}(D) \subset \Sigma_{Ly}(A, U). \tag{11.1}$$

Proof. (i) Without loss of generality we may assume that $t_0 = 0$ as the Lyapunov exponents satisfy $\lambda(\xi_0, u) = \lambda(\xi(t_0; \xi_0, u), u(t_0 + \cdot))$ where $u(t_0 + \cdot) = (u(t_0), u(t_0 + 1), \ldots)$ is the shifted control. Let $\{t_k\}_{k\in\mathbb{N}} \subset \mathbb{N}$ be an increasing sequence such that

$$\lim_{k\to\infty} \frac{1}{t_k} J(t_k; \xi_0, u) = \lambda(\xi_0, u)\,. \tag{11.2}$$

Taking a subsequence we may assume that

$$\lim_{k\to\infty} \xi(t_k; \xi_0, u) =: \eta \in \omega^+(\xi_0, u) \subset D\,. \tag{11.3}$$

As $\xi_0 \in \operatorname{core}(D)$ it follows that $\eta \in \hat{\mathcal{O}}^-(\xi_0)$, and hence there is a $t \in \mathbb{N}$ and a neighborhood $V(\eta)$ such that $V(\eta) \subset \hat{\mathcal{O}}^-_t(\xi_0)$. For k large enough it holds that $\xi(t_k; \xi_0, u) \in V(\eta)$. By continuous dependence of $\xi(t_k; \xi_0, u)$ on u, the continuous dependence of $J(t_k; \xi_0, u)$ on u, and the genericity of $U^{t_k}_{reg}$, we may choose controls $u_k \in U^{t_k}_{reg}$ such that $\xi(t_k; \xi_0, u_k) \in V(\eta)$ for all k large enough and

$$\lim_{k\to\infty} \frac{1}{t_k} J(t_k; \xi_0, u_k) = \lambda(\xi_0, u)\,.$$

We can therefore find a control $v_k \in \operatorname{int} U^t$ such that $\xi_0 = \xi(t; \xi(t_k; \xi_0, u_k), v_k)$. The Floquet exponent corresponding to the control (u_k, v_k) and ξ_0 is given by

$$\begin{aligned} \lambda(\xi_0, (u_k, v_k)) &= \frac{1}{t_k + t} \left(J(t_k; \xi_0, u_k) + J(t; \xi(t_k; \xi_0, u_k), v_k)\right) \\ &= \lambda(\xi(t_k; \xi_0, u_k), (v_k, u_k)) \in \Sigma_{Fl}(D)\,, \end{aligned} \tag{11.4}$$

by the universal regularity of u_k. Letting $k \to \infty$ the assertion follows after noting that Lemma 10.1 guarantees that the v_k can be chosen so that $|J(t;\xi(t_k;\xi_0,u_k),v_k)|$ is bounded independently of k.

(ii) Let $\lambda^* \in \operatorname{cl}\Sigma_{Fl}(D)$. Let $u_k \in U^{t_k}_{reg}$, $k \in \mathbb{N}$, be a sequence of controls such that

$$\lim_{k\to\infty} \frac{1}{t_k} \log|\lambda_k| = \lambda^*, \tag{11.5}$$

where $\lambda_k \in \sigma(\Phi(t_k,u_k))$ and $\mathbb{P}E(\lambda_k,u_k) \subset \operatorname{core}(D)$. By Corollary 10.4 such a sequence exists and we may assume that $\lambda_k \in \mathbb{R}$ if $\mathbb{K} = \mathbb{R}$. For $k \in \mathbb{N}$ let $\xi_k \in \mathbb{P}E(\lambda_k,u_k)$. Therefore it holds for all $l,k \in \mathbb{N}$ that $\xi(lt_k;\xi_k,(u_k)^l) = \xi_k \in \operatorname{core}(D)$. For all $k \in \mathbb{N}$ there exists a control $v_k \in U^{s_k}$ such that $\xi_{k+1} = \xi(s_k;\xi_k,v_k)$. Let H_k be such that $|J(s;\xi_k,v_k)| < H_k$ for $0 \le s \le s_k$. We construct a control that generates the Lyapunov exponent λ^* as follows: choose $m_1 \in \mathbb{N}$ such that

$$\left|\left(\frac{m_1t_1}{m_1t_1+s_1+t_2}-1\right)\frac{1}{t_1}\log|\lambda_1|\right| < \frac{1}{8}, \tag{11.6}$$

$$\frac{H_1}{m_1t_1} < \frac{1}{8}, \tag{11.7}$$

$$\left|\frac{J(s;\xi_2,u_2)}{m_1t_1}\right| < \frac{1}{8}, \quad 0 \le s \le t_2. \tag{11.8}$$

Let $u_1^* := ((u_1)^{m_1},v_1) \in U^{T_1}$ and $T_1 := m_1t_1+s_1$. Using (11.6) and (11.7) it may be seen that for $0 \le s \le s_1$

$$\begin{aligned}&\left|\frac{1}{m_1t_1+s}J(m_1t_1+s;\xi_1,u_1^*) - \frac{1}{t_1}\log|\lambda_1|\right| \\ &\le \frac{1}{m_1t_1+s}|J(s;\xi_1,v_1)| + \left|\left(\frac{m_1t_1}{m_1t_1+s}-1\right)\frac{1}{t_1}\log|\lambda_1|\right| < \frac{1}{4}.\end{aligned} \tag{11.9}$$

Note also that by (11.8), we obtain as in (11.9) that for $0 \le s \le t_2$ and $v = (u_1^*,u_2)$,

$$\begin{aligned}&\left|\frac{1}{T_1+s}J(T_1+s;\xi_1,v) - \frac{1}{t_1}\log|\lambda_1|\right| \\ &\le \left|\frac{1}{T_1+s}J(m_1t_1;\xi_1,v) - \frac{1}{t_1}\log|\lambda_1|\right| + \frac{H_1}{T_1+s} + \left|\frac{1}{T_1+s}J(s;\xi_2,u_2)\right| < \frac{1}{2}.\end{aligned} \tag{11.10}$$

For $k > 1$ assume that we have constructed u_{k-1}^*, m_{k-1}, and T_{k-1} such that for $-s_{k-1} \le s \le t_k$ it holds that

$$\left|\frac{1}{(T_{k-1}+s)}J(T_{k-1}+s;\xi_1,(u_{k-1}^*,u_k)) - \frac{1}{t_{k-1}}\log|\lambda_{k-1}|\right| < 2^{-(k-1)}. \tag{11.11}$$

Choose $m_k \in \mathbb{N}$ such that

$$\left|\frac{1}{T_{k-1}+m_kt_k}J(T_{k-1};\xi_1,u_{k-1}^*)\right| < 2^{-(k+3)}, \tag{11.12}$$

$$\left|\left(\frac{m_kt_k}{T_{k-1}+m_kt_k+s_k+t_{k+1}}-1\right)\frac{1}{t_k}\log|\lambda_k|\right| < 2^{-(k+3)}, \tag{11.13}$$

$$\frac{H_k}{T_{k-1}+m_kt_k} < 2^{-(k+3)}, \tag{11.14}$$

$$\left|\frac{J(s;\xi_{k+1},u_{k+1})}{m_k t_k}\right| < 2^{-(k+2)}\,, \quad 0 \le s \le t_{k+1}. \tag{11.15}$$

Set $u_k^* := (u_{k-1}^*, (u_k)^{m_k}, v_k)$ and $T_k := T_{k-1} + m_k t_k + s_k$. For $T_{k-1} + m_k t_k \le t \le T_k$ we obtain with (11.12), (11.13), and (11.14) that

$$\begin{aligned}
&\left|\frac{1}{t}J(t;\xi_1,u_k^*) - \frac{1}{t_k}\log|\lambda_k|\right| \\
&\le \left|\frac{1}{t}J(T_{k-1};\xi_1,u_{k-1}^*)\right| + \left|\left(\frac{m_k t_k}{t} - 1\right)\frac{1}{t_k}\log|\lambda_k|\right| + \frac{1}{t}|J(t - T_{k-1} - m_k t_k;\xi_k,v_k)| \\
&< 2^{-(k+3)} + 2^{-(k+3)} + 2^{-(k+3)} < 2^{-(k+1)}\,.
\end{aligned}$$

Analogously to (11.10), it may be seen from (11.13) and (11.15) that for $0 \le s \le t_{k+1}$ and $v = (u_k^*, u_{k+1})$,

$$\left|\frac{1}{T_k + s}J(T_k + s;\xi_1,v) - \frac{1}{t_k}\log|\lambda_k|\right| < 2^{-k}\,. \tag{11.16}$$

For the control u^* that is recursively defined via $u^*_{[0,T_k]} = u_k^*$ we claim that

$$\lambda(\xi_1,u^*) = \lim_{k\to\infty}\frac{1}{t_k}\log|\lambda_k| = \lambda^*\,. \tag{11.17}$$

We have shown that for $k > 1$ and $T_{k-1} + m_k t_k \le t \le T_k + t_{k+1}$, it holds that

$$\left|\frac{1}{t}J(t;\xi_1,u^*) - \frac{1}{t_k}\log|\lambda_k|\right| < 2^{-k}.$$

Thus our claim follows if we can show that for $t = T_{k-1},\dots,T_{k-1} + (m_k - 1)t_k$ the following relation holds:

$$\left|\frac{1}{t}J(t;\xi_1,u_k^*) - \frac{1}{t_k}\log|\lambda_k|\right| \ge \left|\frac{1}{t + t_k}J(t + t_k;\xi_1,u_k^*) - \frac{1}{t_k}\log|\lambda_k|\right|\,, \tag{11.18}$$

because this means that the sequence behaves in a monotonic way, at least if viewed at every t_kth step. For $l = 0,\dots,m_k - 1$ and $t = T_{k-1} + l t_k$ this is clear by

$$\begin{aligned}
\left|\frac{1}{t}J(t;\xi_1,u_k^*) - \frac{1}{t_k}\log|\lambda_k|\right| &= \left|\frac{1}{t}(J(T_{k-1};\xi_1,u_k^*) + l\log|\lambda_k|) - \frac{1}{t_k}\log|\lambda_k|\right| \\
&= \frac{1}{t}\left|J(T_{k-1};\xi_1,u_k^*) - \frac{T_{k-1}}{t_k}\log|\lambda_k|\right|\,.
\end{aligned}$$

The other cases can be treated using the same argument, with the modification that the time from which periodicity is used is not T_{k-1} but $T_{k-1}+s$ for some $0 \le s \le t_k - 1$. This proves the assertion. □

A further question of interest, especially if stabilization and robust stability questions are considered, concerns the lower and upper bounds of the spectral sets that we have defined. For a general discrete inclusion given by a bounded set $\Sigma \subset \mathbb{K}^{n\times n}$ and

$$x(t+1) \in \{Ax(t)\,;\ A \in \Sigma\}\,, \quad t \in \mathbb{N}, \tag{11.19}$$

this has been studied in [8], [15], [38], [27]. In particular, the latter three references study the relation between the generalized spectral radius

$$\bar\rho(\Sigma) := \limsup_{t\to\infty}\bar\rho_t(\Sigma)^{1/t}\,,$$

where

$$\bar\rho_t(\Sigma) := \sup\{r(A_{t-1}\cdot\ldots\cdot A_0); A_s\in\Sigma, s=0,\ldots,t-1\},$$

and the joint spectral radius

$$\hat\rho(\Sigma) := \limsup_{t\to\infty}\hat\rho_t(\Sigma)^{1/t},$$

where

$$\hat\rho_t(\Sigma) := \sup\{\|A_{t-1}\cdot\ldots\cdot A_0\|; A_s\in\Sigma, s=0,\ldots,t-1\}.$$

Theorem IV in [15] states that for every bounded set Σ we have $\bar\rho(\Sigma)=\hat\rho(\Sigma)$. Although Berger and Wang restrict themselves to the real case, it is clear that they also prove the complex case, which may be seen via identification of $\mathbb{C}^{n\times n}$ with $\mathbb{R}^{2n\times 2n}$. Note that these definitions correspond to our definitions but for the fact that we have introduced the logarithm. Thus it is easy to see that

$$\log(\bar\rho(A(U))) = \sup\Sigma_{Fl}(A,U) \tag{11.20}$$

and

$$\log(\hat\rho(\Sigma)) = \limsup_{t\to\infty}\sup_{u\in U^{\mathbb{N}},\,\xi\in\mathbb{P}^{n-1}_{\mathbb{K}}}\frac{1}{t}J(t;\xi,u). \tag{11.21}$$

We therefore immediately obtain the following corollaries, where we do not have to make our usual forward accessibility assumption. In order to conform to our previously introduced notation we will still think of the discrete inclusion to be given by an analytic map A and a set U. Note, however, that if we drop Assumption 2.1, then any bounded set of matrices may be represented in this way.

COROLLARY 11.2. *Let $\mathbb{K}=\mathbb{R},\mathbb{C}$, and consider system* (2.2)*. Assume that $A(U)$ is bounded. Then*

$$\sup\Sigma_{Fl}(A,U)=\sup\Sigma_{Ly}(A,U)=\sup\Sigma_{Bo}(A,U)=\limsup_{t\to\infty}\sup_{u\in U^{\mathbb{N}},\,\xi\in\mathbb{P}^{n-1}_{\mathbb{K}}}\frac{1}{t}J(t;\xi,u).$$

Using this result we can also prove the following statements on the infima of the spectra.

PROPOSITION 11.3. *Let $\mathbb{K}=\mathbb{R},\mathbb{C}$, and consider system* (2.2)*. Assume that $A(U)$ is bounded. Then*

$$\inf\Sigma_{Fl}(A,U)=\inf\Sigma_{Ly}(A,U)=\liminf_{t\to\infty}\inf_{u\in U^{\mathbb{N}},\,\xi\in\mathbb{P}^{n-1}_{\mathbb{K}}}\frac{1}{t}J(t;\xi,u). \tag{11.22}$$

Proof. Obviously, it holds that

$$\inf\Sigma_{FL}(A,U)\geq\inf\Sigma_{Ly}(A,U)\geq\liminf_{t\to\infty}\inf_{u\in U^{\mathbb{N}},\,\xi\in\mathbb{P}^{n-1}_{\mathbb{K}}}\frac{1}{t}J(t;\xi,u).$$

If there exists a $u\in\operatorname{cl}U$ such that $\det(A(u))=0$ the claim is trivially true as both infima are given by $-\infty$. If this is not the case we may consider the time-reversed system

$$\begin{aligned} x(t+1) &= A(u(t))^{-1}x(t),\ t\in\mathbb{N},\\ x(0) &= x_0\in\mathbb{K}^n,\\ u(t) &\in U,\ t\in\mathbb{N}. \end{aligned} \tag{11.23}$$

Denote the Floquet spectrum of the time-reversed system by $\Sigma_{Fl}^{-}(A,U)$. It is immediate that $\sup \Sigma_{Fl}^{-}(A,U) = -\inf \Sigma_{Fl}(A,U)$. Note also that

$$\inf_{x\in\mathbb{K}^n,\|x\|=1} \log\|\Phi_u(t,0)x\| = -\sup_{x\in\mathbb{K}^n,\|x\|=1} \log\|\Phi_u(t,0)^{-1}x\|$$

and therefore

$$\liminf_{t\to\infty} \inf_{u\in U^{\mathbb{N}},\ \xi\in\mathbb{P}_{\mathbb{K}}^{n-1}} \frac{1}{t}J(t;\xi,u) = -\limsup_{t\to\infty} \sup_{u\in U^{\mathbb{N}},\xi\in\mathbb{P}_{\mathbb{K}}^{n-1}} \frac{1}{t}J^{-}(t;\xi,u),$$

where $J^{-}(t;\xi,u) = \log \frac{\|\Phi_u(t,0)^{-1}x\|}{\|x\|}$ for $\xi = \mathbb{P}x$. The assertion now follows by applying Corollary 11.2. □

Barabanov [9] proved that to each discrete inclusion given by a bounded set of matrices there exists a trajectory that realizes the maximal Lyapunov exponent. The following statement brings this in relation to the control structure of system (2.5).

PROPOSITION 11.4. *Let $\mathbb{K} = \mathbb{R},\mathbb{C}$, let Assumption 2.1 hold, and assume that (2.5) is forward accessible. Then*

(i) *there exist $u \in U^{\mathbb{N}}$, $\xi \in C$ such that $\lambda(\xi,u) = \beta(u) = \sup \Sigma_{Ly}(A,U)$;*

(ii) *there exist $v \in U^{\mathbb{N}}$, $\eta \in C^{-}$ such that $\lambda(\eta,v) = \inf \Sigma_{Ly}(A,U)$.*

Proof. (i) (resp., (ii)) follows from Corollary 11.2 (resp., Proposition 11.3, Remark 7.4(ii), and Theorem 11.1 (ii)). □

If the finiteness conjecture holds as discussed by Lagarias and Wang [38], then the previous result can be restated in terms of the Floquet spectrum; i.e., it would be possible to realize maximal and minimal Floquet exponents via some periodic control sequence u. This is the topic of ongoing research.

Let us also note that Gurvits [27] has shown that for discrete inclusions given by finitely many matrices, the indices $\inf \Sigma_{Fl,n}(A,U)$ and $\inf \Sigma_{Bo}(A,U)$ coincide. It remains to be investigated how this result may be carried over to our case.

Acknowledgment. The author wishes to express his thanks to F. Colonius, D. Hinrichsen, and W. Kliemann for a number of helpful discussions and remarks and an anonymous referee for supplying important references.

REFERENCES

[1] E. AKIN, *The General Topology of Dynamical Systems*, Grad. Stud. Math. 1, American Mathematical Society, Providence, RI, 1993.

[2] F. ALBERTINI, *Controllability of Discrete-Time Nonlinear Systems and Some Related Topics*, Ph.D. thesis, Rutgers University, New Brunswick, NJ, 1993.

[3] F. ALBERTINI AND E. SONTAG, *Some Connections between Chaotic Dynamical Systems and Control Systems*, Report SYCON-90-13, Rutgers Center for Systems and Control, New Brunswick, NJ, 1990.

[4] F. ALBERTINI AND E. SONTAG, *Discrete-time transitivity and accessibility: Analytic systems*, SIAM J. Control Optim., 31 (1993), pp. 1599–1622.

[5] L. ARNOLD AND W. KLIEMANN, *Qualitative theory of stochastic systems*, in Probabilistic Analysis and Related Topics, A. Bharucha-Reid, ed., Academic Press, New York, pp. 11–79.

[6] L. ARNOLD, W. KLIEMANN, AND E. OELJEKLAUS, *Lyapunov exponents of linear stochastic systems*, in Lyapunov Exponents, L. Arnold and V. Wihstutz, eds., Lecture Notes in Math. 1186, Springer-Verlag, Berlin, 1983, pp. 85–125.

[7] L. ARNOLD AND L. SAN MARTIN, *A control problem related to the Lyapunov spectrum of stochastic flows*, Mat. Apl. Comp., 5 (1986), pp. 31–64.

[8] N. E. BARABANOV, *Lyapunov indicator of discrete inclusions.* I, Automat. Remote Control, 49 (1988), pp. 152–157.

[9] N. E. Barabanov, *Lyapunov indicator of discrete inclusions.* II, Automat. Remote Control, 49 (1988), pp. 283–287.
[10] N. E. Barabanov, *Lyapunov indicator of discrete inclusions.* III, Automat. Remote Control, 49 (1988), pp. 558–565.
[11] N. E. Barabanov, *Method for the computation of the Lyapunov exponent of a differential inclusion*, Automat. Remote Control, 50 (1989), pp. 475–479.
[12] C. J. B. Braga and L. A. B. San Martin, *On the Number of Control Sets on Projective Spaces*, Systems Control Lett., 29 (1996), pp. 2–26.
[13] H. Baumgärtel, *Analytic Perturbation Theory for Matrices and Operators*, Oper. Theory Adv. Appl. 15, Birkhäuser, Basel, 1985.
[14] P. H. Baxendale and R. Z. Has'minski, *Stability index for products of random transformations*, Advances in Applied Probability (1998), to appear.
[15] M. A. Berger and Y. Wang, *Bounded semigroups of matrices*, Linear Algebra Appl., 166 (1992), pp. 21–27.
[16] P. Bohl, *Über Differentialgleichungen*, J. Reine Angew. Math., 144 (1913), pp. 284–313.
[17] F. Colonius, *Asymptotic behaviour of optimal control systems with low discount rates*, Math. Oper. Res., 14 (1989), pp. 309–316.
[18] F. Colonius and W. Kliemann, *Infinite time optimal control and periodicity*, Appl. Math. Optim., 20 (1989), pp. 113–130.
[19] F. Colonius and W. Kliemann, *Stability radii and Lyapunov exponents*, in Proc. Workshop Control of Uncertain Systems, Bremen 1989, D. Hinrichsen and B. Mårtensson, eds., Prog. Systems Control Theory 6, Birkhäuser, Basel, 1990, pp. 19–55.
[20] F. Colonius and W. Kliemann, *Linear control semigroups acting on projective space*, J. Dynam. Differential Equations, 5 (1993), pp. 495–528.
[21] F. Colonius and W. Kliemann, *Maximal and minimal Lyapunov exponents of bilinear control systems*, J. Differential Equations, 101 (1993), pp. 232–275.
[22] F. Colonius and W. Kliemann, *The Lyapunov spectrum of families of time varying matrices*, Trans. Amer. Math. Soc., 348 (1996), pp. 4389–4408.
[23] F. Colonius, W. Kliemann, and S. Krull, *Stability and Stabilization of Linear Uncertain Systems—A Lyapunov Exponents Approach*, Report 372, Schwerpunktprogramm der Deutschen Forschungsgemeinschaft "Anwendungsbezogene Optimierung und Steuerung," Universität Augsburg, Germany, 1992.
[24] J. L. Daleckii and M. G. Krein, *Stability of Solutions of Differential Equations in Banach Spaces*, Transl. Math. Monographs 43, AMS, Providence, RI, 1974.
[25] G. Floquet, *Sur les équations différentielles linéaires à coefficients périodiques*, Ann. École Norm. Supérieure, 12 (1883), pp. 47–89.
[26] L. Grüne, *Numerical stabilization of bilinear control systems*, SIAM J. Control Optim., 34 (1996), pp. 2024–2050.
[27] L. Gurvits, *Stability of discrete linear inclusions*, Linear Algebra Appl., 231 (1995), pp. 47–85.
[28] R. Z. Has'minskii, *Necessary and sufficient conditions for the asymptotic stability of linear stochastic systems*, Theory Probab. Appl., 12 (1967), pp. 144–147.
[29] D. Hinrichsen, A. Ilchmann, and A. J. Pritchard, *Robustness of stability of time-varying linear systems*, J. Differential Equations, 82 (1989), pp. 219–250.
[30] D. Hinrichsen and A. J. Pritchard, *Real and complex stability radii: A survey*, in Control of Uncertain Systems, D. Hinrichsen and B. Mårtensson, eds., Prog. Systems Control Theory 6, Birkhäuser, Basel, 1990, pp. 119–162.
[31] D. Hinrichsen and A. J. Pritchard, *Destabilization by output feedback*, Differential Integral Equations, 5 (1992), pp. 357–386.
[32] P. Homblé, *Ergodicity conditions for nonlinear discrete time stochastic dynamical systems with Markovian noise*, Stochast. Anal. Appl., 11 (1993), pp. 513–568.
[33] P. Homblé, *Lyapunov Spectrum of Discrete Linear Stochastic Systems Perturbed by Noise*, unpublished manuscript, 1994.
[34] B. Jakubczyk and E. Sontag, *Controllability of nonlinear discrete time systems: A Lie algebraic approach*, SIAM J. Control Optim., 28 (1990), pp. 11–33.
[35] R. Johnson, K. Palmer, and G. Sell, *Ergodic properties of linear dynamical systems*, SIAM J. Appl. Math., 18 (1987), pp. 1–33.
[36] L. Kaup and B. Kaup, *Holomorphic Functions of Several Variables: An Introduction to the Fundamental Theory*, de Gruyter Stud. Math. 3, de Gruyter, New York, 1983.
[37] W. Kliemann, *Recurrence and invariant measures for degenerate diffusions*, Ann. Probab., 15 (1987), pp. 690–707.
[38] J. C. Lagarias and Y. Wang, *The finiteness conjecture for the generalized spectral radius of a set of matrices*, Linear Algebra Appl., 214 (1995), pp. 17–42.

[39] P. Lancaster and M. Tismenetsky, *The Theory of Matrices*, 2nd ed., Academic Press, Boston, MA, 1985.

[40] A. M. Lyapunov, *Problème général de la stabilité du mouvement*, Ann. Fac. Sci. Toulouse, 9 (1907), pp. 203–474. Translation of the original paper published in 1893 in Comm. Soc. Math. Kharkow and reprinted as Ann. Math Studies 17, Princeton University Press, Princeton, NJ, 1949.

[41] S. P. Meyn and P. E. Caines, *Asymptotic behavior of stochastic systems possessing Markovian realizations*, SIAM J. Control Optim., 29 (1991), pp. 535–561.

[42] R. Narasimhan, *Introduction to the Theory of Analytic Spaces*, Lecture Notes in Math. 25, Springer-Verlag, Berlin, 1966.

[43] K. M. Przyłuski, *On asymptotic stability of linear time-varying infinite-dimensional systems*, Systems Control Lett., 6 (1985), pp. 147–152.

[44] K. M. Przyłuski, *Remarks on the stability of linear infinite-dimensional discrete-time systems*, J. Differential Equations, 72 (1988), pp. 189–200.

[45] K. M. Przyłuski, *Stability of linear infinite-dimensional systems revisited*, Internat. J. Control, 48 (1988), pp. 513–523.

[46] K. M. Przyłuski and S. Rolewicz, *On stability of linear time-varying infinite-dimensional discrete-time systems*, Systems Control Lett., 4 (1984), pp. 307–315.

[47] R. J. Sacker and G. R. Sell, *A spectral theory for linear differential systems*, J. Differential Equations, 37 (1978), pp. 320–358.

[48] E. Sontag and Y. Wang, *Orders of input/output differential equations and state-space dimensions*, SIAM J. Control Optim., 33 (1995), pp. 1102–1126.

[49] E. D. Sontag, *Universal nonsingular controls*, Systems Control Lett., 19 (1992), pp. 221–224. Errata, ibid., 20 (1993), p. 77.

[50] E. D. Sontag and F. R. Wirth, *Remarks on Universal Nonsingular Controls for Discrete-Time Systems*, Systems Control Lett., to appear.

[51] H. J. Sussmann, *Single-input observability of continuous-time systems*, Math. Systems Theory, 12 (1979), pp. 371–393.

[52] H. J. Sussmann, *Real-analytic desingularization and subanalytic sets: An elementary approach*, Trans. Amer. Math. Soc., 317 (1990), pp. 417–461.

[53] F. Wirth, *Robust Stability of Discrete-Time Systems under Time-Varying Perturbations*, Ph.D. thesis, Univ. of Bremen, Germany, 1995.

[54] F. Wirth, *Universal controls for homogeneous discrete-time systems*, in Proc. 34th IEEE CDC, New Orleans, LA, 1995, pp. 25–26.

[55] F. Wirth, *The Calculation of Real Time-Varying Stability Radii*, Internat. J. Robust Nonlinear Control, to appear.

[56] F. Wirth, *An introduction to spectral theory of linear time-varying systems*, J. Math. Systems Estim. Control, to appear.

[57] F. Wirth and D. Hinrichsen, *On stability radii of infinite dimensional time-varying discrete-time systems*, IMA J. Math. Control Inform., 11 (1994), pp. 253–276.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 2, pp. 488–541, March 1998

004

AN OPTIMAL CONTROL THEORY FOR DISCRETE EVENT SYSTEMS*

RAJA SENGUPTA† AND STÉPHANE LAFORTUNE‡

Abstract. In certain discrete event applications it may be desirable to find a particular controller, within the set of acceptable controllers, which optimizes some quantitative performance measure. In this paper we propose a theory of optimal control to meet such design requirements for deterministic systems. The discrete event system (DES) is modeled by a formal language. Event and cost functions are defined which induce costs on controlled system behavior. The event costs associated with the system behavior can be reduced, in general, only by increasing the control costs. Thus it is nontrivial to find the optimal amount of control to use, and the formulation captures the fundamental tradeoff motivating classical optimal control. Results on the existence of minimally restrictive optimal solutions are presented. Communication protocols are analyzed to motivate the formulation and demonstrate optimal controller synthesis. Algorithms for the computation of optimal controllers are developed for the special case of DES modeled by regular languages. It is shown that this framework generalizes some of the existing literature.

Key words. discrete event systems, optimal control, regular languages, dynamic programming

AMS subject classifications. 93A99, 49-XX, 90C27

PII. S0363012994260957

1. Introduction.

This paper presents a new framework for the optimal control of discrete event systems (DESs). The aim is to find methods to handle numerical performance measures in the DES controller design process.

The most influential paradigm for DES control is the supervisory control theory (SCT) suggested by Ramadge and Wonham [9, 8]. SCT makes certain system-theoretic assumptions which are appropriate for DES control problems. SCT as developed in [8] partitions all possible DES behavior into legal or illegal, and then addresses the problem of designing a DES controller that guarantees legal behavior. Here we enrich this view by accepting that some legal behaviors are better than others. For example, for a transaction submitted to a database management system (DBMS), all commit times below a certain threshold may be legal, but a smaller commit time is better. We propose numerical measures on the set of legal behaviors to capture such distinctions. The new problem, then, is to produce a controller that is not only legal but also "good" in the sense of the given numerical performance measures. We present our various findings collectively as a theory of *optimal control for discrete event systems.* It is our hope that this theory lays out the boundaries within which future work on the performance improvement or performance tuning of specific DESs can be attempted.

In the historical development of control theory, optimal control has been considered interesting only after the design and analysis of other control-theoretic concepts such as controllability, stabilizability, etc. have attained some degree of maturity.

*Received by the editors January 3, 1994; accepted for publication (in revised form) January 6, 1997. This research was supported in part by the National Science Foundation under grant ECS-9057967 with additional support from GE and DEC.

http://www.siam.org/journals/sicon/36-2/26095.html

†Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109–2122 (raja@eecs.umich.edu, stephane@eecs.umich.edu).

‡Current address: California PATH, University of California at Berkeley, Richmond Field Station Bldg. 452, 1357 S. 46th Street, Richmond, CA 94804-4698.

These concepts are concerned with the design and analysis of controllers which are *feasible* or *tolerable* by some general specification. In the field of DESs there is a substantial body of literature on the design of controllers which will satisfy logical specifications. The theory of discrete event dynamical systems has evolved to a point where it is meaningful to study control methods that not only satisfy legality specifications but also improve quantitative measures of performance. Relevant published papers on the subject are Passino and Antsaklis [7], Kumar and Garg [6], and Sengupta and Lafortune [11, 12, 14, 13]. The work of Brave and Heymann [2] on the optimal attractor problem is also of interest.

Our view of the system, the controller, and their interaction is similar to SCT [8, 16] in the following manner. We view a DES as a system that usually operates with several concurrent processes. Each process may serve a different user or objective. A typical example is that of multiple transactions, reading and writing different records running concurrently on a database. In general, the concurrent processes have interdependencies, which suggest global level problems relating to correctness, consistency, fairness, etc. SCT conceives of an object called the supervisor that enforces correctness at the overall system level. We pose our problem in this setting of system and supervisor; i.e., we wish to design supervisors that are both correct and optimal. The supervisor acts by disabling events, where the events themselves are assumed to be generated *spontaneously, instantaneously,* and *asynchronously* by the DES itself. Since the demands and disturbances on DES are uncertain, the supervisor should be *minimally restrictive*; i.e., it should allow the DES maximum freedom in its response. It is not acceptable to prevent illegal behavior by preventing all behavior! The notion of a minimally restrictive supervisor guaranteeing a legality specification is formalized in [8] and computed in [16]. We adopt these assumptions on the role of the system and the supervisor. We will address the issue of minimally restrictive supervisors that improve quantitative performance measures.

We assume a formal language representation of a DES. The language is not necessarily regular. The behavior of a DES is constituted from events. The set of possible events is known as the *alphabet* (denoted Σ). In any given time line the DES executes a sequence of events. Each sequence is assumed to be of finite length and is a string in Σ^*, which represents the Kleene closure of Σ. The set of trajectories of the uncontrolled DES is represented by a set of strings, collectively referred to as the plant language (denoted $L \subseteq \Sigma^*$). The language is prefix-closed. This is a very general representation of a DES. Some strings in L are marked and these constitute a marked language $L_m \subseteq L$. A marked string represents a properly completed behavior, i.e., one that fulfills control objectives. We assume the plant is *nonblocking* ($L = \bar{L}_m$) [4]. Of course, many DESs would exhibit blocking behaviors without supervision. Nevertheless the nonblocking plant assumption is made because the synthesis of supervisors guaranteeing this property is well understood [8, 4]. Moreover, we require that any controlled DES must also be nonblocking; i.e., the supervisor must ensure that for any past there exists some future that accomplishes control objectives. Thus, we present our problem as one of finding the optimal nonblocking supervisor given a nonblocking plant. This innocuous requirement is actually extremely fundamental, since without it the optimal solution to many problems is to have the system do nothing at all!

All performance measures are modeled by two functions that we call event and control cost functions. In this paper we present one example to illustrate their use. The event cost is associated with the execution of an event by the plant and the control cost with the disabling of an event by the supervisor. These are the only

two types of actions in an SCT framework. The event and control costs are used to induce a cost on each string the system may generate. One part of this string cost is the sum of the event costs associated with the events in the string. It represents the energies and resources expended by the plant. The other part is the sum of the control costs associated with the various events disabled by the supervisor after each prefix of the string. It represents the energies and resources expended by the controller. The cost associated with a single string lying in two different controlled sublanguages may be different, since the supervisors will be different. The event and control costs are found to be naturally antagonistic as in classical optimal control. It will be shown that trying to reduce the event costs associated with a language by using control can raise the control costs. The overall cost of a string in the language, being the sum of the event and control costs, may increase. This could be undesirable, in which case we have an optimization problem.

It is reasonable to expect that the optimal and legal supervisor synthesis problems be related in the following manner. If a solution is optimal, then it should not allow any illegal traces. Such a view implies that any optimal sublanguage lies within the supremal controllable sublanguage of the legal language (assuming also that disabling uncontrollable events entail infinite cost). Since the synthesis of the supremal controllable sublanguage of a legal language is understood from [16], we will assume that our plant language L_m is legal. If it is not, our problem can be posed on the supremal controllable sublanguage of the legal language, i.e., $(L_m \cap H)^{\uparrow}$, where H is the legal language. $(L_m \cap H)^{\uparrow}$ can be computed as discussed in [16]. Thus the problem of synthesizing an optimal legal supervisor for a regular language DES can be solved as a two-step process. This paper provides the second step.

The prior literature is as follows. The work by Passino and Antsaklis [7] examines optimal control of DESs. Unlike the paradigm of SCT, this is a forced event model where the optimal controller drives the DES along the shortest path. The SCT notion of designing a controller that is robust under a variety of disturbances and user demands is lost in this work. Thus the representation of a supervisor that provides good quantitative performance in an uncertain environment requires a more general concept of optimality than that provided in [7]. Kumar and Garg [6] also study optimal control of DESs. Theirs is a state-based formulation for DESs modeled by finite state machines (FSMs). No intuition is provided into infinite state systems, though many DESs (e.g., queuing systems) are infinite state systems without control protocols or scheduling disciplines. We will study both finite and infinite state systems. However, the control assumptions in [6] are consistent with the generality of the supervisory control philosophy. The cost structure, which assumes payoff and control costs, is interesting. Payoff costs are associated with the set of reachable states, and the control costs, with the disabling of a transition in an FSM. Disabling costs and payoffs are incurred only once regardless of the number of times the state is visited. This is often restrictive, since in most DESs the cost is an explicit function of the dynamic behavior. We will also assume two cost functions, i.e., event costs and control costs, though we relate them to the dynamic behavior in a way that is more like classical optimal control. There are also a variety of interesting modeling and computational issues associated with the control cost function that are exposed in this paper.

The work of Brave and Heymann [2] on the optimal attractor problem is also relevant. They are concerned with the cost of keeping a DES within a given finite state set. Event traces taking the system outside this set are priced, and the supervisor disables events to return the DES to the designated state set as cheaply as possible.

However, unlike [6], disabling is free. As mentioned earlier, we see the control cost function as an important modeling tool that should not be ignored.

This paper presents results on the existence and computation of optimal controllers. In Sengupta and Lafortune [12] we studied a similar problem for DES modeled by finite vertex acyclic directed graphs. Here, the existence theory is developed for DES represented by any formal language defined over a countable alphabet. The computation theory is developed for DES represented by regular languages. Therefore the problem treated in [12] is a special case of the problem studied here.

Section 2 states the problem mathematically and presents examples to illustrate the formulation. Section 3 discusses the main existence and computational results. Proofs of the existence results are presented in section 4. Some additional concepts, required to prove the computational results, are introduced in section 5. The main computability theorems are proved in the same section. Section 6 is an investigation of polynomial-time controller synthesis for DES modeled by cyclic and acyclic FSMs. We present a controller synthesis algorithm together with a proof of correctness and complexity. An example is also included in the section to help the reader follow the different steps in the computation. Section 7 is a concluding comment. An index of notation is included as an appendix.

2. Mathematical formulation of the problem. To state the problem precisely it is necessary to define the plant, the supervisor, and the relationship between them mathematically. We also define the objective function and the set of feasible solutions.

As stated earlier, the uncontrolled system or plant is described by a language L and a marked language $L_m(L = \bar{L}_m)$, defined over an alphabet $\Sigma = \{\sigma_1, \sigma_2 \ldots\}$. The notation Σ^* represents the Kleene closure of Σ, and ε represents the empty string. The alphabet exhaustively represents the various events that can occur in the DES. It is assumed that $\Sigma = \Sigma_c \cup \Sigma_{uc}$ and $\Sigma_c \cap \Sigma_{uc} = \emptyset$; i.e., the alphabet is partitioned into controllable and uncontrollable events. For a language A and string $s \in \bar{A}$, we denote the *active set* at s in A by $\Sigma_A(s) = \{\sigma \in \Sigma : s\sigma \in \bar{A}\}$ and use $A/s = \{t \in \Sigma^* : st \in A\}$ to denote the set of continuations of s in A. A/s is called the suffix language of s in A. For two strings s and t the notation $s \leq t$ denotes that s is a prefix of t.

Our usage of the terms FSM and submachine of an FSM is as usual and is taken from [8] and [4], respectively. A FSM G is a 5-tuple $G = \langle \Sigma, Q, q_0, Q_m, \delta \rangle$, where Σ is the alphabet, Q is a finite set of states, $q_0 \in Q$ is the initial state, $Q_m \subseteq Q$ is a distinguished set of marked states, and $\delta : \Sigma \times Q \to Q$ is the transition function. We also use the extended transition function $\delta^* : \Sigma^* \times Q \to Q$, defined in the usual way by composing the transition function $\delta(.,.)$; i.e., $\delta^*(s, q)$ is defined if and only if there exists a sequence of transitions s in G starting at q. A FSM $A = \langle \Sigma_A, Q_A, q_{0A}, Q_{mA}, \delta_A \rangle$ is a submachine of G if $\Sigma_A \subseteq \Sigma, Q_A \subseteq Q, Q_{mA} \subseteq Q_m$, and wherever the transition function δ_A exists it is equal to the transition function $(\delta(.,.))$ of G.

A few norms and projection functions are used for mathematical convenience. For a string, the symbol $\|.\|$ denotes the length of the string, and for a language, $\|.\|$ denotes the number of equivalence classes. If L is regular, then $\|L\|$ is finite and known as the Myhill congruence index (refer to p. 65 of [5]) of L. It is assumed that $\|\varepsilon\| = 0$. Two projection functions denoted by p and P are defined on strings and languages, respectively. For a string s, $p_j(s)$ represents the prefix of length j. For any string, $p_0(.) = \varepsilon$. For a language A, $P_j(A) = \{s \in A : \|s\| = j\}$; i.e., it is the set of all strings of length j in A.

A supervisor is a disabling control law. We clarify our usage of the terms control law, controlled system, and controllability. The definitions are similar to those in section (iv) of Ramadge and Wonham [9]. A control law π is a map $\pi : \Sigma^* \longrightarrow 2^{\Sigma}$. It specifies the set of events allowed after a string. We use Π to denote the set of all control laws. π together with a string $s \in \bar{L}_m$ specifies a prefix-closed language in the following manner.

DEFINITION 2.1 (language specified by a control law). *For $s \in \bar{L}_m$ and $\pi \in \Pi$ define*

$$\mathcal{L}(\pi, s) = \{t = \sigma_0 \dots \sigma_{\|t\|-1} \in \Sigma^* : st \in \bar{L}_m, \sigma_i \in \pi(sp_i(t)), 0 \leq i \leq \|t\| - 1\},$$
$$\mathcal{L}_m(\pi, s) = \mathcal{L}(\pi, s) \cap L_m/s.$$

If $s = \varepsilon$, then we write $\mathcal{L}(\pi)$ or $\mathcal{L}_m(\pi)$.

Thus, for a control law π and plant L_m, the supervised system is $\mathcal{L}(\pi)$. Note that it is possible to have π, π' such that $\mathcal{L}(\pi, s) = \mathcal{L}(\pi', s)$. Obviously $\mathcal{L}(\pi, s) \subseteq \overline{L_m/s}$. $\mathcal{L}_m(.,.)$ denotes the marked language specified by the control law. The trajectories of a controlled DES should constitute some nonblocking sublanguage of L. The following definition is from [8]. For $A \subseteq L$, A is *nonblocking* iff $\overline{A \cap L_m} = A$. A control law π is *nonblocking* if for all $s \in \mathcal{L}(\pi), \mathcal{L}(\pi, s) = \overline{\mathcal{L}_m(\pi, s)}$. A nonblocking control law generates a nonblocking language. Let Π_{nb} denote the set of all nonblocking control laws.

We wish to construct the optimal nonblocking supervised system. The set of nonblocking control laws is equivalent to the set of L_m-closed sublanguages [8] of L_m in the sense of the following proposition. The proposition also states that the control laws can generate any prefix-closed sublanguage of L_m.

PROPOSITION 2.2. *For the control laws defined, the following are true. Let $t \in \bar{L}_m$.*

(i) $\forall A \subseteq \overline{L_m/t}, A$ *prefix-closed* $, \exists \pi$ *s.t.* $\mathcal{L}(\pi, t) = A$.
(ii) $\{\mathcal{L}_m(\pi, t) : \pi \in \Pi_{nb}\} = \{A \subseteq L_m/t : \bar{A} \cap L_m/t = A\}$.

The proofs are trivial and omitted. The condition $\bar{A} \cap L_m/t = A$ is the L_m-closure condition. It simply says that if a marked string is in A, then all marked prefixes of the string are also in A. This is not a serious restriction for reasons explained after defining the optimization problem. Our usage of controllability is standard and from [9].

A sublanguage $A \subseteq \bar{L}_m$ is controllable iff $\bar{A}\Sigma_{uc} \cap \bar{L}_m \subseteq \bar{A}$.

As expected, our definitions have the property that a sublanguage is controllable iff it can be specified by a control law that disables only controllable events. We next propose the numerical measures required in our framework. We start by defining two nonnegative real-valued functions

$$c_e : \Sigma \to \mathbb{R}^+ \cup \{0\},$$
$$c_c : \Sigma \to \mathbb{R}^+ \cup \{0, \infty\}$$

on the alphabet. They are known as the event and control cost functions, respectively. The event cost is incurred whenever the DES generates an event, and the control cost is incurred whenever the supervisor disables an event. Note that we do not allow infinite event costs. Infinite event costs are conceivably a good way of pricing illegal strings. However, for reasons explained in the introduction, it is assumed that L_m is a legal language. It is assumed that if $\sigma \in \Sigma_{uc}$ then $c_c(\sigma) = \infty$. Thus uncontrollable

events should not be disabled. We also denote the maximum and minimum values of these functions by $\bar{c}_e, \bar{c}_c$ and $\underline{c}_e, \underline{c}_c$, respectively. The event and control cost functions are used to induce a cost on the trajectories of a supervised system.

DEFINITION 2.3 (string cost function). *Let $t \in \bar{L}_m$ be a past behavior and π be a nonblocking control law. We define*

(i) *a one-stage cost function $\bar{c} : \bar{L}_m \times \Pi \times \Sigma \longrightarrow \mathbb{R}^+ \cup \{0, \infty\}$ for an event $\sigma \in \pi(t)$ to be*

$$\bar{c}(t, \pi(t), \sigma) = c_e(\sigma) + \sum_{e' \in \Sigma_{L_m}(t) - \pi(t)} c_c(e');$$

(ii) *string cost functions for a string $s = \sigma_0 \dots \sigma_{\|s\|-1} \in \mathcal{L}(\pi, t)$ to be*

$$c(t, \pi, s) = \sum_{j=0}^{j=\|s\|-1} \bar{c}(tp_j(s), \pi(tp_j(s)), \sigma_j) + \bar{c}(ts, \pi(ts), \phi),$$

where ϕ is a stopping event having zero event cost. We also define for any $s \in \mathcal{L}(\pi, t)$,

$$\hat{c}(t, \pi, s) = \sum_{j=0}^{j=\|s\|-1} \bar{c}(tp_j(s), \pi(tp_j(s)), \sigma_j);$$

i.e., the last term in $c(.,.,.)$ is left out.

Observe that the cost of a string is dependent upon the sublanguage in which it lies, or in other words, it is influenced by the control law. The first part of the one-stage cost function is an event cost. This part is independent of the control law or sublanguage. The second part is the sum of the control costs associated with all events that must be disabled after the string t is executed. This part is determined by the control law. Thus the cost of a trajectory or string in a supervised system is the sum of the event costs of the events in the string and the control costs of the control actions taken by the supervisor during that trajectory. The term $\bar{c}(ts, \pi(ts), \phi)$ accounts for the termination cost. If the system stops with behavior $ts \in L_m$, there is no event cost associated with the stopping but there may be control costs arising from $\pi(ts)$.

Two control laws can generate the same language. However, the costs remain unaffected in the sense of the following proposition. It is stated without proof.

PROPOSITION 2.4. *Let $t \in \bar{L}_m$ and $\pi, \pi' \in \Pi_{nb}$ be such that $\mathcal{L}(\pi, t) = \mathcal{L}(\pi', t) = A$. Then for all $s \in A$,*

$$c(t, \pi, s) = c(t, \pi', s).$$

Thus we can use the notation $c(t, A, .), A \subseteq \overline{L_m/t}$, with the understanding that A is generated by any one of the suitable control laws. The objective function is as follows.

DEFINITION 2.5 (objective function). *For a language $A \subseteq L_m/t, t \in \bar{L}_m$, the objective function is*

$$c_{\sup}(A, t) = \sup_{s \in A/t} c(t, A, s).$$

If the second argument in $c_{\sup}(.,.)$ is ε, it will be omitted and we write $c_{\sup}(A)$. This objective function costs a supervisor by the worst-case behavior it allows in the face of uncertain demands and disturbances. Therefore this formulation is consistent with the view that the supervisor will generate not one but any of a set of possible behaviors. This is also a weaker notion of optimality than an average or expected value criterion. However, since in DES control the supervisor is expected to interfere as little as possible, the notion of interfering only in the worst case is appropriate. Moreover, as our subsequent results show, this problem has useful and interesting features.

2.1. The optimal control problem. The most general class of nonblocking supervised systems possible is the set of systems represented by the nonblocking sublanguages of the plant language. This set is the feasible space for the optimization problem. Since $L = \bar{L}_m$, the set of nonblocking sublanguages of L is related to the sublanguages of L_m by

$$\{A \subseteq L : A = \bar{A}, \overline{A \cap L_m} = A\} = \{\bar{A}_m : A_m \subseteq L_m\}.$$

Thus a good description of the feasible space is the set of sublanguages of L_m.

The optimal control problem is to find a sublanguage A_{om} of L_m such that

$$c_{\sup}(A_{om}) \quad = \quad \min_{A_m \subseteq L_m} c_{\sup}(A_m) < \infty.$$

Such an A_{om} is an optimal sublanguage of L_m. The empty set is not admissible as a solution unless $L_m = \emptyset$. $\{\varepsilon\}$ is admissible as a solution if it is a nonblocking sublanguage of L_m. Note that the problem requires us to minimize over both the L_m-closed and non-L_m-closed sublanguages of L_m, whereas control laws as defined can only specify L_m-closed sublanguages of L_m (Proposition 2.2). Fortunately, the cost structure ensures that every non-L_m-closed language is contained in an L_m-closed language of equal cost. Due to this observation and Proposition 2.2 we get the following equivalence that we state as a proposition. Its proof is omitted.

PROPOSITION 2.6.

$$\min_{A_m \subseteq L_m} c_{\sup}(A_m) = \min_{\pi \in \Pi_{nb}} c_{\sup}(\mathcal{L}_m(\pi)).$$

Thus, equivalently, we can minimize over the set of nonblocking control laws.

We conclude this subsection with an example and a few remarks. First, though the event and control cost functions are defined on the alphabet, the case where they are state dependent is easily accommodated by expanding the alphabet. Second, since $\sigma \in \Sigma_{uc}$ implies $c_c(\sigma) = \infty$, we have the following proposition.

PROPOSITION 2.7. *Let $L \subseteq L_m$. If $c_{\sup}(L) < \infty$, then L is controllable.*

Consequently, any optimal language is controllable.

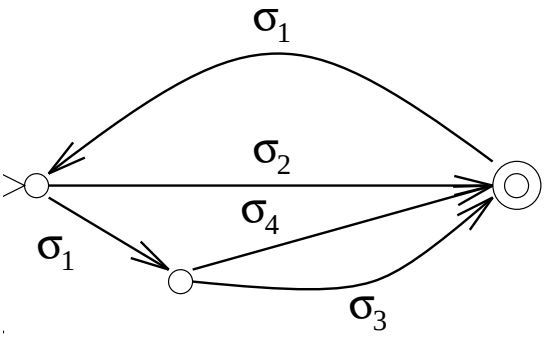
Example 2.1.1. This example illustrates the problem formulation. Let the plant language be

$$L_m = (\sigma_2 + \sigma_1(\sigma_3 + \sigma_4))(\sigma_1(\sigma_2 + \sigma_1(\sigma_3 + \sigma_4)))^*.$$

We assume all events are controllable and hence $\Sigma = \Sigma_c = \{\sigma_i : i = 1, 2, 3, 4\}$. A generator for this language and the cost functions defined on Σ are as shown in Figure 2.1.

Since $c_e(\sigma_1) > 0$ it is obvious that $c_{\sup}(L_m) = \infty$. Hence for any optimal solution the control law must set

$$\pi(\sigma_2) = \pi(\sigma_1\sigma_4) = \pi(\sigma_1\sigma_3) = \emptyset.$$



$$\begin{array}{llllll} c_e(\sigma_1) & = & 1, & c_c(\sigma_1) & = & 1 \\ c_e(\sigma_2) & = & 5, & c_c(\sigma_2) & = & 5 \\ c_e(\sigma_3) & = & 1, & c_c(\sigma_3) & = & 1 \\ c_e(\sigma_4) & = & 3, & c_c(\sigma_4) & = & 1 \end{array}$$

FIG. 2.1. *The plant.*

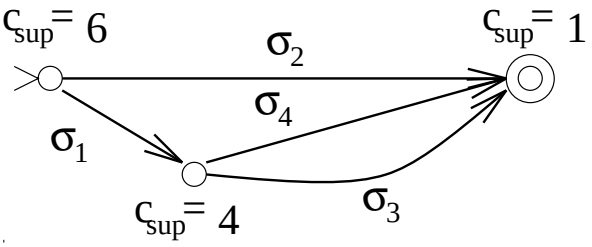


FIG. 2.2. *An optimal solution.*

Thus the optimal solution is contained in $A_m = \{\sigma_2, \sigma_1\sigma_4, \sigma_1\sigma_3\}$. The worst case of this language is set by σ_2. Since $\Sigma_{L_m}(\sigma_2) = \{\sigma_1\}$ and $\pi(\sigma_2) = \emptyset$, we obtain

$$c_{\sup}(A_m) = c(\varepsilon, A_m, \sigma_2) = c_e(\sigma_2) + c_c(\sigma_1) = 6.$$

It is easily shown by examining the six other marked sublanguages of A_m that A_m is actually optimal. It has minimum worst-case cost. The generator is as shown in Figure 2.2.

This optimization problem is also closely related to others in the literature. If the control costs are all zero, then for a language $A \subseteq L_m$, $c_{\sup}(A)$ is generated by the longest path, insofar as one exists. If all events are controllable, then a solution to the optimization problem is any shortest path in L_m. If uncontrollable events are allowed, but still no control costs, then we obtain a problem similar to the optimal attractor problem. The use of event and control costs is also standard in stochastic optimal control. Since L_m is a collection of strings of finite length defined over a countable alphabet, it is a countable set. An optimal sublanguage of L_m is like an optimal control for a Markov chain with a countable state space; i.e., the strings of the optimal language correspond to the set of state transition sequences of finite length having positive measure under the optimal control. However, this is a worst-case optimization (min-max) problem formulated on a deterministic system representation. The differences give this problem certain unique advantages and disadvantages with respect to the optimal control of Markov chains. We say more about this when presenting the specific results of this paper. For a detailed exposition the interested reader is referred to chapter 5 of [10].

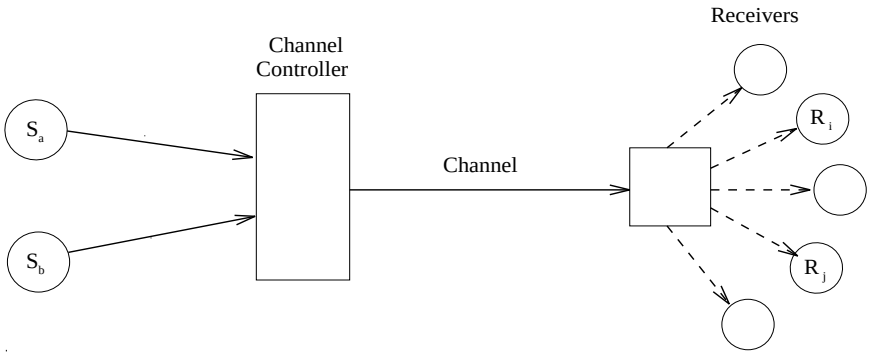


FIG. 2.3. *System configuration.*

We believe the event and control cost functions can be used in a variety of interesting ways. One use of the control costs is to represent the costs of impacting on an external environment which is not modeled. A large number of systems are embedded in very complex environments which cannot be modeled in detail. Control costs suffice as a crude way of recognizing the existence of this larger world. Event costs, on the other hand, represent physical resources such as time and energy that may be used by the plant to execute events. We present an example protocol design problem to motivate the modeling.

2.2. Example: Developing a protocol using delay priorities. Our objective is to provide tools for supervisor design. In this section we use the problem formulation to design supervisors for a channel in a communication network under three different design priorities. The example is similar to that of a CI (computer interconnect) bus connected to multiple nodes in a VAX cluster system (refer to [3]). For simplicity it is assumed that the channel has two users or senders connected to it. The task is to design a supervisor which arbitrates the channel allocation. We assume that the channel controller has complete knowledge of the demands of the senders but imprecise knowledge of conditions at the receiver end. It is also assumed that the channel controller has no control over events at the receiver end. Figure 2.3 illustrates the configuration.

The aim of the exercise is to show that if design priorities are specified by the cost functions, then the theory algorithmically develops the appropriate control laws. The system events are

t_i = sender i transmits on the channel,
y_i = sender i's transmission is positively acknowledged by the receiver,
n_i = sender i's transmission is negatively acknowledged by the receiver,

where $i \in \{a, b\}$. The set of all possible behaviors is generated by the FSM in Figure 2.4.

To complete the model we need to define the marked language. It is assumed that any behavior which gives both a and b at least one successful transmission is marked. Such a marking makes sense if both a and b have something to transmit. The new system model is as in Figure 2.5. Let L'_m denote the marked language generated by the FSM in Figure 2.5.

It is assumed that the channel controller can deny channel access to a sender. Thus the events t_a and t_b are controllable. Since receiver conditions are unknown

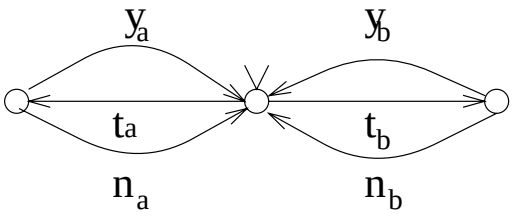


FIG. 2.4. *FSM generating all possible behaviors.*

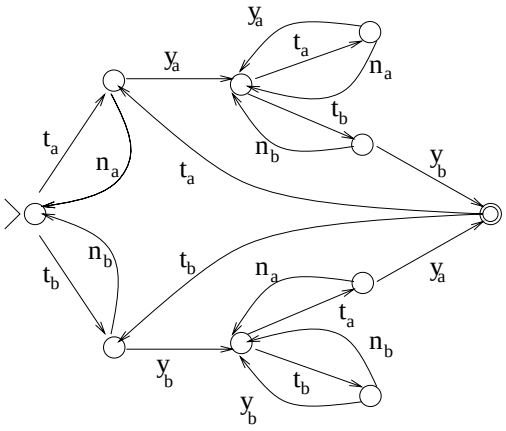


FIG. 2.5. *Plant FSM.*

the event t_i may be generally followed by either y_i or n_i. These events are assumed to be uncontrollable; i.e., $c_c(y_i) = c_c(n_i) = \infty, c_c(t_i) < \infty$. We also assume $c_e(y_i) = c_e(n_i) = 0$. Since transmission represents the use of channel time which is a network resource, it is assumed that $c_e(t_i) > 0$. It is also being assumed that all packet transmission times are the same, so that the same t_i and $c_e(t_i)$ are used for all packets.

From Figure 2.5, we see that $c_{\sup}(L'_m) = \infty$, and $c_{\sup}(L) = \infty$ for any $L \subseteq L'_m$. This is because successful transmission in finite time can only be guaranteed by disabling of uncontrollable events. Appropriately, this system has no solution under worst-case analysis. To allow the possibility of guaranteeing finite termination of transmission processes we introduce a *timeout* event and associate it with two modes of operation.

Assumption 1. If a sender accesses the channel three times in succession, then the third transmission attempt will be followed by a timeout.

Assumption 2. If a sender accesses the channel every alternate transmission, then the third transmission attempt will be followed by a timeout.

Observe that these two modes require the ordering and counting of transmissions. The alphabet is enriched to make this possible. The event t_i is superscripted as t_i^n to represent the transmission, by user i, of a packet which has already waited n transmission times in the network. This makes the alphabet Σ countably infinite. The new alphabet is more extensive than required by these two assumptions. However, the distinction between older and newer packets, introduced by the superscript n, allows us to reflect different design priorities. The timeout event represents the assumption that regardless of positive or negative acknowledgment, the job is dropped. We do not use a new event for the timeout but simply model it as a case of necessarily successful

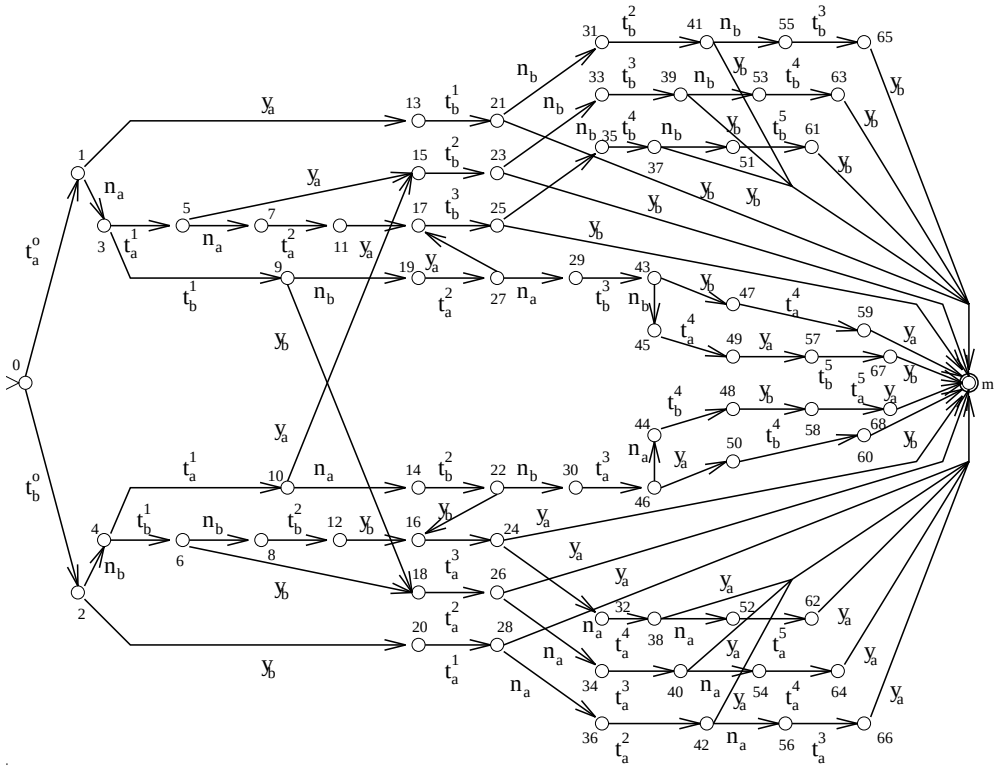


FIG. 2.6. *The FSM G_s.*

transmission. Thus a timeout state is one at which only a y_i is defined. There is no n_i.

The assumptions ensure that at least two modes of operation will terminate in a finite number of transmissions. This is not necessarily true of other modes. The FSM generating plant behavior under these two protocols is complex. Unlike the previous plant model there are now some states at which events y_i are possible but not the events n_i. These states are the timeout states. This model defines the plant. The machine will be denoted by G_m, and its generated marked language, will be denoted by L_m. It is a submachine of the superscripted version of the FSM in Figure 2.5. We will not say more about these complexities. Instead, we show a simpler way of approaching this problem, by means of the submachine of G_m, in Figure 2.6. The submachine is denoted by G_s and its generated language is L_s. We present a proof to show that all our optimization problems are reducible to equivalent problems on the FSM in Figure 2.6.

G_s is formed by disabling events in G_m. The required control actions are given in Table 2.1. Note that all the disabled events are controllable. Since all the sequences in G_s are of finite length and $c_c(t_i^n) < \infty$, G_s has finite worst-case cost. It actually generates only the fastest terminating sequences created by Assumptions 1 and 2, i.e., termination with three sequential transmissions and three alternating transmissions.

We make one more assumption. It is reasonable that controllers should clear everybody's transmission requests as fast as possible. To delay a particular user is plausible only if not doing so will hinder other more important users. The next assumption is one simple way to assure such behavior.

TABLE 2.1
Control law for the machine G_s.

State	Disabled events
7	t_b^2
13, 15, 17	t_a^0
19	t_b^2
29	t_a^3
31, 33, 35	t_a^1
45	t_b^4
51, 53, 55	t_a^2
57	t_a^0

State	Disabled events
8	t_a^2
16, 18, 20	t_b^0
14	t_a^2
30	t_b^3
32, 34, 36	t_b^1
44	t_a^4
52, 54, 56	t_b^2
58	t_b^0

Assumption 3. It is assumed that the cost $c_e(t_i) \equiv c_e$ and c_e is much greater than the control costs. More specifically,

$$c_e >> c_c(t_i), \quad \text{or in the case in which the events are denoted by } t_i^n,$$
$$c_e >> c_c(t_i^n), \quad 1 \leq n \leq 4.$$

Without this assumption, for decreasing control cost functions characterized by

$$c_c(t_i^n) > c_c(t_i^{n+1}),$$

the optimal solution may be such that the channel is unused and packets are kept waiting just because the control costs will reduce as the packets get older. The assumed high event cost will allow the examination of decreasing control cost functions without having them produce such strange solutions.

We formulate three optimal control problems (OCPs) with three types of cost structures.

OCP 1. $c_e(t_i^n) \equiv c_e$ and $c_c(t_i^n) \equiv c_c$. In other words the cost of delaying old or new packets is the same; i.e., all users, whether old or new, have equal priority.

OCP 2. $c_e(t_i^n) \equiv c_e$ and $c_c(t_i^n) = n^2, n \geq 0$. In other words the cost of delaying old packets is higher than that of delaying new packets; i.e., old users have higher priority than new users.

OCP 3. $c_e(t_i^n) \equiv c_e$ and $c_c(t_i^n) = \frac{1}{(n+1)^2}$. In other words the cost of delaying new packets is higher than that of delaying old packets; i.e., new users have higher priority than old users.

The analysis is structured into the following propositions.

PROPOSITION 2.8. *Any finite cost sublanguage of L_m must contain a string with at least six transmission events.*

Proof. It is assumed that both senders A and B have packets to transmit. Consequently a string is marked if and only if it represents successful transmission by both senders. By Assumptions 1 and 2 this can be guaranteed, without disabling uncontrollable events, only by three consecutive transmissions or by three alternating transmissions of each sender. The proposition follows. □

The following proposition establishes that our OCPs are reducible to an equivalent problem on the language generated by the FSM G_s shown in Figure 2.6.

PROPOSITION 2.9. *For all three cost functions the optimal solution is a sublanguage of L_s.*

Proof. Upon examination of Figure 2.6 and Table 2.1, we obtain that the worst-case control cost associated with a string in L_s is of the form

$$c_c(t_i^x) + c_c(t_j^y) + c_c(t_k^z) + c_c(t_l^w), \text{ with } i, j, k, l \in \{a, b\} \text{ and } 0 \leq x, y, z, w \leq 4.$$

Since by Assumption 3

$$c_e(t_i^n) = c_e > c_c(t_i^x) + c_c(t_j^y) + c_c(t_k^z) + c_c(t_l^w),$$

any sublanguage L of L_m containing a string with more than six transmission events in it, even if the string has zero control costs, will have $c_{\sup}(L) > c_{\sup}(L_s)$. Therefore, no optimal solution has more than six transmission events in any string belonging to it. This fact, together with Proposition 2.8, implies that the longest string in any optimal solution must have exactly six transmission events in it. Note that G_s represents the minimally restrictive system that is guaranteed to terminate with no more than six transmissions. Relaxing the control action at any state of G_s will force inclusion of strings of L_m containing more than six transmissions.

If G_s or L_s is the minimally restrictive controlled system guaranteeing termination in six transmission events, it follows that the optimal solution must be contained in L_s, since in the worst case it will have exactly six transmissions. □

The original problems are now equivalent to ones on a finite state system with a finite alphabet. The problem formulation is such that the optimal solution is a finitely terminating process. In practice, the protocol would reset to the initial state after reaching the marked state and repeat again.

Note that for all three OCPs $c_e(t_a^n) = c_e(t_b^n)$ and $c_c(t_a^n) = c_c(t_b^n)$. Thus the cost structure is symmetrical in a and b and so is the structure of the machine G_s. Hence, whatever cost analysis is done in one half of the graph is also true of the other half.

Solution to OCP 1. The worst-case cost in the L_s/t_a^0 half of G_s is obviously set by either

$$\begin{aligned} s &= t_a^0\ n_a\ t_a^1\ n_a\ t_a^2\ y_b\ t_b^3\ n_b\ t_b^4\ n_b\ t_b^5\ y_b \quad \text{or} \\ s' &= t_a^0\ n_a\ t_b^1\ n_b\ t_a^2\ n_a\ t_b^3\ n_b\ t_a^4\ y_a\ t_b^5\ y_b \end{aligned}$$

since these two strings involve both the largest number of transmission events and the largest number of control actions. The costs associated with the two strings are

$$\begin{aligned} c(s, L_s) &= c_e(t_a^0) + c_e(t_a^1) + c_c(t_b^2) + c_e(t_a^2) + c_c(t_a^0) + c_e(t_b^3) + c_c(t_a^1) \\ &\quad + c_e(t_b^4) + c_c(t_a^2) + c_e(t_b^5) \\ &= 6c_e + 4c_c \end{aligned}$$

and

$$\begin{aligned} c(s', L_s) &= c_e(t_a^0) + c_e(t_b^1) + c_c(t_b^2) + c_e(t_a^2) + c_c(t_a^3) + c_e(t_b^3) + c_c(t_b^4) \\ &\quad + c_e(t_a^4) + c_c(t_a^0) + c_e(t_b^5) \\ &= 6c_e + 4c_c. \end{aligned}$$

It is easily seen that the two worst-case paths in the L_s/t_b^0 half of G_s are similar and have exactly the same cost. Since any sublanguage of L_s must contain one of these four paths and they all have identical cost, it is clear that disabling any of these four paths can only increase the control costs, while leaving the event costs unchanged (refer to Theorem 3.4 in the following section). Since the cost of a string is the sum of event and control costs, the optimal solution is L_s or G_s itself.

Solution to OCP 2. Once again, the worst case in the L_s/t_a^0 half must be set by either s or s'. The associated costs are, however, different.

$$\begin{aligned} c(s, L_s) &= c_e(t_a^0) + c_e(t_a^1) + c_c(t_b^2) + c_e(t_a^2) + c_c(t_a^0) + c_e(t_b^3) + c_c(t_a^1) \\ &\quad + c_e(t_b^4) + c_c(t_a^2) + c_e(t_b^5) \\ &= 6c_e + 2^2 + 0^2 + 1^2 + 2^2 = 6c_e + 9 \end{aligned}$$

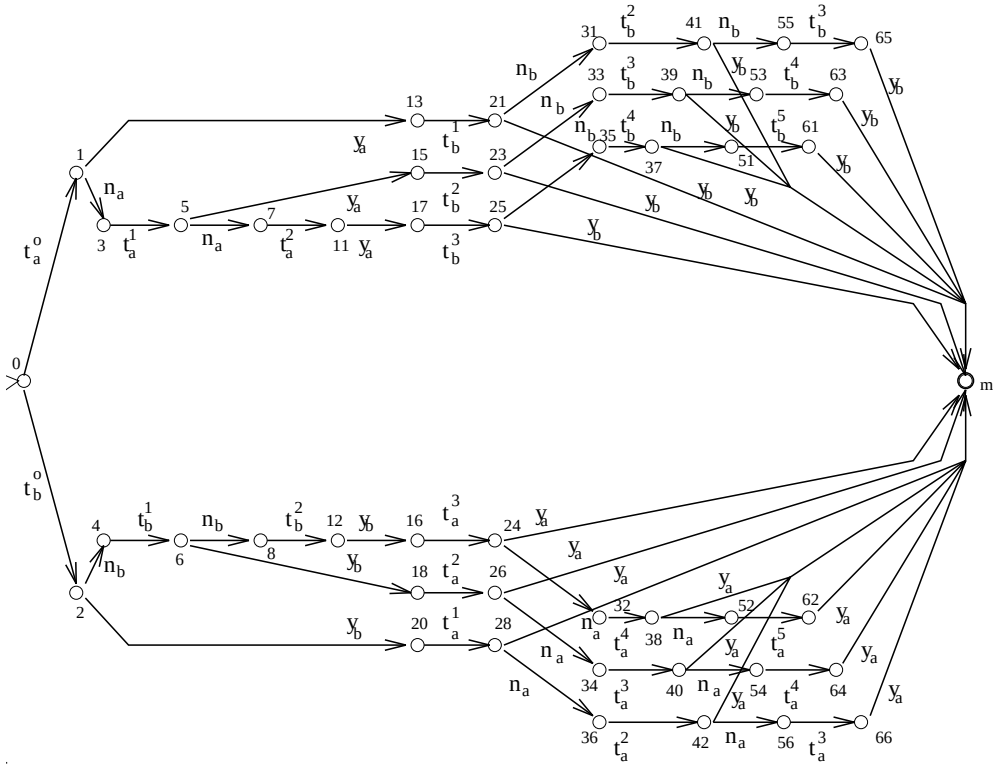


FIG. 2.7. *Solution to OCP* 2.

and

$$\begin{aligned} c(s', L_s) &= c_e(t_a^0) + c_e(t_b^1) + c_c(t_b^2) + c_e(t_a^2) + c_c(t_a^3) + c_e(t_b^3) + c_c(t_b^4) \\ &\quad + c_e(t_a^4) + c_c(t_a^0) + c_e(t_b^5) \\ &= 6c_e + 2^2 + 3^2 + 4^2 + 0^2 = 6c_e + 29. \end{aligned}$$

The analysis for the L_s/t_b^0 half is similar. By disabling the event t_b^1 after $t_a^0 n_a$ we remove the string s'. The worst-case cost is now $c_e + 9 + 1^2 = c_e + 10$. An identical cost will be obtained in the other half of G_s by disabling t_a^1 after $t_b^0 n_b$. The optimal solution is depicted in Figure 2.7. Obviously, the two control actions described above are in addition to those given in Table 2.1. The solution represents the forcing of sequential transmission.

Solution to OCP 3. Again, the worst case in the L_s/t_a^0 half must be set by either s or s'. The associated costs are different.

$$\begin{aligned} c(s, L_s) &= c_e(t_a^0) + c_e(t_a^1) + c_c(t_b^2) + c_e(t_a^2) + c_c(t_a^0) + c_e(t_b^3) + c_c(t_a^1) \\ &\quad + c_e(t_b^4) + c_c(t_a^2) + c_e(t_b^5) \\ &= 6c_e + \frac{1}{3^2} + \frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} = 6c_e + 1.472 \end{aligned}$$

and

$$\begin{aligned} c(s', L_s) &= c_e(t_a^0) + c_e(t_b^1) + c_c(t_b^2) + c_e(t_a^2) + c_c(t_a^3) + c_e(t_b^3) + c_c(t_b^4) \\ &\quad + c_e(t_a^4) + c_c(t_a^0) + c_e(t_b^5) \\ &= 6c_e + \frac{1}{3^2} + \frac{1}{4^2} + \frac{1}{5^2} + \frac{1}{1^2} = 6c_e + 1.213. \end{aligned}$$

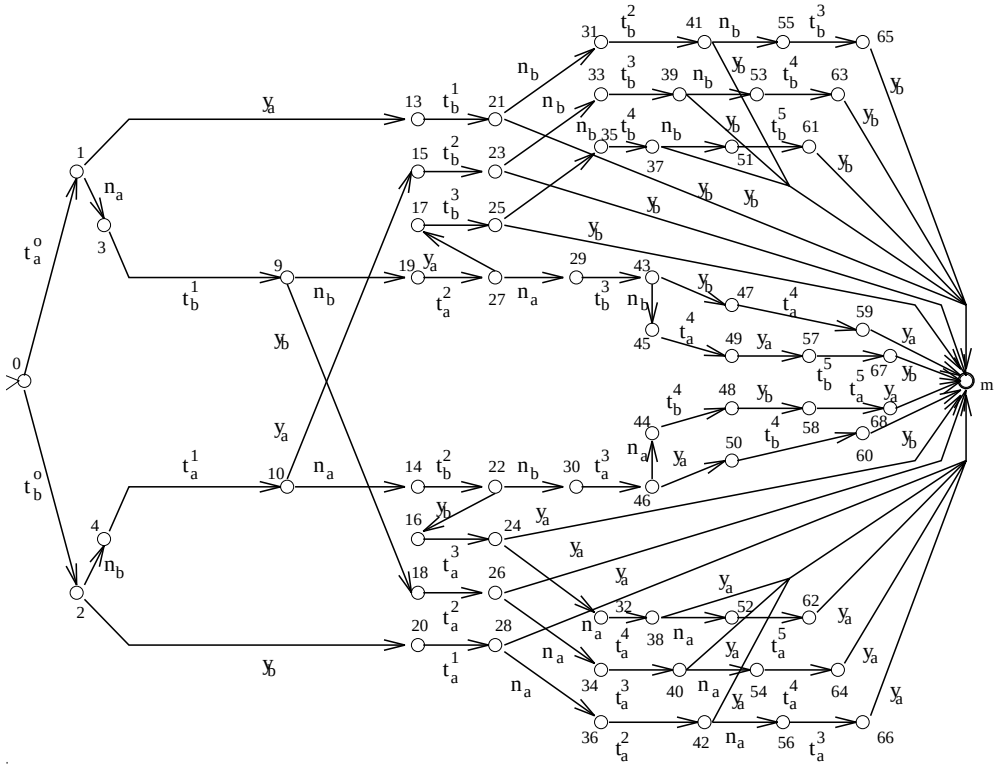


FIG. 2.8. *Solution to OCP* 3.

The analysis for the L_s/t_b^0 is similar. By disabling the event t_a^1 after $t_a^0 n_a$ we remove the string s. The worst-case cost is now $c_e + 1.213 + \frac{1}{2^2} = c_e + 1.463$. An identical cost will be obtained in the other half of G_s by disabling t_b^1 after $t_b^0 n_b$. The optimal solution is as depicted in Figure 2.8. Obviously, the two control actions described above are in addition to those given in Table 2.1. The solution represents the forcing of alternating transmission.

The clear connection between design priorities and the selected protocols can be seen if the results are summarized as follows.

OCP	**Cost assumption**	**Protocol**
1	Old and new users have same priority	Sequential or alternating protocol
2	Old users have higher priority	Force sequential protocol
3	New users have higher priority	Force alternating protocol

This example illustrates important features of our approach. The control cost of disabling a transmission event exists because of the obstruction of unmodeled sender activities; e.g., the sender may be executing programs which wait on the successful transmission of a packet. The event cost associated with spontaneously executing a transmission exists because it entails the use of channel time, which is a network resource. The control costs are used to represent the cost of impacting on an external environment which is not modeled. Event costs, on the other hand, represent physical resources such as time and energy that may be used by the plant to execute events.

We also note that, in this example, once design priorities are stated in the cost functions, then in all three cases there exist interesting and well-defined optimal solu-

tions. The optimal solutions are also minimally restrictive and satisfy the principle of dynamic programming; i.e., they have optimal substructure. The purpose of the subsequent results is to make these ideas clear and to characterize, in a general manner, problems for which we may expect these specific types of optimal solutions.

3. The principal results. For the convenience of the reader, we present all major results in this section. The section is divided into three subsections covering general existence results, the role of dynamic programming, and computational results, respectively. The proofs of the results in the first two subsections are covered in section 4. The proofs of the computational results are covered in sections 5 and 6. We present some general computational results and also investigate conditions under which optimal solutions are polynomially computable.

3.1. General existence results. The theorems in this section clarify conditions for the existence of optimal solutions and minimally restrictive optimal solutions. An optimal solution exists if the infimum of $c_{\sup}(.)$ over the sublanguages of L_m is finite and if it is realized by some sublanguage of L_m. In other words, the infimum must be a minimum that is a real number. The first existence theorem is stated for DES represented by a finite alphabet. It asserts that the existence of a sublanguage with bounded cost is sufficient for the existence of an optimum.

THEOREM 3.1. *Let $|\Sigma| < \infty$. An optimal sublanguage exists iff there exists some $A \subseteq L_m$ such that $c_{\sup}(A) < \infty$.*

The necessity of boundedness is trivial. For the sufficiency proof, it is immediate that the infimum exists. It remains to be argued that it is realized. Note that by Proposition 2.7 controllability is a necessary condition for a bounded cost sublanguage.

For the case of DES defined over a countable alphabet we present the following result.

THEOREM 3.2. *For all $s \in \bar{L}_m$ let $|\Sigma_{L_m}(s)| < \infty$. Let $c_e(.) > \delta > 0$. There exists an optimal sublanguage of L_m iff there exists $A \subseteq L_m$ such that $c_{\sup}(A) < \infty$.*

We assume that the event costs are greater than some positive number, however small. We also assume that the active event set after any string in $\bar{L}_m$ is finite. Under these assumptions we get a result similar to the first theorem. This theorem implies that optimal solutions to the problems in the example of section 2.2 exist. $\mathcal{L}(G_s)$ is a bounded cost sublanguage, and if we treat $t_i^n y_i$ and $t_i^n n_i$ as composite events with cost $c_e(t_i^n)$, then the conditions of the theorem are satisfied for all three control cost functions. The next theorem is stated for the case in which L_m is a regular language.

THEOREM 3.3. *Let L_m be a regular language. An optimal solution exists iff there exists $A_m \subseteq L_m, A_m$ regular, controllable, and having the following property for any $n \in \mathbb{N}$:*

$$\forall s = tu^*v \subseteq A_m, \quad \hat{c}(t, \bar{A}_m, u^n) = 0.$$

Intuitively the theorem says that optimal solutions exist when there are controllable sublanguages of L_m in which all cycles have zero-cost. Thus either cycles with positive cost must be broken in such a way that all the reachable broken prefixes can complete to marked states, or cycles with positive cost are altogether unreachable. Thus a system is not allowed to cycle around using resources (positive event cost) without finishing its task (completing to a marked string). The controllability condition ensures that the positive cost cycles can be broken using controllable events alone.

As a consequence of Theorem 3.3 all nonterminating behaviors of regular language DESs must be zero-cost. If the behavior of interest is itself a positive cost process,

e.g., steps involved in the manufacture of a part, then the formulation should be used to model only one cycle of the repetitive process which can then be optimized. A second step of optimization may then be to model the optimized single step-process as a repetitive zero-cost process and then use the formulation to optimize the deviation from this desired behavior, i.e., associate positive event costs on any events that disturb the process from its zero-cost trajectory. However, since such modeling approaches are domain-specific we will not discuss them further.

The next theorem characterizes the interaction of event and control costs.

THEOREM 3.4. *Let* $s \in A_m \subseteq B_m \subseteq L_m/t, t \in \bar{L}_m$. *Then* $c(t, A_m, s) \geq c(t, B_m, s)$.

The cost associated with a string lying in a nonincreasing sequence of languages is nondecreasing. Smaller languages entail more control and hence more control costs. As the language containing a string is made to shrink, the control cost associated with the string tends to rise. The event costs are independent of the language. The purpose of contracting a language is to remove strings with high event costs. Since this process is accompanied by rising control costs, we have an optimization problem. This tradeoff is similar to classical optimal control. Observe that the uncontrolled plant language has no control actions and no control costs. Its worst case is the longest path as obtained from the event costs. The use of control to disable this longest path may reduce the event cost, but only at the expense of additional control costs. The sum of event costs and control costs for the worst-case path in the new controlled system may be greater. This is the fundamental tradeoff that has made classical optimal control meaningful, and when captured as in section 2, it makes the same tradeoff interesting in the control of formal languages. This essence of optimal control is a powerful theoretical motivation for our formulation.

It can be established from Theorem 3.4 that the union of optimal solutions is optimal. As a language grows, the event costs associated with its strings do not change, but the control costs decrease. Therefore, the worst-case costs, in the union of two languages, cannot be worse than the worst of the two. Standard set-theoretic arguments lead to the existence of a unique supremal or minimally restrictive element in the class of optimal solutions. This is stated in the following theorem.

THEOREM 3.5. *If an optimal solution exists, then the unique supremal optimal solution exists.*

The supremal optimal sublanguage of L_m will be denoted by $L_o^{\uparrow}$. Note that $L_o^{\uparrow}$ should not be confused with the supremal controllable sublanguage ($L_m^{\uparrow}$) defined in [16] as a solution to the supervisory control problem defined under legality specifications. On the basis of this theorem we define an operator

$$\begin{aligned} \mathcal{L}_o^{\uparrow} &: \bar{L}_m \longrightarrow 2^{\Sigma^*}, \\ \mathcal{L}_o^{\uparrow}(s) &= \text{supremal optimal sublanguage of } L_m/s, \end{aligned}$$

if it exists. The operator is undefined otherwise. By the theorem, if an optimal solution exists in the post-language L_m/s, then the operator $\mathcal{L}_o^{\uparrow}(s)$ is well defined. Furthermore, from the definition of the Nerode equivalence relation[1] [5], if $s \equiv_{L_m} t$, where $\equiv_L$ represents the standard Nerode equivalence relation on a language $\bar{L}$, then $L_m/s = L_m/t$. This implies $\mathcal{L}_o^{\uparrow}(s) = \mathcal{L}_o^{\uparrow}(t)$. Hence it is meaningful to define

$$\mathcal{L}_o^{\uparrow} : \bar{L}_m/\equiv_{L_m} \longrightarrow 2^{\Sigma^*},$$

[1] $s, t \in L$ are Nerode equivalent iff $\{u : su \in L\} = \{v : tv \in L\}$.

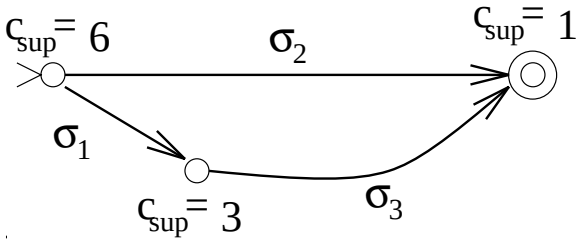


FIG. 3.1. *A DP-optimal solution.*

and accordingly, we will refer to $\mathcal{L}_o^{\uparrow}([s])$ or $\mathcal{L}_o^{\uparrow}([t])$ in the subsequent development, where $[s]$ and $[t]$ are the Nerode equivalence classes of s and t, respectively, and $\bar{L}_m/\equiv_{L_m}$ represents the set of Nerode equivalence classes of L_m. Note that $\mathcal{L}_o^{\uparrow}([\varepsilon]) = L_o^{\uparrow}$.

3.2. The principle of dynamic programming. There are interesting issues connected with the principle of dynamic programming in this problem of finding the worst-case optimal supervisor. The optimal solution to this problem is not unique. Moreover, all the optimal solutions do not structurally have optimal subsolutions (refer to [1]); i.e., they do not satisfy the principle of dynamic programming. This fact is later demonstrated by an example. Unlike the optimal control of Markov chains, the principle of dynamic programming is only a sufficient condition in this min-max problem. It is not necessary. However, we will prove that in the case of the finite alphabet, if optimal solutions exist, then solutions having optimal substructure also exist. We call this latter type a *DP-optimal* solution and define it as follows.

DEFINITION 3.6. *$A_{DO} \subseteq L_m$ is a DP-optimal solution iff it is optimal, and for all $s \in \bar{A}_{DO}$, A_{DO}/s is an optimal sublanguage of L_m/s.*

This type of optimal solution also has a unique physical significance. It guarantees the best possible future behavior, given that the system has already executed some prefix spontaneously in the past. Since this is a min-max problem, the post-languages of all optimal solutions do not have this property. It is shown later that if L_m is a regular language, then a computational specification of the supremal DP-optimal sublanguage can be derived in polynomial time by algorithms based on dynamic programming. Since we are unable to make the same claim for the supremal optimal solution, the DP-optimal solution is a crucial component of these investigations. We illustrate the distinctions by two examples.

Example 3.2.1. We illustrate the distinctions by referring back to Example 2.1.1. A_m and L_m are as previously defined. We see from Figure 2.2 that $\{\sigma_3, \sigma_4\}$ is not an optimal sublanguage of L_m/σ_1. Thus A_m does not have optimal substructure; i.e., it is optimal but not DP-optimal. We see from Figure 2.2 that $c_{\sup}(A_m/\sigma_1, \sigma_1) = 4$, whereas the optimal solution in the class L_m/σ_1 is $\{\sigma_3\}$ with $c_{\sup}(.) = 3$. We use this to construct a DP-optimal sublanguage of L_m. This language is $\{\sigma_2, \sigma_1\sigma_3\}$. A generator for this solution is shown in Figure 3.1.

Example 3.2.2. This example presents a DES with an infinite alphabet that has a supremal optimal solution but no DP-optimal solution. The plant state machine is as in Figure 3.2. The plant language is $L_m = \{b\} \cup \{a^i d_i : i \in \mathbb{N}\}$.

We assume $c_c(a) = c_c(b) = 100$ and $c_c(d_n) = 0$. The event costs are $c_e(a) = 0, c_e(b) = 1$, and $c_e(d_n) = \frac{1}{n}$. The supremal optimal sublanguage of this language is the language itself. However, in the post-language L_m/a, there is no optimal

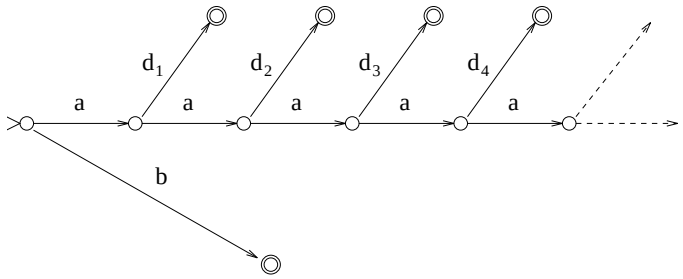


FIG. 3.2. *The plant.*

sublanguage. To see this consider the sequence of languages

$$L_0 = L_m/a, \qquad L_n = L_{n-1} - \{a^{n-1}d_n\}.$$

Then $\lim_{n\to\infty} c_{\sup}(L_n) = 0$, but there is no zero-cost nonblocking sublanguage in L_m/a. Consequently though an optimal sublanguage of L_m exists, no DP-optimal sublanguage exists.

The example reveals that the case of the infinite alphabet is complicated. However, for DES modeled by a finite alphabet, we are able to state the following theorem.

THEOREM 3.7. *Let $|\Sigma| < \infty$. If an optimal solution exists, then a DP-optimal solution exists, and furthermore the unique supremal DP-optimal solution, denoted by $L^{\uparrow}_{DO}$, exists.*

The union of DP-optimal solutions is also DP-optimal, and once again, a minimally restrictive DP-optimal solution exists. The theorem is proved by a construction detailed in section 4.2.

Our next theorem is a useful relation between the Nerode equivalence on the plant language and that on the supremal DP-optimal sublanguage.

THEOREM 3.8. *If $s, t \in L^{\uparrow}_{DO}$ and $s \equiv_{L_m} t$, then $s \equiv_{L^{\uparrow}_{DO}} t$.*

The theorem says that if two prefixes of the supremal DP-optimal sublanguage are Nerode equivalent in L_m, then they are also equivalent in $L^{\uparrow}_{DO}$.

Theorems 3.1, 3.2, 3.3, 3.5, and 3.7 constitute the existence theory. Theorems 3.1, 3.2, and 3.3 present existence conditions for different types of problems. Theorem 3.5 shows that minimally restrictive solutions are meaningful within this problem formulation and that they exist if the problem has any solution at all. Theorem 3.7 says that, for finite alphabet systems, if optimal solutions exist, then solutions having optimal substructure also exist, and there is such a unique minimally restrictive solution.

3.3. Computational results. We now discuss the main results on the computation of optimal solutions. It is possible to argue from Theorem 3.8 that if L_m is regular then $L^{\uparrow}_{DO}$ is also regular. This fact leads to the following computability theorem for DES modeled by regular languages.

THEOREM 3.9. *Let G be a FSM generating L_m and let an optimal sublanguage of L_m exist.*

(i) *There exists a unique submachine of G generating $L^{\uparrow}_{DO}$.*

(ii) *$L^{\uparrow}_{DO}$ is a regular language.*

$L^{\uparrow}_{DO}$ is exceptional in the fact that it is regular. Not all optimal solutions are necessarily regular. Theorem 3.9 tells us that in the finite set of trim submachines of G, there is a FSM generating $L^{\uparrow}_{DO}$. However, the theorem gives no way of identifying the

required submachine. In section 5 we show how to identify the FSM generating $L^{\uparrow}_{DO}$ by setting up an appropriate optimization problem on the set of trim submachines of G. It is shown that a specific type of solution to this new optimization problem generates $L^{\uparrow}_{DO}$. The time complexity is exponential in the number of states of G.

The results pertaining to polynomial-time complexity optimal controller synthesis are as follows. For DES modeled by cyclic FMSs we make additional assumptions. For DES modeled by acyclic FSMs no additional assumptions are required. For cyclic FSMs it is assumed that all event costs are positive and that the FSM has only one marked state; i.e., all strings in L_m are Nerode equivalent. The existence of a general polynomial algorithm for the acyclic case makes us suspect that it may be possible to relax the single marked state assumption for the cyclic case. However, this has not been proved. We discuss it further in section 6. We state the two main theorems.

THEOREM 3.10. *Let L_m be regular and such that all marked strings are equivalent in the sense of Nerode. Let all event costs be positive. If an optimal solution exists then, given a generator of L_m with n states, a generator for the supremal DP-optimal sublanguage is computable in time*

$$O(n^2|\Sigma|\log(|\Sigma|) + n^3|\Sigma|).$$

THEOREM 3.11. *Let the plant language L_m be generated by a trim acyclic FSM G having n states. A generator for the supremal DP-optimal solution is computable in time:*

$$O(n|\Sigma|\log(|\Sigma|)).$$

4. Proofs of the existence results. This section presents the proofs of theorems 3.1, 3.2, 3.3, 3.4, 3.5, 3.7, and 3.8. Theorems 3.9, 3.10, 3.11 require some additional concepts. They are proved in sections 5 and 6. Theorems are presented in this section in the order in which their proofs are developed. No earlier theorem uses a later theorem. The proof of Theorem 3.3 uses Theorem 3.8, and so Theorem 3.3 is proved last. All other theorems are proved in the order stated in section 3. We use the notation $\Pi_0(\pi(s)) = \Sigma_{L_m}(s) - \pi(s)$ to denote the set of active events disabled by the control law, i.e., events that determine the control cost.

4.1. Existence of optimal solutions. The following is the proof of the existence theorem for DES modeled by a finite alphabet.

THEOREM 3.1. *Let $|\Sigma| < \infty$. An optimal sublanguage exists iff there exists some $A \subseteq L_m$ such that $c_{\sup}(A) < \infty$.*

Proof. The necessity of the existence of a bounded cost sublanguage is trivial. We consider the sufficiency. The existence of $A_m \subseteq L_m$ such that $c_{\sup}(A_m) < \infty$ implies that the infimum in the optimal control problem exists. In particular, consider the set $X = \{A \subseteq L_m : c_{\sup}(A) \leq c_{\sup}(A_m)\}$. Then

$$\inf_{A \subseteq L_m} c_{\sup}(A) = \inf_{A \subseteq X} c_{\sup}(A) = x_0 < \infty,$$

where x_0 is a nonnegative real number. Pick a sequence $\langle A_n \rangle \subseteq X$ such that $\lim_{n\to\infty} c_{\sup}(A_n) = x_0$. We will establish that the elements of this sequence can take only finitely many values.

As a first step we show that for any $A \subseteq L_m, c_{\sup}(A) < c_{\sup}(A_m)$, there exists $u \in A$ such that $c(\varepsilon, A, u) = c_{\sup}(A)$. Consider a sequence $\langle u_k \rangle \subseteq A$ such that

$\lim_{k\to\infty} c(\varepsilon, A, u_k) = c_{\sup}(A)$. Now, by definition,

$$c(\varepsilon, A, u_k) = \sum_{j=0}^{j=\|u_k\|-1} \bar{c}(p_j(u_k), \pi_A(p_j(u_k)), \sigma_j^{(u_k)}) + \bar{c}(u_k, \pi_A(u_k), \phi),$$

where $u_k = \sigma_0^{(u_k)} \dots \sigma_{\|u_k\|-1}^{(u_k)}$. Since the alphabet is finite, $c_e(.)$ and $c_c(.)$ take only finitely many values, which implies that $\bar{c}(.,.,.)$ only takes values in some finite set Z. Define $\delta > 0$ such that $y \in Z$ implies that $y > \delta$ or $y = 0$. Pick M such that $M\delta \le c_{\sup}(A_m)$ but $(M+1)\delta > c_{\sup}(A_m)$. Then, in any u_k, there are not more than M events with positive one-stage costs. Thus $c(\varepsilon, A, u_k)$ takes one of $|Z|^M + 1$ possible values for all k. Since this is a finite set there exists u_k such that $c(\varepsilon, A, u_k) = c_{\sup}(A)$.

By the above property we can construct a sequence $\langle u_n \rangle, u_n \in A_n, c(\varepsilon, A_n, u_n) = c_{\sup}(A_n)$ such that $\lim_{n\to\infty} c(\varepsilon, A_n, u_n) = x_0$. But $c(\varepsilon, A_n, u_n)$ takes only values in a finite set of cardinality $|Z|^M + 1$, which implies that there exists n such that $c(\varepsilon, A_n, u_n) = c_{\sup}(A_n) = x_0$. Then A_n is an optimal sublanguage. □

The following is the proof of the existence theorem for the countable alphabet. It uses the condition $c_e(.) > \delta > 0$ to establish an equivalent finite alphabet problem. The theorem is then immediate.

THEOREM 3.2. *For all $s \in \bar{L}_m$ let $|\Sigma_{L_m}(s)| < \infty$. Let $c_e(.) > \delta > 0$. There exists an optimal sublanguage of L_m iff there exists $A_m \subseteq L_m$ such that $c_{\sup}(A_m) < \infty$.*

Proof. Let A_m be as in the hypothesis of the theorem. The necessity of the existence of a bounded cost language is trivial. We consider the sufficiency. Once again, pick $M > 0$ such that $M\delta \le c_{\sup}(A_m)$ but $(M+1)\delta > c_{\sup}(A_m)$. Then for any $A \subseteq L_m$ and $u \in A$ with $c_{\sup}(A) < c_{\sup}(A_m)$, there are at most M events in u. Consider $X = \{s \in L_m : \|s\| \le M\}$. Then

$$c_{\sup}(A) \le c_{\sup}(A_m) \Longrightarrow A \subseteq X.$$

Let $|\Sigma_{L_m}(u)|$ be finite for all $u \in \bar{L}_m$. Then $|X|$ is finite, which implies

$$\inf_{A\subseteq L_m} c_{\sup}(A) = \inf_{A\subseteq X} c_{\sup}(A) = \min_{A\subseteq X} c_{\sup}(A)$$

since $|X|$ is finite. Thus the minimum in the set X is an optimal sublanguage. □

The following is the proof of the monotonicity of the cost functions.

THEOREM 3.4. *Let $s \in A_m \subseteq B_m \subseteq L_m/t, t \in \bar{L}_m$. Then $c(t, A_m, s) \ge c(t, B_m, s)$.*

Proof. Let $s = \sigma_0 \dots \sigma_{\|s\|-1}$. By Definition 2.3

$$c(t, A_m, s) = \sum_{j=0}^{j=\|s\|-1} c_e(\sigma_j) + \sum_{j=0}^{j=\|s\|-1} \sum_{\sigma\in\Pi_0(\pi_{A_m}(p_j(s)))} c_c(\sigma) + \bar{c}(ts, \pi_{A_m}(ts), \phi).$$

But

$$\begin{array}{rrcl}
 & A_m & \subseteq & B_m \\
\Longrightarrow & \pi_{A_m}(tp_j(s)) & \subseteq & \pi_{B_m}(tp_j(s)) \\
\Longrightarrow & \Pi_0(\pi_{B_m}(p_j(s))) & \subseteq & \Pi_0(\pi_{A_m}(p_j(s))) \\
\Longrightarrow & \displaystyle\sum_{\sigma\in\Pi_0(\pi_{B_m}(tp_j(s)))} c_c(\sigma) & \le & \displaystyle\sum_{\sigma\in\Pi_0(\pi_{A_m}(tp_j(s)))} c_c(\sigma) \\
\Longrightarrow & c(t, B_m, s) & \le & c(t, A_m, s). \quad \square
\end{array}$$

The next proof establishes the existence of a unique minimally restrictive optimal solution.

THEOREM 3.5. *If an optimal solution exists, then the unique supremal optimal solution exists.*

Proof. Let L_o and L'_o be two optimal sublanguages of $L_m/t, t \in \bar{L}_m$. By Theorem 3.4, for any $s \in L_o$,

$$c(t, L_o \cup L'_o, s) \leq c(t, L_o, s).$$

Since the same argument applies for $s \in L'_o$ we have that

$$c_{\sup}(L_o \cup L'_o, t) \leq c_{\sup}(L_o, t) = c_{\sup}(L'_o, t).$$

Thus $L_o \cup L'_o$ is optimal, or in general, the union of optimal solutions is optimal. Hence

$$L_o^{\uparrow} = \bigcup_{\substack{L_o \subseteq L_m/t \\ L_o \text{ optimal}}} L_o$$

is the unique supremal optimal solution. □

4.2. Existence of the supremal DP-optimal solution. We prove the existence of the supremal DP-optimal solution by construction. The symbol "$\circ$" will denote the concatenation of a string with a language. If s is a string and A is a language, then

$$s \circ A = \{su : u \in A\}.$$

Consider the following nonincreasing sequence of sets having the supremal optimal solution as its first element.

$$\begin{aligned} & K_0 = L_o^{\uparrow}, \\ & K_n = A_n \cup B_n, \\ \text{where} \quad & A_n = \left(\bigcup_{\omega \in P_n(\bar{K}_{n-1})} \omega \circ \mathcal{L}_o^{\uparrow}([\omega]) \right), \\ & B_n = B_{n-1} \cup (P_n(A_n) \cap L_m), \\ \text{and} \quad & B_0 = \begin{cases} \{\varepsilon\} \text{ if } \varepsilon \in L_m, \\ \emptyset \text{ otherwise.} \end{cases} \end{aligned}$$

The two sets A_n and B_n contain the strings of K_n which are of length greater than n and less than n, respectively. The strings of length n are contained in both sets. It is easily shown that K_n is a sublanguage of L_m.

The intuition of the construction is easily seen from the definition of a DP-optimal solution. The nth step of the construction replaces all post-languages of prefixes of length n with supremal optimal sublanguages. It will be shown that for all $s \in \bar{K}_n, \|s\| \leq n$, K_n/s is an optimal sublanguage. K_n is in a sense "*$(n-DP)$-optimal*" and the limit may be expected to be DP-optimal. The supremality will come from the use of the supremal optimal sublanguage at each step.

We first establish that

$$K = \bigcap_{j=0}^{j=\infty} K_j$$

is nonempty whenever the supremal optimal solution (K_0) exists. The proof proceeds by showing that in the supremal optimal solution (K_0), among the strings realizing the worst-case behavior, there exists at least one whose costs cannot be reduced by optimizing at successive prefixes as is done in the construction of the sequence K_n. Intuitively, this must be the the case, because if the costs of all strings in $\mathcal{L}_o^{\uparrow}(L_m) = K_0$ could be reduced, then the supremal optimal solution would not be optimal. Lemma 4.3 (ii) proves the existence of such a string and (iii) establishes that it survives in all the K_n. Once we have established that K is nonempty, we prove that it is supremal DP-optimal. This is done by arguing that the jth element of the sequence $\langle K_n \rangle$ is such that for any $s \in \bar{K}_j$ with $\|s\| \leq j$, the language K_j/s is optimal. This is the precise meaning of the sense of ($j-$DP)-optimal. The relevant result is Lemma 4.5. Thereafter, we prove that K is supremal DP-optimal by establishing that any DP-optimal sublanguage is contained by it.

The proof is broken into five lemmas. Lemmas 4.1 and 4.2 establish certain properties of our definitions and the construction $\langle K_n \rangle$. Lemmas 4.3 and 4.5 are as described above. Lemma 4.4 is essentially a convergence result establishing that the prefix-closures of K_n converge to the prefix-closure of K. We start with Lemma 4.1, which states two simple properties of our definitions. It is stated without proof.

LEMMA 4.1.

(i) $\forall n \in \mathbb{N}, P_n(A \cup B) = P_n(A) \cup P_n(B)$.

(ii) *For all* $s \in \bar{A}_m, \sigma \in \Pi_0(\pi_{A_m}(s))$ *iff* $s\sigma \in P_{\|s\|+1}(\bar{L}_m) - P_{\|s\|+1}(\bar{A}_m)$.

The next lemma states some properties of the sequence $\langle K_n \rangle$. The first part states that the set B_n is made of strings with length not greater than n. The second part states that the sequence is a nested one. The third part says that the prefixes of length less than n of K_n remain in the prefix-closure of all subsequent languages of the sequence.

LEMMA 4.2.

(i) $\forall n \in \mathbb{N}, s \in B_n \Rightarrow \|s\| \leq n$.

(ii) $\forall n \in \mathbb{N}, P_n(\bar{K}_n) = P_n(\bar{K}_{n-1})$.

(iii) $\forall n \in \mathbb{N}, K_n \subseteq K_{n-1} \subseteq L_m$.

(iv) $\forall j, n, n'$ *where* $j \leq n \leq n'$, $P_j(\bar{K}_n) = P_j(\bar{K}_{n'})$.

Proof.

(i) By definition, $B_0 = \{\varepsilon\}$ or $\emptyset$. If $B_0 = \emptyset$, then $\|s\| \leq 0$ trivially. If $B_0 = \{\varepsilon\}$, then $\|\varepsilon\| = 0$ by definition. Hence B_0 satisfies the hypothesis. Let the result be true for $n-1$. By definition, $B_n = (P_n(A_n) \cap L_m) \cup B_{n-1}$. If $s \in B_{n-1}$, then by assumption, $\|s\| \leq n-1$. If $s \in P_n(A_n) \cap L_m$ then $\|s\| = n$. The result follows.

(ii)

$$
\begin{array}{rcll}
P_n(\bar{K}_n) & = & P_n(\overline{A_n \cup B_n}) & \\
& = & P_n(\bar{A}_n \cup \bar{B}_n) & \\
& = & P_n(\bar{A}_n) \cup P_n(\bar{B}_n) & \text{[by (i)]} \\
& = & P_n(\bar{A}_n) \cup P_n(\bar{B}_{n-1}) \cup P_n(\overline{P_n(A_n) \cap L_m}) & \text{[by (i)]} \\
& = & P_n(\bar{A}_n) \cup (P_n(A_n) \cap L_m) & \text{[by (i), } P_n(B_{n-1}) = \emptyset\text{]} \\
& = & P_n(\bar{A}_n) & [P_n(A_n) \subseteq P_n(\bar{A}_n)] \\
& = & P_n(\bar{K}_{n-1}) & \text{[by definition of } K_n\text{]}.
\end{array}
$$

(iii) We first establish that $A_n \subseteq A_{n-1}$.

Let $s = p_{n-1}(s)\sigma_n t \in A_n$. Then, by definition of A_n, $t \in \mathcal{L}_o^{\uparrow}([p_n(s)])$ and $p_n(s) = p_{n-1}(s)\sigma_n \in P_n(\bar{K}_{n-1})$, which implies $p_{n-1}(s)\sigma_n \in \bar{K}_{n-1}$, $p_{n-1}(s)\sigma_n \in$

$\bar{A}_{n-1}$ (by (1)), and $\sigma_n \in \overline{\mathcal{L}_o^{\uparrow}([p_{n-1}(s)])}$. Thus

$$c_{\sup}(\mathcal{L}_o^{\uparrow}([p_{n-1}(s)])) \geq c_e(\sigma_n) + \sum_{\sigma \in \Pi_0\left(\pi_{\mathcal{L}_o^{\uparrow}([p_{n-1}(s)])}(\varepsilon)\right)} c_c(\sigma) + c_{\sup}(\mathcal{L}_o^{\uparrow}([p_{n-1}(s)])/\sigma_n)$$

$$\geq c_e(\sigma_n) + \sum_{\sigma \in \Pi_0\left(\pi_{\mathcal{L}_o^{\uparrow}([p_{n-1}(s)])}(\varepsilon)\right)} c_c(\sigma) + c_{\sup}(\mathcal{L}_o^{\uparrow}([p_n(s)]))$$

by the optimality of $\mathcal{L}_o^{\uparrow}([p_n(s)])$. Since $\mathcal{L}_o^{\uparrow}([p_{n-1}(s)])$ is supremal optimal, this implies that $\sigma_n \circ \mathcal{L}_o^{\uparrow}([p_n(s)]) \subseteq \mathcal{L}_o^{\uparrow}([p_{n-1}(s)])$. Thus

$$s = p_{n-1}(s)\sigma_n t \in p_{n-1}(s)\sigma_n \mathcal{L}_o^{\uparrow}([p_n(s)]) \subseteq p_{n-1}(s) \circ \mathcal{L}_o^{\uparrow}([p_{n-1}(s)]) \subseteq A_{n-1},$$

since $p_{n-1}(s) \in P_{n-1}(\bar{A}_{n-1})$. This proves $A_n \subseteq A_{n-1}$.

The inclusion $K_n \subseteq K_{n-1}$ is now easy. Let $s \in K_n = A_n \cup (P_n(A_n) \cap L_m) \cup B_{n-1} = A_n \cup B_{n-1}$. If $s \in A_n$, then $s \in A_{n-1} \subseteq K_{n-1}$. If $s \in B_{n-1}$ then $s \in K_{n-1}$ by definition. The inclusion $K_n \subseteq L_m$ follows from $K_0 \subseteq L_m$ and an inductive application of $K_n \subseteq K_{n-1}$.

(iv) By part (iii)

$$K_{n+1} \subseteq K_n \Rightarrow \bar{K}_{n+1} \subseteq \bar{K}_n \Rightarrow P_j(\bar{K}_{n+1}) \subseteq P_j(\bar{K}_n).$$

Let $s \in P_j(\bar{K}_n), j \leq n$. Then $\|s\| = j$ and $\exists s'$ such that $ss' \in K_n$. The argument may be considered in two cases.

Case 1. $ss' \in B_n$. By definition of B_{n+1}, $ss' \in B_{n+1}$, which implies $s \in P_j(\bar{K}_{n+1})$.

Case 2. $ss' \notin B_n$. Then $ss' \in A_n$, but $ss' \notin P_n(A_n)$. Hence $p_{n+1}(ss')$ exists, which implies $p_{n+1}(ss') \in P_{n+1}(\bar{K}_n) = P_{n+1}(\bar{K}_{n+1})$. Since $j \leq n$ and $\|s\| = j$, we get $s \in P_j(\bar{K}_{n+1})$. By induction, $s \in P_j(\bar{K}_{n'})$ for all $n' \geq n$. □

We prove next that K is nonempty in three steps. First we state the simple property that if a string is not in the supremal optimal sublanguage, then any language containing that string is nonoptimal. Next we show that there exists a string in K_0 that realizes the supremum, and the post-languages corresponding to all prefixes of this string are optimal. Finally, it is established that this particular string must be contained in K.

LEMMA 4.3.

(i) *Let $s \in L_m/u$ and $s \notin \mathcal{L}_o^{\uparrow}(L_m/u)$. Then, for any $L \subseteq L_m/u$ and $s \in L$, we have*

$$c_{\sup}(L) > c_{\sup}(\mathcal{L}_o^{\uparrow}(L_m/u)).$$

(ii) *There exists $s \in K_0 = \mathcal{L}_o^{\uparrow}(L_m)$ such that $K_0/p_i(s)$ is optimal for all $i, 0 \leq i \leq \|s\|$.*

(iii) *Let there exist s as in part* (ii). *Then $s \in K$.*

Proof.

(i) Let the hypotheses be true and $c_{\sup}(L) \leq c_{\sup}(\mathcal{L}_o^{\uparrow}(L_m/u))$. Then L is optimal. This implies $s \in L \subseteq \mathcal{L}_o^{\uparrow}(L_m/u)$, which contradicts the hypothesis. The result follows.

(ii) Suppose that for all $s \in K_0$ there exists i such that $c_{\sup}(K_0/p_i(s), p_i(s)) > c_{\sup}(\mathcal{L}_o^{\uparrow}(L_m/p_i(s)), p_i(s))$. Let $i(s)$ denote the smallest such i for s. If $s \leq t$ then $i(s) = i(t)$. Note that $K_0 = \mathcal{L}_o^{\uparrow}(L_m)$. Define

$$L = \bigcup_{t \in \overline{\mathcal{L}_o^{\uparrow}(L_m)}} p_{i(t)}(t) \circ \mathcal{L}_o^{\uparrow}(L_m/p_{i(t)}(t)).$$

Let $u \in L$. Then there exists $s \in K_0$, such that $u = p_{i(s)}(s)v$. Then we have

$$\begin{aligned} c(\varepsilon, L, u) &= c(\varepsilon, L, p_{i(s)}(s)v) \\ &\leq \bar{c}(\varepsilon, L, p_{i(s)}(s)) + c_{\sup}(L/p_{i(s)}(s), p_{i(s)}(s)) \\ &= \bar{c}(\varepsilon, L, p_{i(s)}(s)) + c_{\sup}(\mathcal{L}_o^{\uparrow}(L_m/p_{i(s)}(s)), p_{i(s)}(s)) \\ &< \bar{c}(\varepsilon, L, p_{i(s)}(s)) + c_{\sup}(K_0/p_{i(s)}(s), p_{i(s)}(s)). \end{aligned}$$

Next we claim that for all j, such that $0 \leq j \leq i(s) - 1$,

$$p_j(s)\sigma \in \bar{K}_0 \Leftrightarrow p_j(s)\sigma \in \bar{L}.$$

The case $p_j(s)\sigma = p_{j+1}(s)$ is trivial. The following argument is for the case $p_j(s)\sigma \neq p_{j+1}(s)$. Let $p_j(s)\sigma \in \bar{K}_0$. Then there exists $w = p_j(s)\sigma v \in K_0$ and $i(w)$ such that $K_0/p_{i(w)}(w)$ is not optimal. By definition of $i(s)$, $i(w) > j$. This is true since $i(w) \leq j$ implies that $i(w) < i(s)$, which contradicts the definition of $i(s)$ as the smallest such i for the string s. Thus $p_{i(w)}(w) \circ \mathcal{L}_o^{\uparrow}(L_m/p_{i(w)}(w)) \subseteq L$, which implies $p_j(s)\sigma \in \bar{L}$. For the reverse inclusion assume $p_j(s)\sigma \in \bar{L}$. Then there exists v such that $p_j(s)\sigma v \in L$, and there exists $w \in K_0$ such that $p_{i(w)}(w) \in \bar{K}_0$ and $p_j(s)\sigma v \in p_{i(w)}(w) \circ \mathcal{L}_o^{\uparrow}(L_m/p_{i(w)}(w))$. By definition of $i(s)$, $i(w) > j$, which implies $p_j(s)\sigma \in \bar{K}_0$. This proves the claim. The claim implies $\bar{c}(\varepsilon, L, p_{i(s)}(s)) = \bar{c}(\varepsilon, K_0, p_{i(s)}(s))$, from which

$$c(\varepsilon, L, u) < \bar{c}(\varepsilon, K_0, p_{i(s)}(s)) + c_{\sup}(K_0/p_{i(s)}(s), p_{i(s)}(s)) \leq c_{\sup}(K_0)$$

for all $u \in L$. Next we use the finiteness of the alphabet to show that there exists $u \in L$ such that $c_{\sup}(L) = c(\varepsilon, L, u)$. Define a sequence $\langle u_n \rangle$ in L such that

$$\lim_{n \to \infty} c(\varepsilon, L, u_n) = c_{\sup}(L).$$

Since the alphabet, which is the domain of definition of c_e, and c_c, is finite, the range of these functions is also finite. This, together with the fact that $c_{\sup}(L) < \infty$, implies that $c(\varepsilon, L, .)$ can take only finitely many values. Thus there exists $N \in \mathbb{N}$ such that $c(\varepsilon, L, u_n) = c_{\sup}(L)$ for $n \geq N$. Hence,

$$c_{\sup}(L) = c(u_n, L) < c_{\sup}(K_0),$$

which contradicts the optimality of K_0. The result follows.

(iii) Let $s \notin k$. Get the smallest n such that $s \notin K_n$. Then $s \in K_{n-1}$, which implies that $p_n(s) \in \bar{K}_n$, by definition of K_n. Let $s = p_n(s)v_n$. Then $v_n \notin \bar{K}_n/p_n(s)$, and by definition of K_n, $v_n \notin \mathcal{L}_o^{\uparrow}(L_m/p_n(s))$. But $v_n \in L_m/p_n(s)$ and $v_n \in K_0/p_n(s) \subseteq L_m/p_n(s)$. By part (i)

$$c_{\sup}(K_0/p_n(s), p_n(s)) > c_{\sup}(\mathcal{L}_o^{\uparrow}(L_m/p_n(s)), p_n(s)).$$

This contradicts the hypothesis on s. Hence $s \in K_n$ for all n, implying $s \in K$. ∎

Observe that part (ii) of the lemma says that the hypothesis of part (iii) is not vacuous. The two parts taken together then establish that K is nonempty. The next lemma states that the sequence K_n develops K incrementally.

LEMMA 4.4. $\forall j \leq n \in \mathbb{N}, P_j(\bar{K}) = P_j(\bar{K}_n)$.

Proof. We prove $\bar{K} = \cap_{n=0}^{n=\infty} \bar{K}_n$. The inclusion $\bar{K} \subseteq \cap_{n=0}^{n=\infty} \bar{K}_n$ is obvious.

Let $s \in \cap_{n=0}^{n=\infty} \bar{K}_n$. Then for all n there exists v_n such that $sv_n \in \bar{K}_n$. In particular, consider $K_{\|s\|}/s = \mathcal{L}_o^{\uparrow}([s])$. We use Lemma 4.3 with K_0 replaced by $\mathcal{L}_o^{\uparrow}([s]) = H_0$ and

L_m replaced by L_m/s. Denote the analogous sequence by $\langle H_n \rangle$. It is easy to show that $H = K/s$. Then by Lemma 4.3, there exists $v \in H = K/s$. This implies $sv \in K$ and hence $s \in \bar{K}$. This proves the reverse inclusion.

Now by the above and Lemma 4.2 (iv),

$$P_j(\bar{K}) = \bigcap_{n=0}^{n=\infty} P_j(\bar{K}_n) = P_j(\bar{K}_i)$$

for any $i \geq j$. The result follows. □

The next lemma states that the worst-case costs associated with the post-language of a prefix s do not change after $\|s\|$ iterations, and moreover, that the post-language becomes and remains optimal.

LEMMA 4.5.

(i) $\forall s \in K_j, c_{\sup}(K_j/s, s) = c_{\sup}(K_i/s, s) = c_{\sup}(\mathcal{L}_o^{\uparrow}([s])), j \geq i \geq \|s\|$.

(ii) $\forall s \in \bar{K}, c_{\sup}(K/s, s) = c_{\sup}(K_j/s, s) = c_{\sup}(\mathcal{L}_o^{\uparrow}([s])), j \geq \|s\|$.

Proof.

Claim 4.5.1. $\forall s \in \bar{K}_i, \|s\| \leq i \leq j, 0 \leq k \leq \|s\|$,

$$\Pi_0(\pi_{K_i}(p_k(s))) = \Pi_0(\pi_{K_j}(p_k(s))).$$

Proof. First consider the case $k < i \leq j$.

$$\begin{array}{rrcll} & \sigma & \in & \Pi_0(\pi_{K_i}(p_k(s))) & \\ \Leftrightarrow & p_k(s)\sigma & \in & P_{k+1}(\bar{L}_m) - P_{k+1}(\bar{K}_i) & \text{[Lemma 4.1 (ii)]} \\ \Leftrightarrow & p_k(s)\sigma & \in & P_{k+1}(\bar{L}_m) - P_{k+1}(\bar{K}_j) & \text{[Lemma 4.2 (iv)]} \\ \Leftrightarrow & \sigma & \in & \Pi_0(\pi_{K_j}(p_k(s))) & \text{[Lemma 4.1 (ii)].} \end{array}$$

Next consider the case $k = i < j$.

$$\begin{array}{rrcll} & \sigma & \in & \Pi_0(\pi_{K_i}(p_i(s))) & \\ \Leftrightarrow & p_i(s)\sigma & \in & P_{i+1}(\bar{L}_m) - P_{i+1}(\bar{K}_i) & \text{[Lemma 4.1 (ii)]} \\ \Leftrightarrow & p_i(s)\sigma & \in & P_{i+1}(\bar{L}_m) - P_{i+1}(\bar{K}_{i+1}) & \text{[Lemma 4.2 (ii)]} \\ \Leftrightarrow & p_i(s)\sigma & \in & P_{i+1}(\bar{L}_m) - P_{i+1}(\bar{K}_j) & \text{[Lemma 4.2 (iv)]} \\ \Leftrightarrow & \sigma & \in & \Pi_0(\pi_{K_j}(p_i(s))) & \text{[Lemma 4.1 (ii)].} \end{array}$$

The case $k = i, j = i$ is trivial. □

Next we show that $c_{\sup}(K_{i+1}/s, s) \leq c_{\sup}(K_i/s, s)$ for all $s \in \bar{K}_i, i \geq \|s\|$.

$$\begin{aligned} c_{\sup}(K_{i+1}/s, s) &= c_{\sup}((A_{i+1} \cup B_{i+1})/s, s) \\ &= c_{\sup}((A_{i+1} \cup (P_{i+1}(A_{i+1}) \cap L_m) \cup B_i)/s, s) \\ &= c_{\sup}((A_{i+1} \cup B_i)/s, s) \\ &= c_{\sup}(A_{i+1}/s \cup B_i/s, s) \\ &= c_{\sup}((\cup_{\omega \in P_{i+1}(\bar{K}_i)} \omega \circ \mathcal{L}_o^{\uparrow}([\omega]))/s \cup B_i/s, s). \end{aligned}$$

If $u \in B_i/s$ then $c(s, K_{i+1}/s, u) = c(s, K_i/s, u)$ by the following argument:

$$u \in B_i/s \Rightarrow su \in B_i \Rightarrow \|su\| \leq i \quad \text{[Lemma 4.2 (i)]},$$

from which by Claim 4.5.1,

$$\Pi_0(\pi_{K_i}(p_k(su))) = \Pi_0(\pi_{K_{i+1}}(p_k(su))).$$

From the definition of the cost of a string in a language it is now evident that

$$\begin{array}{rrcl} & c(\varepsilon, K_{i+1}, su) & = & c(\varepsilon, K_i, su) \\ \Rightarrow & c(s, K_{i+1}/s, u) & = & c(s, K_i/s, u) \\ \Rightarrow & c(s, K_{i+1}/s, u) & \leq & c_{\sup}(K_i/s, s). \end{array}$$

This takes care of the case $u \in B_i/s$. The other possible case is $u \in A_{i+1}/s$ with $u = t\omega$ and $st \in P_{i+1}(\bar{K}_i)$. Let $t = \sigma_0 \ldots \sigma_{\|t\|-1}$. Observe that $\hat{c}(s, K_i/s, t) = \hat{c}(s, K_{i+1}/s, t)$ since

$$\begin{array}{rcl} \hat{c}(s, K_i/s, t) & = & \displaystyle\sum_{l=0}^{l=\|t\|-1} c_e(\sigma_l) + \sum_{k=\|s\|}^{k=\|st\|-1} \sum_{\sigma \in \Pi_0(\pi_{K_i}(p_k(st)))} c_c(\sigma) \\ & = & \displaystyle\sum_{l=0}^{l=\|t\|-1} c_e(\sigma_l) + \sum_{k=\|s\|}^{k=\|st\|-1} \sum_{\sigma \in \Pi_0(\pi_{K_{i+1}}(p_k(st)))} c_c(\sigma) \quad \text{[Claim 4.5.1]} \\ & = & \hat{c}(s, K_{i+1}/s, t), \end{array}$$

where Claim 4.5.1 is used with $p_{\|st\|-1}(st)$ substituted for s, $\|st\| - 1$ substituted for i, $\|st\|$ substituted for j, and k such that $\|s\| \leq k \leq \|st\| - 1$. Equipped with this observation on the nature of $\hat{c}(.,.,.)$, the following argument is made.

$$\begin{array}{rcl} c(s, K_{i+1}/s, u) & = & c(s, K_{i+1}/s, t\omega) \\ & = & \hat{c}(s, K_{i+1}/s, t) + c(st, K_{i+1}/st, \omega) \\ & = & \hat{c}(s, K_i/s, t) + c(st, K_{i+1}/st, \omega) \\ & \leq & \hat{c}(s, K_i/s, t) + c_{\sup}(K_{i+1}/st, st) \\ & = & \hat{c}(s, K_i/s, t) + c_{\sup}(\mathcal{L}_o^{\uparrow}([st]), st) \\ & \leq & \hat{c}(s, K_i/s, t) + c_{\sup}(K_i/st, st) \\ & \leq & c_{\sup}(K_i/s, s). \end{array}$$

Since u is arbitrary we obtain from the two cases that $c_{\sup}(K_{i+1}/s, s) \leq c_{\sup}(K_i/s, s)$. By induction, it follows that $\forall j \geq i \geq \|s\|$,

$$c_{\sup}(K_j/s, s) \leq c_{\sup}(K_i/s, s).$$

In particular, if $i = \|s\|$, then

$$c_{\sup}(K_j/s, s) \leq c_{\sup}(K_{\|s\|}/s, s) = c_{\sup}(\mathcal{L}_o^{\uparrow}([s]), s)$$

by definition of K_i. Now the optimality of $\mathcal{L}_o^{\uparrow}([s])$ gives $c_{\sup}(K_j/s, s) = c_{\sup}(\mathcal{L}_o^{\uparrow}([s]), s)$. Thus part (i) is proved.

The proof of part (ii) is similar. We begin with an analogous claim.

Claim 4.5.2. $\forall s \in \bar{K}, \|s\| \leq j, 0 \leq k \leq \|s\|$,

$$\Pi_0(\pi_K(p_k(s))) = \Pi_0(\pi_{K_j}(p_k(s))).$$

Proof. Consider first the case $k < \|s\|$.

$$\begin{array}{rrcll} & \sigma & \in & \Pi_0(\pi_K(p_k(s))) & \\ \Leftrightarrow & p_k(s)\sigma & \in & P_{k+1}(\bar{L}_m) - P_{k+1}(\bar{K}) & \text{[Lemma 4.1 (ii)]} \\ \Leftrightarrow & p_k(s)\sigma & \in & P_{k+1}(\bar{L}_m) - P_{k+1}(\bar{K}_j) & \text{[Lemma 4.4]} \\ \Leftrightarrow & \sigma & \in & \Pi_0(\pi_{K_j}(p_k(s))) & \text{[Lemma 4.1 (ii)].} \end{array}$$

For the case $k = \|s\|$,

$$\begin{array}{rrcll} & \sigma & \in & \Pi_0(\pi_K(s)) & \\ \Leftrightarrow & s\sigma & \in & P_{\|s\|+1}(\bar{L}_m) - P_{\|s\|+1}(\bar{K}) & \text{[Lemma 4.1 (ii)]} \\ \Leftrightarrow & s\sigma & \in & P_{\|s\|+1}(\bar{L}_m) - P_{\|s\|+1}(\bar{K}_{\|s\|+1}) & \text{[Lemma 4.4]} \\ \Leftrightarrow & s\sigma & \in & P_{\|s\|+1}(\bar{L}_m) - P_{\|s\|+1}(\bar{K}_{\|s\|}) & \text{[Lemma 4.2 (ii)]} \\ \Leftrightarrow & \sigma & \in & \Pi_0(\pi_{K_{\|s\|}}(p_{\|s\|}(s))) & \text{[Lemma 4.1 (ii)]} \\ \Leftrightarrow & \sigma & \in & \Pi_0(\pi_{K_j}(s)) & \text{[Claim 4.5.1].} \quad \square \end{array}$$

This claim is used to prove part (ii). Let $t = \sigma_0 \dots \sigma_{\|t\|-1}$. By definition,

$$\begin{array}{rcll} c(s, K/s, t) & = & \displaystyle\sum_{k=0}^{k=\|t\|-1} c_e(\sigma_k) + \sum_{k=\|s\|}^{k=\|st\|} \sum_{\sigma \in \Pi_0(\pi_K(p_k(st)))} c_c(\sigma) & \\ & = & \displaystyle\sum_{k=0}^{k=\|t\|-1} c_e(\sigma_k) + \sum_{k=\|s\|}^{k=\|st\|} \sum_{\sigma \in \Pi_0(\pi_{K_{\|st\|}}(p_k(st)))} c_c(\sigma) & \text{[Claim 4.5.2]} \\ & = & c(s, K_{\|st\|}/s, t) & \\ & \leq & c_{\sup}(K_{\|st\|}/s, s) & \\ & = & c_{\sup}(K_j/s, s), \quad j \geq \|s\| & \text{[part (i)]} \\ & = & c_{\sup}(\mathcal{L}_o^{\uparrow}([s]), s). & \end{array}$$

Thus $c_{\sup}(K/s, s) \leq c_{\sup}(K_j/s, s), j \geq \|s\|$, and by the optimality of $\mathcal{L}_o^{\uparrow}([s])$

$$c_{\sup}(K/s, s) = c_{\sup}(K_j/s, s) = c_{\sup}(\mathcal{L}_o^{\uparrow}([s]), s)$$

for all $j \geq \|s\|$. $\square$

We now have all properties of the sequence $\langle K_n \rangle$ necessary to prove existence of the supremal DP-optimal solution.

THEOREM 3.7. *Let $|\Sigma| < \infty$. If an optimal solution exists, then a DP-optimal solution exists, and furthermore, the unique supremal DP-optimal solution, denoted by $L_{DO}^{\uparrow}$, exists.*

Proof. We claim that K is the supremal DP-optimal solution. Note first that by Lemma 4.1 (iii), K is a sublanguage of L_m. By Lemma 4.3, K is nonempty. The DP-optimality of K is immediate from Lemma 4.5 (ii) since for all $s \in \bar{K}$

$$c_{\sup}(K/s, s) = c_{\sup}(\mathcal{L}_o^{\uparrow}([s]), s),$$

which is the definition of a DP-optimal sublanguage of L_m.

The supremality of K is established as follows. Let A_{DO} be any DP-optimal sublanguage of L_m and $s = \sigma_0 \dots \sigma_{\|s\|-1} \in A_{DO} \subseteq L_o^{\uparrow}$. Since $L_o^{\uparrow}(= K_0)$ exists, we get $\varepsilon \in \bar{K}_0$, whence $\varepsilon \in P_0(\bar{K}_0)$. This is the base case for an inductive argument. Assume $p_j(s) \in P_j(\bar{K}_j)$. Since $A_{DO}/p_j(s)$ is optimal and $K_j/p_j(s) = \mathcal{L}_o^{\uparrow}([p_j(s)])$ it implies that

$$\begin{array}{rrcll} & A_{DO}/p_j(s) & \subseteq & K_j/p_j(s) & \\ \Rightarrow & \sigma_j & \in & \overline{A_{DO}/p_j(s)} \subseteq \overline{K_j/p_j(s)} & \\ \Rightarrow & p_{j+1}(s) & \in & \bar{K}_j & \\ \Rightarrow & p_{j+1}(s) & \in & P_{j+1}(\bar{K}_j) = P_{j+1}(\bar{K}_{j+1}) & \text{[by Lemma 4.2 (ii)].} \end{array}$$

Thus, by induction, $p_{\|s\|}(s) = s \in P_{\|s\|}(\bar{K}_{\|s\|})$. Since $s \in L_m$ and $\mathcal{L}_o^{\uparrow}([s])$ exists, we get $\varepsilon \in \mathcal{L}_o^{\uparrow}([s])$. This implies $s \in A_{\|s\|}$ and so

$$s \in P_{\|s\|}(A_{\|s\|}) \cap L_m \subseteq B_{\|s\|} \subseteq K.$$

Since s was arbitrary we have $A_{DO} \subseteq K$. Thus K contains all DP-optimal solutions, and being itself DP-optimal, it is the unique supremal DP-optimal solution. □

The next theorem proves that if two strings are Nerode equivalent in the plant language, they are Nerode equivalent in the supremal DP-optimal sublanguage.

THEOREM 3.8. *If $s, t \in L^{\uparrow}_{DO}$ and $s \equiv_{L_m} t$ then $s \equiv_{L^{\uparrow}_{DO}} t$.*

Proof. For $s \in \bar{L}_m$ let $\mathcal{L}^{\uparrow}_{DO}(s)$ represent the supremal DP-optimal sublanguage in the post-language L_m/s if it exists. Since $s, t \in L^{\uparrow}_{DO}$ and it is supremal DP-optimal we have

$$L^{\uparrow}_{DO}/s = \mathcal{L}^{\uparrow}_{DO}(s),$$
$$L^{\uparrow}_{DO}/t = \mathcal{L}^{\uparrow}_{DO}(t)$$

because by definition of DP-optimality the post-languages are DP-optimal, and it is easily shown that if the post-languages are not supremal, then $L^{\uparrow}_{DO}$ is not supremal either. But $s \equiv_{L_m} t$ implies that $L_m/s = L_m/t$. This, together with the uniqueness of the supremal DP-optimal sublanguage, gives

$$\mathcal{L}^{\uparrow}_{DO}(s) = \mathcal{L}^{\uparrow}_{DO}(t) \Rightarrow L^{\uparrow}_{DO}/s = L^{\uparrow}_{DO}/t \Rightarrow s \equiv_{L^{\uparrow}_{DO}} t. \qquad \square$$

The last theorem in this section is the existence theorem for DESs represented by regular languages. The theorem uses the existence of the unique supremal DP-optimal sublanguage, Theorems 3.8 and 3.1.

THEOREM 3.3. *Let L_m be a regular language. An optimal solution exists iff there exists $A_m \subseteq L_m, A_m$ regular, controllable, and having the following property for any $n \in \mathbb{N}$:*

$$\forall s = tu^*v \subseteq A_m, \quad \hat{c}(t, \bar{A}_m, u^n) = 0.$$

Proof. If an optimal solution exists, then $L^{\uparrow}_{DO}$ exists. By Theorem 3.8, $\|L^{\uparrow}_{DO}\| \leq \|L_m\|$. This implies that $L^{\uparrow}_{DO}$ is regular. By Proposition 2.7, $L^{\uparrow}_{DO}$ is also controllable. For $tu^*v \subseteq L^{\uparrow}_{DO}, \hat{c}(t, L^{\uparrow}_{DO}, u^n) \geq n\hat{c}(t, L^{\uparrow}_{DO}, u)$, which implies $c_{\sup}(L^{\uparrow}_{DO}) \geq n\hat{c}(t, L^{\uparrow}_{DO}, u)$. Then necessarily, $\hat{c}(t, L^{\uparrow}_{DO}, u) = 0$. This establishes that the given conditions are necessary for the existence of an optimal solution.

The sufficiency is argued as follows. Let A_m be as in the hypothesis of the theorem. Let $s = \sigma^s_0 \ldots \sigma^s_{\|s\|-1} \in A_m$. The condition on $\hat{c}(.,.,.)$ implies the following. For $i \leq j$,

$$p_i(s) \sim_{A_m} p_j(s) \Longrightarrow \hat{c}(p_i(s), \pi_{A_m}, \sigma^s_i \ldots \sigma^s_{j-1}) = 0.$$

Thus,

$$\bar{c}(p_k(s), \pi_{A_m}(p_k(s)), \sigma^s_k) > 0 \Longrightarrow [p_k(s)\sigma^s_k]_{A_m} \neq [p_i(s)]_{A_m}$$

for all $i \leq k$. Since L_m is a regular language $|\{k : [p_k(s)\sigma^s_k]_{A_m} \neq [p_i(s)]_{A_m}, i \leq k\}| \leq \|A_m\|$. Thus

$$\begin{array}{rrcl} & c(\varepsilon, A_m, s) & \leq & \|A_m\|(\bar{c}_e + |\Sigma|\bar{c}_c) \\ \Longrightarrow & c_{\sup}(A_m) & \leq & \|A_m\|(\bar{c}_e + |\Sigma|\bar{c}_c). \end{array}$$

Thus $A_m \subseteq L_m$ is a bounded cost sublanguage defined over a finite alphabet. The existence of an optimal sublanguage is immediate from Theorem 3.1. □

All the existence results stated in section 3 are now proven.

5. Computability. We have developed the existence theory and some of the structural properties of optimal supervisors for DESs represented by any formal language consisting of strings of finite length. This section is concerned with the additional developments possible for DESs represented by regular languages only; i.e., we are able to develop controller synthesis algorithms having polynomial complexity. The algorithms involve several complex manipulations of FSMs and are unfortunately more complicated than the algorithms synthesizing legal supervisors. Accordingly, we have tried to be careful and rigorous in arguing the correctness and complexity of our synthesis algorithms. Intuitive explanations, showing similarities with shortest path algorithms for finite vertex-directed graphs, are provided. However, this controller synthesis problem computes an optimal submachine of a FSM or, in graph-theoretic terms, an optimal subgraph of a directed graph. It requires additional stages of processing not seen in the shortest path problem.

In general, there are infinitely many sublanguages of L_m, and therefore, infinitely many candidate solutions. It is assumed in the subsequent development that an optimal sublanguage, i.e., one that realizes the infimum, exists. We show how to synthesize a FSM generating the supremal DP-optimal sublanguage ($L_{DO}^{\uparrow}$). This is done in two steps. First it is shown that a FSM generating the supremal DP-optimal solution is contained within the set of submachines of any FSM (G) generating the plant language. This is Theorem 3.9. While this theorem characterizes a finite set within which the solution may be found, it gives no way of identifying the submachine of interest. We solve this problem by setting up an appropriate optimization problem on the set of submachines of G. The submachine of interest is an optimal solution to this problem. Thus, a FSM generating an optimal sublanguage can at least be found by evaluating the objective function for every submachine of G and then finding one that realizes the minimum. This is the import of Theorem 5.3.

The following additional notation is used in this section. We define for any FSM A and $q \in Q_A$,

$$\begin{aligned} \mathcal{T}(A) &= \{(\sigma, q, q') : \sigma \in \Sigma, q \in Q_A, \delta_A(\sigma, q) = q'\}, \\ \mathcal{T}(A, q) &= \{(\sigma, q, q') : \sigma \in \Sigma, \delta_A(\sigma, q) = q'\}. \end{aligned}$$

These two functions represent the transitions in the machine A and the transitions defined at each state of A, respectively. $\Sigma_A(q)$ will denote the active event set at state q of machine A. The projection functions π_1, π_2, π_3 are used to represent the first, second, and third components of the 3-tuple (σ, q, q'), respectively. Recall the definition of a submachine of a FSM from section 2. It is immediate that $\mathcal{T}(A) \subseteq \mathcal{T}(G)$. The statement "$A \subseteq G$" denotes that A is a submachine of G. We also say A is a submachine of G at q whenever $q_{0A} = q \in Q$ and $A \subseteq G$.

We are particularly interested in *trim* [8] submachines of G. Trim submachines of G at q_0 generate nonblocking sublanguages of L_m. If G is accessible with respect to q_0 and co-accessible with respect to Q_m then it is *trim*. The notation $\mathcal{M}(G, q) = \{A \subseteq G : A \text{ trim}, q_{0A} = q\}$ represents the set of trim submachines of G at q. The set $\mathcal{M}(G, q)$ has a maximal element, in the sense that all other elements of $\mathcal{M}(G, q)$ are submachines of the maximal element. The maximal element is denoted by $M(G, q)$. The language generated by an FSM A is denoted by $\mathcal{L}(A)$, and the marked language, by $\mathcal{L}_m(A)$. If A is trim then $\mathcal{L}(A) = \overline{\mathcal{L}_m(A)}$. We reserve the symbol G for a trim generator of L_m. Since L_m is nonblocking,

$$\overline{\mathcal{L}_m(G)} = \bar{L}_m = L = \mathcal{L}(G).$$

The following general property of regular languages gives us an important implication of Theorem 3.8. Its proof is omitted. The interested reader is referred to [10].

LEMMA 5.1. *Let $A_m \subseteq L_m$ have the property,*

$$((s, t \in \bar{A}_m) \wedge (s \equiv_{L_m} t)) \Rightarrow s \equiv_{A_m} t.$$

Let $G = \langle \Sigma, Q, q_o, Q_m, \delta \rangle$ be an FSM generating L_m. Then there exists a submachine of G that generates A_m.

It is evident from Theorem 3.8 that $L^{\uparrow}_{DO}$ and L_m satisfy the preconditions of Lemma 5.1. We can now prove Theorem 3.9.

THEOREM 3.9. *Let G be an FSM generating L_m and let an optimal sublanguage of L_m exist.*

(i) *There exists a unique submachine of G generating $L^{\uparrow}_{DO}$.*

(ii) *$L^{\uparrow}_{DO}$ is a regular language.*

Proof. By Theorem 3.7, $L^{\uparrow}_{DO}$ exists. Part (i) follows from Theorem 3.8 and Lemma 5.1. The uniqueness property follows from the determinism of G. (ii) is immediate from (i). □

To identify the required submachine we define a new optimization problem on the set of trim submachines of G as follows.

For all $q \in Q, A_o \in \mathcal{M}(G, q)$ is an optimal submachine if

$$c^g_{\sup}(A_o) = \min_{A \in \mathcal{M}(G,q)} c^g_{\sup}(A) < \infty.$$

The notation $c^g_{\sup}(A)$ represents the worst-case behavior that is possible in submachine A. Its mathematical definition is

$$c^g_{\sup}(A) = \sup_{s \in \mathcal{L}_m(A)} c^g(q_{0A}, A, s),$$

where $c^g(q_{0A}, A, s)$ is the cost of a string s, which starts at q_{0A} and is generated by A. For any submachine A and state $q \in Q_A$ and string $s = \sigma^s_0 \sigma^s_1 \ldots \sigma^s_{\|s\|-1}$, such that $\delta^*_A(s, q)$ exists, the mathematical definition of $c^g(.,.,.)$ is

$$\begin{aligned} c^g(q, A, s) &= \sum_{j=0}^{j=\|s\|-1} c_e(\sigma^s_j) + \sum_{j=0}^{j=\|s\|} \sum_{e' \in \mathcal{T}(G, \delta^*(p_j(s), q)) - \mathcal{T}(A, \delta^*_A(p_j(s), q))} c_c(e') \\ &= \sum_{j=0}^{j=\|s\|-1} \bar{c}^g(\delta^*_A(p_j(s), q), A, \sigma^s_j) + \bar{c}^g(\delta^*_A(s, q), A, \phi), \end{aligned}$$

where $\bar{c}^g(.,.,.)$ is a one-stage cost function. The lower limit of the control cost summation represents the set of transitions in G that are disabled in A. The one-stage cost function is defined for any submachine A, state $q' \in Q_A$, and $\sigma \in \Sigma_A(q')$ by

$$\bar{c}^g(q', A, \sigma) = c_e(\sigma) + \sum_{\tau \in \mathcal{T}(G, q') - \mathcal{T}(A, q')} c_c(\pi_1(\tau)).$$

Note that ϕ is a dummy symbol having zero event cost. The term containing ϕ has only control costs associated with the end of the string. We also define, for mathematical convenience, the function

$$\hat{c^g}(q, A, s) = \sum_{j=0}^{j=\|s\|-1} \bar{c}^g(\delta^*_A(p_j(s), q), A, \sigma^s_j),$$

with the last term of $c^g(.,.,.)$ missing.

Note the similarity of $c^g(.,.,.)$ with the function $c(.,.,.)$ denoting the cost of a string occurring in a sublanguage of L_m. The cost of a transition σ generated by A at state q', $(\bar{c}^g(q', A, \sigma))$, is the event cost of the transition plus the control cost of all events disabled at state q' in submachine A. The event cost of a transition is independent of the submachine. This is not true of the control cost. A transition in a smaller submachine may have more control costs associated with it than the same transition in a bigger submachine. This is because smaller submachines imply more disabling actions. The following result, similar to Theorem 3.4, is useful. Its proof is omitted.

LEMMA 5.2. *Let $A \subseteq B \subseteq G$. Then for all $s \in \Sigma^*$ and $q \in Q_A$ such that $\delta_A^*(s, q)$ is defined, we have*

$$\bar{c}^g(q, A, s) \geq \bar{c}^g(q, B, s),$$
$$c^g(q, M(A, q), s) \geq c^g(q, M(B, q), s) \quad \text{if } \delta_A^*(s, q) \in Q_{mA}.$$

The new optimization problem on FSMs has been set up to get the following equalities. If $s \in \mathcal{L}_m(A)$ and t is such that $\delta^*(t, q_0) = q_{0A}$, then $c^g(q_{0A}, A, s) = c(t, \mathcal{L}_m(A), s)$. Obviously, then, for any submachine A of G,

$$c^g_{\sup}(A) = c_{\sup}(\mathcal{L}_m(A)). \tag{5.1}$$

This is the desired relationship between the objective function of our new FSM optimization problem and the original language optimization problem. This relationship does not in itself imply that an optimal solution to the FSM problem generates an optimal solution to the language problem. The complication is due to the fact that in general most sublanguages of L_m are not generated by submachines of G. However, Theorem 3.9 together with the above equality are enough to establish the equivalence of the FSM and the language optimization problems. To state the theorem we require one more concept. This is the definition of a DP-optimal submachine. It is analogous to the definition of a DP-optimal sublanguage.

A submachine $A_{DO} \in \mathcal{M}(G, q)$ is DP-optimal iff it is optimal and for all $q' \in Q_{A_{DO}}, M(A_{DO}, q')$ is an optimal submachine in $\mathcal{M}(G, q')$.

Note that the statement "A is an optimal submachine of G" will imply $q_{0A} = q_0$ unless stated otherwise. If a particular DP-optimal FSM includes all other DP-optimal FSMs as submachines of itself, then we call it the *maximal DP-optimal submachine.* The following theorem asserts the existence of a unique maximal DP-optimal submachine of G and that it generates the supremal DP-optimal sublanguage of L_m. The theorem concludes our investigation of computability since the maximal DP-optimal submachine can be found by exhaustively searching the finite set of trim submachines of G. Recall that the notation $\mathcal{L}^{\uparrow}_{DO}(.)$ represents the supremal DP-optimal sublanguage of the language (.).

THEOREM 5.3. *Assume that an optimal sublanguage of L_m exists. Then an optimal submachine of G exists, and the unique maximal DP-optimal submachine $(G^{\uparrow}_{DO})$ of G also exists. Moreover, $\mathcal{L}_m(G^{\uparrow}_{DO}) = \mathcal{L}^{\uparrow}_{DO}(\mathcal{L}_m(G))$; i.e., the maximal DP-optimal submachine generates the supremal DP-optimal sublanguage of $\mathcal{L}_m(G)$.*

Proof. Since an optimal sublanguage exists, by Theorem 3.7, $L^{\uparrow}_{DO}$ exists and $c_{\sup}(L^{\uparrow}_{DO}) < \infty$. From this, by (5.1), Theorem 3.9, and the finiteness of the set of submachines of G, an optimal submachine of G exists. By Theorem 3.9 there exists a submachine of G that generates $L^{\uparrow}_{DO}$. Let this FSM be denoted by $G^{\uparrow}_{DO} \subseteq G$. We show that $G^{\uparrow}_{DO}$ is the maximal DP-optimal submachine of G.

Claim 5.3.1. $G^{\uparrow}_{DO}$ is a DP-optimal submachine of G.

Proof. Let there exist $A_q \in \mathcal{M}(G,q), q \in Q_{G^{\uparrow}_{DO}}$, such that

$$c^g_{\sup}(A_q) < c^g_{\sup}(M(G^{\uparrow}_{DO},q)).$$

Pick some s such that $\delta^*_{G^{\uparrow}_{DO}}(s,q_0) = q$. By (5.1)

$$c_{\sup}(\mathcal{L}_m(A_q),s) < c_{\sup}(\mathcal{L}_m(M(G^{\uparrow}_{DO},q)),s) = c_{\sup}(L^{\uparrow}_{DO}/s,s).$$

This contradicts the DP-optimality of $L^{\uparrow}_{DO}$. Thus $M(G^{\uparrow}_{DO},q)$ is optimal. Since q is arbitrary, $G^{\uparrow}_{DO}$ is a DP-optimal submachine of G. □

Claim 5.3.2. Let A be a DP-optimal submachine of G. Then $\mathcal{L}_m(A)$ is a DP-optimal sublanguage of $\mathcal{L}_m(G)$; i.e., for all $s \in \overline{\mathcal{L}_m(A)}, \mathcal{L}_m(A)/s$ is optimal.

Proof. Pick any $s \in \overline{\mathcal{L}_m(A)}$ such that $\delta^*_A(s,q_0) = q$. $M(A,q)$ is a DP-optimal submachine in $\mathcal{M}(G,q)$. Thus, by (5.1), $c^g_{\sup}(M(A,q)) = c_{\sup}(\mathcal{L}_m(A)/s,s) < \infty$. This implies that an optimal solution exists in $\mathcal{L}_m(G)/s$. By Theorem 3.9 there exists a submachine in the set $\mathcal{M}(G,q)$ which generates $\mathcal{L}^{\uparrow}_{DO}(\mathcal{L}_m(G)/s)$. Let this submachine be represented by $A^{\uparrow}_{DOq}$. By Claim 5.3.1, $A^{\uparrow}_{DOq}$ is a DP-optimal submachine of $M(G,q)$. $M(A,q)$ is DP-optimal by hypothesis, which gives

$$c_{\sup}(\mathcal{L}_m(A)/s,s) = c^g_{\sup}(M(A,q)) = c^g_{\sup}(A^{\uparrow}_{DOq}) = c_{\sup}(\mathcal{L}^{\uparrow}_{DO}(\mathcal{L}_m(G)/s),s).$$

Thus $\mathcal{L}_m(A)/s$ is optimal in $\mathcal{L}_m(G)/s$. Since s was any string in $\overline{\mathcal{L}_m(A)}$, $\mathcal{L}_m(A)$ is a DP-optimal sublanguage. □

Claim 5.3.3. A maximal DP-optimal submachine of G exists. It is $G^{\uparrow}_{DO}$.

Proof. Let A be a DP-optimal submachine of G. Pick any $s \in \mathcal{L}_m(A)$. By Claim 5.3.2, $\mathcal{L}_m(A)$ is DP-optimal. It must be a sublanguage of $\mathcal{L}^{\uparrow}_{DO}(\mathcal{L}_m(G))$, which by hypothesis is generated by $G^{\uparrow}_{DO}$. Hence $s \in \mathcal{L}_m(G^{\uparrow}_{DO})$. Thus $\mathcal{L}_m(A) \subseteq \mathcal{L}_m(G^{\uparrow}_{DO})$, and since both are submachines of the deterministic machine G, we obtain $A \subseteq G^{\uparrow}_{DO}$. Thus $G^{\uparrow}_{DO}$ is the maximal DP-optimal submachine of G. □

The proof of the theorem is immediate from Claim 5.3.3. □

The maximal DP-optimal submachine of a machine G at q will be denoted by $M^o_D(G,q)$. Theorem 5.3 establishes that if an optimal sublanguage exists, then the maximal DP-optimal solution to the FSM problem will generate the supremal DP-optimal solution to the language problem. Observe that the only assumptions made are that the plant language is regular and the costs are nonnegative. We have been unable to prove this method for other optimal solutions. It is easy to show from Theorem 5.3 and (5.1) that any optimal FSM of G will generate an optimal sublanguage of $\mathcal{L}_m(G)$. It is the converse that is not obvious. In particular, we do not know whether the supremal optimal sublanguage is generated by an optimal submachine of G or, for that matter, by any FSM at all.

On the basis of Theorem 5.3 the computational complexity is exponential. However, we will make some additional assumptions in order to address the issue of polynomial computability in section 6.

6. Polynomial computation of optimal solutions. In this section we investigate the polynomial computability of optimal solutions. Additional assumptions are required to establish the results in this section. We treat languages modeled by cyclic and acyclic FSMs separately. This is because the computational complexities

in the two cases, and also the required assumptions, are significantly different. The synthesis algorithms together with proofs of correctness and complexity are stated for cyclic systems. Acyclic systems are easier to deal with than cyclic systems. They are discussed informally.

6.1. DP-optimal solutions: The cyclic case. As stated before, two additional assumptions are required for polynomial computability. They are

(i) $\forall \sigma \in \Sigma, c_e(\sigma) > 0$,
(ii) $|Q_m| = 1$.

In other words all event costs are positive, and there is only one marked state in G. This implies that all strings of $\mathcal{L}_m(G)$ are equivalent since they go to the same state. Note that this is not necessarily true of $\mathcal{L}(G)$.

We are relatively comfortable with the first assumption since it appears likely that in practice the execution of any event will entail the use of some system resource and accordingly some positive cost. Of course, from a theoretical standpoint, this cost is allowed to be any arbitrarily small positive number. The critical property realized by this assumption is that though plant FSMs may contain cycles, the optimal submachines must be acyclic. In a way we have a type of decyclization problem on a finite vertex digraph. Unfortunately the general decyclization problem is intractable. Here we exploit properties of the cost structure to do this in polynomial time.

The second assumption seems more severe. Note that connecting multiple marked states to a hypothetical marked state does not work. If the algorithm given below is started at this hypothetical marked state, it is easy to construct cases for which the algorithm will be incorrect. We explain this further after presenting the algorithm. For acyclic FSMs the single marked state assumption is not required, which might suggest the existence of a polynomial-time algorithm in the general case. However, based on our investigations of the multiple marked state case, we conjecture that any algorithm accommodating multiple marked states will be of higher complexity than the algorithm presented here.

An intuitive explanation of the algorithm follows this formal statement. It is basically a DP-algorithm that recurses backwards state by state. At each state it solves a "local" optimization problem. We formalize this local object as a one-step submachine. It is any FSM comprised of a state in $q \in Q$ and some of the transitions defined out of it, i.e., any subset of $\mathcal{T}(G, q)$. Any state of Q_m is allowed to be a one-step submachine by itself.

DEFINITION 6.1. *A is a one-step submachine of G at q if*

$$A = \langle \Sigma, Q_A, q_{0A}, Q_{mA}, \delta_A \rangle$$

satisfies

(i) $q_{0A} = q$,
(ii) $\mathcal{T}(A) \subseteq \mathcal{T}(G, q)$,
(iii) $\mathcal{T}(A) = \emptyset \Rightarrow q_{0A} \in Q_m$,
(iv) $Q_A = \{q' \in Q : \exists \tau \in \mathcal{T}(A), (\pi_3(\tau) = q')\} \cup \{q\}$,
(v) $Q_{mA} = Q_m \cap Q_A$.

Note that A is not generally trim. The transitions of A are some subset of those defined out of q in G. Condition (iii) says that only in the case $q \in Q_m$ is the trivial submachine

$$\langle \Sigma, \{q\}, q, \{q\}, \delta|_{\Sigma \times \emptyset} \rangle$$

a valid one-step submachine of G at q. We denote the set of all one-step submachines of G at q by $\mathcal{M}_1(G,q)$. This set also has a maximal element in the sense of containing all other elements as submachines. It is denoted by $M_1(G,q)$.

The Algorithm. All new notation adopted in this section is defined below for easy reference. The states of the plant FSM are assumed to be stored in a data structure having pointers to parent states and children states. Note that $|Q| = n$. Complexity will be stated in terms of n and $|\Sigma|$.

(i) SL = solved list of states (also sometimes called the closed list).

(ii) $\mathcal{T}_{opt}(q)$ = the set of transitions of $M_1(M_D^o(G,q),q)$, the maximal one-step submachine of $M_D^o(G,q)$ rooted at q.

(iii) $CL = \{(q, c_{\sup}(M_D^o(G,q))) : q \in Q_G\}$. This is the cost list maintained by the algorithm for the recursive computation of the cost associated with a particular submachine.

The variables CL^{temp} and $\mathcal{T}_{opt}^{temp}(.)$ are also used for temporary storage of the same quantities.

(iv) C = set of states to be processed in the current iteration.

(v) $P_f(C) = \{q \in Q : \exists \tau \in \mathcal{T}(G), (\pi_2(\tau) = q) \wedge (\pi_3(\tau) \in C)\}$.

(vi) $S_f(C) = \{q \in Q : \exists \tau \in \mathcal{T}(G), (\pi_3(\tau) = q) \wedge (\pi_2(\tau) \in C)\}$.

In other words $P_f(C)$ is the set of parent states of set C, and $S_f(C)$, the set of children of states of C. A function denoted by $c_{\max}(E)$, where E is some set of transitions of a one-step submachine, will be used in the algorithm. This is appropriate since each step of the algorithm only computes some one-step object and assumes that the rest has been correctly computed in prior iterations. The function represents worst-case cost.

For the convenience of modularity we present the algorithm as a main program and two subprograms referred to as Optimize and One-Step Optimize, respectively. The main program primarily orders the backward recursive search and updates C and SL based on data provided by the subprogram Optimize. The optimization at each state is described in the subprogram Optimize, which in turn calls the subprogram One-Step Optimize.

The Main Program.

(i) Input: $\Sigma, Q, q_o, q_m, \delta, c_e, c_c$.

(ii) Initialize: $C = \{q_m\}, SL = \emptyset, CL = \emptyset, CL^{temp} = \emptyset$. If there exists $\sigma \in \Sigma_{uc}$ and $q \in Q$ such that $(\sigma, q_m, q) \in \mathcal{T}(G)$, then STOP since no optimal solution exists. Otherwise set

$$\hat{E_0}(q_m) = E_0(q_m) = \emptyset.$$

(iii) Optimize: Call subprogram Optimize with argument C.

(iv) Compute

$$A = \{q_d \in C : c_{\max}(\mathcal{T}_{opt}^{temp}(q_d)) = \min_{q \in C} c_{\max}(\mathcal{T}_{opt}^{temp}(q))\}.$$

(Note: $c_{\max}(.)$ is computed in the subprogram One-Step Optimize.)

(v)

$$\begin{aligned} \forall q \in A \quad \mathcal{T}_{opt}(q) &= \mathcal{T}_{opt}^{temp}(q), \\ CL &\rightarrow CL \cup \{(q, c_{\max}(\mathcal{T}_{opt}(q))\}, \\ SL &\rightarrow SL \cup A, \\ CL^{temp} &= \emptyset. \end{aligned}$$

(vi) Termination condition: Is $q_o \in SL$? If yes then STOP. Otherwise continue.

(vii) Computation of the states to be optimized in the next iteration: Compute the following.

$$\begin{array}{rrcl} & P_f(SL) & & \\ & A & = & P_f(SL) - SL, \\ \forall q \in A, & \hat{E}_0(q) & = & \{\tau \in \mathcal{T}(G,q) : \pi_3(\tau) \notin SL\}, \\ & E_0(q) & = & \mathcal{T}(G,q) - \hat{E}_0(q), \\ & B & = & \{q \in A : \exists \tau \in \hat{E}_0(q), \pi_1(\tau) \in \Sigma_{uc}\}, \\ & C & = & A - B. \end{array}$$

If $C = \emptyset$, then STOP since no optimal solution exists.

(viii) GOTO (iii).

Complexity: Steps (i) and (viii) are independent of n. Step (ii) is linear in $|\Sigma|$. Steps (iv), (v), and (vi) are linear in n. Step (vii) is $O(n^2|\Sigma|)$. This is explained in Remark 6.1.1 made after the statement of this algorithm. Let the complexity of (iii) be $O(x)$. Because of step (viii) the complexity of the main program is $O(n(x+n^2|\Sigma|))$.

Optimize.

(i) Input: $C, \{\hat{E}_0(q) : q \in C\}, \{E_0(q) : q \in C\}$.

(ii) Pick any $q_d \in C$.

(iii) Update C: $C = C - \{q_d\}$.

(iv) If $E_0(q_d) \neq \emptyset$ order $E_0(q_d)$ such that

$$i < j \Rightarrow c_e(\pi_1(\tau_i)) + c^g_{\sup}(M^o_D(G, q_i)) \leq c_e(\pi_1(\tau_j)) + c^g_{\sup}(M^o_D(G, q_j)),$$

where $\pi_3(\tau_k) = q_k$ for $k = i, j$.

(v) If $E_0(q_d) = \emptyset$ then set $c_{\max}(E_0(q_d)) = \sum_{\tau \in \mathcal{T}(G,q_d)} c_c(\pi_1(\tau))$. Else set

$$c_{\max}(E_0(q_d)) = c_e(\pi_1(\tau_n)) + c^g_{\sup}(M^o_D(G, \pi_3(\tau_n))) + \sum_{\tau \in \hat{E}_0(q_d)} c_c(\pi_1(\tau)).$$

(vi) Call subprogram One-Step Optimize.

(vii) Termination condition: Is $C = \emptyset$? If yes then return to the main program. Otherwise GOTO (iii).

Complexity: Steps (i), (ii), (iii), (v), and (vii) are independent of n. Step (iv) is $O(|\Sigma| \log(|\Sigma|))$. Let the complexity of (vi) or One-Step Optimize be $O(y)$. Then since $|C|$ is $O(n)$, the complexity of Optimize is $O(n(|\Sigma| \log(|\Sigma|) + y))$.

One-Step Optimize. We use the notation $c_{\max}(E)$ where $E = \{\tau_1, \ldots, \tau_j\}$ to denote the following calculation:

$$c_{\max}(E) = c_e(\pi_1(\tau_j)) + c^g_{\sup}(M^o_D(G, \pi_3(\tau_j))) + \sum_{i=j+1}^{i=n} c_c(\pi_1(\tau_i)) + \sum_{\tau \in \hat{E}_0(q_d)} c_c(\pi_1(\tau)),$$

where $\hat{E}_0(.)$ is precalculated in the main program.

(i) Input: $q_d, E_0(q_d), \hat{E}_0(q_d), c_{\max}(E_0)$.

(ii) Initialize: $E = E_0(q_d)$, $E' = E_0(q_d)$, $CMAX = c_{\max}(E_0(q_d)), \mathcal{T}^{temp}_{opt}(q_d) = \emptyset$.

If $E' = \emptyset$ then goto (iv).

Note that the elements of E' are always kept in the same order as $E_0(q_d)$.

(iii) Compute $E' \leftarrow E' - \{\tau_i : i = \max_{\tau_j \in E'} j\}$. If $E' \neq \emptyset$ then set

$$c_{\max}(E') = c_e(\pi_1(\tau_{i-1})) + c^g_{\sup}(M^o_D(G, \pi_3(\tau_{i-1}))) + \sum_{k=i}^{k=n} c_c(\pi_1(\tau_k)) + \sum_{\tau \in \hat{E_0}(q_d)} c_c(\pi_1(\tau)).$$

(iv) Termination condition: Is $E' = \emptyset$?
If yes, then set

$$\mathcal{T}^{temp}_{opt}(q_d) \leftarrow \mathcal{T}^{temp}_{opt}(q_d) \cup E,$$
$$CL^{temp} \leftarrow CL^{temp} \cup \{(q_d, c_{\max}(E))\}$$

and return to Optimize. Otherwise continue.

(v) Recursion condition: Is $c_{\max}(E') < CMAX$? If not then GOTO (iii). If yes, then continue.

(vi) Set $E = E', CMAX = c_{\max}(E')$.

(vii) GOTO (iii).

Complexity: All steps here are independent of n. Since E_0 is of order $O(|\Sigma|)$ the complexity of One-Step Optimize is $O(|\Sigma|)$.

Remark 6.1.1. The following pieces of pseudocode compute step (vii) of the main program.

```
Nextiter (A)
        q = first (A)
        trantest (q, T(G,q), A)
        C ← C ∪ {q}
        Nextiter (A − {q})
trantest (q, T(G,q), A)
        τ = first (T(G,q))
        if π3(τ) ∉ SL then
                if π1(τ) ∈ Σuc
                   then Nextiter (A − {q})
                else Ê0(q) ← Ê0(q) ∪ {τ}
                     trantest (q, T(G,q) − {τ}, A)
        else E0(q) ← E0(q) ∪ {τ}
             trantest (q, T(G,q) − {τ}, A)
```

The function Nextiter (.) computes the vertices to be optimized in the next iteration. It stores the new list in C. The function first (.) returns the first element of a set. It is assumed that the set is structured as a linked list. The function trantest (.) checks transitions for the conditions stated in (vii) and computes $E_0(.)$ and $\hat{E}_0(.)$. The function trantest (.) may recurse at most $|\Sigma|$ levels, and at each level the greatest number of computations is $|SL|$. Thus trantest (.) is of order $O(n|\Sigma|)$. Nextiter (.) recurses at most $|A|$ times. Since the cardinality of A is $O(n)$ the overall complexity of *Nextiter (.)* is $O(n^2|\Sigma|)$. Note that the computation of A itself is $O(n^2)$ and that of $P_f(SL)$ is $O(n)$. Hence the dominant term is $O(n^2|\Sigma|)$.

The overall complexity of the algorithm can now be calculated as follows. Since $O(y) = O(|\Sigma|)$ the complexity of Optimize is

$$O(x) = O(n(|\Sigma| + |\Sigma| \log(|\Sigma|))) = O(n|\Sigma| + n|\Sigma| \log(|\Sigma|)) = O(n|\Sigma| \log(|\Sigma|)).$$

The complexity of the main program is obtained by substituting for x. This gives

$$O(n(n|\Sigma| \log(|\Sigma|) + n^2|\Sigma|)) = O(n^2|\Sigma| \log(|\Sigma|) + n^3|\Sigma|).$$

In general the alphabet is expected to be small. In such cases the complexity is $O(n^3)$. The complexity is set by the computation of the set C in the main program. The updating of SL creates the $O(n^2|\Sigma|\log(|\Sigma|))$ term in the expression. Without the ordering of $E_0(.)$ this step would have become exponential.

It is important to note that the algorithm as stated will generally produce a submachine that is not trim. In particular, it will have states not accessible from q_0. However, the trimming of a submachine is a standard problem, and consequently we shall not dwell on it any more.

We draw attention to the connection between the acyclic nature of the optimal solution and the stopping criteria in (ii) and (vi) of the main program. Since there are no cycles and only one marked state, if there is an uncontrollable transition defined out of q_m, then it must execute a path returning to q_m. This is because the optimal solution is also trim. This path is a cycle and we have a contradiction which is resolved by concluding that the optimal solution does not exist. If the condition in (vi) fails to evaluate, then once again the acyclic nature of the optimal solution guarantees that the set C in (vii) will be nonempty (refer to the proof of Theorem 6.7 presented later in this section).

The algorithm bears some procedural similarity to a backward recursive shortest path algorithm. The shortest path algorithm will start at the terminal state and place it on the solved list (SL). At each step it will develop the set of parents of SL (denoted by $P_f(SL)$). From this set the algorithm identifies the next state to be added to SL. This identification is $O(n)$, and since the identification process is repeated each time a state is added to SL, the overall complexity is $O(n^2)$.

Our algorithm also starts at the terminal state (q_m) and places it in SL. At each step it develops $P_f(SL)$ (main program, (vii); complexity $O(n^2|\Sigma|)$) and identifies the next state to be added to SL (main program, (iv)). Unlike the path algorithm the complexity of the identification process is $O(n|\Sigma|\log(|\Sigma|))$. Since the process is repeated each time a state is added to SL the overall complexity is $O(n^2|\Sigma|\log(|\Sigma|) + n^3|\Sigma|)$.

The next state to be added to SL is determined as follows. Develop the set C and pick some $q \in C$. In general, q has transitions leading to children which are not in SL. Disable all such transitions and consider the one-step FSM constituted of the remaining transitions (main program, (vii)). This is the set $E_0(q)$. Sort these transitions (Optimize, (iv)) as required by Theorem 6.8 and construct the sequence of submachines $\langle M'_j \rangle$. This sort is $O(|\Sigma|\log(|\Sigma|))$. Next find the minimum cost M_j (One-Step Optimize). Let it be denoted by M_q. M_q must be found for each $q \in C$. The states realizing $\min_{q \in C} c^g_{\sup}(M_q)$ may be added to SL (main program, (iv), (v)) with the corresponding $\mathcal{T}_{\mathcal{T}}(q)$ representing $M_1(M^o_D(G, q), q)$ (main program, (v)). This completes one iteration of of the algorithm. It continues until $q_0 \in A$ (main program, (vi)). For a clearer understanding of the working of the algorithm, refer to the example presented in section 6.

Before passing to the proofs of correctness we give a brief explanation of why the introduction of a hypothetical marked state does not work. The correctness of the algorithm, as will be proven in the subsequent results, rests upon the nonexistence of cycles in the optimal solution and the property that the point at which the algorithm starts is the one and only one point at which all strings terminate in the optimal solution. No string that reaches this state ever goes anywhere again. The hypothetical marked state would not have this property, since if two marked states are connected to it a behavior might reach one marked state and hence the hypothetical state, but then continue onto the other marked state and hence to the hypothetical marked state

again. Thus, to make the algorithm correct, the hypothetical marked state should only be identified with those marked states which in the optimal solution have no transitions going out of them. How does one find this subset of Q_m? Exhaustive examination of all subsets of Q_m is exponential in the cardinality of Q_m. It is obvious that new properties of the structure must be discovered and exploited to isolate this subset.

It now remains to prove correctness. Since the algorithm computes by incrementally constructing bigger submachines out of smaller submachines, we define an algebraic operation called *merge* that combines FSMs.

DEFINITION 6.2 (merge operation). *Let A, B be FSMs and $q_{0C} \in Q_A \cup Q_B$ be some state. Then*

$$\begin{array}{rcccl} A \oplus B & = & C & = & \langle \Sigma_C, Q_C, q_{0C}, Q_{mC}, \delta_C \rangle, \\ & & \Sigma_C & = & \Sigma_A \cup \Sigma_B, \\ & & Q_C & = & Q_A \cup Q_B, \\ & & Q_{mC} & = & Q_{mA} \cup Q_{mB}, \\ & & \delta_C(\sigma, q) & = & \begin{cases} \delta_A(\sigma, q) \text{ if it exists,} \\ \delta_B(\sigma, q) \text{ if it exists,} \\ \text{undefined otherwise.} \end{cases} \end{array}$$

Observe that the merge is defined not just by A and B but also by the state q_{oC}. The merge of A and B produces a machine C whose transitions are the union of the transitions of A and B. It is apparent from this definition that the merge of two trim FSMs is not necessarily a trim FSM. It is also true that the transition function is not necessarily well defined. Fortunately, the next lemma allows us to avoid these pitfalls in this specific problem. Its proof is straightforward.

LEMMA 6.3. *Let $A, B \subseteq G$. Then the transition function for $C = A \oplus B$ in Definition 6.2 is well defined. Moreover, if $q_{0A} \in Q_B$ and A, B are trim, then*

$$C = \langle \Sigma_C, Q_C, q_{0B}, Q_{mC}, \delta_C \rangle$$

is a trim submachine of G.

The requirement that A, B be submachines of G ensures that if $\delta_A(\sigma, q)$ exists and $\delta_B(\sigma, q)$ exists, then $\delta_A(\sigma, q) = \delta_B(\sigma, q)$.

For all $A, B \in \mathcal{M}(G, q)$ this lemma implies that $A \oplus B \in \mathcal{M}(G, q)$, whence $\mathcal{M}(G, q)$ has a maximal element in the sense that all others in the set are submachines of it. This maximal trim submachine of G at q will be denoted by $M(G, q)$. Similarly, for all $A, B \in \mathcal{M}_1(G, q)$ it can be shown that $A \oplus B \in \mathcal{M}_1(G, q)$, and consequently, there exists a maximal element $M_1(G, q)$. Obviously $\mathcal{T}(M_1(G, q)) = \mathcal{T}(G, q)$. In the subsequent development it is implicitly assumed that all FSMs are trim unless mentioned otherwise.

The following theorem establishes that the merge operation also preserves DP-optimality.

THEOREM 6.4. *Let A be a DP-optimal submachine in the set $\mathcal{M}(G, q_{0A})$ and B be DP-optimal in $\mathcal{M}(G, q_{0B})$. Furthermore, let $q_{0B} \in Q_A$. Then $A \oplus B$ is DP-optimal in $\mathcal{M}(G, q_{0A})$.*

Proof. Observe that by Lemma 6.3 $A \oplus B$ is a trim submachine of G. It lies in the set $\mathcal{M}(G, q_{0A})$. Let $s \in \Sigma^*$ be such that $\delta^*_{A \oplus B}(s, q)$ exists for some $q \in Q_A \cup Q_B$. Assume $q \in Q_A$. The case $q \in Q_B$ is identical. By the definition of "$\oplus$" there exists a decomposition of s such that $s = u_1 \dots u_n$, where for all i, $1 \leq i \leq n$, $\delta^*_{A \oplus B}(u_1 \dots u_i, q) = q_i, \{q_1, \dots q_{n-1}\} \subseteq Q_A \cap Q_B$ and $\delta^*_{A \oplus B}(u_i, q_{i-1}) =$

$\delta_A^*(u_i, q_{i-1})$ or $\delta_B^*(u_i, q_{i-1})$. Note that $q_0 = q$. From Lemma 5.2, if $\delta_A^*(u_i, q_{i-1})$ exists then $\hat{c^g}(q_{i-1}, A \oplus B, u_i) \leq \hat{c^g}(q_{i-1}, A, u_i)$. The case when $\delta_B^*(u_i, q_{i-1})$ exists is similar. We also note that for all $q_i,\ 1 \leq i \leq n-1$,

$$c^g_{\sup}(M(A, q_i)) = c^g_{\sup}(M(B, q_i)).$$

This follows from the DP-optimality of A and B. The rest of the proof proceeds by induction. Let

$$c^g(u_{i+1} \dots u_n, M(A \oplus B, q_i)) \leq c^g_{\sup}(M(A, q_i)) = c^g_{\sup}(M(B, q_i)).$$

Consider a case with $\delta_B^*(u_i, q_{i-1})$ defined. The case with $\delta_A^*(u_i, q_{i-1})$ defined is identical.

$$\begin{array}{lll} & c^g(u_i \dots u_n, M(A \oplus B, q_{i-1})) & \\ = & \hat{c^g}(q_{i-1}, A \oplus B, u_i) + c^g(u_{i+1} \dots u_n, M(A \oplus B, q_i)) & \text{[by definition of } c^g(.)] \\ \leq & \hat{c^g}(q_{i-1}, B, u_i) + c^g_{\sup}(M(B, q_i)) & \text{[Theorem 5.2]} \\ \leq & c^g_{\sup}(M(B, q_{i-1})) & [u_i \circ \mathcal{L}_m(M(B, q_i)) \subseteq \mathcal{L}_m(M(B, q_{i-1}))] \\ = & c^g_{\sup}(M(A, q_{i-1})) & \end{array}$$

The base case is argued as follows. Assume $\delta_A^*(u_n, q_{n-1})$ exists. The case where $\delta_B^*(u_n, q_{n-1})$ exists is similar. By Theorem 5.2,

$$c^g(u_n, M(A \oplus B, q_{n-1})) \leq c^g(u_n, M(A, q_{n-1})) \leq c^g_{\sup}(M(A, q_{n-1})).$$

By induction

$$c^g(s, M(A \oplus B, q)) = c^g(u_1 \dots u_n, M(A \oplus B, q)) \leq c^g_{\sup}(M(A, q)).$$

Since s was arbitrary we obtain $c^g_{\sup}(M(A \oplus B, q)) \leq c^g_{\sup}(M(A, q))$, whence $M(A \oplus B, q)$ is optimal. Since q was arbitrary and the case $q \in Q_B$ can be argued similarly, we conclude that $A \oplus B$ is DP-optimal. □

We are now ready to establish correctness. The first theorem establishes the relevance of the DP-equation.

THEOREM 6.5.

$$c^g_{\sup}(M^o_D(G, q_d)) = \min_{A' \in \mathcal{M}_1(G, q_d)} \left[\max_{\tau \in \mathcal{T}(A')} [\bar{c}^g(q_d, A', \pi_1(\tau)) + c^g_{\sup}(M^o_D(G, \pi_3(\tau)))] \right],$$
$$M^o_D(G, q_d) = M' \oplus \left(\oplus_{\tau \in \mathcal{T}(M')} M^o_D(G, \pi_3(\tau)) \right),$$

where M' is the maximal minimizing A' above.

Proof. The first equation is an application of the principle of dynamic programming. The second equation is a consequence of the maximality of $M^o_D(G, q)$ and Theorem 6.4. □

Thus a maximal DP-optimal solution is constituted of other maximal DP-optimal solutions. If the other maximal DP-optimal solutions are known, then it is only necessary to find the largest minimizing A' in the DP-equation. While this is a substantial simplification, observe that there are $O(2^{|\Sigma|})$ candidate one-step submachines. In view of the large number of times this equation is solved, exhaustive examination is computationally prohibitive. Consequently, we have sought methods to solve the equation in polynomial time. We will prove that this can be done in $O(|\Sigma| \log(|\Sigma|))$ computations.

Some new notation is adopted for convenience. $SL(i)$ denotes the value of the set SL in the ith iteration of the algorithm. For each $q \in P_f(SL) - SL$ let $B_1(q, SL)$ represent the maximal one-step submachine of G rooted at q with all transitions disabled other than those leading into SL. Define

$$B(q_d, SL) = B_1(q_d, SL) \oplus \left(\oplus_{\tau \in \mathcal{T}(B_1(q_d, SL))} M_D^o(G, \pi_3(\tau))\right).$$

The next result establishes that in each iteration of the algorithm there exists a parent state q of $SL(i)$ for which the DP-optimal solution exists. The lemma also says that the solution will be a submachine of $B(q, SL)$. Lemmas 6.7 and 6.8 show how to find one or more such q. The three lemmas together show that if an optimal solution exists, then SL will keep growing until q_0 is one of the q's.

LEMMA 6.6. *If an optimal sumachine of G exists and $q_0 \notin SL(i)$, then there exists $q \in P_f(SL(i)) - SL(i)$ such that $M_D^o(G, q) \in \mathcal{M}(B(q, SL(i)), q)$.*

Proof.

Claim 6.6.1. Let $q \notin SL(i)$. Then for all $s \in \Sigma^*$ such that $\delta^*(s, q)$ is defined, there exists

$$q' \in P_f(SL(i)) - SL(i) \text{ and } t, u \in \Sigma^*$$

such that $s = tu$ and $\delta^*(t, q) = q'$.

Proof. From the description of the algorithm (main program, (v)), we have $SL(i) \subseteq SL(i+1)$. The proof is immediate from $q_m \in SL(i)$, the co-accessibility of G with respect to q_m and $q \notin SL(i)$. □

Claim 6.6.2. If an optimal submachine of G exists and $q_0 \notin SL(i)$, then there exists $q \in (P_f(SL(i)) - SL(i))$ such that $M_D^o(G, q)$ exists.

Proof. Pick any $s \in \mathcal{L}_m(M_D^o(G, q_0))$. Since $\mathcal{L}_m(M_D^o(G, q_0)) \subseteq \mathcal{L}_m(G)$, Claim 6.6.1 implies that there exists $q \in P_f(SL(i)) - SL(i), tu \in \Sigma^*$ such that $s = tu$ and $q = \delta^*(t, q_0)$. Thus q is a state of $M_D^o(G, q_0)$, which implies that an optimal submachine exists at q. Hence $M_D^o(G, q)$ exists for at least one $q \in P_f(SL(i)) - SL(i)$. □

Let $P'_f(SL(i)) = \{q \in P_f(SL(i)) - SL(i) : M_D^o(G, q) \text{ exists}\}$. By the prior claim, $P'_f(SL(i)) \neq \emptyset$. We now prove the lemma by contradiction.

For all $q \in P'_f(SL(i))$ let $M_D^o(G, q) \notin \mathcal{M}(B(q, SL(i)), q)$. Then, by Theorem 6.5,

$$M_1(M_D^o(G, q), q) \notin \mathcal{M}_1(B(q, SL(i)), q),$$

which implies that there exists $\tau \in \mathcal{T}(M_D^o(G, q), q), \pi_3(\tau) \notin SL(i)$. Pick such a τ. By Claim 6.6.1 there exists $q'' \in P_f(SL(i)) - SL(i)$ and $u \in \Sigma^*$ such that $\delta^*(\pi_1(\tau)u, q) = q''$ and $\pi_1(\tau)u \in \mathcal{L}(M_D^o(G, q))$.

Define $R(q) = \{q' \in P_f(SL(i)) - SL(i) : \exists s \in \mathcal{L}(M_D^o(G, q)), \delta^*(s, q) = q', q \neq q'\}$. Obviously $q'' \in R(q)$ and $R(q) \neq \emptyset$ for all $q \in P'_f(SL(i))$. Consider the following inductive construction.

(i) Base step: Pick any $q \in P'_f(SL(i))$. Set $q_1 = q$.

(ii) Induction step: $q_{j+1} = q, q \in R(q_j)$.

Since $R(q) \neq \emptyset, q_{j+1}$ always exists and the construction never terminates. Since no cycles are possible in an optimal solution,

$$\begin{array}{rrcl} & q_1, \ldots q_j, q_{j+1} & \notin & R(q_{j+1}) \\ \Rightarrow & 0 < |R(q_{j+1})| & \leq & |P'_f(SL(i))| - (j+1) \\ \Rightarrow & j+1 & \leq & |P'_f(SL(i))| \end{array}$$

for all j. By the finiteness of $P'_f(SL(i))$ this is a contradiction. Thus the lemma is proved. □

The next lemma indicates how q as in the prior lemma can be found. From the construction of $B(q, SL(i))$ and the fact that $M_D^o(G, q')$ exists for all $q' \in SL(i)$, it is evident that if $M_D^o(G, q) \in \mathcal{M}(B(q, SL(i)), q)$ then $c^g_{\sup}(B(q, SL(i))) < \infty$. By Theorem 5.3 the maximal DP-optimal submachine $\hat{M}_D^o(G, q)$ of $B(q, SL(i))$ also exists. These entities, well defined by this argument, are referred to in the next lemma.

LEMMA 6.7. *For all $q \in P_f(SL(i)) - SL(i)$ such that $c^g_{\sup}(B(q, SL(i))) < \infty$, let $\hat{M}_D^o(G, q)$ be the maximal DP-optimal submachine in the set $\mathcal{M}(B(q, SL(i)), q)$. In particular, let $q_d \in P_f(SL(i)) - SL(i)$ be such that*

$$c^g_{\sup}(\hat{M}_D^o(G, q_d)) = \min_{\substack{q \in P_f(SL(i)) - SL(i) \\ c^g_{\sup}(B(q, SL(i))) < \infty}} c^g_{\sup}(\hat{M}_D^o(G, q)).$$

Then $M_D^o(G, q_d) = \hat{M}_D^o(G, q_d)$.

Proof. Let $c^g_{\sup}(M_D^o(G, q_d)) < c^g_{\sup}(\hat{M}_D^o(G, q_d))$. Then

$$M_D^o(G, q_d) \notin \mathcal{M}(B(q_d, SL(i)), q_d).$$

Set $q_1 = q_d$. We define an inductive construction with this as the base case.

Inductive step: Let $q_j \in P_f(SL(i)) - SL(i), M_D^o(G, q_j) \notin \mathcal{M}(B(q_j, SL(i)), q_j)$, and

$$c^g_{\sup}(M_D^o(G, q_1)) > \cdots > c^g_{\sup}(M_D^o(G, q_j)).$$

We will now show the construction of q_{j+1} having all the same attributes as q_j. Since $M_D^o(G, q_j) \notin \mathcal{M}(B(q_j, SL(i)), q_j)$, by Theorem 6.5 there exists $\tau \in \mathcal{T}(M_D^o(G, q_j))$ such that $\pi_3(\tau) \notin SL(i)$. By Claim 6.6.1 there exists $q'' \in P_f(SL(i)) - SL(i), uv \in \Sigma^*$ such that

$$\pi_1(\tau)uv \in \mathcal{L}_m(M_D^o(G, q_j)) \wedge \delta^*(\pi_1(\tau)u, q_j) = q''.$$

Assume q'' is such that $M_D^o(G, q'') \in \mathcal{M}(B(q'', SL(i)), q'')$. By the above properties of q'',

$$\begin{array}{rcll} c^g_{\sup}(M_D^o(G, q_j)) & > & c^g_{\sup}(M_D^o(G, q'')) & [\delta^*(\pi_1(\tau)u, q) = q'', \pi_1(\tau)u \in \mathcal{L}(M_D^o(G, q_j))] \\ & = & c^g_{\sup}(\hat{M}_D^o(G, q'')) & [M_D^o(G, q'') \in \mathcal{M}(B(q, SL(i)), q)] \\ & \geq & c^g_{\sup}(\hat{M}_D^o(G, q_d)) & [\text{hypothesis of theorem}] \\ & > & c^g_{\sup}(M_D^o(G, q_d)) & [\text{by assumption}] \\ & = & c^g_{\sup}(M_D^o(G, q_1)) & [\text{base case } q_d = q_1] \\ & > & c^g_{\sup}(M_D^o(G, q_j)) & [\text{induction hypothesis}], \end{array}$$

which is absurd. Thus $M_D^o(G, q'') \notin \mathcal{M}(B(q'', SL(i)), q'')$. Moreover

$$c^g_{\sup}(M_D^o(G, q'')) < c^g_{\sup}(M_D^o(G, q_j)) < \cdots < c^g_{\sup}(M_D^o(G, q_1)).$$

Set $q_{j+1} = q''$. The state q_{j+1} has all the properties of q_j. Observe that the positivity of the event cost function has been explicitly used in the above argument.

This is obviously a nonterminating construction. From the uniqueness of the maximal DP-optimal solution (Theorem 5.3 and the strict inequality on the cost $M_D^o(G, q_j)$) we have that no two members of the sequence $\langle q_j \rangle_{j=1}^{j=\infty}$ are equal. But for any given j,

$$\{q_1, \ldots, q_j\} \subseteq P_f(SL(i)) - SL(i).$$

Since the set $P_f(SL(i)) - SL(i)$ has finite cardinality, this is a contradiction. The lemma is immediate. □

The next lemma shows how the one-step submachine of $\hat{M}_D^o(G,q)$ may be constructed efficiently, i.e., $O(|\Sigma|\log(|\Sigma|))$, though in general there could be $2^{|\Sigma|}$ candidate solutions.

LEMMA 6.8. *Let $q_d \in Q$ be such that $M_D^o(G,q_d)$ exists. Consider the following construction. Let*

$$\hat{Q} = \{q \in Q : (\exists \tau \in \mathcal{T}(G,q_d), \pi_3(\tau) = q) \wedge (M_D^o(G,q) \text{ exists})\} = \{q_1, \ldots, q_{|\hat{Q}|}\}.$$

For all i such that $1 \leq i \leq |\hat{Q}|$ let

$$M_i = M_i' \oplus \left(\oplus_{j=1}^{j=i} M_D^o(G,q_j)\right),$$
$$\text{where } \mathcal{T}(M_i') = \{\tau_1, \ldots, \tau_i\} \subseteq \mathcal{T}(G,q_d), \tau_i = (\sigma_i, q_d, q_i), \quad \text{and}$$
$$c_e(\pi_1(\tau_i)) + c_{\sup}^g(M_D^o(G,q_i)) \leq c_e(\pi_1(\tau_{i+1})) + c_{\sup}^g(M_D^o(G,q_{i+1})).$$

Then $M_D^o(G,q_d) \in \{M_i : 1 \leq i \leq |\hat{Q}|\}$.

Proof. For each q_i define the one-step submachine

$$A_i = \langle \Sigma, \{q_d, q_i\}, q_d, q_m, \delta_{A_i} \rangle,$$
$$\delta_{A_i}(\sigma_i, q_d) = q_i.$$

We assume that δ_{A_i} is undefined for all other states and events. Since A_i has only one transition and $M_D^o(G,q_i)$ is trim, $A_i \oplus M_D^o(G,q_i)$ is a trim submachine of G at q_d. Furthermore

$$M_i = \oplus_{j=1}^{j=i}(A_j \oplus M_D^o(G,q_j)),$$

whence by Lemma 6.3, M_i is trim for all i. We prove that for all $j \leq i$,

$$A_i \oplus M_D^o(G,q_i) \subseteq M_D^o(G,q_d) \Rightarrow A_j \oplus M_D^o(G,q_j) \subseteq M_D^o(G,q_d).$$

The latter inclusion is equivalent to the statement

$$M_D^o(G,q_d) \oplus A = M_D^o(G,q_d),$$

where $A = A_j \oplus M_D^o(G,q_j)$. The following argument proves this statement.

We use the notation $\mathcal{Q}_f(.)$ to represent the set of states of the machine $(.)$. We first show that $M(A,q)$ is optimal for all $q \in \mathcal{Q}_f(A), q \neq q_d$. By definition of "$\oplus$" $\mathcal{Q}_f(M_D^o(G,q_d) \oplus A) = \mathcal{Q}_f(M_D^o(G,q_d)) \cup Q_A$. Consider the case $q \in \mathcal{Q}_f(M_D^o(G,q_d))$. Let

$$X = \mathcal{Q}_f(M(M_D^o(G,q_d),q)) \cap \mathcal{Q}_f(M_D^o(G,q_j)).$$

Then we obtain the following decomposition of $M_D^o(G,q_d)$.

$$\begin{aligned} M(M_D^o(G,q_d) \oplus A, q) &= M(M_D^o(G,q_d),q) \oplus (\oplus_{q' \in X} M(M_D^o(G,q_j),q')) \\ &= M_D^o(G,q) \oplus (\oplus_{q' \in X} M_D^o(G,q')) \qquad \text{[Theorem 5.3]} \\ &= M_D^o(G,q) \qquad \text{[Theorem 6.4].} \end{aligned}$$

The case $q \in \mathcal{Q}_f(M_D^o(G,q_j))$ can be treated similarly by interchanging $M_D^o(G,q_d)$ and $M_D^o(G,q_j)$ in the preceeding argument. Thus for all $q \neq q_d$, $M(M_D^o(G,q_d) \oplus A, q)$

is optimal. We now prove the case $q = q_d$ by splitting the argument into two cases. Consider any $s \in \mathcal{L}_m(M_D^o(G, q_d) \oplus A)$. The first case covers all s such that the first event of s lies in $M_D^o(G, q_d)$, whereas the second case covers all s such that the first event lies in A.

Case 1. $s = \sigma_j u$.

$$\begin{aligned}
c^g(s, M_D^o(G, q_d) \oplus A) &= \bar{c}^g(q_d, M_D^o(G, q_d) \oplus A, \sigma_j) + c^g(q_j, M(M_D^o(G, q_d) \oplus A, u)) \\
&= c_e(\sigma_j) + \textstyle\sum_{\tau \in \mathcal{T}(G,q_d) - (\mathcal{T}(M_D^o(G,q_d),q_d) \cup \mathcal{T}(A,q_d))} c_c(\pi_1(\tau)) \\
&\quad + c^g(q_j, M_D^o(G, q_j), u) \\
&\leq c_e(\sigma_j) + \textstyle\sum_{\tau \in \mathcal{T}(G,q_d) - \mathcal{T}(M_D^o(G,q_d),q_d)} c_c(\pi_1(\tau)) \\
&\quad + c^g_{\sup}(M_D^o(G, q_j)) \\
&\leq c_e(\sigma_i) + \textstyle\sum_{\tau \in \mathcal{T}(G,q_d) - \mathcal{T}(M_D^o(G,q_d),q_d)} c_c(\pi_1(\tau)) \\
&\quad + c^g_{\sup}(M_D^o(G, q_i)) \\
&= \bar{c}^g(q_d, M_D^o(G, q_d), \sigma_i) + c^g_{\sup}(M_D^o(G, q_i)) \\
&= \bar{c}^g(q_d, M_D^o(G, q_d), \sigma_i) + c^g_{\sup}(M(M_D^o(G, q_d) q_i)) \\
&\leq c^g_{\sup}(M_D^o(G, q_d)).
\end{aligned}$$

Note that the second-to-last inequality follows from the ordering hypothesis.

Case 2. $s = \sigma_k u,\ \sigma_k \neq \sigma_j$.

$$\begin{aligned}
c^g(s, M_D^o(G, q_d) \oplus A) &= \bar{c}^g(q_d, M_D^o(G, q_d) \oplus A, \sigma_k) + c^g(q_k, M(M_D^o(G, q_d) \oplus A, u)) \\
&= c_e(\sigma_k) + \textstyle\sum_{\tau \in \mathcal{T}(G,q_d) - (\mathcal{T}(M_D^o(G,q_d),q_d) \cup \mathcal{T}(A,q_d))} c_c(\pi_1(\tau)) \\
&\quad + c^g(q_k, M_D^o(G, q_k), u) \\
&\leq c_e(\sigma_k) + \textstyle\sum_{\tau \in \mathcal{T}(G,q_d) - \mathcal{T}(M_D^o(G,q_d),q_d)} c_c(\pi_1(\tau)) \\
&\quad + c^g_{\sup}(M_D^o(G, q_k)) \\
&= \bar{c}^g(q_d, M_D^o(G, q_d), \sigma_k) + c^g_{\sup}(M_D^o(G, q_k)) \\
&= \bar{c}^g(q_d, M_D^o(G, q_d), \sigma_k) + c^g_{\sup}(M(M_D^o(G, q_d) q_k)) \\
&\leq c^g_{\sup}(M_D^o(G, q_d)).
\end{aligned}$$

Observe that in Case 2 we have the following advantage:

$$\sigma_k \circ M_D^o(G, q_k) \subseteq M_D^o(G, q_d).$$

The same cannot be said of j in Case 1. The extra steps in Case 1 take us from j to i, the i being equivalent to k in Case 2.

Thus for all $s \in \mathcal{L}_m(M_D^o(G, q_d) \oplus A)$ we have

$$c^g(q_d, M_D^o(G, q_d) \oplus A, s) \leq c^g_{\sup}(M_D^o(G, q_d)),$$

and consequently, $c^g_{\sup}(M_D^o(G, q_d) \oplus A) \leq c^g_{\sup}(M_D^o(G, q_d))$. Thus $M_D^o(G, q_d) \oplus A$ is optimal. Since we have already shown that for all $q \neq q_d$

$$M(M_D^o(G, q_d) \oplus A, q) = M_D^o(G, q),$$

it follows that $M_D^o(G, q_d) \oplus A$ is DP-optimal. Since the maximal DP-optimal solution is unique this implies

$$M_D^o(G, q_d) \oplus A \subseteq M_D^o(G, q_d).$$

The reverse inclusion being obvious, we conclude that

$$M_D^o(G, q_d) \oplus A = M_D^o(G, q_d) \Rightarrow A_j \oplus M_D^o(G, q_j) \subseteq M_D^o(G, q_d).$$

The lemma is immediate. □

The correctness proof is a simple argument based on these results.

THEOREM 6.9. *The algorithm is correct. It computes the maximal DP-optimal FSM* $M_D^o(G, q_0)$ *with worst-case complexity*

$$O(n^2|\Sigma|\log(|\Sigma|) + n^3|\Sigma|).$$

Proof. We prove correctness inductively. As a base case we show that $\mathcal{T}_{opt}(q_m)$ is computed correctly. From step (ii) of the main program, $C = \{q_m\}$. Since $c_e(.) > 0$ there are no cycles in any optimal solution. Also, there exists only one marked state. These two facts taken together imply that if there exists $(\sigma, q_m, q) \in \mathcal{T}(M_D^o(G, q_0))$ such that $\sigma \in \Sigma_{uc}$, then all controllable trim submachines (submachines constructed by disabling only controllable events) must contain some cycle passing through q_m. All optimal submachines, if they exist, must lie within the class of controllable submachines, and consequently, no optimal solution exists. Thus the testing condition in step (ii) is correct. If an optimal solution exists, then, q_m being the only marked state, $M_D^o(G, q_m)$ must exist and $\mathcal{T}(M_D^o(G, q_m)) = \emptyset$. Note that $\mathcal{T}_{opt}^{temp}(q_m) = \emptyset$ by step (iv) of One-Step Optimize. In step (v) of the main program, $A = \{q_m\}$ and consequently $\mathcal{T}_{opt}(q_m) = \emptyset$, as theoretically expected. Thus the computation for q_m is correct.

Assume that for all $q \in SL$, $\mathcal{T}_{opt}(q), c_{\max}(\mathcal{T}_{opt}(q))$, and consequently $M_D^o(G, q)$ and $c_{\max}(M_D^o(G, q))$ are known correctly. By the DP-equation (Theorem 6.5), for any $q_d \in P_f(SL) - SL$, it is only necessary to find the maximal $A_1 \in \mathcal{M}_1(G, q_d)$ which solves the DP-equation.

Statement (vi) of the main program tests the hypothesis of Lemma 6.6. If the program does not terminate here and an optimal solution exists, then the hypothesis of Lemma 6.6 is satisfied and it guarantees that $C \neq \emptyset$, since otherwise there will not exist any $B(q, SL)$, $q \in P_f(SL) - SL$ with finite cost, and consequently no $M_D^o(G, q) \subseteq B(q, SL)$. Thus C in the main program is the set of all $q \in P_f(SL) - SL$ for which $c_{\sup}^g(B(q, SL))$ is finite. Next, by Lemma 6.7 it is necessary to construct $M_1(\hat{M}_D^o(G, q))$ for each $q \in C$, and by Lemma 6.8 this can be done by sorting and ordering the transitions of $B_1(q, SL)$. The required sorting and ordering is done in step (iv) of Optimize, and the minimum of the ordered set is computed in One-Step Optimize, which returns the transitions of $M_1(\hat{M}_D^o(G, q))$ in $\mathcal{T}_{opt}^{temp}(q)$ and its cost in $CMAX$. The minimization required in the hypothesis of Lemma 6.7 is done in step (iv) of the main program. Consequently, by Lemma 6.7, $\mathcal{T}_{opt}(q)$ in step (v) correctly represents the transitions of $M_1(M_D^o(G, q))$ and $c_{\max}(\mathcal{T}_{opt}(q)) = c_{\sup}^g(M_D^o(G, q))$. Thus $\mathcal{T}_{opt}(q)$ solves the DP-equation. This proves the correctness of the algorithm.

The complexity is immediate from remarks made during the statement of the algorithm and from Remark 6.1.1. □

Theorem 3.10 summarizes the computational theory for cyclic systems. It is an aggregation of Theorems 5.3 and 6.9.

THEOREM 3.10. *Let* L_m *be regular and such that all marked strings are equivalent in the sense of Nerode. Let all event costs be positive. If an optimal solution exists, then, given a generator of* L_m *with* n *states, a generator for the supremal DP-optimal sublanguage is computable in time*

$$O(n^2|\Sigma|\log(|\Sigma|) + n^3|\Sigma|).$$

This concludes the examination of cyclic FSMs.

6.2. DP-optimal solutions: The acyclic case. In the subsequent development the plant FSM (G) is assumed to be acyclic. The notation used is as defined in

the prior subsection. The positivity assumption on the event costs and that on the marked states are relaxed. Thus we return to the premise that the cost functions are nonnegative and G is co-accessible with respect to the set Q_m. Observe that since $c_e(.)$ is always finite, the existence issues are trivial for the following reasons. The FSM is acyclic and there exists a path of maximum length. All event costs being finite this implies $c^g_{\sup}(M(G,q)) < \infty$ for all $q \in Q$. The set of possible submachines (sublanguages) is also finite. Thus optimal solutions always exist at all states of G. The complexity result is as follows.

THEOREM 3.11. *Let the plant language L_m be generated by a trim acyclic FSM G having n states. A generator for the supremal DP-optimal solution is computable in*

$$O(n|\Sigma|\log(|\Sigma|)).$$

The proof is similar to the cyclic case. We discuss it informally. Since the FSM G is acyclic it is possible to order the states of G so that any state in the sequence is connected only to states to the right of it. This can be done by a topological sort (refer to Leiserson, Cormen, and Rivest [15, section 23.4, p. 485]). This sort is of complexity

$$O(n + |\mathcal{T}(G)|) = O(n + n|\Sigma|) = O(n|\Sigma|).$$

The rightmost states in this order are the marked states of zero outdegree, since they are connected to nothing at all. The backward recursive algorithm will start at these states. Note that since G is nonblocking there always exists at lease one marked state having zero outdegree. The controller synthesis algorithm should proceed on the sequence of states from right to left, starting with the marked states of zero outdegree. For each state it must compute the maximal DP-optimal submachine at that state, using the maximal DP-optimal submachines at states to the right of it. This can be done by sorting the edges of the maximal one-step submachine rooted at the state, as in step (iv) of Optimize. This is of complexity $O(|\Sigma|\log(|\Sigma|))$. The maximal DP-optimal submachine is then computed by One-Step Optimize exactly as in the cyclic case. Since there are n states in the sequence, the overall time complexity is $O(n|\Sigma|\log(|\Sigma|))$. The leftmost state in the sequence will always be the initial state if G is trim. The algorithm will terminate when it reaches the initial state.

6.3. Examples. The following example is constructed to illustrate the essential features of the algorithm stated in section 6.1. The plant model and costs are as in Figure 6.1 and Table 6.1, respectively.

For each run of the algorithm we present the values of the variables

$$C, \mathcal{T}^{temp}_{opt}(.), c_{\max}(\mathcal{T}^{temp}_{opt}(.)), A, \mathcal{T}_{opt}(.), CL, SL.$$

The working of Optimize and One-Step Optimize is detailed only in the fourth iteration or Run 4 of the algorithm. These routines, being simpler for the other iterations, are omitted. The algorithm is initialized with $C = \{m\}$. The six figures given next (Figures 6.2–6.7) represent the submachines computed by the six consecutive runs required to find the optimal solution.

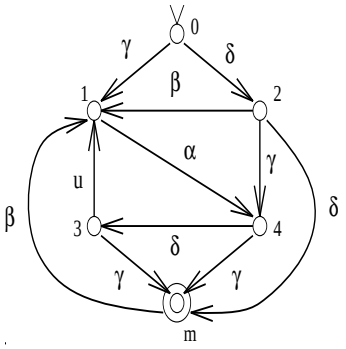


FIG. 6.1. *The plant machine G.*

TABLE 6.1
Costs and controllability of G.

Event	$c_e(.)$	$c_c(.)$	Remarks
α	0.5	2	Controllable
β	1	1	Controllable
γ	1	2	Controllable
δ	1	2	Controllable
u	1	∞	Uncontrollable

FIG. 6.2. *Run* 1.

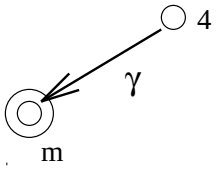


FIG. 6.3. *Run* 2.

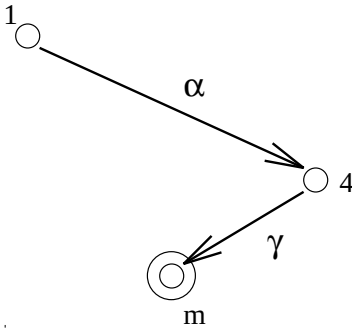


FIG. 6.4. *Run* 3.

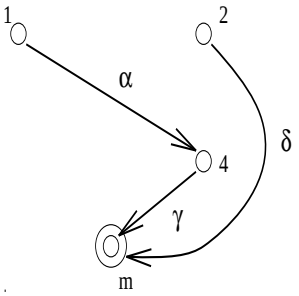


FIG. 6.5. *Run* 4.

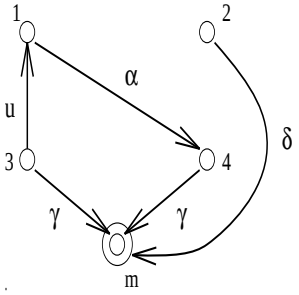


FIG. 6.6. *Run* 5.

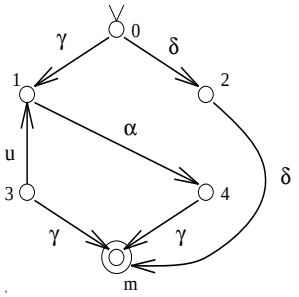


FIG. 6.7. *Run* 6.

Run 1:

$$\begin{aligned}
\mathcal{T}_{opt}^{temp}(m) &= \emptyset, \\
c_{\max}(\mathcal{T}_{opt}^{temp}(m)) &= 1, \\
\mathcal{T}_{opt}(m) &= \emptyset, \\
c_{\max}(\mathcal{T}_{opt}(m)) &= 1, \\
CL &= \{(m,1)\}, \\
SL &= \{m\}, \\
C &= \{2,4\}.
\end{aligned}$$

Note that $1 \notin C$ since u is uncontrollable and 1 is not in SL (refer to step (vii) of the main program).

Run 2:

$$\begin{aligned}
\mathcal{T}_{opt}^{temp}(2) &= \{(\delta, 2, m)\},\\
c_{\max}(\mathcal{T}_{opt}^{temp}(2)) &= 5,\\
\mathcal{T}_{opt}^{temp}(4) &= \{(\gamma, 4, m)\},\\
c_{\max}(\mathcal{T}_{opt}^{temp}(4)) &= 4,\\
\mathcal{T}_{opt}(4) &= \{(\gamma, 4, m)\},\\
c_{\max}(\mathcal{T}_{opt}(4)) &= 4,\\
CL &= \{(m, 1)(4, 4)\},\\
SL &= \{m, 4\},\\
C &= \{1, 2\}.
\end{aligned}$$

Note that the figure is drawn by using the data in the array $[\mathcal{T}_{opt}(m)\ \mathcal{T}_{opt}(4)]$.

Run 3:

$$\begin{aligned}
\mathcal{T}_{opt}^{temp}(1) &= \{(\alpha, 1, 4)\},\\
c_{\max}(\mathcal{T}_{opt}^{temp}(1)) &= 4.5,\\
\mathcal{T}_{opt}^{temp}(2) &= \{(\delta, 2, m)\},\\
c_{\max}(\mathcal{T}_{opt}^{temp}(2)) &= 5,\\
\mathcal{T}_{opt}(1) &= \{(\alpha, 1, 4)\},\\
c_{\max}(\mathcal{T}_{opt}(1)) &= 4.5,\\
CL &= \{(m, 1)(4, 4)(1, 4.5)\},\\
SL &= \{m, 4, 1\},\\
C &= \{0, 2, 3\}.
\end{aligned}$$

Run 4:

$$\begin{aligned}
\mathcal{T}_{opt}^{temp}(0) &= \{(\gamma, 0, 1)\},\\
c_{\max}(\mathcal{T}_{opt}^{temp}(0)) &= 7.5,\\
\mathcal{T}_{opt}^{temp}(2) &= \{(\delta, 2, m)\}\text{(refer to computations below)},\\
c_{\max}(\mathcal{T}_{opt}^{temp}(2)) &= 5,\\
\mathcal{T}_{opt}^{temp}(3) &= \{(u, 3, 1), (\gamma, 3, m)\},\\
c_{\max}(\mathcal{T}_{opt}^{temp}(3)) &= 5.5,\\
\mathcal{T}_{opt}(2) &= \{(\delta, 2, m)\},\\
c_{\max}(\mathcal{T}_{opt}(2)) &= 5,\\
CL &= \{(m, 1)(4, 4)(1, 4.5)(2, 5)\},\\
SL &= \{m, 4, 1, 2\},\\
C &= \{0, 3\}.
\end{aligned}$$

The working of Optimize and One-Step Optimize in the computation of $\mathcal{T}_{opt}^{temp}(2)$ is as follows. Observe that $E_0(2) = \{(\beta, 2, 1), (\gamma, 2, 4), (\delta, 2, m)\}$. When $q_d = 2$ in Optimize the set is reordered by step (iv) to $E_0(2) = \{(\delta, 2, m), (\gamma, 2, 4), (\beta, 2, 1)\}$. The required data is obtained from CL in Run 2, e.g., $c_{\sup}^{g}(M_D^o(G, 1)) = 4.5$. The event costs are of course known a priori. Finally, $c_{\max}(E_0(2)) = 5.5$. In One-Step Optimize the sequence of computation is

(i)

$$\begin{aligned}
E' &= \{(\delta, 2, m), (\gamma, 2, 4), (\beta, 2, 1)\},\\
c_{\max}(E') &= 5.5,\\
CMAX &= 5.5.
\end{aligned}$$

(ii)

$$\begin{aligned} E' &= \{(\delta, 2, m), (\gamma, 2, 4)\}, \\ c_{\max}(E') &= 6, \\ CMAX &= 5.5. \end{aligned}$$

(iii)

$$\begin{aligned} E' &= \{(\delta, 2, m)\}, \\ c_{\max}(E') &= 5, \\ CMAX &= 5. \end{aligned}$$

Thus, finally, we obtain $E = \{(\delta, 2, m)\}$ and $\mathcal{T}_{opt}^{temp}(2), c_{\max}(\mathcal{T}_{opt}^{temp}(2))$ as already stated.

Run 5:

$$\begin{aligned} \mathcal{T}_{opt}^{temp}(0) &= \{(\gamma, 0, 1)(\delta, 0, 2)\}, \\ c_{\max}(\mathcal{T}_{opt}^{temp}(0)) &= 6, \\ \mathcal{T}_{opt}^{temp}(3) &= \{(u, 3, 1), (\gamma, 3, m)\}, \\ c_{\max}(\mathcal{T}_{opt}^{temp}(3)) &= 5.5, \\ \mathcal{T}_{opt}(3) &= \{(u, 3, 1), (\gamma, 3, m)\}, \\ c_{\max}(\mathcal{T}_{opt}(3)) &= 5.5, \\ CL &= \{(m, 1)(4, 4)(1, 4.5)(2, 5)(3, 5.5)\}, \\ SL &= \{m, 4, 1, 2, 3\}, \\ C &= \{0\}. \end{aligned}$$

Run 6:

$$\begin{aligned} \mathcal{T}_{opt}^{temp}(0) &= \{(\gamma, 0, 1)(\delta, 0, 2)\}, \\ c_{\max}(\mathcal{T}_{opt}^{temp}(0)) &= 6, \\ \mathcal{T}_{opt}(0) &= \{(\gamma, 0, 1), (\delta, 0, 2)\}, \\ c_{\max}(\mathcal{T}_{opt}(0)) &= 6, \\ CL &= \{(m, 1)(4, 4)(1, 4.5)(2, 5)(3, 5.5)(0, 6)\}, \\ SL &= \{m, 4, 1, 2, 3, 0\}. \end{aligned}$$

The information used to construct Figure 6.7 is in the array $[\mathcal{T}_{opt}(.)]$. Observe that this FSM is not trim. It can now be trimmed by standard methods to obtain the trim maximal DP-optimal submachine depicted in Figure 6.8.

7. Conclusion. In this paper we have introduced numerical performance measures in supervisory control theory. The DES is represented by a formal language, and the measures are represented by event and control cost functions. The two costs are associated with the generation of events by the DES and the disabling of events by the supervisor, respectively. Using these quantitative measures, we have examined the problem of minimizing the worst-case behavior of a DES. DESs operate in uncertain environments and it is desirable that a DES supervisor deal with uncertainty in a minimally restrictive manner. Thus computing a DES supervisor is more than computing a trajectory. This gives some interesting and unique features to the problem of defining and computing an optimal supervisor. We have presented an example in section 2, to motivate an interesting use of the cost function and DES models. Another example may be found in [10].

The investigation in this paper shows us that optimal supervisory control is related to two important domains. One is the area of path problems on a directed graph and

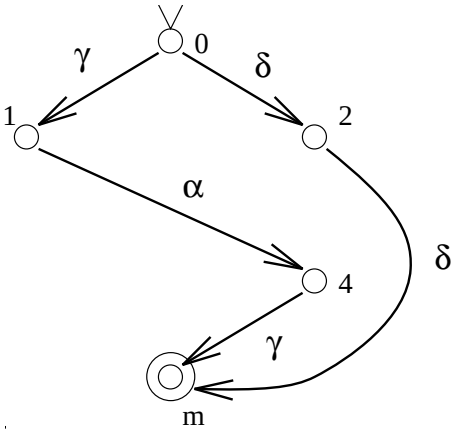


FIG. 6.8. *Trim maximal DP-optimal submachine of G.*

the other is the optimization of Markov decision processes. For further details on the latter, we refer to [10]. If there are no control costs, then an optimal supervisor is one that disables everything other than the shortest path, and a minimally restrictive supervisor would be one that allows all shortest paths. At the other end of the spectrum we have the class of optimization problems on Markov chain models. In stochastic control it is assumed that the system is uncertain, though one has stochastic information about the uncertainty. The system may execute any of a set of transition sequences, and control consists of altering the probability distribution on the set of future behaviors. Like our formulation, this involves control costs. The concept of disabling used in SCT can be viewed as altering the probability distribution so as to reduce the probabilities associated with some events to zero. The modeling and control assumptions are remarkably similar.

The principle of dynamic programming plays an important role. An interesting feature of the worst-case problem, which it shares with other min-max problems, is that all optimal solutions do not necessarily solve a DP-equation, though if optimal solutions exist, the DP-equations also have solutions. These solutions represent the DP-optimal solutions. Moreover, in this class of solutions there exists a minimally restrictive DP-optimal solution. The minimally restrictive DP-optimal solution guarantees the best possible future regardless of the past.

Theorems 3.1–3.7 constitute the existence theory. Theorem 3.1 deals with DES defined over a finite alphabet and asserts that the existence of a bounded-cost supervisor is sufficient for the existence of an optimal supervisor. Theorem 3.2 asserts the same for DES defined over a countable alphabet, but assumes positive event costs and a finite active event set. Theorem 3.3 covers specifically the case of DES modeled by regular languages. Theorems 3.5 and 3.7 deal with the existence of specific types of optimal solutions, when it is already known that an optimal solution exists. The two types of solutions are the supremal optimal solution and the supremal DP-optimal solution, respectively.

Aside from the interesting property that the supremal DP-optimal solution always guarantees the best possible worst-case behavior for a particular past in a minimally restrictive manner, it also has a suboptimal structure that we are able to exploit to develop two polynomially computable algorithms. The first is for DES modeled by FSMs having cycles, and the second is for DES modeled by acyclic FSMs. The implications of this suboptimal substructure are developed in section 5. Theorem 3.8

implies that the number of Nerode equivalence classes of the supremal DP-optimal sublanguage is no more than that of the plant language. Moreover, Theorem 3.9 says that any deterministic FSM generating the plant language has a submachine generating the supremal DP-optimal sublanguage. The next step is to find the relevant submachine. This is accomplished by formulating an equivalent (in the sense of equation (5.1)) optimization problem on the set of submachines of G (generator of the plant language). Then Theorem 5.3 establishes that the required submachine is the unique maximal DP-optimal submachine of G.

The polynomial-time algorithms presented in section 6 involve a few stages of processing. They are backward recursive dynamic programming algorithms. We summarize the acyclic case first. The algorithm starts at the set of terminal marked states and recurses backwards. For each state a DP-equation has to be solved over the set of subsets of the set of child states (Theorem 6.5). This is more complex than computing a shortest path or supremal controllable sublanguage, which require search operations only over the set of child states. Although the set of subsets of a state is not polynomially related to the alphabet, it is possible to use certain structural properties of the DP-optimal solution to sort the set of child states (Lemma 6.8). Using this sorted set the DP-equation can be solved with linear complexity. All these operations collectively result in an overall complexity of $O(n|\Sigma|\log(|\Sigma|))$ (Theorem 3.11).

For DES modeled by FSMs having cycles, the additional complication lies in the ordering of the vertices for the backward recursion. In the acyclic case this ordering is straightforward and obtained by starting at the set of terminal marked vertices. The DP-equation can be solved at all parents of this set, and the new set of solved vertices obtained. Once again, the entire set of parent states can be computed, the DP-equation can be solved at all of them, and the solved state list can be augmented as before. This process can be repeated until the initial state is seen. However, if the plant model has cycles in it, then the DP-equation cannot necessarily be correctly solved for all parent states of the set of solved states. Once again, a structural property of the supremal DP-optimal solution, i.e., that it is acyclic (hence the assumption that event costs are positive), can be used to extract a subset of the parent states for which the DP-equation can be solved correctly (Lemmas 6.6 and 6.7). This limited set of states can be added to the solved list and the new process repeated until the initial state is encountered. This additional processing of the set of parent states increases the complexity and makes the overall complexity rise to $O(n^2|\Sigma|\log(|\Sigma|) + n^3|\Sigma|)$ (Theorems 6.9 and 3.10). We have included an example, for the cyclic case, to help the reader understand the different stages in the algorithms.

The controller synthesis process involves several complex manipulations of FSMs. Accordingly, we have been mathematically rigorous in ascertaining the correctness of the algorithms. To the best of our knowledge the different stages of processing are unavoidable. It is desirable that an efficient controller synthesis algorithm be developed for cyclic DES having multiple marked states.

Appendix. List of notation.

$\|.\|$	=	Myhill congruence index of a language or the length of a string,
$\equiv_{L_m}$	=	the Nerode equivalence relation on the language $\bar{L}_m$,
$[.]$	=	Nerode equivalence class of a string,
$\bar{L}_m/\equiv_{L_m}$	=	set of equivalence classes of the language $\bar{L}_m$ defined by the Nerode equivalence relation.
ε	=	empty event,

$\delta_A(.,.)$ = transition function of machine A,
$\delta_A^*(.,.)$ = extension of δ_A to strings,
σ_j = jth event in the alphabet,
Π = set of control laws,
Π_{nb} = set of nonblocking control laws,
$\Pi_0(.)$ = set of active events disabled by the control law after a particular string,
$\pi(.)$ = control law defined on a language,
π_{A_m} = control law generating sublanguage A_m,
Σ = plant alphabet or event set,
Σ_c = set of controllable events,
Σ_{uc} = set of uncontrollable events,
$\Sigma_A(q)$ = active event set at state q in machine A,
$\Sigma_L(s)$ = active event set after string s in language L,
Σ^* = Kleene star closure of Σ.

$c_c(.)$ = control cost,
$c_e(.)$ = event cost,
$c(.,.)$ = cost of a string in a language,
$\hat{c}(.,.,.)$ = cost of a string less control costs after the last event,
$\bar{c}(.,.,.)$ = one-stage cost of an event in a language,
$c_{\sup}(.)$ = worst-case cost of a language,
$c^g(.,.,.)$ = cost of a string in a FSM,
$\bar{c}^g(.,.)$ = one-stage cost of an event in a FSM,
$c^g_{\sup}(.)$ = worst-case cost of a FSM,
$p_j(.)$ = prefix of length j of a string,
q_0 = initial state of FSM G,
q_{0A} = initial state of submachine A,
$\bar{s}$ = set of prefixes of the string s,
u^* = $\{u^n : n \in \mathbb{N}\}$.

G = plant FSM,
$G^{\uparrow}_{DO}$ = maximal DP-optimal submachine of G,
L = language,
$\bar{L}$ = prefix closure of L,
L_m/s = post-language of $s \in \bar{L}_m = \{t \in \Sigma^* : st \in L_m\}$,
$L^{\uparrow}_o$ = supremal optimal sublanguage,
$L^{\uparrow}_{DO}$ = supremal DP-optimal sublanguage,
L_m = the plant language,
$M(A,q)$ = the maximal trim submachine of A at $q \in Q_A$,
$M_1(A,q)$ = the maximal one-step submachine of A at $q \in Q_A$,
$M^o_D(G,q)$ = maximal DP-optimal submachine of G at q,
$P_f(q)$ = set of parent states of state q,
$P_j(.)$ = set of strings of length j in a language,
Q_A = set of states of FSM A,
Q_{mA} = set of marked states of submachine A,
$S_f(q)$ = set of children states of state q.

$\mathcal{L}_o^{\uparrow}([s])$ = supremal optimal solution in L_m/s,
$\mathcal{L}_{DO}^{\uparrow}([s])$ = supremal DP-optimal solution in L_m/s,
$\mathcal{M}(A,q)$ = set of trim submachines of A at $q \in Q_A$,
$\mathcal{M}_1(A,q)$ = set of one-step submachines of G at $q \in Q_A$,
$\mathbb{R}^+$ = positive reals (excluding zero).

Acknowledgment. We thank the reviewers for their useful comments.

REFERENCES

[1] F. J. Beutler and K. W. Ross, *Optimal policies for controlled markov chains with a constraint*, J. Math. Anal. Appl., 112 (1985), pp. 236–252.
[2] Y. Brave and M. Heymann, *On optimal attraction of discrete-event processes*, Inform. Sciences, 67 (1993), pp. 245–276.
[3] X.-R. Cao, *Ci bus arbitration performance in a vaxcluster system*, Digital Tech. J., (1987), pp. 93–103.
[4] E. Chen and S. Lafortune, *Dealing with blocking in supervisory control of discrete event systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 724–735.
[5] J. E. Hopcroft and J. D. Ullman, *Introduction to Automata Theory, Languages, and Computation*, Addison-Wesley, Reading, MA, 1979.
[6] R. Kumar and V. Garg, *Optimal supervisory control of discrete event dynamical systems*, SIAM J. Control Optim., 33 (1995), pp. 419–39.
[7] K. M. Passino and P. J. Antsaklis, *On the optimal control of discrete event systems*, in Proc. 28th IEEE Conf. on Decision and Control, Tampa, FL, 1989, pp. 2713–2718.
[8] P. J. Ramadge and W. M. Wonham, *Supervisory control of a class of discrete event processes*, SIAM J. Control Optim., 25 (1987), pp. 206–230.
[9] P. J. Ramadge and W. M. Wonham, *The control of discrete event systems*, Proc. IEEE, 77 (1989), pp. 81–98.
[10] R. Sengupta, *Optimal Control of Discrete Event Systems*, Ph.D. thesis, University of Michigan, Ann Arbor, MI, 1995.
[11] R. Sengupta and S. Lafortune, *Optimal control of a class of discrete event systems*, in Preprints, IFAC Int. Symp. on Distributed Intelligence Systems, Arlington, VA, 1991, pp. 25–30.
[12] R. Sengupta and S. Lafortune, *A graph-theoretic optimal control problem for terminating discrete event processes*, Journal of Discrete Event Dynamic Systems: Theory and Applications, 2 (1992), pp. 139–172.
[13] R. Sengupta and S. Lafortune, *A deterministic optimal control theory for discrete event systems*, in Proc. 32nd IEEE Conf. on Decision and Control, San Antonio, TX, 1993, pp. 1182–1187.
[14] R. Sengupta and S. Lafortune, *Extensions to the theory of optimal control of discrete event systems*, in Discrete Event Systems: Modeling and Control – Proceedings of a Joint Workshop on Discrete Event Systems, S. Balemi, P. Kozák, and R. Smedinga, eds., Birkhäuser Verlag, Basel, 1993, pp. 153–160.
[15] C. E. Leiserson, T. H. Cormen, and R. L. Rivest, *Introduction to Algorithms*, MIT Press and McGraw-Hill, Cambridge, MA, 1990.
[16] W. M. Wonham and P. J. Ramadge, *On the supremal controllable sublanguage of a given language*, SIAM J. Control Optim., 25 (1987), pp. 637–659.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 2, pp. 542–578, March 1998

005

CLASSES OF NONLINEAR PARTIALLY OBSERVABLE STOCHASTIC OPTIMAL CONTROL PROBLEMS WITH EXPLICIT OPTIMAL CONTROL LAWS*

CHARALAMBOS D. CHARALAMBOUS† AND ROBERT J. ELLIOTT‡

Abstract. This paper introduces certain nonlinear partially observable stochastic optimal control problems which are equivalent to completely observable control problems with finite-dimensional state space. In some cases the optimal control laws are analogous to linear-exponential-quadratic-Gaussian and linear-quadratic-Gaussian tracking problems. The problems discussed allow nonlinearities to enter the unobservable dynamics as gradients of potential functions. The methodology is based on explicit solutions of a modified Duncan–Mortensen–Zakai equation.

Key words. stochastic control, risk-sensitive, nonlinear filtering, sector-bounded nonlinearities, exact optimal controls

AMS subject classifications. 93E20, 93E03, 93E11

PII. S036301299528732 65

1. Introduction.

An important concept associated with closed loop control laws for noisily observed linear systems is the so-called "*separation principle*." This principle allows one to solve an estimation problem first, and then solve a completely observable control problem whose state is the estimate (observer state). For linear-quadratic-Gaussian (LQG) tracking problems the observer dynamics are given by the conditional mean and error covariance equations (see [1, 2]); for linear-exponential-quadratic-Gaussian (LEQG) tracking problems the observer dynamics are given by a variant of the conditional mean and error covariance equations (see [2, 3, 4, 5, 6]). Thus, the problem of optimally controlling the dynamics of the plant is equivalent, for both the LQG and the LEQG regulator problems, to a standard completely observable optimal control problem with a new state which is either the conditional mean or a variant of the conditional mean, respectively.

However, when the dynamics or observations are nonlinear in the unobservable state, the optimal control laws are infinite dimensional and, consequently, the classical separation principle discussed above does not apply. In this paper we identify classes of partially observed optimal control problems which are equivalent to completely observed control problems having a finite-dimensional state space. This allows us to apply the separation principle, as in LEQG/LQG problems. We then state sufficient conditions that enable us to compute the optimal control laws explicitly. Further, we describe techniques which compute suboptimal control laws in closed form.

Our results are also applicable in evaluating Feynman–Kac integrals for partially observable systems, such as the ones arising in risk-sensitive filtering. In addition, from the duality between estimation and control problems, explicit solution of the es-

*Received by the editors June 8, 1995; accepted for publication (in revised form) January 6, 1997. http://www.siam.org/journals/sicon/36-2/28732.html

†Department of Electrical Engineering, McGill University, Montreal, PQ, Canada H3A 2A7 (chadcha@cim.mcgill.ca). The research of this author was supported by the Measurement and Control Research Center of Idaho State University.

‡Department of Mathematical Sciences, University of Alberta, Edmonton, AB, Canada T6G 2G1 (relliott@gpu.sru.ualberta.ca). Present address: Department of Applied Mathematics, University of Adelaide, Adelaide, SA 5000, Australia (relliott@maths.adelaide.edu.au). The research of this author was supported by NSERCC grant 47964.

timation problem translates into explicit solution of completely observable, stochastic optimal control problems (see [7]).

The classes of problems which we shall treat involve an $\Re^n$-valued unobservable state process $x(\cdot)$ given by the stochastic differential equation

$$dx(t) = f(t, x(t))dt + B(t, u(t, y))dt + G(t)dw(t), \qquad x(0) \in \Re^n, \quad 0 \le t \le T. \tag{1.1}$$

This is observed through an $\Re^d$-valued process $y(\cdot)$, which satisfies the stochastic differential equation

$$dy(t) = h(t, x(t))dt + N(t)^{\frac{1}{2}} db(t), \qquad y(0) = 0 \in \Re^d. \tag{1.2}$$

$y(\cdot)$ is called the observation process. Here, $w(\cdot), b(\cdot)$ are, respectively, $\Re^n$- and $\Re^d$-valued independent Wiener processes, independent of the random variable $x(0)$, $u(\cdot)$ is the control process, and $T \in \Re$ is fixed and finite. The cost function to be minimized over the controls u (which are nonanticipating functionals of the observations y) is of the general form

$$\begin{aligned} J_G^\theta(u(\cdot)) = E^u \Big\{ \int_0^T \ell_2(t, x(t), u(t, y)) \exp\theta \Big(\int_0^t \ell_1(s, x(s), u(s, y))ds \Big) dt \\ + \varphi_2(T, x(T)) \exp\theta \Big(\int_0^T \ell_1(t, x(t), u(t, y))dt + \varphi_1(T, x(T)) \Big) \Big\}, \quad \theta > 0. \end{aligned} \tag{1.3}$$

Here $\ell_i, \varphi_i, i = 1, 2$ are real-valued functions and E^u denotes expectation with respect to a certain probability measure P^u. The precise assumptions on the coefficients of (1.1)–(1.3) are stated under Assumptions 2.1. This cost criterion appears to be quite general, as it includes both the integral and the exponential-of-integral cost criteria: the integral cost criterion can be found by considering

$$J_I^0(u(\cdot)) \doteq \left\{ J_G^\theta(u(\cdot)); \quad \theta = 0 \right\}, \tag{1.4}$$

while the exponential-of-integral cost criterion can be found as

$$J_{EI}^\theta(u(\cdot)) \doteq \left\{ J_G^\theta(u(\cdot)); \quad \ell_2 = 0, \varphi_2 = 1 \right\}. \tag{1.5}$$

The approach is based on an "information state" formulation which recasts the problem as a completely observable control problem with an infinite-dimensional state space, and control laws which are functionals of this quantity. The information state associated with the usual integral cost criterion is the unnormalized conditional distribution; this satisfies the Duncan–Mortensen–Zakai (DMZ) equation (see [8]). The information state for the exponential-of-integral cost criterion is a modified version of the unnormalized conditional distribution; this satisfies a variant of the DMZ equation [3, 4, 5, 9]. To distinguish between the two we refer to the former as the information state and to the latter as the Feynman–Kac information state.

The results obtained in this paper are extensions of recent related work pursued independently by Charalambous, Naidu, and Moore [9], Bensoussan and Elliott [10], and Charalambous [6].

In section 2, we state the main assumptions, identify an "information state," and present an equivalent formulation of the partially observable problem (1.1)–(1.3), which, although completely observable, has an infinite-dimensional state space.

In section 3, we restrict the coefficients in the unobservable dynamics, observations, and cost (see (1.1)–(1.3)) to forms (for simplicity we often write x_t instead of $x(t)$, etc.)

$$f(t, x, u) = F_t x + g(t, x) + f_t + B_t u, \quad h(t, x) = H_t x + h_t, \quad \ell_2 = 0, \tag{1.6}$$

$$2\ell_1(t,x,u) = Q_t x.x + R_t x.x + 2m_t x + 2n_t u + \tilde{\ell}_1(t,x,u), \qquad 2\varphi_1(T,x) = Q_T x.x + 2m_T x. \tag{1.7}$$

Here the notation "$\alpha.\beta \doteq \alpha'\beta$" is used throughout the paper, where $(\cdot)'$ denotes the transpose of a matrix. We show that if

$$g(t,x) = G_t G_t' \frac{\partial}{\partial x}\phi(x), \tag{1.8}$$

$$\tilde{\ell}_1(t,x,u) = \tfrac{1}{\theta}\left\{|G_t^{-1}(F_t x + f_t + B_t u + g(t,x))|^2 - |G_t^{-1}(F_t x + f_t + B_t u)|^2 + Tr(D_x g(t,x))\right\}, \tag{1.9}$$

then the sufficient statistics are similar to those of an LEQG tracking problem and, consequently, the optimal control laws are finite-dimensional. Moreover, if in addition

$$\varphi_2(T,x) = \exp(-\phi(x)), \tag{1.10}$$

then the optimal control law is precisely that of the LEQG tracking problem.

When $\tilde{\ell}_1(t,x,u) \equiv \tilde{\ell}_1(x,u) = \frac{1}{\theta}V(x,u)$, which is a quadratic function of x and u and (without loss of generality) $G_t = I_n$ (an identity matrix), using $g(x) = D_x\phi(x)$, then (1.9) is reduced to the controlled Riccati equation

$$Tr\left(\frac{\partial}{\partial x}g(x)\right) + |g(x)|^2 + 2(Fx + f + Bu).g(x) = V(x,u). \tag{1.11}$$

Solutions of (1.11) yield finite-dimensional controllers. Notice that, when $\tilde{\ell}_1(t,x,u) \equiv \tilde{\ell}_1(x,u) = \frac{2}{\theta}(Fx+f+Bu).g(x) + \frac{1}{\theta}\tilde{V}(x)$, where $\tilde{V}(\cdot)$ is a quadratic function of x, then (1.9) reduces to the Riccati equation

$$Tr\left(\frac{\partial}{\partial x}g(x)\right) + |g(x)|^2 = \tilde{V}(x), \tag{1.12}$$

encountered in identifying finite-dimensional solutions of the DMZ equation by Benes in [11].

In section 3.1.1, we also show that if $\varphi_2 = 1$ and the nonlinear drift terms $g(\cdot)$ satisfy sector criteria (see A14), then suboptimal linear feedback control laws are found by employing simple upper and lower bounds on the terminal cost. Finite-dimensional sufficient statistics are found when the coefficients in the observations have the form $h(t,x) = \frac{1}{2}x.\tilde{H}_t x + H_t x + h_t$. Analogous results for stochastic control problems with cost (1.4) are derived in section 4. In section 5, we discuss the use of dynamic programming to formally derive verification theorems for nonlinear stochastic control problems which emerge from solving the above Riccati equations.

The optimal solutions of two examples that emerge from the developments of this paper are now presented.

Example 1.1. Consider the stochastic optimal control analog of Benes's filter (see [11]):

$$dx_t = \tanh(x_t)dt + u(t,y)dt + dw_t, \qquad x(0) = 0 \in \Re,$$
$$dy_t = x_t dt + db_t, \qquad y(0) = 0 \in \Re.$$

The objective is to determine the optimal control law u^* that minimizes the cost function

$$J^\theta(u(\cdot)) = E^u\left\{\exp\tfrac{\theta}{2}\left(\int_0^T\left[Qx_t^2 + Ru(t,y)^2\right]dt + Q_T x_T^2\right)\right.$$
$$\left.\times\exp\left(\int_0^T u(t,y)\tanh(x_t)dt\right)\times\tfrac{1}{\cosh(x_T)}\right\},$$

where $Q_T, Q \geq 0, R > 0$.

Example 1.2. Consider a stochastic optimal control problem with cubic nonlinearity in the unobservable dynamics:

$$dx_t = -\alpha x_t^{2p+1}dt + u(t,y)dt + dw_t, \qquad x(0) = 0 \in \Re, \qquad \alpha > 0, \quad p = 1,3,5,\ldots,$$
$$dy_t = x_t dt + db_t, \qquad y(0) = 0 \in \Re.$$

The objective is to determine the optimal control law u^* that minimizes the cost function

$$J^\theta(u(\cdot)) = E^u\left\{\exp\tfrac{\theta}{2}\left(\int_0^T\left[Qx_t^2 + Ru(t,y)^2\right]dt + Q_T x_T^2\right)\right.$$
$$\times\exp\left(\int_0^T\tfrac{1}{2}\left[|-\alpha x_t^{2p+1} + u(t,y)|^2 - u(t,y)^2 + \tfrac{\partial}{\partial x}(-\alpha x_t^{2p+1})\right]dt\right.$$
$$\left.\left.+\tfrac{\alpha}{2p+2}x_T^{2p+2}\right)\right\},$$

where $Q_T, Q \geq 0, R > 0$. Completing the squares in x and u ensures that the integrand in the exponent is bounded below.

The solution of Example 1.1 is an application of Theorem 5.1; the solution of Example 1.2 is an application of Theorem 3.4 (using (1.8)–(1.10)). The optimal control law and optimal sensitivity parameter for the above two problems are identical and are:

$$u^*(t) = -R^{-1}\Sigma_t r_t = -R^{-1}(1-\theta S_t P_t)^{-1}S_t r_t,$$

$$\theta^* = \sup\left\{\theta; \quad P_t \geq 0,\ S_t \geq 0,\ (1-\theta P_t S_t) > 0 \quad \forall t \in [0,T]\right\}.$$

Here $\Sigma(\cdot), S(\cdot)$ are the solutions of the Riccati differential equations:

$$\dot\Sigma_t + 2\theta P_t Q - \Sigma(R^{-1} - \theta P_t^2)\Sigma_t = 0, \qquad \Sigma_T = (1-\theta Q_T P_T)^{-1}Q_T,$$
$$\dot S_t - S_t(R^{-1} - \theta)S_t + Q = 0, \qquad S_T = Q_T, \tag{1.13}$$

and $r(\cdot), P(\cdot)$ are solutions of the observer dynamics:

$$dr_t = \theta P_t Q r_t dt + u^*(t,r)dt + P_t(dy_t - r_t dt), \qquad r(0) = 0,$$
$$\dot P = -P_t(1-\theta Q)P_t + 1, \qquad P(0) = 0.$$

Setting $R = Q = 1, Q_T = 0$, one can verify that the control Riccati differential equations and observer dynamics correspond to the following $\mathcal{H}^\infty$, or robust, control problem:

$$\dot x_t = u(t,y) + w_t, \qquad x(0) = 0 \in \Re,$$
$$y_t = x_t + b_t, \qquad y(0) = 0 \in \Re,$$
$$J_{\mathcal{H}^\infty}(u^*(\cdot)) = \inf_u \sup_{(w,b)} \int_0^T \frac{1}{2}\left[x_t^2 + u(t,y)^2 - \frac{1}{\theta}\left(w_t^2 + b_t^2\right)\right]dt,$$

where $w(\cdot), b(\cdot) \in L^2([0,T];\Re)$. In addition, from [12, pp. 131–132], we know that, for $\theta > 1$, there exist unique solutions

$$P_t = \frac{\tan((\sqrt{\theta-1})t)}{\sqrt{\theta-1}}, \qquad S_t = \frac{\tan(\sqrt{\theta-1}(T-t))}{\sqrt{\theta-1}},$$

provided $\theta < \frac{4T^2+\pi^2}{4T^2}$. Also, for $T = \frac{\pi}{2}$ the optimal risk-sensitive parameter θ^* is $\theta^* \approx 1.3763$. Therefore, for $1 < \theta < 1.3763$ the optimal control law $u^*(\cdot)$ exists for a family of optimal controllers. An important observation concerning the exact solution of Examples 1.1 and 1.2 is the linearity of the observer dynamics, despite the nonlinearity of the unobservable dynamics. This feature of the observer is also present in the Benes's filter in [11].

2. Problem formulation.

2.1. Dynamics. We start with a reference probability space $(\Omega, \mathcal{A}, P)$ with a complete filtration $\{\mathcal{F}_t; t \in [0,T]\}$, two $\{\mathcal{F}_t; t \in [0,T]\}$-adapted Wiener processes $\{w(t); t \in [0,T]\}$, $\{b(t); t \in [0,T]\}$, and an $\mathcal{F}_0$-measurable random variable $x(0)$ such that:

$$w : [0,T] \times \Omega \to \Re^n \quad \text{is a standard Wiener process independent of } b(\cdot),$$

$$b : [0,T] \times \Omega \to \Re^d \quad \text{is a standard Wiener process independent of } w(\cdot),$$

$$x(0) : \Omega \to \Re^n \quad \text{is a random variable independent of } w(\cdot), b(\cdot).$$

Further, suppose an observation process $y(\cdot)$ is given by

$$dy(t) = N(t)^{\frac{1}{2}} db(t), \qquad y(0) = 0 \in \Re^d. \tag{2.14}$$

The assumptions concerning (1.1)–(1.3) are now given; some of these assumptions will be weakened at a later stage depending on the nature of the optimization problem under consideration.

ASSUMPTION 2.1. *$|\cdot|^2$ denotes the Euclidean norm and $\mathcal{L}(V_1; V_2)$ denotes the space of linear transformations of a vector space V_1 into a vector space V_2.*

A1. *$\mathcal{U}$ is a nonempty subset of $\Re^m$.*

A2. *$f : [0,T] \times \Re^n \to \Re^n$ is continuous in t, continuous differentiable in x, and*

$$|f(t,x) - f(t,z)| \le K|x-z|, \qquad |f(t,x)| \le K(1+|x|). \tag{2.15}$$

A3. *$B : [0,T] \times \mathcal{U} \to \Re^n$ is Borel measurable, continuous in t, and*

$$|B(t,u)| \le K(1+|u|).$$

A4. *$h : [0,T] \times \Re^n \to \Re^d$ is continuous in t, once continuously differentiable in t, twice continuously differentiable in x, and*

$$|h(t,x)| \le K(1+|x|). \tag{2.16}$$

A5. *$N : [0,T] \to \mathcal{L}(\Re^d; \Re^d)$, $N = N^{\frac{1}{2}} N^{\frac{1}{2}\prime}$, $\exists\, \beta_1 > 0$ such that $N \ge \beta_1 I_d$.*

A6. *$G : [0,T] \to \mathcal{L}(\Re^n; \Re^n)$, $\exists\, \beta_2 > 0$ such that $G \ge \beta_2 I_n$.*

A7. *$\ell_i : [0,T] \times \Re^n \times \mathcal{U} \to \Re$, $\varphi_i : [0,T] \times \Re^n \to \Re, i = 1,2$ are Borel measurable, continuous in t, and*

$$|\ell_i(t,x,u)| \le K(1+|x|+|u|)^{l_i}, \quad |\varphi_i(t,x)| \le K(1+|x|)^{k_i}, \quad i = 1,2, \tag{2.17}$$

where l_i, k_i are positive constants.

A8. *The distribution* $\Pi_0(\cdot)$ *of* $x(0)$ *has a density* $q_0^\theta(\cdot)$ *in the space*

$$\mathcal{M}_k \doteq \left\{ p \in L^1(\Re^n); \quad ||p||_k \doteq \int_{\Re^n} (1+|x|^k)|p(x)|dx < \infty \right\} \quad \textit{for all } k \geq 1.$$

A9. N, G *are continuous in* t.

We write $\{\mathcal{F}_t^y; t \in [0,T]\}$ for the complete filtration generated by the observation σ-algebras $\sigma\{y(s); 0 \leq s \leq t \leq T\}$, and we denote by E^u (resp., $E, \hat{E}^u$), the expectation with respect to measure P^u (resp., $P, \hat{P}^u$).

DEFINITION 2.2. *Denote by* $L_y^2\left([0,T];\Re^k\right)$ *the set of square integrable stochastic processes adapted to* $\{\mathcal{F}_t^y, t \in [0,T]\}$ *with values in* $\Re^k$. *The set of admissible controls denoted by* $\hat{\mathcal{U}}$ *is defined by*

$$\hat{\mathcal{U}} \doteq \left\{ u(\cdot) \in L_y^2([0,T];\Re^k); \quad u(t,y) \in \mathcal{U}, \textit{ a.e. } t \, P\text{—a.s.} \right\}.$$

For the system $(\Omega, \mathcal{A}, P; \mathcal{F}_t)$ and for $u \in \hat{\mathcal{U}}$, consider the diffusion process $x(\cdot)$ satisfying the Ito equation

$$dx(t) = f(t,x(t))dt + B(t,u(t,y))dt + G(t)dw(t), \qquad x(0) \in \Re^n. \tag{2.18}$$

Under Assumptions 2.1 we have $B(\cdot, u(\cdot,y)) \in L_y^2([0,T];\Re^n)$, and there exists a unique solution $x(\cdot) \in L^2(\Omega, \mathcal{F}_t, P; C([0,T];\Re^n))$ of (2.18). For $u \in \hat{\mathcal{U}}$ define the $\{\mathcal{F}_t; t \in [0,T]\}$-adapted process $\Lambda^u(\cdot)$ by

$$\Lambda^u(t) \doteq \exp\left\{ \int_0^t h(s,x(s)).N(s)^{-1}dy(s) - \frac{1}{2}\int_0^t h(s,x(s)).N(s)^{-1}h(s,x(s))ds \right\}.$$

For $u \in \hat{\mathcal{U}}$, in view of Assumptions 2.1 we deduce that there exists some $\delta > 0$ such that $\sup_{t\in[0,T]} E\left\{\exp\left(\delta \int_0^t |N(s)^{-\frac{1}{2}}h(s,x(s))|^2 ds\right)\right\} < \infty$, and thus we have $E[\Lambda^u(t)] = 1$ (as in [13, Theorems 4.7, 6.1]). Therefore, for $u \in \hat{\mathcal{U}}$ a new measure P^u can be defined through the Radon–Nikodým derivative

$$\frac{dP^u}{dP}|_{\mathcal{F}_T} \doteq \Lambda^u(T).$$

Then, Girsanov's theorem states that P^u is a probability measure on $(\Omega, \mathcal{A}; \mathcal{F}_t)$ and that if the stochastic processes $w^u(\cdot), b^u(\cdot)$ are defined by

$$dw^u(t) \doteq dw(t), \qquad db^u(t) \doteq db(t) - N(t)^{-\frac{1}{2}}h(t,x(t))dt,$$

then $b^u(\cdot), w^u(\cdot)$ are independent standard Wiener processes on $(\Omega, \mathcal{A}, P^u; \mathcal{F}_t)$. Furthermore, for each $u \in \hat{\mathcal{U}}$, under P^u, $(x(\cdot), y(\cdot))$ is a unique weak solution of

$$dx(t) = f(t,x(t))dt + B(t,u(t,y))dt + G(t)dw^u(t), \qquad x(0) \in \Re^n, \tag{2.19}$$

$$dy(t) = h(t,x(t))dt + N(t)^{\frac{1}{2}}db^u(t), \qquad y(0) = 0 \in \Re^d. \tag{2.20}$$

Now let $w(\cdot) \equiv w^u(\cdot), b(\cdot) \equiv b^u(\cdot)$; then (2.19), (2.20) correspond, respectively, to the stochastic equations (1.1), (1.2).

2.2. Cost criterion. The problem consists of controlling the evolution of the state process $\{x(t); t \in [0,T]\}$ using the control process $\{u(t,y); t \in [0,T]\}$, which is a function of the data $\{y(t); t \in [0,T]\}$. The objective is to find an optimal control, denoted by u^*, such that

$$J_G^\theta(u^*(\cdot)) = \inf_{u \in \hat{\mathcal{U}}} J_G^\theta(u(\cdot)), \tag{2.21}$$

where $J_G^\theta(u(\cdot))$ is given by (1.3). Under the reference probability measure P, (2.21) has the equivalent representation

$$\begin{aligned} J_G^\theta(u(\cdot)) = E\Big\{ & \Lambda^u(T) \int_0^T \ell_2(t,x(t),u(t,y)) \exp\theta\left(\int_0^t \ell_1(s,x(s),u(s,y))ds\right) dt \\ & + \Lambda^u(T)\varphi_2(T,x(T)) \exp\theta\left(\int_0^T \ell_1(t,x(t),u(t,y))dt + \varphi_1(T,x(T))\right)\Big\}. \end{aligned} \tag{2.22}$$

2.3. Feynman–Kac information state. In this section we shall introduce the Feynman–Kac information state associated with the stochastic control problem (2.19)–(2.21) (or equivalently, (1.1)–(1.3)).

For each $u \in \hat{\mathcal{U}}$ and $\phi \in C_b^2(\Re^n)$ consider the (backward) differential operator

$$A\phi \doteq \frac{1}{2} Tr\left(GG' \frac{\partial^2}{\partial x^2}\phi\right) + (f+B).\frac{\partial}{\partial x}\phi = \frac{1}{2} Tr\left(GG' D_x^2 \phi\right) + (f+B).D_x\phi,$$

whose formal adjoint is denoted by A^*. Write

$$\chi_t^{u,\theta} = \exp\left\{\theta \int_0^t \ell_1(s,x(s),u(s))ds\right\}, \qquad (\alpha,\beta) = \int_{\Re^n} \alpha(z)\beta(z)dz.$$

A consequence of the above formulation and Assumptions 2.1 is the following theorem.

THEOREM 2.3. *Suppose Assumptions* 2.1 *hold with* $k_1 = 2$, $l_1 = 0$, *and* $\mathcal{U}$ *in* $A1$ *is replaced by a compact subset of* $\Re^m$. *For some* $0 < \theta \le \theta^*$ *there exists an* $\mathcal{F}_t^y$*-measurable positive function* $q^\theta(x,t) \equiv q^\theta(x,\{y(s); 0 \le s \le t\},t)$ *satisfying the Feynman–Kac stochastic evolution equation*

$$\begin{aligned} dq^\theta(x,t) &= \left(A(t)^* + \theta\ell_1(t,x,u(t))\right) q^\theta(x,t)dt + h(t,x)q^\theta(x,t).N(t)^{-1}dy(t), \\ q^\theta(x,0) &= q_0^\theta(x), \end{aligned} \tag{2.23}$$

P*—a.s. for any* $u \in \hat{\mathcal{U}}$, *which is unique among the functions with exponential growth condition in the space variable.*

For any bounded continuous function $\varphi : \Re^n \to \Re$ *with compact support we have*

$$\rho_t^\theta(\varphi) \doteq E^u\left[\varphi(x(t))\chi_t^{u,\theta} | \mathcal{F}_t^y\right] = \int_{\Re^n} \rho^\theta(z,t)\varphi(z)dz = \frac{\int_{\Re^n} q^\theta(z,t)\varphi(z)dz}{\int_{\Re^n} q^\theta(z,t)dz}, \tag{2.24}$$

where

$$\begin{aligned} d\rho^\theta(x,t) &= \left(A(t)^* + \theta\ell_1(t,x,u(t))\right)\rho^\theta(x,t)dt - \theta\rho_t^\theta(\ell_1)\rho^\theta(x,t)dt \\ &\quad + [h(t,x) - \rho_t^\theta(h)].N(t)^{-1}d\hat{b}^u(t), \quad d\hat{b}^u(t) \doteq dy(t) - \rho_t^\theta(h)dt, \\ \rho_t^\theta(x,0) &= q_0^\theta(x), \end{aligned} \tag{2.25}$$

P^u*—a.s. for any* $u \in \hat{\mathcal{U}}$.

Furthermore, for $u \in \hat{\mathcal{U}}$, the total cost given in section 2.2 has the equivalent representation

$$J_G^\theta(u(\cdot)) = E\left\{\int_0^T \left(\ell_2(t,\cdot,u(t)), q^\theta(t)\right) dt + \left(\varphi_2 \exp\theta(\varphi_1), q^\theta(T)\right)\right\}. \tag{2.26}$$

When $l_1 = 2$, similar results hold for sufficiently small values of $\theta > 0$, that ensure $\theta\ell_1(t,x,u) - \frac{1}{2}|h(t,x)|^2 < 0\ \forall (t,x,u) \in [0,T] \times \Re^n \times \mathcal{U}$.

Proof. The evolution of the Feynman–Kac information state is established in [4] (and in [9, 5] when the signal and observation noises are correlated). When $\theta = 0$ this theorem is derived in [14], by first establishing existence and uniqueness results for (2.23) indirectly, using results from parabolic PDEs (see [15]). Following the derivation in [14], for a fixed sample path of the observation process $y(\cdot,\omega) \in C([0,T];\Re^d)$ and, consequently, a fixed sample path of the control process, we introduce the gauge transformation:

$$\hat{q}_t^\theta = \exp(-h(t,x).N_t^{-1}y_t)q_t^\theta, \qquad 0 \le t \le T. \tag{2.27}$$

Using (2.23) (see also [4, 16]), we see that $\hat{q}_t^\theta$ satisfies the following robust, or pathwise, version of the Feynman–Kac information state equation:

$$\frac{\partial}{\partial t}\hat{q}_t^\theta = \hat{A}(t)^*\hat{q}_t^\theta + e(t,x,y_t)\hat{q}_t^\theta, \qquad \hat{q}^\theta(x,0) = q^\theta(x,0). \tag{2.28}$$

Here

$$\hat{A} = A - D_x.GD_x(h.N^{-1}y_t), \tag{2.29}$$

and

$$e(\cdot,x,y_t) = \frac{1}{2}|G'D_x(h.N^{-1}y_t)|^2 - N^{-1}y_t.\left(\frac{\partial}{\partial t}h + Ah\right) - \frac{1}{2}|N^{-\frac{1}{2}}h|^2 + \theta\ell_1. \tag{2.30}$$

This is a parabolic PDE with $y(\cdot,\omega) \in C([0,T];\Re^d)$ entering parametrically through the coefficients. It has a bounded diffusion term, a linear growth drift term, and a quadratic growth potential term (because $l_1 = 0$). Therefore, for sufficiently small θ, there exists a unique positive fundamental solution $\hat{\Gamma}(x,t;z,s), 0 \le s \le t \le T$ (see [15, Theorem 4.5]) of (2.28) satisfying $|D_x^m\hat{\Gamma}(x,t;z,s)| < c(t-s)^{-(r+|m|)/2}\exp[-\mu\frac{|x-z|^2}{t-s}]$, $0 \le |m| \le 2$. Here c, μ are constants depending on $y(\cdot,\omega)$. Then, (2.28) has the classical solution $\hat{q}^\theta(\cdot) \in C_{x,t}^{2,1}(\Re^n \times [0,T])$ given by

$$\hat{q}^\theta(x,t) = \int_{\Re^n} \hat{\Gamma}(x,t;z,0)q^\theta(z,0)dz, \tag{2.31}$$

which is unique among the class of functions bounded above by $\alpha_1 \exp\left(-\alpha_2|x|^2\right), \alpha_1 > 0, \alpha_2 > 0$. Since the gauge transformation (2.27) is invertible, one easily verifies that $q_t^\theta = \exp(h(t,x).N_t^{-1}y_t)\hat{q}_t^\theta$ solves (2.23) P—a.s. The results translate directly into corresponding results for $q^\theta(\cdot)$, as claimed.

If we now set $\rho^\theta(x,t) = q^\theta(x,t)(\int_{\Re^n} q^\theta(x,t)dx)^{-1}$ and apply the Ito differential rule, (2.25) follows.

The derivation of (2.24) is established by introducing the adjoint backward version of (2.28) and following the derivation in [14], which treats the case $\theta = 0$. Moreover,

by Assumptions 2.1, with $l_1 = 0$, for $y(\cdot,\omega) \in C([0,T];\Re^d)$, we have

$$\begin{aligned} &\rho_t^\theta \in \mathcal{M}_k, \quad A(t)^*\rho_t^\theta \in \mathcal{M}_{k-1}, \quad \left[\ell_1 - \theta\rho_t^\theta(\ell_1)\right]\rho_t^\theta \in \mathcal{M}_k, \\ &\left[h - \rho_t^\theta(h)\right].N^{-1}\left[dy_t - \rho_t^\theta(h)dt\right]\rho_t^\theta \in \mathcal{M}_{k-1} \end{aligned} \tag{2.32}$$

for all $0 < t \le T$. This allows us to establish (2.24) for functions $\varphi(\cdot)$ which are continuous and satisfy the growth condition $|\varphi(T,x)| \le \exp\beta(|x|^2), \beta > 0$. Consequently, for sufficiently small $\theta > 0$, (2.26) is established. □

From Theorem 2.3 we formally deduce, setting $\theta = 0$, the evolution equation for the information state known as the DMZ equation:

$$\begin{aligned} dq^0(x,t) &= A(t)^*q^0(x,t)dt \\ &+h(t,x)q^0(x,t).N(t)^{-1}dy(t), \quad q^0(x,0) = q_0^0(x) = q_0^\theta(x). \end{aligned} \tag{2.33}$$

The new stochastic control problem derived in Theorem 2.3, although fully observable, has an infinite-dimensional state $q^\theta(\cdot)$, because the process $q^\theta(\cdot)$ is determined by the PDE (2.23). If we are able to describe the state process $q^\theta(\cdot)$ by finite-dimensional parameters, then we might be able to convert the infinite-dimensional control problem of Theorem 2.3 to a standard, finite-dimensional control problem. This is done by seeking an explicit solution of the equation governing the information state and its Feynman–Kac version in terms of the solutions of a finite number of ordinary differential equations which form finite-dimensional sufficient statistics for the estimation problem. If these statistics are also sufficient for the control problem, then by carrying out the integration of inner product terms $(\cdot,\cdot)$ present in (2.26) (whenever possible), we recover a cost function which is expressed in terms of the sufficient statistics. Unfortunately, one cannot in general expect the state space of the Feynman–Kac information state to evolve on a finite-dimensional manifold unless some restrictions on the vector fields $f(t,\cdot,u), h(t,\cdot)$ are imposed.

3. Finite-dimensional Feynman–Kac information states.

3.1. Nonlinear dynamics linear sensor problem. First consider the following family of nonlinear control systems.

Control system (Σ_G^1). Suppose Assumptions 2.1 hold with $l_1 = 2, k_1 = 2$. Suppose the dynamics, observations, and cost criterion are given by:

$$dx_t = (F_tx_t + g_t(x) + f_t)\,dt + B_tu(t,y)dt + G_tdw_t, \qquad x(0) \in \Re^n, \tag{3.34}$$

$$dy_t = (H_tx + h_t)\,dt + N_t^{\frac{1}{2}}db_t, \qquad y(0) = 0 \in \Re^d, \tag{3.35}$$

$$\begin{aligned} J^\theta_{\Sigma_G^1}(u(\cdot)) = E^u\Big\{&\varphi_2(T,x_T)\exp\tfrac{\theta}{2}\Big(\int_0^T[Q_tx_t.x_t + R_tu_t.u_t \\ &+ 2m_tx_t + 2n_tu_t + \tilde{\ell}_1(t,x_t,u_t)]dt + (Q_Tx_T.x_T + 2m_Tx_T)\Big)\Big\}. \end{aligned} \tag{3.36}$$

Further, suppose the following additional assumptions hold.

A10. The nonlinear drift term is the gradient of some potential function; that is,

$$g(t,x) = G_tG_t'D_x\phi(x), \qquad \phi \in C_x^2(\Re^n), \tag{3.37}$$

where $\phi(\cdot)$ has at most quadratic growth in the space variable.

A11. The nonlinear function $\tilde{\ell}_1(\cdot)$ is related to the nonlinear drift term by

$$\begin{aligned}\tilde{\ell}_1(t,x,u) &= \tfrac{1}{\theta}\left\{2(F_tx+f_t+B_tu).(G_tG_t')^{-1}g(t,x)\right.\\ &\qquad \left.+\, g(t,x).(G_tG_t')^{-1}g(t,x)+Tr(D_xg(t,x))\right\}\\ &= \tfrac{1}{\theta}\left\{|G_t^{-1}\left(F_tx+f_t+B_tu+g(t,x)\right)|^2\right.\\ &\qquad \left.-\,|G_t^{-1}\left(F_tx+f_t+B_tu\right)|^2+Tr\left(D_xg(t,x)\right)\right\}.\end{aligned} \tag{3.38}$$

A12. The density of random variable $x(0)$ is

$$q_0^\theta(x)=\exp(\phi(x))\times\tilde{q}_0^\theta(x),\qquad \tilde{q}_0^\theta(x)=\frac{\exp(-P_0^{-1}(x-\xi).(x-\xi))}{(2\pi)^{\frac{n}{2}}|P_0|^{\frac{1}{2}}}, \tag{3.39}$$

where $P_0=P_0'\geq 0$.

A13. $Q_t=Q_t'\geq 0,\quad R_t=R_t'>0\,\forall t\in[0,T]$.

Remark 3.1. Suppose $m_t=0, n_t=0, f_t=0\ \forall t\in[0,T]$. From A11 it is easily verified that when $Tr(D_xg(x,t))\geq 0\ \forall(t,x)\in[0,T]\times\Re^n$, there exist matrices $Q_t\geq 0, R_t>0, t\in[0,T]$, such that the integrand in the exponential of (3.36) is nonnegative $\forall(t,x,u)\in[0,T]\times\Re^n\times\mathcal{U}$.

For $u\in\hat{\mathcal{U}}$ the Feynman–Kac information state equation associated with system Σ_G^1 is

$$\begin{aligned}dq_t^\theta &= \tfrac{1}{2}Tr\left(G_tG_t'D_x^2q_t^\theta\right)dt-D_x.\left(q_t^\theta\left(F_tx+g_t(x)+f_t+B_tu\right)\right)dt\\ &\quad+\tfrac{\theta}{2}\left(Q_tx.x+R_tu.u+2m_tx+2n_tu+\tilde{\ell}_1(t,x,u)\right)q_t^\theta dt\\ &\quad+\left(H_tx+h_t\right).q_t^\theta N_t^{-1}dy_t,\qquad q^\theta(x,0)=q_0^\theta(x).\end{aligned} \tag{3.40}$$

We now show that (3.40) is, after a gauge transformation, equivalent to the Feynman–Kac information state of a LEQG tracking problem.

Introduce the gauge transformation

$$\tilde{q}_t^\theta=\exp(-\phi(x))q_t^\theta. \tag{3.41}$$

Using (3.40) and A10–A12 we derive the following stochastic PDE for $\tilde{q}^\theta(\cdot)$:

$$\begin{aligned}d\tilde{q}_t^\theta &= \tfrac{1}{2}Tr\left(G_tG_t'D_x^2\tilde{q}_t^\theta\right)dt-D_x.\left(\tilde{q}_t^\theta\left(F_tx+f_t+B_tu\right)\right)dt\\ &\quad+\tfrac{\theta}{2}\left(Q_tx.x+R_tu.u+2m_tx+2n_tu\right)\tilde{q}_t^\theta\\ &\quad+\left(H_tx+h_t\right).\tilde{q}_t^\theta N_t^{-1}dy_t,\qquad \tilde{q}^\theta(x,0)=\tilde{q}_0^\theta(x).\end{aligned} \tag{3.42}$$

Now, (3.42) is the Feynman–Kac information state equation corresponding to the LEQG tracking problem specified by (3.51)–(3.53) (see [3, 5]). Moreover, the solution of (3.42) corresponding to setting $\theta=0$, and denoted by $\tilde{q}^0(\cdot)$, is the solution of the DMZ equation corresponding to the LQG tracking problem. As the coefficients of (3.42) are continuous in t, we can infer that there exists a unique solution of (3.42) among the class of Gaussian density functions [5, 6]. Using the invertibility of the Gauge transformation (3.41) and the quadratic growth assumption on $\phi(\cdot)$, the existence and uniqueness results for $\tilde{q}^\theta(\cdot)$ translate directly into corresponding results for $q^\theta(\cdot)$ satisfying (3.40).

From the solution of (3.42) and the gauge transformation (3.41) we have the following lemma.

LEMMA 3.2. *Suppose there exists a* $0 < \theta \leq \theta^*$ *such that* $H_t' N_t^{-1} H_t - \theta Q_t \geq 0 \; \forall t \in [0,T]$. *The solution of* (3.40) *is given by*

$$q^\theta(x,t) = \exp(\phi(x)) \times \frac{\exp\left(-\frac{1}{2} P_t^{-1}(x-r_t).(x-r_t)\right)}{(2\pi)^{\frac{n}{2}} |P_t|^{\frac{1}{2}}} \times \exp(c_t + \lambda_t). \tag{3.43}$$

Here $P = P' : [0,T] \to \mathcal{L}(\Re^n; \Re^n)$ *is the solution of the Riccati equation*

$$\dot{P}_t = F_t P_t + P_t F_t' - P_t \left(H_t' N_t^{-1} H_t - \theta Q_t\right) P_t + G_t G_t', \quad P(0) = P_0, \tag{3.44}$$

and $r : [0,T] \times \Omega \to \Re^n$ *is the observer state satisfying*

$$\begin{aligned} dr_t = {} & (F_t + \theta P_t Q_t)\, r_t dt + f_t dt + B_t u(t,y) dt \\ & + \theta P_t m_t' dt + P_t H_t' N_t^{-1} (dy_t - H_t r_t dt - h_t dt), \quad r(0) = \xi. \end{aligned} \tag{3.45}$$

Moreover, $c : [0,T] \times \Omega \to \Re, \lambda : [0,T] \times \Omega \to \Re$ *are given by*

$$c_t = \int_0^t (H_s r_s + h_s).N_s^{-1} dy_s - \frac{1}{2} \int_0^t |N_s^{-\frac{1}{2}} (H_s r_s + h_s)|^2 ds, \tag{3.46}$$

$$\begin{aligned} \lambda_t = {} & \tfrac{\theta}{2} \textstyle\int_0^t [Q_s r_s.r_s + R_s u(s,y).u(s,y) + 2 m_s r_s \\ & + 2 n_s u(s,y) + Tr(P_s Q_s)]\, ds. \end{aligned} \tag{3.47}$$

Proof. This follows from the explicit solution of (3.42) (see, for example, [3, 5]) and the gauge transformation (3.41); it is also a special case of the results derived in Theorem 3.7. □

Rewriting the total cost (3.36) of system Σ_G^1 using (2.26) (by setting $\ell_2 = 0$) and then substituting (3.43), the resulting total cost is

$$\begin{aligned} J^\theta_{\Sigma_G^1}(u(\cdot)) &= E\left\{ \textstyle\int_{\Re^n} \varphi_2(T,x) \times \exp\left(\phi(z) + \frac{\theta}{2}(Q_T z.z + 2 m_T z)\right) \right. \\ &\qquad \left. \times \frac{1}{(2\pi)^{\frac{n}{2}} |P_T|^{\frac{1}{2}}} \exp\left(-\tfrac{1}{2} P_T^{-1}(z-r_T).(z-r_T)\right) dz \; \times \exp(c_T + \lambda_T) \right\} \\ &= \hat{E}^u \left\{ \textstyle\int_{\Re^n} \varphi_2(T,x) \times \exp\left(\phi(z) + \frac{\theta}{2}(Q_T z.z + 2 m_T z)\right) \right. \\ &\qquad \left. \times \frac{1}{(2\pi)^{\frac{n}{2}} |P_T|^{\frac{1}{2}}} \exp\left(-\tfrac{1}{2} P_T^{-1}(z-r_T).(z-r_T)\right) dz \; \times \exp(\lambda_T) \right\}. \end{aligned} \tag{3.48}$$

In the second equality the expectation is with respect to measure $\hat{P}^u$ and is established as follows. Define $\hat{\Lambda}_t^u \doteq \exp(c_t)$. Since $(Hr_t + h_t)$ is a Gaussian random variable for $0 < \theta \leq \theta^*$ we have $E[\hat{\Lambda}_t^u] = 1 \; \forall t \in [0,T]$. Therefore, for $u \in \hat{\mathcal{U}}$ we define a new measure $\hat{P}^u$ through the Radon–Nikodým derivative $\frac{d\hat{P}^u}{dP}|_{\mathcal{F}_T^y} \doteq \hat{\Lambda}_T^u$. (Note that this is different from that defined in section 2.1 because for each $t \in [0,T]$, $\hat{\Lambda}_t^u$ is now an $\mathcal{F}_t^y$-measurable random variable.) By Girsanov's theorem $\hat{P}^u$ is a probability measure on $(\Omega, \mathcal{A}; \mathcal{F}_t^y)$ and the second equality in (3.48) is established. If we define $d\hat{b}_t^u \doteq dy_t - (Hr_t + h_t)dt$, then $\hat{b}^u(\cdot)$ is a Wiener process corresponding to the innovations process with correlation $N(\cdot)$ on $(\Omega, \mathcal{A}, \hat{P}^u; \mathcal{F}_t^y)$. Consequently, the observer state $r(\cdot)$ satisfies, under measure $\hat{P}^u$, the following equation:

$$\begin{aligned} dr_t = {} & (F_t + \theta P_t Q_t)\, r_t dt + f_t dt + B_t u(t,y) dt \\ & + \theta P_t m_t' dt + P_t H_t' N_t^{-1} d\hat{b}_t^u, \quad r(0) = \xi, \quad d\hat{b}_t^u \doteq dy_t - (Hr_t + h_t)dt. \end{aligned} \tag{3.49}$$

Clearly, the optimal stochastic control problem associated with the family of control systems Σ_G^1 is equivalent to the standard, completely observable, stochastic optimal control problem of minimizing (3.48) subject to the observer state satisfying the linear stochastic differential equation (3.49).

We have thus established the following recipe for constructing partially observable stochastic optimal control problems equivalent to LEQG optimal control problems.

THEOREM 3.3. *Consider the family of control system Σ_G^1, and assume*

$$\varphi_2(T,x) = \exp(-\phi(x)). \tag{3.50}$$

Then the optimal control law corresponding to Σ_G^1 is precisely given by the optimal control law of the following LEQG tracking problem:

$$J_{EQ}^{\theta}(u^*(\cdot)) = \min_{u\in\hat{\mathcal{U}}} E^u \left\{ \exp\tfrac{\theta}{2}\left(\int_0^T [Q_t x_t.x_t + R_t u(t,y).u(t,y) + 2m_t x_t + 2n_t u(t,y)]\,dt + (Q_T x_T.x_T + 2m_T x_T)\right)\right\}, \tag{3.51}$$

subject to

$$dx_t = (F_t x_t + f_t)dt + B_t u(t,y)dt + G_t dw_t, \qquad x(0)\in\Re^n, \tag{3.52}$$

$$dy_t = (H_t x_t + h_t)dt + N_t^{\frac{1}{2}}\,db_t, \qquad y(0) = 0 \in \Re^d. \tag{3.53}$$

Here $x(0)$ is a Gaussian random variable with density $\tilde{q}_0^{\theta}$ given in A12.

Proof. The proof follows from the above construction. The solution of (3.51)–(3.53) is derived in [3] using the method of completing the squares, in [6], using dynamic programming, and in [5], using a maximum principle; it is also given in section 5.1. □

3.1.1. Nonlinear systems Σ_G^1 with sector-bounded nonlinearities. Consider now the family of nonlinear systems Σ_G^1 with $\varphi_2(T,x) = 1$. In this case the corresponding terminal cost of (3.48), which results from carrying out the integration against the space variable, is not an exponential-of-quadratic function of the observer state r, because $\phi(\cdot)$ is generally a nonquadratic function of x. Consequently, the optimal control minimizing (3.48) with $\varphi_2(T,x) = 1$ subject to (3.49) cannot be computed explicitly, as in Theorem 3.3. However, when such a situation arises one can derive suboptimal control laws for the important class of nonlinear drift terms known as sector-bounded nonlinearities (see [17, Chapter 5]). These are sometimes known as first and third quadrant nonlinearities. This family of control systems is defined by Σ_G^1 and the following assumption.

A14. Suppose $\varphi_2(T,x) = 1$, $\phi(x) = \sum_{j=1}^{\ell}\int_0^{y_j}\tilde{g}_j(\sigma)d\sigma$, $y_j = C_j x$, $C_j \in (\Re^n)'$, $1\le j\le \ell$, where the $\{\tilde{g}\}_{j=1}^{\ell}$ satisfy the "sector criterion"

$$k_j^-|y_j|^2 \le \tilde{g}_j(C_j x)C_j x \le k_j^+|y_j|^2, \quad 0\le k_j^-\le k_j^+,\ 1\le j\le \ell. \tag{3.54}$$

(Condition (3.54) ensures that the graph of $y_j \to \tilde{g}_j(y_j)$ lies in the first and third quadrants.)

Note that for the class of systems Σ_G^1 which satisfy A14, the martingale problem is well posed, and therefore (3.34) has a unique weak solution (see [18, Theorem 10.2.2, p. 255]). Moreover, by A10 and A14 we have

$$g(t,x) = \sum_{j=1}^{\ell} G_t G_t' C_j' \tilde{g}_j(y_j), \qquad \frac{1}{2}\sum_{j=1}^{\ell} k_j^-|y_j|^2 \le \phi(x) \le \frac{1}{2}\sum_{j=1}^{\ell} k_j^+|y_j|^2. \tag{3.55}$$

Substituting (3.55) into (3.48), the upper and lower bounds of $\phi(\cdot)$ translate into the following upper and lower bounds for $J^{\theta}_{\Sigma^1_G}(u(\cdot))$, respectively:

$$J^{\theta,-}_{\Sigma^1_G}(u^-(\cdot)) \leq J^{\theta}_{\Sigma^1_G}(u(\cdot)) \leq J^{\theta,+}_{\Sigma^1_G}(u^+(\cdot)), \tag{3.56}$$

where

$$\begin{aligned} J^{\theta,-}_{\Sigma^1_G}(u^-(\cdot)) = \hat{E}^u \Bigg\{ & \int_{\Re^n} \exp\left(\tfrac{\theta}{2}\left[Q^{\theta,-}_T z.z + 2m_T z\right]\right) \\ & \times \tfrac{1}{(2\pi)^{\frac{n}{2}}|P_T|^{\frac{1}{2}}} \exp\left(-\tfrac{1}{2}|P_T^{-\frac{1}{2}}(z-r_T)|^2\right) dz \times \exp\left(\lambda_T\right) \Bigg\}, \end{aligned} \tag{3.57}$$

$$\begin{aligned} J^{\theta,+}_{\Sigma^1_G}(u^+(\cdot)) = \hat{E}^u \Bigg\{ & \int_{\Re^n} \exp\left(\tfrac{\theta}{2}\left[Q^{\theta,+}_T + 2m_T z\right]\right) \\ & \times \tfrac{1}{(2\pi)^{\frac{n}{2}}|P_T|^{\frac{1}{2}}} \exp\left(-\tfrac{1}{2}|P_T^{-\frac{1}{2}}(z-r_T)|^2\right) dz \times \exp\left(\lambda_T\right) \Bigg\}. \end{aligned} \tag{3.58}$$

Also

$$Q^{\theta,-}_T \doteq Q_T + \frac{1}{\theta}\sum_{j=1}^{\ell} k_j^- C_j' C_j \quad \text{and} \quad Q^{\theta,+}_T \doteq Q_T + \frac{1}{\theta}\sum_{j=1}^{\ell} k_j^+ C_j' C_j,$$

where $P(\cdot), r(\cdot)$ are solutions of (3.44), (3.49), respectively, and $\lambda(\cdot)$ is given by (3.47). Moreover, the optimal control laws $u^{-,*}(\cdot), u^{+,*}(\cdot)$ resulting from minimizing (3.57), (3.58), respectively, subject to (3.49), are linear feedback, reminiscent of the optimal control law of the LEQG tracking problem of Theorem 3.3. Hence, the optimal cost of the family of control systems Σ^1_G satisfying A14 is bounded from above and from below by the optimal cost of an LEQG tracking problem.

3.1.2. Nonlinear systems Σ^1_G with polynomial nonlinearities. We now wish to relax the linear growth assumption on the nonlinear drift term $g(\cdot)$ associated with the family of nonlinear control systems Σ^1_G, to consider the situation when the nonlinear drift terms are polynomial functions of the unobservable state as described by the next assumption.

A15. $n = d = m = 1, G_t = 1\ \forall t \in [0,T], k_1 = 2p$, and the nonlinear drift term $g(\cdot)$ is a polynomial of odd degree and stable; that is,

$$g(t,x) = D_x\phi(x) = \sum_{j=1}^{2p-1} F_j x^j, \qquad F_{2p-1} < 0, \quad p \geq 2. \tag{3.59}$$

Notice that the growth condition (2.17) is now specified by substituting (3.59) into (3.38). Conditions on existence and uniqueness of solutions for the robust version of the DMZ equation, (2.33), having strongly unbounded coefficients (i.e., greater than polynomial growth in x), are derived in [19], and include as a special case the situation when the drift, diffusion, and signal terms in (1.1), (1.2) have polynomial growth (see [19, section III, Example 1, pp. 207–210]). Following [19], we first note that the class of state processes $x(\cdot)$ associated with system Σ^1_G and satisfying A15 does not explode in finite time because they satisfy Khas'minskii's test of nonexplosion (see [18]). Second, from [19, section III] we deduce that for each $y(\cdot,\omega) \in C([0,T];\Re^d)$ and for some $\theta > 0$, the robust version of the Feynman–Kac equation of systems Σ^1_G satisfying A15 has a unique positive solution among those having exponential growth

$\alpha_1 \exp(-\alpha_2|x|^{2p}), \alpha_1 > 0, \alpha_2 > 0$, because the dominant part of $q^\theta(x,0)$, as $|x| \to \infty$, in the exponent is $F_{2p-1}\frac{|x|^{2p}}{2p}$. Alternatively, we can establish existence and uniqueness of solutions of the Feynman–Kac information state equation corresponding to system Σ_G^1 and A15, by invoking the gauge transformation (3.41) given by

$$q_t^\theta = \exp\left(\sum_{j=1}^{2p-1}\int_0^x F_j\sigma^j\, d\sigma\right)\tilde{q}_t^\theta, \tag{3.60}$$

where $\tilde{q}^\theta(\cdot)$ is the unique Gaussian density function satisfying (3.42), and $q^\theta(\cdot)$ is the information state of system Σ_G^1 and A15. As $|x| \to \infty$, its dominant part in the exponent is $F_{2p-1}\frac{|x|^{2p}}{2p}$. Thus, we establish for each $y(\cdot,\omega) \in C([0,T];\Re^d)$ the existence and uniqueness results stated above.

THEOREM 3.4. *Suppose there exists a* $0 < \theta \le \theta^*$ *such that* $H_tN_tH_t' - \theta Q_t \ge 0\ \forall t \in [0,T]$. *The Feynman–Kac information state corresponding to system* Σ_G^1 *and A15 is*

$$\begin{aligned} q^\theta(x,t) &= \exp\left(\textstyle\sum_{j=1}^{2p-1}\int_0^x F_j\sigma^j\, d\sigma\right) \\ &\quad \times \tfrac{1}{(2\pi)^{\frac{1}{2}}|P_t|^{\frac{1}{2}}}\exp\left(-\tfrac{1}{2}|P_t^{-\frac{1}{2}}(x-r_t)|^2\right) \times \exp\left(c_t+\lambda_t\right), \end{aligned} \tag{3.61}$$

where P, r, c, λ *are given in Lemma* 3.2.

Moreover, suppose

$$\varphi_2(t,x) = \exp\left(-\sum_{j=1}^{2p-1}\int_0^x F_j\sigma^j\, d\sigma\right). \tag{3.62}$$

Then the optimal control law corresponding to system Σ_G^1 *and* A15 *is given by the optimal control law of the LEQG tracking problem* (3.51)–(3.53).

Proof. This follows from the above construction, Lemma 3.2, and Theorem 3.3. □

Example 1.2 presented in the introduction is an application of Theorem 3.4.

Remark 3.5. In many practical applications one is usually interested in minimizing exponential-of-quadratic integral cost functions of the form (3.51). Such cost functions can be incorporated into our earlier framework by requiring $\tilde{\ell}_1(\cdot)$ to be quadratic in x, u. For example, suppose

$$\tilde{\ell}_1(t,x,u) = \frac{1}{\theta}\left(\tilde{V}(t,x) + \tilde{R}_tu.u + 2\tilde{n}u\right), \quad \tilde{V}(t,x) \doteq \Lambda_tx.x + 2x.\sigma_t + \delta_t, \tag{3.63}$$

where $\Lambda = \Lambda' : [0,T] \to \mathcal{L}(\Re^n;\Re^n), \sigma : [0,T] \to \Re^n, \delta : [0,T] \to \Re, \tilde{R} = \tilde{R}' : [0,T] \to \mathcal{L}(\Re^m;\Re^m), \tilde{n} : [0,T] \to (\Re^m)'$. In this case, the potential functions $\phi(\cdot)$ related to $g(\cdot)$ by A10 should be smooth classical solutions of the PDE (obtained from (3.38)), namely,

$$\begin{aligned} &\tfrac{1}{2}Tr\left(G_tG_t'D_x^2\phi(x)\right) + \tfrac{1}{2}D_x\phi(x).G_tG_t'D_x\phi(x) + \left(F_tx + f_t + B_tu\right).D_x\phi(x) \\ &\qquad = \tfrac{1}{2}\left(\tilde{V}(t,x) + \tilde{R}_tu.u + 2\tilde{n}_tu\right). \end{aligned} \tag{3.64}$$

If (3.63) takes the form $\tilde{\ell}_1(t,x,u) = \frac{2}{\theta}\left(F_tx + f_t + B_tu\right).(G_tG_t')^{-1}g(x) + \frac{1}{\theta}\tilde{V}(t,x)$ and we set $\tilde{R} = 0, \tilde{n} = 0$, then (3.64) is reduced to the well-known Riccati equation

$$Tr\left(D_xg(x)\right) + |G_t^{-1}g(x)|^2 = \tilde{V}(t,x),$$

first introduced in [20] for evaluating Feynman–Kac-type Wiener integrals, and in [11] for identifying finite-dimensional nonlinear filtering examples.

3.2. Nonlinear dynamics quadratic sensor problem. Now consider the generalized family of nonlinear control systems described by allowing the observations to be quadratic functions of the unobservable state.

Control system (Σ_G^2). Suppose Assumptions 2.1 hold with $l_1 = 4, k_1 = 2$, and $\mathcal{U}$ in A1 replaced by a compact subset of $\Re^m$. A4 is replaced by a quadratic growth condition, and the dynamics, observations, and cost criterion are given by

$$dx_t = (F_t x_t + g_t(x) + f_t)\,dt + B(t, u(t,y))dt + G_t dw_t, \qquad x(0) \in \Re^n, \tag{3.65}$$

$$dy_t = \left(\frac{1}{2} x_t' \tilde{H}_t x_t + H_t x_t + h_t\right) dt + N_t^{\frac{1}{2}} db_t, \qquad y(0) = 0 \in \Re, \tag{3.66}$$

$$J^{\theta}_{\Sigma_G^2} \doteq J^{\theta}_G(u(\cdot)) = (1.3). \tag{3.67}$$

Furthermore, suppose the following additional assumptions hold.

A16. $g(t,x) = G_t G_t D_x \phi(x,t)$, $\phi(\cdot) \in C^{2,1}_{x,t}(\Re^n \times [0,T])$, $\phi(\cdot)$ has at most quadratic growth in the space variable uniformly in t, $Fx + g(t,x)$ is stable, and the initial density of $x(0)$ is $q^{\theta}(x,0) = \exp(\phi(x,t)) \times \tilde{q}^{\theta}(x,0)$, where $\tilde{q}^{\theta}(\cdot)$ is a Gaussian density function (see A12).

A17. $2\ell_1(t,x,u) = \tilde{Q}(t,u)x.x + \tilde{R}(t,u)u.u + 2\tilde{m}(t,u)x + 2\tilde{n}(t,u)u + \tilde{\ell}_1(t,x,u)$.

A18. $\tilde{Q} : [0,T] \times \mathcal{U} \to \mathcal{L}(\Re^n; \Re^n), \tilde{R} : [0,T] \times \mathcal{U} \to \mathcal{L}(\Re^m; \Re^m), \tilde{m} : [0,T] \times \mathcal{U} \to (\Re^n)', \tilde{n} : [0,T] \times \mathcal{U} \to (\Re^m)', \tilde{\ell}_1 : [0,T] \times \Re^n \times \mathcal{U} \to \Re, \tilde{Q} = \tilde{Q}' \geq 0, \tilde{R} = \tilde{R}' > 0$.

Whenever $\tilde{H} = 0$ we assume $y : [0,T] \times \Omega \to \Re^d$.

A derivation of the sufficient statistics associated with system Σ_G^2, when $g = 0$, is given in [9], using the Fisk–Stratonovich version of the Feynman–Kac information state equation.

First, we point out that by the stability of $Fx + g(t,x)$ and the quadratic growth of $h(\cdot)$ (as in [19], at least for $n = 1$) we deduce that for each $y \in C([0,T]; \Re)$, and for some $\theta > 0$, there exists a unique solution of the robust version of the Feynman–Kac information state equation of system Σ_G^2, among the class of functions bounded above by $\alpha_1 \exp(-\alpha_2 |x|^2), \alpha_1 > 0, \alpha_2 > 0$.

Remark 3.6. Notice that for $\tilde{H} \neq 0$, the observation process associated with system Σ_G^2 is defined by $y : [0,T] \times \Omega \to \Re$ (i.e., is a real-valued function), although we believe that one can generalize the results of this section to multidimensional observations of the form

$$dy_t = \left(\frac{1}{2}\tilde{H}_t(x)x + H_t x_t + h_t\right) dt + N_t^{\frac{1}{2}} db_t,$$

where

$$\tilde{H}_t(x) = \sum_{i=1}^n x_i H_t^i, H^i : [0,T] \to \mathcal{L}(\Re^n; \Re^d), H : [0,T] \to \mathcal{L}(\Re^n; \Re^d), h : [0,T] \to \Re^d.$$

For $u \in \hat{\mathcal{U}}$ the Feynman–Kac information state equation associated with control system Σ_G^2 is given by

$$\begin{aligned} dq_t^{\theta} &= \tfrac{1}{2} Tr\left(G_t G_t' D_x^2 q_t^{\theta}\right) dt - \tfrac{\partial}{\partial x}\left(q_t^{\theta}\left(F_t x + g_t(x) + f_t + B(t,u)\right)\right) dt \\ &\quad + \tfrac{\theta}{2}\left(\tilde{Q}_t(u)x.x + \tilde{R}_t(u)u.u + 2\tilde{m}_t(u)x + 2\tilde{n}_t(u)u + \tilde{\ell}_1(t,x,u)\right) q_t^{\theta} dt \\ &\quad + q_t^{\theta}\left(\tfrac{1}{2} x' \tilde{H}_t x + H_t x + h_t\right) . N_t^{-1} dy_t \doteq RHS, \qquad q^{\theta}(x,0) = q_0^{\theta}(x). \end{aligned} \tag{3.68}$$

We derive solutions of (3.68) through an alternative technique, by seeking a solution of the form

$$q^\theta(x,t) = \exp\left(\phi(x,t) - \frac{1}{2}\tilde{P}_t x.x + \tilde{r}_t.x + \tilde{\rho}_t\right). \tag{3.69}$$

Here we suppose

$$\tilde{P} : [0,T]\times\Omega \to \mathcal{L}(\Re^n;\Re^n), \quad \tilde{P} = \tilde{P}', \quad \tilde{r} : [0,T]\times\Omega \to \Re^n, \quad \tilde{\rho} : [0,T]\times\Omega \to \Re,$$

and the random processes $\tilde{P}(\cdot), \tilde{r}(\cdot), \tilde{\rho}(\cdot)$ satisfy the following stochastic differential equations:

$$d\tilde{r}_t = k_t dt + H_t' N_t^{-1} dy_t, \quad k : [0,T]\times\Omega \to \Re^n, \tag{3.70}$$

$$d\tilde{P}_t = Z_t dt - \tilde{H}_t' N_t^{-1} dy_t, \quad Z : [0,T]\times\Omega \to \mathcal{L}(\Re^n;\Re^n), \tag{3.71}$$

$$d\tilde{\rho}_t = \mu_t dt + h_t' N_t^{-1} dy_t, \quad \mu : [0,T]\times\Omega \to \Re. \tag{3.72}$$

From (3.69) we have

$$D_x q_t^\theta = \left(D_x\phi_t - \tilde{P}_t x + \tilde{r}_t\right) q_t^\theta, \quad Tr\left(D_x^2 q_t^\theta\right) = |D_x\phi_t - \tilde{P}_t x + \tilde{r}|^2 q_t^\theta + Tr\left(D_x^2\phi_t - \tilde{P}_t\right) q_t^\theta.$$

Since $q^\theta(x,t) \equiv q^\theta(x,\tilde{r},\tilde{P},\tilde{\rho},t)$, an application of the Ito differential rule yields

$$\begin{aligned} dq_t^\theta = \Big\{ & \tfrac{\partial\phi_t}{\partial t}dt + x.\left(k_t dt + H_t.N_t^{-1}dy_t\right) + \left(\mu_t dt + h_t.N_t^{-1}dy_t\right) \\ & - \tfrac{1}{2}x.\left(Z_t dt - \tilde{H}_t.N_t^{-1}dy_t\right)x + \tfrac{1}{2}\left(H_t x.N_t^{-1}H_t x + h_t.N_t^{-1}h_t\right. \\ & \left. + \tfrac{1}{4}\tilde{H}_t x.xN_t^{-1}x'\tilde{H}_t x + \ 2H_t x.N_t^{-1}h_t + H_t x.N_t^{-1}x'\tilde{H}_t x + x'\tilde{H}_t x.N_t^{-1}h_t\right) dt\Big\} q_t^\theta. \end{aligned}$$

From (3.68) we have

$$\begin{aligned} RHS = \Big\{ & \Big[\tfrac{1}{2}|G_t'(D_x\phi_t - \tilde{P}_t x + \tilde{r}_t)|^2 \\ & + \tfrac{1}{2}\left(G_t G_t'\left(D_x^2\phi_t - \tilde{P}_t\right)\right) - B(t,u).\left(D_x\phi_t - \tilde{P}_t x + \tilde{r}_t\right) \\ & - Tr(F_t) - Tr(D_x g_t(x)) - (g_t(x) + F_t x + f_t).\left(D_x\phi_t - \tilde{P}x + \tilde{r}_t\right) \\ & + \tfrac{\theta}{2}\left(\tilde{Q}_t(u)x.x + \tilde{R}_t(u)u.u + 2\tilde{m}_t(u)x + 2\tilde{n}_t(u)u + \tilde{\ell}_1(t,x,u)\right) \\ & - \ B(t,u).\left(D_x\phi_t - \tilde{P}_t x + \tilde{r}_t\right)\Big] dt + \left(\tfrac{1}{2}x'\tilde{H}_t x + H_t x + h_t\right).N_t^{-1}dy_t\Big\} q_t^\theta. \end{aligned}$$

Equating dq_t^θ to RHS, the stochastic integral terms cancel; therefore, we deduce the following equation for $\phi(\cdot)$:

$$\begin{aligned} & \tfrac{\partial\phi_t}{\partial t} - \tfrac{1}{2}x.Z_t x + x.k_t + \mu_t + \tfrac{1}{2}|N_t^{-\frac{1}{2}}\left(\tfrac{1}{2}x.\tilde{H}_t x + Hx + h_t\right)|^2 \\ & - \tfrac{1}{2}|G_t'(D_x\phi_t - \tilde{P}_t x + \tilde{r}_t)|^2 - \tfrac{Tr}{2}\left(G_t G_t'\left(D_x^2\phi_t - \tilde{P}_t\right)\right) \\ & + (g_t(x) + F_t x + f_t).\left(D_x\phi_t - \tilde{P}_t x + \tilde{r}_t\right) + Tr(D_x g_t(x)) + Tr(F_t) \\ & - \tfrac{\theta}{2}\left(\tilde{Q}_t(u)x.x + \tilde{R}_t(u)u.u + 2\tilde{m}_t(u)x + 2\tilde{n}_t(u)u\right) \\ & - \tfrac{\theta}{2}\tilde{\ell}_1(t,x,u) + B(t,u).\left(D_x\phi_t - \tilde{P}_t x + \tilde{r}_t\right) = 0. \end{aligned}$$

Now isolate the coefficients of powers of x^2, x^1, x^0, respectively, as follows:

$$x.\left(-\tfrac{1}{2}Z_t - \tfrac{1}{2}\tilde{P}_t G_t G_t' \tilde{P}_t - F_t'\tilde{P}_t - \tfrac{\theta}{2}\tilde{Q}_t(u)\right)x,$$

$$\begin{aligned} &x.\Big(k_t + \tilde{P}_t G_t G_t' D_x\phi_t + \tilde{P}_t G_t G_t'\tilde{r}_t - \tilde{P}_t g_t(x) \\ &\quad + F_t'\tilde{r}_t - \tilde{P}_t f_t - \tilde{P}_t B(t,u) - \theta\tilde{m}_t'(u)\Big), \end{aligned} \tag{3.73}$$

$$\begin{aligned} &\mu_t - \tfrac{1}{2}\tilde{r}_t' G_t G_t' - D_x\phi_t.G_t G_t'\tilde{r}_t + \tfrac{1}{2}Tr\left(G_t G_t'\tilde{P}_t\right) + g_t(x).\tilde{r}_t \\ &\quad + f_t.\tilde{r}_t - \tfrac{\theta}{2}\tilde{R}_t(u)u.u - \theta\tilde{n}_t(u)u + B(t,u).\tilde{r}_t. \end{aligned}$$

Since $g(t,x) = G_t G_t' D_x\phi_t$, we have $D_x(g_t(x)) = G_t G_t' D_x^2\phi_t$, $g_t(x).D_x\phi_t = G_t G_t' \times D_x\phi_t.D_x\ \phi_t$. Therefore, the coefficients of powers of x^2, x^1, x^0 are independent of $g(\cdot), D_x\phi(\cdot)$.

Now introduce the matrix-valued, vector-valued, and scalar-valued functions $\gamma(\cdot)$, $\alpha(\cdot), \beta(\cdot)$ as follows:

$$\gamma_t \doteq -\frac{1}{2}Z_t - \frac{1}{2}\tilde{P}_t G_t G_t'\tilde{P}_t - F_t'\tilde{P}_t - \frac{\theta}{2}\tilde{Q}_t(u), \tag{3.74}$$

$$\alpha_t \doteq k_t + \tilde{P}_t G_t G_t'\tilde{r}_t + F_t'\tilde{r}_t - \theta\tilde{m}_t'(u) - \tilde{P}_t B(t,u) - \tilde{P}_t f_t, \tag{3.75}$$

$$\begin{aligned} \beta_t \doteq{}& \mu_t - \tfrac{1}{2}\tilde{r}_t' G_t G_t'\tilde{r}_t + \tfrac{1}{2}Tr\left(G_t G_t'\tilde{P}_t\right) - \tfrac{\theta}{2}\tilde{R}_t(u)u.u - \theta\tilde{n}_t(u)u \\ &+ B(t,u).\tilde{r}_t + f_t.\tilde{r}_t. \end{aligned} \tag{3.76}$$

Using (3.74)–(3.76) in the equation for $\phi(\cdot)$ we see

$$\begin{aligned} &\tfrac{\partial\phi_t}{\partial t} + \tfrac{1}{2}Tr\left(G_t G_t' D_x^2\phi_t\right) + \tfrac{1}{2}D_x\phi_t.G_t G_t' D_x\phi_t + F_t x.D_x\phi_t + \tfrac{1}{2}|N_t^{-\frac{1}{2}}(\tfrac{1}{2}x'\tilde{H}_t x + Hx + h_t)|^2 \\ &\quad = -Tr(F_t) + x.\left(-\alpha_t\right) + \tfrac{1}{2}x.\left(-2\gamma_t\right)x - \beta_t + \tfrac{\theta}{2}\tilde{\ell}_1(t,x,u) - B(t,u).D_x\phi_t. \end{aligned}$$

Rearranging the terms of this equation and setting

$$\Lambda_t = -2\gamma_t - H_t' N_t^{-1} H_t - \tilde{H}_t' N_t^{-1} h_t, \quad \Lambda_t = \Lambda_t', \tag{3.77}$$

$$\sigma_t = -\alpha_t - H_t' N_t^{-1} h_t, \tag{3.78}$$

$$\delta_t = -\beta_t - \frac{1}{2}h_t' N_t^{-1} h_t - Tr(F_t), \tag{3.79}$$

we have

$$\begin{aligned} &\tfrac{\partial\phi_t}{\partial t} + \tfrac{Tr}{2}\left(G_t G_t' D_x^2\phi_t\right) + \tfrac{1}{2}D_x\phi_t.G_t G_t' D_x\phi_t + \left(F_t x + f_t\right).D_x\phi_t = \tfrac{1}{2}x.\Lambda_t x + x.\sigma_t + \delta_t \\ &\quad + \tfrac{\theta}{2}\tilde{\ell}_1(t,x,u) - B(t,u).D_x\phi_t - \tfrac{1}{2}|N_t^{-\frac{1}{2}}\tfrac{1}{2}x'\tilde{H}_t x|^2 - \tfrac{1}{2}x'\tilde{H}_t x N_t^{-1} H_t x. \end{aligned}$$

Using (3.74)–(3.76) and (3.77)–(3.79), the equations satisfied by the functions $\tilde{P}(\cdot)$, $\tilde{r}(\cdot), \tilde{\rho}(\cdot)$ are given by (see (3.70)–(3.72))

$$
\begin{aligned}
d\tilde{P}_t + \left(\tilde{P}_t F_t + F_t'\tilde{P}_t + |G_t\tilde{P}_t|^2\right)dt &= \left(\Lambda_t + H_t' N_t^{-1} H_t\right)dt \\
&\quad + \tilde{H}_t' N_t^{-1} h_t dt - \theta\tilde{Q}_t(u)dt - \tilde{H}_t' N_t^{-1} dy_t, \\
d\tilde{r}_t + \left(F_t' + \tilde{P}_t G_t G_t'\right)\tilde{r}_t dt - \tilde{P}_t f_t dt + \sigma_t dt - \tilde{P}_t B(t,u)dt + H_t' N_t^{-1} h_t dt & \\
-\theta\tilde{m}_t'(u)dt = H_t' N_t^{-1} dy_t, & \\
d\tilde{\rho}_t + \left(-\tfrac{1}{2}|G_t'\tilde{r}_t|^2 + \delta_t + \tfrac{1}{2}|N_t^{-\frac{1}{2}} h_t|^2 + \tfrac{1}{2}Tr\left(G_t G_t'\tilde{P}_t + 2F_t\right)\right)dt + B(t,u).\tilde{r}_t dt & \\
+ f_t.\tilde{r}_t - \tfrac{\theta}{2}\left(\tilde{r}_t.\tilde{Q}_t(u)\tilde{r}_t + \tilde{R}_t(u)u.u + 2\tilde{n}_t(u)u\right)dt = h_t' N_t^{-1} dy_t. &
\end{aligned}
$$

Hence, we obtain the following theorem.

THEOREM 3.7. *Consider system Σ_G^2 and suppose for $u \in \hat{\mathcal{U}}$ there exist functions $\phi \in C_{x,t}^{2,1}(\Re^n \times [0,T])$, which are independent of the paths of y, which satisfy the PDE*

$$
\begin{aligned}
&\tfrac{\partial\phi_t}{\partial t} + \tfrac{1}{2}Tr\left(G_t G_t' D_x^2\phi_t\right) + \tfrac{1}{2}D_x\phi_t.G_t G_t' D_x\phi + \left(F_t x + f_t\right).D_x\phi_t = \tfrac{1}{2}x.\Lambda_t x + x.\sigma_t + \delta_t \\
&\quad + \left(\tfrac{\theta}{2}\tilde{\ell}_1(t,x,u) - B(t,u).D_x\phi_t - \tfrac{1}{2}|\tfrac{1}{2}N_t^{-\frac{1}{2}} x'\tilde{H}_t x|^2 - \tfrac{1}{2}x'\tilde{H}x.N_t^{-1}H_t x\right).
\end{aligned}
\tag{3.80}
$$

Here the function $\tilde{\ell}_1(\cdot)$ is free to be chosen so that (3.80) *yields explicit solutions. Then, a Feynman–Kac information state corresponding to system Σ_G^2 and satisfying* (3.68) *is given by*

$$
q^\theta(x,t) = \exp\left(\phi(x,t) + \tilde{c}_t + \tilde{\lambda}_t + \lambda_t\right) \times \frac{\exp\left(-\tfrac{1}{2}P_t^{-1}(x - r_t).(x - r_t)\right)}{(2\pi)^{\frac{n}{2}}|P_t|^{\frac{1}{2}}},
\tag{3.81}
$$

where

$$
r : [0,T]\times\Omega \to \Re^n, \quad P = P' : [0,T]\times\Omega \to \mathcal{L}(\Re^n;\Re^n), \quad \tilde{c}, \tilde{\lambda}, \lambda : [0,T]\times\Omega \to \Re
$$

are given by the following equations:

$$
\begin{aligned}
dr_t &= \left\{F_t - P_t\left(\tilde{H}_t' N_t^{-1} h_t - \theta\tilde{Q}_t(u) + \Lambda_t\right)\right\} r_t dt + \left(f_t - P_t\sigma_t + B(t,u)\right)dt \\
&\quad + \left(P_t\tilde{H}_t' N_t^{-1} P_t H_t' + \theta P_t\tilde{m}_t'(u)\right)dt + P_t H_t' N_t^{-1}\left(dy_t - H_t r_t dt - h_t dt\right) \\
&\quad + P_t\tilde{H}_t' N_t^{-1}\left(r_t dy_t - P_t\tilde{H}_t r_t dt\right), \quad r(0) = \xi,
\end{aligned}
\tag{3.82}
$$

$$
\begin{aligned}
dP_t &= \left\{F_t P_t + P_t F_t' - P_t\left(H_t' N_t^{-1} H_t + \tilde{H}_t' N_t^{-1} h_t + \Lambda_t - \theta\tilde{Q}_t(u)\right)P_t\right\}dt \\
&\quad + \left(P_t\tilde{H}_t' P_t N_t^{-1}\tilde{H}_t P_t + G_t G_t'\right)dt + P_t\tilde{H}_t' N_t^{-1} P_t dy_t, \qquad P(0) = P_0,
\end{aligned}
\tag{3.83}
$$

$$
\begin{aligned}
\lambda_t &= \tfrac{\theta}{2}\int_0^t \left\{[\tilde{Q}_s(u) - \tfrac{\Lambda_s}{\theta}]r_s.r_s + Tr\left(P_s[\tilde{Q}_s(u) - \tfrac{\Lambda_s}{\theta}]\right)\right\}ds \\
&\quad + \tfrac{\theta}{2}\int_0^t \left(\tilde{R}_s(u)u_s.u_s + 2r_s.[\tilde{m}_s(u)' - \tfrac{\sigma_s}{\theta}] + 2[\tilde{n}_s(u)u_s - \tfrac{\delta_s}{\theta}]\right)ds,
\end{aligned}
\tag{3.84}
$$

$$
\begin{aligned}
\tilde{\lambda}_t &= \tfrac{1}{2}\int_0^t \left\{|\tfrac{1}{2}N_s^{-\frac{1}{2}} r_s'\tilde{H}_s r_s|^2 + r_s'\tilde{H}_s r_s.N_s^{-1}H_s r_s + r_s.(-3\tilde{H}_s' N_s^{-1} P_s\tilde{H}_s)r_s\right. \\
&\quad \left. + r_s.\tilde{H}_s' P_s N_s^{-1} H_s'\right\}ds + \tfrac{1}{2}\int_0^t Tr\left\{-P_s\tilde{H}_s' N_s^{-1} h_s + P_s\tilde{H}_s' N_s^{-1} P_s\tilde{H}_s\right. \\
&\quad \left. - \tfrac{1}{2}\tilde{H}_s' N_s^{-1}\tilde{H}_s\right\}ds + \tfrac{1}{2}\int_0^t Tr\left(P_s\tilde{H}_s' N_s^{-1}\right)dy_s, \\
\tilde{c}_t &= \int_0^t \left(H_s r_s + h_s + \tfrac{1}{2}r_s'\tilde{H}_s r_s\right).N_s^{-1} dy_s \\
&\quad - \tfrac{1}{2}\int_0^t |N_s^{-\frac{1}{2}}\left(H_s r_s + h_s + \tfrac{1}{2}r_s'\tilde{H}_s r_s\right)|^2 ds.
\end{aligned}
$$

The above results will be valid whenever there exists a $0 < \theta \leq \theta^*$ *such that*

$$H_t' N_t^{-1} H_t + \tilde{H}_t' N_t^{-1} h_t + \Lambda_t - \theta \tilde{Q}_t(u) \geq 0 \quad \forall (t,u) \in [0,T] \times \mathcal{U}.$$

Proof. Define the functions $P(\cdot), r(\cdot), \rho(\cdot)$ by

$$P_t \doteq \tilde{P}_t^{-1}, \quad r_t \doteq \tilde{P}_t^{-1} \tilde{r}_t, \quad \rho_t = -2\tilde{\rho}_t,$$

and seek a representation of $q^\theta(\cdot)$ in the form

$$q^\theta(x,t) = \exp\left(\phi(x,t) - \frac{1}{2} P_t^{-1}(x - r_t).(x - r_t) + \tilde{\mu}_t \right), \qquad \tilde{\mu}_t = \frac{1}{2} P_t^{-1} r_t . r_t - \frac{1}{2} \rho_t.$$

To this end define

$$\begin{aligned} \tilde{c}_t \doteq & \int_0^t \left(H_s r_s + h_s + \tfrac{1}{2} r_s' \tilde{H}_s r_s \right) . N_s^{-1} dy_s \\ & - \tfrac{1}{2} \int_0^t | N_s^{-\frac{1}{2}} \left(H_s r_s + h_s + \tfrac{1}{2} r_s' \tilde{H}_s r_s \right) |^2 ds. \end{aligned} \tag{3.85}$$

An application of the Ito differential rule yields (3.82), (3.83), and

$$\begin{aligned} d\tilde{\mu}_t = & \, d\tilde{c}_t + \tfrac{1}{2} \left(\theta \tilde{Q}_t(u) - \Lambda_t \right) r_t . r_t dt \\ & + \left(-r_t . \sigma_t - \delta_t + \tfrac{1}{2} | N_t^{-\frac{1}{2}} r_t' \tilde{H}_t r_t |^2 + \tfrac{1}{2} r_t' \tilde{H}_t r_t . N_t^{-1} H_t r_t \right) dt \\ & + \Big(-\tfrac{3}{2} r_t' \tilde{H}_t' N_t^{-1} P_t \tilde{H}_t r_t + \tfrac{1}{2} \tilde{H}_t r_t . P_t N_t^{-1} H_t' + \tfrac{1}{2} H_t P_t N_t^{-1} H_t' \\ & \quad - \tfrac{1}{2} Tr \left(2F_t + G_t' P_t^{-1} G_t \right) \Big) dt + \tfrac{\theta}{2} \left(\tilde{R}_t(u) u . u + 2 \tilde{m}_t(u) r_t + 2 \tilde{n}_t(u) u \right) dt, \\ \tilde{\mu}(0) = & \, \tfrac{1}{2} P(0)^{-1} r(0) . r(0) - \tfrac{1}{2} \rho(0). \end{aligned}$$

Writing (3.83) in the form

$$\begin{aligned} dP_t = & \left\{ F_t + P_t F_t' P_t^{-1} + P_t \tilde{H}_t' N_t^{-1} P_t \tilde{H}_t + G_t G_t' P_t^{-1} \right. \\ & \left. - P_t \left(H_t' N_t^{-1} H_t + \tilde{H}_t' N_t^{-1} h_t + \Lambda_t - \theta \tilde{Q}_t(u) \right) \right\} P_t dt + \left(P_t \tilde{H}_t' N_t^{-1} \right) P_t dy_t, \end{aligned}$$

we deduce

$$\begin{aligned} d(\log |P_t| = & \, Tr \left(F_t + P_t F_t' P_t^{-1} + P_t \tilde{H}_t' N_t^{-1} P_t \tilde{H}_t + G_t G_t' P_t^{-1} \right) dt + Tr \left(P_t \tilde{H}_t N_t^{-1} \right) dy_t \\ & + Tr \{ -P_t \left(H' N_t^{-1} H_t + \tilde{H}_t' N_t^{-1} h_t + \Lambda_t - \theta \tilde{Q}_t(u) \right) - \tfrac{1}{2} \tilde{H}_t' N_t^{-1} \tilde{H}_t \} dt. \end{aligned}$$

Hence,

$$\begin{aligned} & Tr \left(2F_t + G_t G_t' P_t^{-1} \right) = d \left(\log |P_t| \right) + Tr(-P_t \tilde{H}_t' P_t N_t^{-1} \tilde{H}_t + \tfrac{1}{2} \tilde{H}_t' N_t^{-1} \tilde{H}_t) dt \\ & + Tr \left\{ P_t \left(H_t' N_t^{-1} H_t + \tilde{H}_t' N_t^{-1} h_t + \Lambda_t - \theta \tilde{Q}_t(u) \right) \right\} dt - Tr \left(P_t \tilde{H}_t' N_t^{-1} \right) dy_t. \end{aligned}$$

If we set

$$\rho(0) = r(0)' P(0)^{-1} r(0) + \log[(2\pi)^n |P(0)|],$$

and then substitute into the equation of $\tilde{\mu}(\cdot)$ we have

$$\begin{aligned}
d\tilde{\mu}_t = {} & d\tilde{c}_t + \tfrac{1}{2} r_t . \left(\tilde{H}_t' P_t N_t^{-1} H_t' - 2\sigma_t \right) dt + \tfrac{\theta}{2} \left(\tilde{R}_t(u) u.u + \tilde{m}_t(u) r_t + 2\tilde{n}_t(u) u \right) dt \\
& + \tfrac{1}{2} r_t \left(\theta \tilde{Q}_t(u) - \Lambda_t - 3\tilde{H}_t' N_t^{-1} P_t \tilde{H}_t \right) . r_t dt \\
& + \tfrac{1}{2} \left(|\tfrac{1}{2} N_t^{-\frac{1}{2}} r_t' \tilde{H}_t r_t|^2 + r_t' \tilde{H}_t r_t . N_t^{-1} H_t r_t \right) dt \\
& + \tfrac{1}{2} Tr \left\{ P_t \left(\theta \tilde{Q}_t(u) - \Lambda_t - \tilde{H}_t' N_t^{-1} h_t \right) + P \tilde{H}_t' N_t^{-1} P_t \tilde{H}_t - \tfrac{1}{2} \tilde{H}_t' N_t^{-1} \tilde{H}_t \right\} dt \\
& + \tfrac{1}{2} Tr \left(P_t \tilde{H}_t' N_t^{-1} \right) dy_t - \delta_t dt - \tfrac{1}{2} d \left(\log \left\{ (2\pi)^n |P_t| \right\} \right), \\
\tilde{\mu}(0) = {} & \log \left\{ (2\pi)^n |P(0)| \right\}^{-\frac{1}{2}} .
\end{aligned}$$

This yields the desired results. □

Remark 3.8. Theorem 3.7 implies that whenever (3.80) admits an explicit solution, $q^\theta(\cdot)$ is described by finite-dimensional parameters. If we set $\phi(\cdot) = 0$, $\Lambda(\cdot) = 0, \sigma(\cdot) = 0, \delta(\cdot) = 0$, $\tilde{H}(\cdot) = 0$, we recover the solution of the Feynman–Kac information state equation corresponding to the LEQG tracking problem given in [3, 5, 6], while if, in addition, we set $\theta = 0$, we recover the conditional density of the LQG tracking problem.

3.2.1. Examples of nonlinear drift terms for Σ_G^2. In this section we present specific examples of nonlinear systems Σ_G^2 that admit explicit solutions of (3.80) and so, finite-dimensional representations of the information state $q^\theta(\cdot)$, for various nonlinear drift terms $g(\cdot)$.

We first turn (3.80) into a linear second-order PDE by introducing the transformation

$$W(x,t) = \exp \phi(x,t).$$

The equation governing $W_t \equiv W(x,t)$ is given by

$$\begin{aligned}
\tfrac{\partial W_t}{\partial t} + \tfrac{Tr}{2} \left(G_t G_t' D_x^2 W_t \right) + \left(F_t x + f_t + B(t,u) \right) . D_x W_t = {} & W_t \left\{ \tfrac{1}{2} x . \Lambda_t x + x . \sigma_t + \delta_t \right\} \\
& + W_t \{ \tfrac{\theta}{2} \tilde{\ell}_1(t,x,u) - \tfrac{1}{2} |\tfrac{1}{2} N_t^{-\frac{1}{2}} x' \tilde{H}_t x|^2 - \tfrac{1}{2} x' \tilde{H}_t x . N_t^{-1} H_t x \}.
\end{aligned} \tag{3.86}$$

Thus, we seek solutions of (3.86). We shall present two alternative methods for solving this equation, each leading to different classes of nonlinear control systems Σ_G^2. In the first method, we choose the function $\tilde{\ell}_1(\cdot)$ to cancel the control-dependent term $B(t,u).D_x W(\cdot)$. This implies that the function $g(\cdot)$ entering the unobservable dynamics is independent of the control $u(\cdot)$. In the second method, we allow $g(\cdot)$ to depend on the control parameter $u(\cdot)$ and hence on the paths of $y(\cdot)$. It is important to note that, from the family of nonlinear systems Σ_G^2, the class of Benes-type [11] nonlinearities emerges from the first method but not the second method. This observation will be made precise through examples. Moreover, the second method might yield finite-dimensional states which are not sufficient for the control, in the sense that the information state depends on the control directly, not indirectly through the finite-dimensional sufficient statistics.

THEOREM 3.9 (uncontrolled classes). *Suppose $u \in \hat{\mathcal{U}}$ and there exists $0 < \theta \leq \theta^*$ such that*

$$H_t' N_t^{-1} H_t + \tilde{H}_t' N_t^{-1} h_t + \Lambda_t - \theta \tilde{Q}_t(u) \geq 0 \qquad \forall (t,u) \in [0,T] \times \mathcal{U}.$$

Define

$$\Gamma_2(t,x) \doteq \frac{1}{2} \Delta_t x . x + x . \zeta_t + \eta_t,$$

where $\Delta(\cdot), \zeta_t(\cdot), \eta(\cdot)$ *are deterministic functions shortly to be made precise, and set*

$$\tilde{\ell}_1(t,x,u) = \frac{2}{\theta}\left\{ B(t,u).D_x\phi(x,t) + \frac{1}{2}\left(\left| \frac{1}{2} N_t^{-\frac{1}{2}} x' \tilde{H}_t x \right|^2 + x'\tilde{H}_t x.N_t^{-1} H_t x \right) \right\}.$$

The Feynman–Kac information state $q^\theta(\cdot)$ *given in Theorem* 3.7 *is a density function, at least for the following two classes of nonlinear drift terms* $\phi(\cdot), (g(\cdot) = GG'D_x\phi(\cdot))$.

Class 1 *(rational nonlinearities). Suppose* $\Gamma_2(t,x) > 0\ \forall (t,x) \in [0,T] \times \Re^n$. *A solution of* (3.80) *is*

$$\phi_{R_2}(x,t) = \log W_1(x,t), \qquad W_1(x,t) = \Gamma_2(x,t),$$

which implies that the nonlinear drift term $g(\cdot)$ *should be of the form*

$$g_t(x) = G_t G_t' D_x \phi_{R_2}(x,t) = \frac{G_t G_t'}{\frac{1}{2}\Delta_t x.x + x.\zeta_t + \eta_t} \left(\Delta_t x + \zeta_t \right).$$

Here

$$\begin{aligned}
&\dot{\Delta}_t + F_t'\Delta_t + \Delta_t F_t = \delta_t \Delta_t,\\
&\dot{\zeta}_t + F_t'\zeta_t + \Delta_t f_t = \delta_t \zeta_t,\\
&\dot{\eta}_t + \frac{1}{2} Tr\left(G_t G_t' \Delta_t \right) + f_t.\zeta_t = \delta_t \eta_t,\\
&\Lambda_t = 0, \quad \sigma_t = 0, \quad \delta_t = \textit{arbitrary}.
\end{aligned}$$

Moreover, if $\Delta_t > 0, \eta_t - \frac{1}{2}\zeta_t.\Delta_t^{-1}\zeta_t > 0\ \forall t \in [0,T]$, *then* $\Gamma_2(t,x) > 0\ \forall (t,x) \in [0,T] \times \Re^n$, *and the nonlinear drift term* $g(t,x)$ *is nonsingular* $\forall (t,x) \in [0,T] \times \Re^n$.

Class 2 *(exponential nonlinearities). A solution of* (3.80) *is*

$$\phi_{E_2}(x,t) = \log W_2(x,t), \qquad W_2(x,t) = \gamma_t^1 \exp\left(\Gamma_2(x,t)\right) + \gamma_t^2 \exp\left(-\Gamma_2(x,t)\right).$$

This implies that the nonlinear drift term $g(\cdot)$ *should be of the form*

$$g_t(x) = G_t G_t' D_x \phi_{E_2}(x,t) = \frac{\gamma_t^1 \exp\left(\Gamma_2(x,t)\right) - \gamma_t^2 \exp\left(-\Gamma_2(x,t)\right)}{\gamma_t^1 \exp\left(\Gamma_2(x,t)\right) + \gamma_t^2 \exp\left(-\Gamma_2(x,t)\right)} G_t G_t' \left(\Delta_t x + \zeta_t \right),$$

where

$$\begin{aligned}
&\dot{\Delta}_t + F_t'\Delta_t + \Delta_t F_t = 0,\\
&\dot{\zeta}_t + F_t'\zeta_t + \Delta_t f_t = 0,\\
&\dot{\eta}_t + \frac{1}{2} Tr\left(G_t G_t' \Delta_t \right) + f_t.\zeta_t = \frac{1}{2}\frac{d}{dt}\left(\log \frac{\gamma_t^1}{\gamma_t^2} \right),\\
&\Lambda_t = \Delta_t G_t G_t' \Delta_t, \quad \sigma_t = \Delta_t G_t G_t' \zeta_t,\\
&\delta_t = \frac{1}{2}\zeta_t' G_t G_t' \zeta_t + \frac{1}{2}\frac{d}{dt}\left(\log \gamma_t^1 \gamma_t^2 \right).
\end{aligned}$$

Class 3 (*combination of classes* 1 *and* 2). *We can take* $\phi(x,t)$ *to be linear combinations of the logarithm of solutions* $W_1(x,t), W_2(x,t)$.

Proof. Follow the derivation given in [2], or substitute the solutions into the evolution equation of $W(\cdot)$ or $\phi(\cdot)$. The last class follows from the linearity of (3.86). ∎

Example 3.10 (rational nonlinearities). Here, we wish to construct specific examples of nonlinear drift terms $g(\cdot)$ using the results of Theorem 3.9 stated under Class 1.

Case 1. Suppose that F, G, f, Γ_2 are independent of time and $F = \frac{\alpha}{2} I_n, \alpha \in \Re$. The nonlinear drift term is

$$g(x) = \frac{GG'}{\frac{1}{2}\Delta x.x + x.\zeta + \eta}(\Delta x + \zeta),$$

where Δ is arbitrary, $\delta = \alpha, \zeta = \frac{2}{\alpha}\Delta f, \eta = \frac{1}{2\alpha} Tr(GG'\Delta) + f.\frac{2}{\alpha^2}\Delta f$. Moreover, $g(x)$ is nonsingular $\forall x \in \Re^n$ provided $\Delta > 0$ and $\alpha > 0$.

Case 2. Suppose F, G, Γ_2 are independent of time and $F = \frac{\alpha}{2} I_n, \alpha \in \Re, f = 0$. The nonlinear drift term is

$$g(x) = \frac{GG'}{\frac{1}{2}\Delta x.x + \eta}\Delta x,$$

where Δ is arbitrary, $\delta = \alpha, \zeta = 0, \eta = \frac{1}{2\alpha} Tr(GG'\Delta)$. Clearly, in this case $g(x)$ is nonsingular $\forall x \in \Re^n$ provided $\Delta > 0$ and $\alpha > 0$.

Case 3. Suppose $F_t = 0, f = 0$. The nonlinear drift term is

$$g(t, x) = \frac{G_t G_t'}{\frac{1}{2}\Delta_t x.x + \eta_t}\Delta_t x,$$

where $\dot{\Delta}_t = \delta_t \Delta_t$, $\zeta_t = 0, \dot{\eta}_t + \frac{1}{2} Tr(G_t G_t' \Delta_t) = \delta_t \eta_t$, $\delta(\cdot)$ is arbitrary. Moreover, $g(t, x)$ is nonsingular $\forall (t, x) \in [0, T] \times \Re^n$ provided $\Delta_t > 0$ and $\eta_t > 0\ \forall t \in [0, T]$.

The functions $\Delta(\cdot), \zeta(\cdot), \eta(\cdot), \Lambda(\cdot), \sigma(\cdot), \delta(\cdot)$ are measurable functions of t. Thus they do not depend on the control $u \in \hat{\mathcal{U}}$, and hence on the paths of $y(\cdot)$. However, an important disadvantage of the results of Theorem 3.9 is the presence of the term $\frac{2}{\theta} B(t, u).D_x \phi_t$ as part of $\tilde{\ell}_1(t, x, u)$. We can overcome this disadvantage by allowing the functions $\phi(\cdot)$, and thus $g(\cdot)$, to be pathwise-dependent on the observations $y(\cdot)$ through the control $u(t, y)$. This modification leads to the additional classes of nonlinear control systems presented in the next theorem.

THEOREM 3.11 (controlled classes). *Suppose $u \in \hat{\mathcal{U}}$ and there exists a $\theta \leq \theta^*$ such that*

$$H_t' N_t^{-1} H_t + \tilde{H}_t' N_t^{-1} h_t + \Lambda_t - \theta \tilde{Q}_t(u) \geq 0 \qquad \forall (t, u) \in [0, T] \times \mathcal{U}.$$

Define

$$\Gamma_2^u(t, x) \doteq \frac{1}{2}\Delta_t x.x + x.\zeta_t^u + \eta_t^u,$$

where $\Delta(\cdot), \zeta_t^u(\cdot), \eta^u(\cdot)$ will be made precise shortly. Set

$$\tilde{\ell}_1(t, x, u) = \frac{2}{\theta}\left\{\frac{1}{2}\left(\left|\frac{1}{2} N_t^{-\frac{1}{2}} x' \tilde{H}_t x\right|^2 + x' \tilde{H}_t x.N_t^{-1} H_t x\right)\right\}.$$

Suppose (3.80) *has a Borel measurable solution $\phi^u : \Re^n \times \mathcal{U} \times [0, T] \to \Re$. The Feynman–Kac information state $q^\theta(\cdot)$ given in Theorem* 3.7 *is a density function at least for the following two classes of control-dependent, nonlinear drift terms $\phi^u(\cdot)$ $(g^u(\cdot) = GG' D_x \phi^u(\cdot))$.*

Class 1 *(rational nonlinearities). Suppose* $\Gamma_2^u(t,x) > 0 \; \forall (t,x,u) \in [0,T] \times \Re^n \times \mathcal{U}$. *A control-dependent solution of* (3.80) *is*

$$\phi_{R_2}^u(x,t) = \log W_1^u(x,t), \qquad W_1^u(x,t) = \Gamma_2^u(x,t).$$

This implies that the nonlinear drift term $g(\cdot)$ *should be of the form*

$$g_t^u(x) = G_t G_t' D_x \phi_{R_2}^u(x,t) = \frac{G_t G_t'}{\frac{1}{2}\Delta_t x.x + x.\zeta_t^u + \eta_t^u} \left(\Delta_t x + \zeta_t^u\right),$$

where

$$\begin{aligned}
&\dot{\Delta}_t + F_t'\Delta_t + \Delta_t F_t = \delta_t \Delta_t,\\
&\dot{\zeta}_t^u + F_t'\zeta_t^u + \Delta_t f_t + \Delta_t B(t,u) = \delta_t \zeta_t^u,\\
&\dot{\eta}_t^u + \frac{1}{2} Tr\left(G_t G_t' \Delta_t\right) + f_t.\zeta_t^u + B(t,u).\zeta_t^u = \delta_t \eta_t^u,\\
&\Lambda_t = 0, \quad \sigma_t = 0, \quad \delta_t = \textit{arbitrary}.
\end{aligned}$$

The functions $\zeta^u(\cdot) \equiv \zeta(\cdot,u), \eta^u(\cdot) \equiv \eta(\cdot,u)$ *are measurable in* t, *and the function* $\phi_{R_2}^u(\cdot) \equiv \phi(\cdot,u,\cdot)$ *is pathwise-dependent on the observations* $y(\cdot)$ *through the control* $u(t,y)$.

Moreover, if $\Delta_t > 0, \eta_t^u - \frac{1}{2}\zeta_t^u.\Delta_t^{-1}\zeta_t^u > 0 \; \forall (t,u) \in [0,T] \times \mathcal{U}$, *then* $\Gamma_2^u(t,x) > 0 \; \forall (t,x,u) \in [0,T] \times \Re^n \times \mathcal{U}$, *and the nonlinear drift term* $g^u(t,x)$ *is nonsingular* $\forall (t,x,u) \in [0,T] \times \Re^n \times \mathcal{U}$.

Class 2 *(exponential nonlinearities). A control-dependent solution of* (3.80) *is*

$$\phi_{E_2}^u(x,t) = \log W_2^u(x,t), \qquad W_2^u(x,t) = \gamma_t^1 \exp\left(\Gamma^u(x,t)\right) + \gamma_t^2 \exp\left(-\Gamma_2^u(x,t)\right),$$

which implies that the nonlinear drift term $g(\cdot)$ *should be of the form*

$$g_t^u(x) = G_t G_t' D_x \phi_{E_2}^u(x,t) = \frac{\gamma_t^1 \exp\left(\Gamma_2^u(x,t)\right) - \gamma_t^2 \exp\left(-\Gamma_2^u(x,t)\right)}{\gamma_t^1 \exp\left(\Gamma_2^u(x,t)\right) + \gamma_t^2 \exp\left(-\Gamma_2^u(x,t)\right)} G_t G_t' \left(\Delta_t x + \zeta_t^u\right),$$

where

$$\begin{aligned}
&\dot{\Delta}_t + F_t'\Delta_t + \Delta_t F_t = 0,\\
&\dot{\zeta}_t^u + F_t'\zeta_t^u + \Delta_t f_t + \Delta_t B(t,u) = 0,\\
&\dot{\eta}_t^u + \frac{1}{2} Tr\left(G_t G_t' \Delta_t\right) + f_t.\zeta_t^u + B(t,u).\zeta_t^u = \frac{1}{2}\frac{d}{dt}\left(\log \frac{\gamma_t^1}{\gamma_t^2}\right),\\
&\Lambda_t = \Delta_t G_t G_t' \Delta_t, \quad \sigma_t^u = \Delta_t G_t G_t' \zeta_t^u,\\
&\delta_t^u = \frac{1}{2}\zeta_t^u.G_t G_t' \zeta_t^u + \frac{1}{2}\frac{d}{dt}\left(\log \gamma_t^1 \gamma_t^2\right).
\end{aligned}$$

The functions $\zeta^u(\cdot) \equiv \zeta(\cdot,u), \eta^u(\cdot) \equiv \eta(\cdot,u), \delta^u(\cdot) \equiv \delta(\cdot,u), \sigma^u(\cdot) \equiv \sigma(\cdot,u)$ *are measurable in* t, *the nonlinear function* $\phi_{E_2}^u(\cdot) \equiv \phi(\cdot,u,\cdot)$ *is pathwise dependent on the observations* $y(\cdot)$ *through the control* $u(t,y)$.

Proof. Follow the derivation given in [2], or substitute the solutions into the evolution equation for $W(\cdot)$ or $\phi(\cdot)$ to verify the results. □

We now demonstrate through specific examples that whether the function $\phi(\cdot)$ is chosen to depend on the control u or not, leads to different classes of nonlinear control problems.

Example 3.12. Suppose we are interested in the control analog of nonlinear dynamical systems with Benes-type [11] nonlinearities, namely:

$$dx_t = \tanh(x_t)dt + u(t,y)dt + dw_t, \qquad x(0) \in \Re, \tag{3.87}$$

$$dy_t = x_t dt + db_t, \qquad y(0) = 0 \in \Re. \tag{3.88}$$

This is a special case of the nonlinear control systems defined by the class Σ_G^2 (i.e., $F = H = \tilde{H} = f = h = 0, G = 1, B(t,u) = u$). When $u = 0$, the above model is shown in [11] to yield finite-dimensional filters. Here, we wish to determine whether the results of Theorems 3.9, 3.11, stated under Class 2, yield explicit solutions of the Feynman–Kac information state equation for the above control system. This will be possible if the results of Theorems 3.9, 3.11, stated under Class 2, can be specialized to the particular form $\Gamma_2(t,x) = x$, $\Gamma^u(t,x) = x$, respectively. That is, we require $\Delta_t = 0, \zeta_t = 1, \eta_t = 0$. If we set $\gamma_1 = \gamma_2 = 1, \Delta(0) = 0, \eta(0) = 0, \zeta(0) = 1$ in the results of Theorem 3.9, stated under Class 2 we deduce

$$\zeta_t = 1, \quad \Delta_t = 0, \quad \eta_t = 0, \qquad \Lambda_t = 0, \quad \sigma_t = 0, \quad \delta_t = \frac{1}{2}.$$

Hence,

$$g_t(x) = \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)} = \tanh(x).$$

This implies that the Feynman–Kac information state equation associated with the above system admits an explicit solution when $\tilde{\ell}_1 = \frac{2}{\theta} u(t,y) \tan h(x)$ (see Theorem 3.9). On the other hand, one could show that the results of Theorem 3.11, stated under Class 2, do not admit an explicit solution for this Feynman–Kac information state equation because, although we could have $\Delta_t = 0, \zeta_t = 1$, we also have $\dot{\eta}_t + u(t,u) = 0$. This would never yield the desired solution $\eta_t = 0$ (and hence $\Gamma^u(t,x) = x$), unless $u(t,y) = 0$.

Example 3.13. Suppose we are interested in the following scalar problem:

$$dx_t = \frac{\Delta_t x_t + \zeta_t}{\frac{1}{2}\Delta_t x_t^2 + \zeta_t x_t + \eta_t} dt + u(t,y)dt + dw_t, \qquad x(0) \in \Re, \tag{3.89}$$

$$dy_t = x_t dt + db_t, \qquad y(0) = 0 \in \Re. \tag{3.90}$$

Setting $F_t = H_t = \tilde{H}_t = f_t = h_t = 0, G_t = 1, B(t,u) = u$ in the results of Theorem 3.9 stated under Class 1, we deduce

$$\dot{\Delta}_t = \delta_t \Delta_t, \quad \dot{\zeta}_t = \delta_t \zeta_t, \quad \dot{\eta}_t + \frac{\Delta_t}{2} = \delta_t \eta_t.$$

By choosing $\delta_t = \delta =$constant, we have

$$\Delta_t = \Delta(0)\exp(\delta t), \quad \zeta_t = \zeta(0)\exp(\delta t), \quad \eta_t = \eta(0)\exp(\delta t) - \frac{1}{2}\Delta(0)\exp(\delta t)t.$$

Therefore, when $\tilde{\ell}_1(t,x,u) = \frac{2}{\theta} u(t,y) \frac{\Delta_t x + \zeta_t}{\Delta_t x^2 + \zeta_t x + \eta_t}$ the Feynman–Kac information state is finite-dimensional and the nonlinear drift term in (3.89) is nonsingular provided $\Delta(0) > 0$ and $\eta(0) - \frac{1}{2}\Delta(0)t - \frac{1}{2}\frac{\zeta(0)^2}{\Delta(0)} > 0$. On the other hand, it is easy to show that

the results of Theorem 3.11, stated under Class 1, do not admit systems of this form because

$$\dot{\Delta}_t = \delta_t \Delta_t, \quad \dot{\zeta}_t^u + \Delta_t u(t,y) = \delta_t \zeta_t^u, \quad \dot{\eta}_t^u + \frac{\Delta_t}{2} + u(t,y)\zeta_t^u = \delta_t \eta_t^u.$$

Thus, $\zeta(\cdot), \eta(\cdot)$ in (3.89) are functionals of u.

Remark 3.14. The last two examples seem to suggest that the results of Theorem 3.9 are better suited for modeling nonlinearities entering the dynamics of the unobservable states. On the other hand, a disadvantage is the presence of the term $\frac{2}{\theta} B(t,u).D_x\phi(x,t)$ as part of $\tilde{\ell}_1(t,x,u)$. This is not present in the results of Theorem 3.11.

3.2.2. Representation of cost function. We shall now convert the family of nonlinear control systems Σ_G^2, which were originally infinite dimensional (see Theorem 2.3), to standard, finite-dimensional, completely observable stochastic control problems. By Theorem 3.7, we know that if a solution of (3.80) exists such that $q^\theta(\cdot)$ is a density function, then the total cost function (2.26) can be expressed in terms of the functions $\phi(\cdot)$ (or $\phi^u(\cdot)$), $P(\cdot), r(\cdot)$, and the differential observation process $dy(\cdot)$. (We shall distinguish between $\phi(\cdot)$ and $\phi^u(\cdot)$ only when referring to specific examples.) To this end we define

$$\breve{\ell}_2(t,x,P,u) \doteq \int_{\Re^n} \ell_2(t,z,u) \frac{\exp\left(\phi(z,t) - \frac{1}{2}P_t^{-1}(z-x).(z-x)\right)}{(2\pi)^{\frac{n}{2}}} dz,$$

$$\breve{\varphi}_2(x,P,u) \doteq \int_{\Re^n} \varphi_2(T,z) \exp\left(\theta\varphi_1(T,z)\right) \frac{\exp\left(\phi(z,T) - \frac{1}{2}P_T^{-1}(z-x).(z-x)\right)}{(2\pi)^{\frac{n}{2}}} dz.$$

Clearly, by incorporating the results of Theorem 3.7, the infinite-dimensional, stochastic control problem given under Theorem 2.3, is now equivalent to a completely observable finite-dimensional stochastic control problem with cost function given by

$$\begin{aligned} J^\theta_{\Sigma_G^2}(u(\cdot)) = E\Bigg\{ & \int_0^T \frac{1}{|P_t|^{\frac{1}{2}}} \breve{\ell}_2(t,r_t,P_t,u) \exp\left(\tilde{c}_t + \tilde{\lambda}_t + \lambda_t\right) dt \\ & + \frac{1}{|P_T|^{\frac{1}{2}}} \breve{\varphi}_2(r_T,P_T) \exp\left(\tilde{c}_T + \tilde{\lambda}_T + \lambda_T\right) \Bigg\}. \end{aligned} \tag{3.91}$$

The functions $\tilde{c}(\cdot), \tilde{\lambda}(\cdot), \lambda(\cdot)$ are defined in Theorem 3.7 and the evolutions of $r(\cdot), P(\cdot)$ are given by (3.82), (3.83), respectively.

4. Finite-dimensional information states.

4.1. Nonlinear dynamics linear sensor problem. Unfortunately, if we consider the quadratic sensor problem the information state equation (2.33) will not evolve on a finite-dimensional manifold. This is not surprising because the Feynman–Kac information state equation contains the additional term $\tilde{\ell}_1$ which has been chosen in Theorems 3.7 and 3.9 in such a way as to cancel nonlinearities in x of specific type. On the other hand, if we set $\tilde{H} = 0$ in the definition of the control system Σ_G^2, we obtain from the results of section 3.2 a finite-dimensional representation for $q^0(\cdot)$. These results are summarized in the next theorem.

Control system (Σ_I^u). Suppose Assumptions 2.1 hold, with $\mathcal{U}$ in A1 replaced by a compact subset of $\Re^m$, and the dynamics and observations are given by

$$\begin{aligned} dx_t &= (F_t x_t + f_t + g(t,x,u(t,y)))\, dt \\ &\quad + B(t,u(t,y))dt + G_t dw_t, \qquad x(0) \in \Re^n, \end{aligned} \tag{4.92}$$

$$dy_t = (H_t x_t + h_t)\, dt + N_t^{\frac{1}{2}} db_t, \qquad y(0) = 0 \in \Re^d, \tag{4.93}$$

$$J^0_{\Sigma_I^u}(u(\cdot)) \doteq J^0_I(u(\cdot)) = (1.4). \tag{4.94}$$

Here $g^u \equiv g : [0,T] \times \Re^n \times \mathcal{U} \to \Re^n$ is Borel measurable, and A10 and A12 hold with $g \to g^u, \phi(x) \to \phi(t,x,u), \tilde{q}_0^\theta(x) \to \tilde{q}_0^0(x)$.

For $u \in \hat{\mathcal{U}}$ the information state associated with the control system Σ_I^u evolves according to the equation

$$dq_t^0 = \tfrac{1}{2} Tr\left(G_t G_t' D_x^2 q_t^0\right) dt - \tfrac{\partial}{\partial x}\left(q_t^0 \left(F_t x + g_t(x,u)\right)\right) dt + \left(H_t x + h_t\right).q_t^0 N_t^{-1} dy_t,$$
$$q^0(x,0) = q_0^\theta(x).$$

The next theorem is a direct consequence of Theorems 3.7 and 3.9. The results corresponding to uncontrolled diffusion processes (i.e., $B = 0$) were first derived in [21].

THEOREM 4.1. *Consider the control system Σ_I^u, suppose $H_t' N_t^{-1} H_t + \Lambda_t^u \geq 0$ $\forall (t,u) \in [0,T] \times \mathcal{U}$, and for $u \in \hat{\mathcal{U}}$, there exist functions $\phi^u \in C^{2,1}_{x,t}(\Re^n \times [0,T])$ satisfying the partial differential equation*

$$\begin{aligned} &\tfrac{\partial \phi_t^u}{\partial t} + \tfrac{1}{2} Tr\left(G_t G_t' D_x^2 \phi_t^u\right) + \tfrac{1}{2} D_x \phi_t^u . G_t G_t' D_x \phi_t^u \\ &+ \left(F_t x + f_t + B(t,u)\right).D_x \phi_t^u = \tfrac{1}{2} x \Lambda_t^u . x + x.\sigma_t^u + \delta_t^u. \end{aligned} \tag{4.95}$$

Here $\Lambda^u(\cdot), \sigma^u(\cdot), \delta^u(\cdot)$ are free to be chosen so that (4.95) *yields explicit solutions. Then*

$$q^0(x,t) = \exp\left(\phi^u(x,t)\right) \frac{\exp\left(-\frac{1}{2} P_t^{-1}(x - r_t).(x - r_t)\right)}{(2\pi)^{\frac{n}{2}} |P_t|^{\frac{1}{2}}} \exp\left(c_t + \mu_t\right).$$

The cost function is

$$J^0_{\Sigma_I^u}(u(\cdot)) = E\left\{ \int_0^T (\ell_2(t,\cdot,u_t), q_t^0) dt + (\varphi_2(T,\cdot), q_T^0) \right\}. \tag{4.96}$$

Here $P(\cdot), r(\cdot), \mu(\cdot), c(\cdot)$ satisfy the following equations:

$$\begin{aligned} dr_t &= \left(F_t - P_t \Lambda_t^u\right) r_t dt + f_t dt - P_t \sigma_t^u dt \\ &\quad + B(t,u) dt + P_t H_t' N_t^{-1} \left(dy_t - H_t r_t dt - h_t dt\right), \quad r(0) = \xi, \\ \dot{P}_t &= F_t P_t + P_t F_t' - P_t \left(H_t' N_t^{-1} H_t + \Lambda_t^u\right) P_t + G_t G_t', \quad P(0) = P_0, \\ d\mu_t &= -\frac{1}{2}\left(r_t.\Lambda_t^u r_t + 2 r_t.\sigma_t^u + 2\delta_t^u + Tr(P_t \Lambda_t^u)\right) dt, \quad \mu(0) = 0, \\ dc_t &= \left(H_t r_t + h_t\right).N_t^{-1} dy_t - \frac{1}{2} |N_t^{-\frac{1}{2}} \left(H_t r_t + h_t\right)|^2 dt, \quad c(0) = 0. \end{aligned}$$

In addition, the information state $q^0(\cdot)$ can be written explicitly for the classes of nonlinear functions $\phi^u(\cdot)$ given in Theorem 3.11.

Proof. The first part of the theorem follows by setting $\theta = 0$ and $\tilde{H} = 0$ in the results of Theorem 3.7. The second part of the theorem follows from Theorem 3.11. ∎

5. Examples of optimal and suboptimal controls for Σ_G^2 with $\tilde{H} = 0$. Note that, when the Feynman–Kac information state is expressed in terms of a finite number of quantities, as in Theorem 3.7, under an appropriate hypothesis one could employ dynamic programming arguments as in [4, 6] to derive a Hamilton–Jacobi (HJ) equation satisfied by the optimal cost-to-go, and then establish a verification theorem. Consequently, in this case, if the optimal control laws exist, they are finite-dimensional.

In the next two theorems we present sufficient conditions for identifying nonlinear partially observable stochastic control problems with $\tilde{H} = 0$, which have exact optimal control laws, reminiscent of LEQG/LQG tracking problems.

THEOREM 5.1. *Suppose Assumptions* 2.1 *hold. Consider the problem of finding a control law $u^* \in \hat{\mathcal{U}}$ minimizing the total cost function*

$$J^\theta(u(\cdot)) = E^u \Big\{ \varphi_2(T, x_T) \exp \tfrac{\theta}{2} \Big(\int_0^T [Q_t x_t . x_t + R_t u(t,y).u(t,y) + 2m_t x_t + 2n_t u(t,y) + \tilde{\ell}_1(t, x_t, u(t,y)]dt + [Q_T x_T . x_T + 2m_T x_T] \Big) \Big\}. \tag{5.97}$$

Here

$$Q = Q' : [0,T] \to \mathcal{L}(\Re^n; \Re^n), \quad R = R' : [0,T] \to \mathcal{L}(\Re^m; \Re^m),$$
$$m : [0,T] \to (\Re^n)', \quad n : [0,T] \to (\Re^m)', \quad Q \geq 0, \quad R > 0,$$

and x, y are subject to dynamics:

$$dx_t = (F_t x_t + g(t, x_t) + f_t)\, dt + B_t u(t,y) dt + G_t dw_t, \quad x(0) \in \Re^n, \tag{5.98}$$

$$dy_t = (H_t x_t + h_t)\, dt + N_t^{\frac{1}{2}} db_t, \quad y(0) = 0 \in \Re^d. \tag{5.99}$$

I. *Suppose the following conditions hold.*

1. *The function $\tilde{\ell}_1(\cdot)$ is defined by*

$$\tilde{\ell}_1(t,x,u) \doteq \frac{2}{\theta} B_t u(t,y).D_x \phi(x,t) + \hat{\ell}_1(t,x,u), \tag{5.100}$$

where $\hat{\ell}_1(\cdot)$ is chosen so that there exists some solution of (3.80) *with $\tilde{H} = 0$.*

2. *The function $\varphi_2(\cdot)$ is defined by*

$$\varphi_2(T,x) \doteq \exp(-\phi(x,T)). \tag{5.101}$$

3. *$\Lambda(\cdot), \sigma(\cdot)$ are functions of t and $\delta(\cdot)$ is either a function of t or $\delta(t,\cdot)$ is a linear function of u.*

Then the optimal control $u^* \in \hat{\mathcal{U}}$ is linear, feedback, as in the LEQG tracking problem.

II. *If conditions* I1, I3 *hold and*

$$\frac{1}{2}\left(\tilde{Q}_T^- x.x + 2\tilde{m}_T^- x + \tilde{\rho}_T^-\right) \leq \phi(x,T) \leq \frac{1}{2}\left(\tilde{Q}_T^+ x.x + 2\tilde{m}_T^+ x + \tilde{\rho}_T^+\right), \quad \varphi 2 = 1, \tag{5.102}$$

where $\tilde{Q}_T^- = \tilde{Q}_T^{-,\prime}, \tilde{Q}_T^+ = \tilde{Q}_T^{+,\prime}$, then the optimal total cost $J(u^*(\cdot))$ is bounded above and below by that of the LEQG tracking problem.

If the term $\frac{2}{\theta} B_t u(t,y).D_x\phi(x,t)$ is removed from (5.100) *one must allow $g(\cdot)$ to be pathwise-dependent on the observations, thus generalizing* (5.98).

Proof. I. From Theorem 3.7, we know that if there exists a function $\phi \in C^{2,1}_{x,t}(\Re^n \times [0,T])$ satisfying (3.80) with $\tilde{H} = 0$, and if we choose the function $\tilde{\ell}_1(\cdot)$ according to condition I1, then the Feynman–Kac information state equation associated with control problem (5.97)–(5.99) is given by

$$q^\theta(x,t) = \exp(\phi(x,t) + c_t + \lambda_t) \times \frac{\exp\left(-\frac{1}{2}P_t^{-1}(x-r_t).(x-r_t)\right)}{(2\pi)^{\frac{n}{2}}|P_t|^{\frac{1}{2}}}\hat{\Lambda}^u_{0,T},$$

where $r(\cdot), P(\cdot)$ satisfy the equations

$$\begin{aligned} dr_t = {} & (F_t - P_t(\Lambda_t - \theta Q_t))\, r_t dt + f_t dt + B_t u(t,y) dt + \theta P_t m_t' dt - P_t \sigma_t dt \\ & + P_t H_t' N_t^{-1}(dy_t - H_t dt - h_t dt), \quad r(0) = \xi, \end{aligned} \tag{5.103}$$

$$\dot{P}_t = F_t P_t + P_t F_t' - P_t(H_t' N_t^{-1} H_t + \Lambda_t - \theta Q_t)P_t + G_t G_t', \quad P(0) = P_0. \tag{5.104}$$

$\hat{\Lambda}^u(\cdot)$ is defined earlier, and $\lambda(\cdot)$ is given by the equation

$$\begin{aligned} \lambda_t = {} & \exp\tfrac{\theta}{2}\left(\int_0^T \left(r_s[Q_s - \tfrac{\Lambda_s}{\theta}].r_s + R_s u(s,y).u(s,y) + Tr\left(P_s[Q_s - \tfrac{\Lambda_s}{\theta}]\right)\right) ds\right) \\ & \times \exp\left(\tfrac{\theta}{2}\int_0^T \left(2r_s.[m_s' - \tfrac{\sigma_s}{\theta}] + 2[n_s u(s,y) - \tfrac{\delta_s}{\theta}]\right) ds\right). \end{aligned} \tag{5.105}$$

If φ_2 is defined according to condition I2 and $I : [0,T] \to \Re$, from Theorem 2.3, we know that the cost function (5.97) admits the representation

$$\begin{aligned} J^\theta(u(\cdot)) = I_{0,T} E\Big\{ & \exp\tfrac{\theta}{2}\left(\int_0^T \left(r_s[Q_s - \tfrac{\Lambda_s}{\theta}].r_s + R_s u(s,y).u(s,y) + Tr\left(P_s[Q_s - \tfrac{\Lambda_s}{\theta}]\right)\right) ds\right) \\ & \times \exp\left(\tfrac{\theta}{2}\int_0^T \left(2r_s.[m_s' - \tfrac{\sigma_s}{\theta}] + 2[n_s u(s,y) - \tfrac{\delta_s}{\theta}]\right) ds\right) \\ & \times \exp\tfrac{\theta}{2}\left(\hat{\varphi}_2(T, r_T)\right) \times \hat{\Lambda}^u_{0,T}\Big\}. \end{aligned} \tag{5.106}$$

Here the function $\hat{\varphi}_2(\cdot)$ is quadratic in r. If condition I3 holds as well, then $\Lambda(\cdot), \sigma(\cdot)$ are deterministic functions of t, and either $\delta(\cdot) : [0,T] \to \Re$ or $\delta(t,\cdot)$ is a linear function of u. Hence, the problem of minimizing (5.97) over $u \in \hat{\mathcal{U}}$, subject to dynamics (5.98), (5.99), is equivalent to the problem of minimizing (5.106) over $u \in \hat{\mathcal{U}}$, subject to dynamics (5.103), (5.104). However, the latter problem is equivalent to a completely observable LEQG tracking problem. Therefore, the optimal control is linear, feedback, and of separated form $u^*(t) = u(t,r)$ (see [3, 6]).

II. This follows by substituting (5.102) into (5.106).

A similar derivation holds when the function $\phi(\cdot)$ is pathwise-dependent on the observation y. This then implies that the term $\frac{2}{\theta}B_t u(t,y).D_x\phi^u(x,t)$ is not present in $\tilde{\ell}_1(\cdot)$. Moreover, it is possible to relax the third condition, allowing σ to be a linear function of u and δ to be a quadratic function of u. □

THEOREM 5.2. *Suppose the assumptions corresponding to the family of systems Σ_I^u hold with $\mathcal{U}$ as defined in* A1. *Consider the problem of finding a control law $u^* \in \hat{\mathcal{U}}$ minimizing the total cost function*

$$\begin{aligned} J^0(u(\cdot)) = {} & \tfrac{1}{2}E^u\Big\{\int_0^T \hat{\ell}_2(t, x_t, u(t,y)) \times (Q_t x_t.x_t + R_t u(t,y).u(t,y) \\ & + 2m_t x_t + 2n_t u(t,y))\, dt + \hat{\varphi}_2^u(T, x_T) \times (Q_T x_T.x_T + 2m_T x_T)\Big\}, \end{aligned} \tag{5.107}$$

where Q, R, m, n are specified in Theorem 5.1, *subject to dynamics and observations given by*

$$dx_t = (F_t x_t + g^u(t, x_t) + f_t)\, dt + B_t u(t, y) dt + G_t dw_t, \quad x(0) \in \Re^n, \tag{5.108}$$

$$dy_t = (H_t x_t + h_t)\, dt + N_t^{\frac{1}{2}} db_t, \quad y(0) = 0 \in \Re^d. \tag{5.109}$$

I. *Suppose the following conditions hold.*

1. *There exists some solution $\phi^u(\cdot)$ of* (4.95).
2. *The functions $\hat{\ell}_2, \hat{\varphi}_2^u(\cdot)$ are defined by*

$$\hat{\ell}_2(t, x, u) \doteq \exp(-\phi^u(x, t)), \qquad \hat{\varphi}_2^u(T, x) \doteq \exp(-\phi^u(x, T)). \tag{5.110}$$

3. *$\Lambda^u(\cdot) = 0, \sigma^u(\cdot) = 0$, and $\delta^u(\cdot) = \delta(\cdot)$ is a function of t.*

Then the optimal control $u^* \in \hat{\mathcal{U}}$ is linear, feedback, as in the LQG tracking problem.

Proof. From Theorem 4.1, we know that, if there exists a function $\phi^u \in C_{x,t}^{2,1}(\Re^n \times [0, T])$, satisfying (4.95), and we set $g^u = GG' D_x \phi^u \mathrm{c}$, then the information state equation associated with control problem (5.107)–(5.109) is given by

$$q^0(x, t) = \exp(\phi^u(x, t) + \lambda_t) \times \frac{\exp\left(-\frac{1}{2} P_t^{-1}(x - r_t).(x - r_t)\right)}{(2\pi)^{\frac{n}{2}} |P_t|^{\frac{1}{2}}} \hat{\Lambda}_{0,t}^u.$$

Here $r(\cdot), P(\cdot)$ satisfy the equations

$$\begin{aligned} dr_t = {} & (F_t - P_t \Lambda_t^u)\, r_t dt + f_t dt + B_t u(t, y) dt - P_t \sigma_t^u dt \\ & + P_t H_t' N_t^{-1} (dy_t - H_t dt - h_t dt), \quad r(0) = \xi, \end{aligned} \tag{5.111}$$

$$\dot{P}_t = F_t P_t + P_t F_t' - P_t (H_t' N_t^{-1} H_t + \Lambda_t^u) P_t + G_t G_t', \quad P(0) = P_0. \tag{5.112}$$

$\hat{\Lambda}^u(\cdot)$ is the exponential martingale defined earlier, and $\lambda(\cdot)$ is given by the equation

$$\lambda_t = \exp \frac{1}{2} \left\{ - \int_0^t \left(r_s[\Lambda_s^u].r_s + Tr\left(P_s[\Lambda_s^u]\right) + 2 r_s.[\sigma_s^u] + 2[\delta_s^u] \right) ds \right\}. \tag{5.113}$$

If conditions I1–I3 of the theorem are satisfied, from Theorem 4.1, we know that the cost function (5.107) is represented by

$$\begin{aligned} J^0(u(\cdot)) = E \Big\{ & \tfrac{1}{2} \Big(\textstyle\int_0^T \left(Q_t r_t.r_t + R_t u(t, y).u(t, y) + 2 r_t.m_t' + 2 n_t u(t, y) \right) dt \\ & + \left(Q_T r_T.r_T + 2 r_T.m_T' \right) + \textstyle\int_0^T Tr\left(P_t Q_t\right) dt + Tr\left(P_T Q_T\right) \Big) \hat{\Lambda}_{0,T}^u \Big\} \\ & \times \exp(-\textstyle\int_0^T \delta_t dt). \end{aligned} \tag{5.114}$$

Hence, the problem of minimizing (5.107) over $u \in \hat{\mathcal{U}}$, subject to dynamics (5.108), (5.109), is equivalent to the problem of minimizing (5.114) over $u \in \hat{\mathcal{U}}$ subject to dynamics (5.111), (5.112). However, the latter problem is equivalent to a completely observable LQG tracking problem. Therefore, the optimal control is linear, feedback, and of separated form $u^*(t) = u(t, r)$. □

In the next two subsections we shall present specific examples of general nonlinear partially observable control problems that yield linear observer dynamics and linear feedback optimal control laws.

5.1. Nonlinear dynamics exponential-of-integral cost. Control system $(\Sigma_{E_i}, i = 1, 2)$. We specialize (5.97)–(5.99) of Theorem 5.1 to the following case. The dynamics and observations are given by

$$dx_t = (F_t x_t + g_i(t, x_t) + f_t)\, dt + B_t u(t, y)dt + G_t dw_t, \qquad x(0) \in \Re^n, \qquad i = 1, 2,$$
$$dy_t = (H_t x_t + h_t)\, dt + N_t^{\frac{1}{2}} db_t, \qquad y(0) = 0 \in \Re^d.$$

Here

$$\Gamma_2^i(t, x) = \frac{1}{2}\Delta_t^i x.x + x.\zeta_t^i + \eta_t^i, \quad i = 1, 2,$$

$$g_1(t, x) \doteq G_t G_t' \frac{D_x \Gamma_2^1(t, x)}{\Gamma_2^1(t, x)} = \frac{G_t G_t'}{\frac{1}{2}\Delta_t^1 x.x + x.\zeta_t^1 + \eta_t^1} \left(\Delta_t^1 x + \zeta_t^1\right),$$
$$g_2(t, x) \doteq \frac{\gamma_t^1 \exp(\Gamma_2^2(t, x)) - \gamma_t^2 \exp(-\Gamma_2^2(t, x))}{\gamma_t^1 \exp(\Gamma_2^2(t, x)) + \gamma_t^2 \exp(-\Gamma_2^2(t, x))} G_t G_t' D_x \Gamma_2^2(t, x).$$

The functions $\Delta^i, \zeta^i, \eta^i, i = 1, 2$, satisfy the equations stated in Theorem 3.9 under Classes 1, 2, respectively. Define

$$\ell_Q^i(t, x, u) \doteq Q_t x.x + R_t u.u + 2m_t x + 2n_t u + \tilde{\ell}_1^i(t, x, u), \qquad i = 1, 2,$$

where

$$\tilde{\ell}_1^i(t, x, u) = \frac{2}{\theta} B_t u.\, (G_t G_t')^{-1} g_i(t, x), \qquad i = 1, 2.$$

For $i = 1, 2$ we wish to minimize over $u \in \hat{\mathcal{U}}$ the cost function

$$J_{\Sigma_{E_i}}^{\theta}(u(\cdot)) = E^u \left\{ \varphi_2^i(T, x_T) \exp \frac{\theta}{2} \left(\int_0^T \ell_Q^i(t, x_t, u_t)dt + (Q_T x_T.x_T + 2m_T x_T) \right) \right\},$$

where

$$\varphi_2^1(T, x) \doteq \frac{1}{\Gamma_2^1(t, x)} = \frac{1}{\Delta_t^1 x.x + x.\zeta_t^1 + \eta_t^1}, \tag{5.115}$$
$$\varphi_2^2(T, x) = \left\{ \gamma_T^1 \exp(\Gamma_2^2(t, x)) + \gamma_T^2 \exp(-\Gamma_2^2(t, x)) \right\}^{-1}.$$

In order to determine explicitly the optimal feedback control law corresponding to the control problem associated with systems $\Sigma_{E_i}, i = 1, 2$, we shall need the following equations.

Observer dynamics.

$$dr_t^i = \left\{ F_t - P_t^i \left(\Lambda_t^i - \theta Q_t \right) \right\} r_t^i dt + \left(f_t - P_t^i \sigma_t^i \right) dt + B_t u(t, y)dt \tag{5.116}$$
$$+\theta P_t^i m_t' dt + P_t^i H_t' N_t^{-1} d\hat{b}_t^u, \quad r^i(0) = \xi, \quad \hat{b}_i^u \equiv \text{Wiener process}, \quad i = 1, 2,$$

$$\dot{P}_t^i = F_t P_t^i + P_t^i F_t' - P_t^i \left(H_t' N_t^{-1} H_t - \theta[Q_t - \tfrac{\Lambda_t^i}{\theta}] \right) P_t^i \tag{5.117}$$
$$+ G_t G_t',\ P^i(0) = P_0, \quad i = 1, 2.$$

Control gains.

$$\begin{aligned}
&\dot{\Sigma}_t^i + \Sigma_t^i \left(F_t + \theta P_t^i [Q_t - \tfrac{\Lambda_t^i}{\theta}]\right) + \left(F_t' + \theta [Q_t - \tfrac{\Lambda_t^i}{\theta}] P_t^i\right) \Sigma_t^i + [Q_t - \tfrac{\Lambda_t^i}{\theta}] \\
&\quad - \Sigma_t^i \left\{B_t R_t^{-1} B_t' - \theta P_t^i H_t' N_t^{-1} H_t P_t^i\right\} \Sigma_t^i = 0, \quad i = 1,2, \qquad (5.118) \\
&\Sigma_T^i = \tfrac{1}{2} \left\{ \left(I - \theta Q_T P_T^i\right)^{-1} Q_T + Q_T \left(I - \theta P_T^i Q_T\right)^{-1} \right\}, \quad i = 1,2. \\
&\dot{k}_t^i + k_t^i \left(F_t + \theta P_t^i H_t' N_t^{-1} H_t P_t^i \Sigma_t^i + \theta P_t^i [Q_t - \tfrac{\Lambda_t^i}{\theta}] - B_t R_t^{-1} B_t' \Sigma_t^i\right) \\
&\quad + [m_t - \tfrac{\sigma_t^{i,*}}{\theta}] + \left(f_t' + \theta [m_t - \tfrac{\sigma_t^i}{\theta}] P_t^i - n_t R_t^{-1} B_t'\right) \Sigma_t^i = 0, \quad i = 1,2, \qquad (5.119) \\
&k_T^i = m_T \left(I - \theta P_T^i Q_T\right)^{-1}, \quad i = 1,2. \\
&\dot{\rho}_t^i + Tr\left(P_t^i H_t' N_t^{-1} H_t P_t^i \Sigma_t^i\right) + \theta k_t^i P_t^i H_t' N_t^{-1} H_t P_t^i k_t^{i,\prime} \\
&\quad + 2k_t^i \left(f_t + \theta P_t^i [m_t - \tfrac{\sigma_t^{i,\prime}}{\theta}]\right) - |R_t^{-\frac{1}{2}} \left(B_t' k_t^{i,\prime} + n_t'\right)|^2 = 0, \quad i = 1,2, \qquad (5.120) \\
&\rho_T^i = 0, \quad i = 1,2. \\
&I_{0,T}^i = \tfrac{1}{|I - \theta P_T^i Q_T|^{\frac{1}{2}}} \exp\left\{ \tfrac{\theta^2}{2} n_T \left(I - \theta P_T^i Q_T\right)^{-1} P_T^i n_T' \right. \\
&\quad \left. + \tfrac{\theta}{2} \int_0^T \left(Tr(P_t^i [Q_t - \tfrac{\Lambda_t^i}{\theta}]) - \tfrac{\delta_t^i}{\theta}\right) dt \right\}, \quad i = 1,2. \qquad (5.121)
\end{aligned}$$

Introduce the Riccati differential equation

$$\begin{aligned}
&\dot{S}_t^i + F_t' S_t^i + S_t^i F_t - S_t^i \left(B_t R_t^1 B_t' - \theta G_t G_t\right) S_t^i \\
&\quad + [Q_t - \tfrac{\Lambda_t^i}{\theta}] = 0, \quad S_T^i = Q_T, i = 1,2. \qquad (5.122)
\end{aligned}$$

Denote by $\tilde{\rho}(AB)$ the spectral radius of AB (where A, B are matrix-valued functions), and define

$$\theta^* \doteq \sup \left\{ \theta;\ P_t^i \geq 0, S_t^i \geq 0\ \forall t \in [0,T],\ \tilde{\rho}(P_t^i S_t^i) < \frac{1}{\theta}\ \forall t \in [0,T] \right\}. \qquad (5.123)$$

Whenever the functions $\Lambda^i(\cdot), \sigma^i(\cdot), \delta^i(\cdot)$ are set to zero, the above equations are identical to the equations associated with determining the optimal control for LEQG tracking problems specified by (3.51)–(3.53) (see [3, 6]).

COROLLARY 5.3 (exact optimal control laws). *Suppose $0 < \theta \leq \theta^*$. The optimal control law corresponding to control system $\Sigma_{E_i}, i = 1,2$, is given by*

$$u^{i,*}(t) = -R_t^{-1} B_t' \left(\Sigma_t^i r_t^i + k_t^{i,\prime}\right) - R_t^{-1} n_t', \quad i = 1,2,$$

where $r^i(\cdot) \equiv r^{i,u^*}(\cdot), P^i(\cdot)$ are given by (5.116), (5.117), *respectively, for $i = 1,2$. Furthermore, the optimal total cost associated with system $\Sigma_{E_i}, i = 1,2$, is given by*

$$J_{\Sigma_{E_i}}^{\theta}(u^{i,*}(\cdot)) = I_{0,T}^i \times \exp \frac{\theta}{2} \left(\Sigma^i(0) r^i(0).r^i(0) + 2k^i(0) r^i(0) + \rho^i(0)\right), \quad i = 1,2,$$

respectively.

Proof. This is a special case of Theorem 5.1. □

Remark 5.4. Proceeding along the lines of the derivation of Corollary 5.3, we could derive the analog of this corollary for the classes of nonlinear control systems identified in Theorem 3.11 under Class 1.

Next, we introduce an example with an explicit optimal control law.

Example 5.5 (control problem with rational polynomial nonlinearities). Suppose we are interested in the stochastic optimal control problem emerging from Example 3.10, Case 2 by setting $\Delta = 2, \alpha = 2$, $\eta = \frac{1}{2}$. Namely,

$$\begin{aligned} dx_t &= x_t dt + \frac{2x_t}{x_t^2 + \frac{1}{2}} dt + u(t,y)dt + dw_t, \qquad x(0) = 0 \in \Re, \\ dy_t &= x_t dt + db_t, \qquad y(0) = 0 \in \Re. \end{aligned} \tag{5.124}$$

The objective is to find the optimal control law $u \in \hat{\mathcal{U}}$ that minimizes the cost function

$$\begin{aligned} J^\theta(u(\cdot)) = E^u \Big\{ &\exp \tfrac{\theta}{2} \left(\int_0^T \left[Qx_t^2 + Ru(t,y)^2 \right] dt + Q_T x_T^2 \right) \\ &\times \exp \left(\int_0^T \left[u(t,y) \tfrac{2x_t}{x_t^2 + \frac{1}{2}} \right] dt + \ln(x_T^2 + \tfrac{1}{2})^{-1} \right) \Big\}. \end{aligned}$$

From Corollary 5.3 we deduce that the optimal control law is given by $u^*(t) = -R^{-1}, \Sigma_t r_t = -R^{-1} \left(1 - \theta S_t P_t\right) S_t r_t$, where $\Sigma(\cdot), S(\cdot), r(\cdot)$ satisfy appropriate equations.

5.2. Nonlinear dynamics integral cost. Control system ($\Sigma_{I_1}^u$). Suppose the dynamics and observations are those given under control system Σ_{E_1} defined by

$$\Sigma_{I_1}^u \doteq \left\{ \Sigma_{E_1}; g_1(t,x) \to g_1(t,x,u) \equiv g^u(t,x), \zeta_t^1 \to \zeta_t^u, \eta_t^1 \to \eta_t^u \right\},$$

and the objective is to minimize over $u \in \hat{\mathcal{U}}$ the cost function

$$\begin{aligned} J_{\Sigma_{I_1}^u}(u(\cdot)) = \tfrac{1}{2} E^u \Big\{ &\tfrac{Q_T x_T . x_T + m_T x_T}{\frac{1}{2}\Delta_T x_T . x_T + \zeta_T^u . x_T + \eta_T^u} \\ &+ \int_0^T \left(\tfrac{Q_t x_t . x_t + R_t u_t . u_t + 2 m_t x_t + 2 n_t u_t}{\frac{1}{2}\Delta_t x_t . x_t + \zeta_t^u . x_t + \eta_t^u} \right) dt \Big\}. \end{aligned}$$

As in section 5.1 we shall show, by using the results of section 4, that the control problem associated with $\Sigma_{I_1}^u$ yields an explicit optimal feedback control law. To this end we introduce the following equations:

$$\dot{\Sigma}_t + \Sigma_t F_t + F_t' \Sigma_t - \Sigma_t B_t R_t^{-1} B_t' \Sigma_t + Q_t = 0, \quad \Sigma_T = Q_T, \tag{5.125}$$

$$\dot{k}_t + k_t \left(F_t - B_t R_t^{-1} B_t' \Sigma_t \right) + m_t + \left(f_t' - n_t R_t^{-1} B_t' \right) \Sigma_t = 0, \quad k_T = m_T, \tag{5.126}$$

$$\dot{\rho}_t + Tr \left(P_t H_t' N_t^{-1} H_t P_t \Sigma_t \right) + 2 k_t f_t - |R_t^{-\frac{1}{2}} \left(B_t' k_t' + n_t' \right)|^2 = 0, \quad \rho_T = 0. \tag{5.127}$$

COROLLARY 5.6 (exact optimal control laws). *The optimal control law corresponding to system $\Sigma_{I_1}^u$ is given by*

$$u^*(t) = -R_t^{-1} B_t' \left(\Sigma_t r_t + k_t' \right) - R_t^{-1} n_t'.$$

Here $\Sigma(\cdot), k(\cdot)$ are given by (5.125), (5.126), *respectively, and $r(\cdot) \equiv r^{u^*}(\cdot)$ is given by*

$$dr_t = F_t r_t dt + f_t dt + B_t u^*(t) dt + P H_t' N_t^{-1} d\hat{b}_t, \qquad r(0) = \xi, \qquad \hat{b}_t \doteq \textit{Wiener process.}$$

Proof. This is a special case of Theorem 5.2. ∎

5.3. Nonlinear dynamics exponential-of-quadratic cost. Now, specialize the cost function (3.91) (i.e., (2.26)) to that corresponding to an exponential-of-integral cost by consider the following family of control systems.

Control system (Σ_{EQ}). Consider the dynamics and observations given by

$$\begin{aligned} dx_t &= (F_t x_t + g(t,x_t) + f_t)\,dt + B_t u(t,y)dt + G_t dw_t, \qquad x(0) \in \Re^n, \\ dy_t &= (H_t x_t + h_t)\,dt + N_t^{\frac{1}{2}} db_t, \qquad y(0) = 0 \in \Re^d. \end{aligned} \tag{5.128}$$

Here

$$g(t,x) = \frac{\gamma_t^1 \exp(\Gamma_2(t,x_t)) - \gamma_t^2 \exp(-\Gamma_2(t,x_t))}{\gamma_t^1 \exp(\Gamma_2(t,x_t)) + \gamma_t^2 \exp(-\Gamma_2(t,x_t))} G_t G_t' (\Delta_t x_t + \zeta_t),$$

$$\Gamma_2(t,x) \doteq \frac{1}{2}\Delta_t x.x + x.\zeta_t + \eta_t,$$

and $\Delta(\cdot), \zeta(\cdot), \eta(\cdot)$ satisfy the equations of Theorem 3.9 under Class 2.

The cost function to be minimized over $u \in \hat{\mathcal{U}}$ is

$$J^\theta_{\Sigma_{EQ}}(u(\cdot)) = E^u \Big\{ \exp\tfrac{\theta}{2} \Big(\int_0^T \Big[Q_s x_s.x_s + R_s u_s.u_s + 2m_s x_s + 2n_s u_s + \tilde{\ell}_1(s,x_s,u_s) \Big] ds + (Q_T x_T.x_T + 2m_T x_T) \Big) \Big\}, \qquad \tilde{\ell}_1(t,x,u) = B(t,u).(G_t' G_t)^{-1} g(t,x),$$

and $Q_t = Q_t' \geq 0, R_t = R_t' > 0$.

Define

$$\begin{aligned} \tilde{\varphi}_1^{EQ}(x,T) \doteq \int_{\Re^n} &\big\{ \gamma_T^1 \exp\big(\tfrac{1}{2}\Delta_T z.z + z.\zeta_T + \eta_T\big) + \gamma_T^2 \exp\big(-\tfrac{1}{2}\Delta_T z.z - z.\zeta_T - \eta_T\big) \big\} \\ &\times \exp\tfrac{\theta}{2}(Q_T z.z + 2m_T z) \times \tfrac{1}{(2\pi)^{\frac{n}{2}} |P_T|^{\frac{1}{2}}} \exp\big(-\tfrac{1}{2} P_T^{-1}(z-x).(z-x)\big)\, dz. \end{aligned} \tag{5.129}$$

As before, we know that the partially observable stochastic control problem Σ_{EQ} is equivalent to the following finite-dimensional, completely observable control problem.

Cost function. Minimize over $u \in \hat{\mathcal{U}}$ the cost function

$$\begin{aligned} J^\theta_{\Sigma_{EQ}}(u(\cdot)) = \hat{E}^u \Big\{ &\tilde{\varphi}_1^{EQ}(r_T,T) \times \exp\Big(\tfrac{\theta}{2} \int_0^T Tr\big(P_s[Q_s - \tfrac{\Lambda_s}{\theta}]\big)\, ds \Big) \\ &\times \exp\Big(\tfrac{\theta}{2} \int_0^T \big(r_s[Q_s - \tfrac{\Lambda_s}{\theta}].r_s + R_s u(s,y).u(s,y) \big)\, ds \Big) \\ &\times \exp\Big(\tfrac{\theta}{2} \int_0^T \big(2r_s.[m_s' - \tfrac{\sigma_s}{\theta}] + 2[n_s u(s,y) - \tfrac{\delta_s}{\theta}] \big)\, ds \Big) \Big\} \end{aligned}$$

subject to the following dynamics.

Observer dynamics.

$$\begin{aligned} dr_t = &\{F_t - P_t(\Lambda_t - \theta Q_t)\} r_t dt + (f_t - P_t \sigma_t)\, dt \\ &+ B_t u(t,y)dt + \theta P_t m_t' dt + P_t H_t' N_t^{-1} d\hat{b}_t^u, \quad r(0) = \xi, \end{aligned} \tag{5.130}$$

$$\begin{aligned} \dot{P}_t = &F_t P_t + P_t F_t' - P_t \big(H_t' N_t^{-1} H_t + \Lambda_t - \theta Q_t \big) P_t \\ &+ G_t G_t', \quad P(0) = P_0. \end{aligned} \tag{5.131}$$

Define

$$Q_T^+ \doteq Q_T + \frac{1}{\theta}\Delta_T, \qquad m_T^+ \doteq m_T + \frac{1}{\theta}\zeta_T, \tag{5.132}$$

$$Q_T^- \doteq Q_T - \frac{1}{\theta}\Delta_T, \qquad m_T^- \doteq m_T - \frac{1}{\theta}\zeta_T, \tag{5.133}$$

$$\tilde{\Sigma}_T^+ \doteq \left(I - \theta P_T Q_T^+\right)^{-1}, \quad \tilde{\Sigma}_T^- \doteq \left(I - \theta P_T Q_T^-\right)^{-1}. \tag{5.134}$$

Use the normalization property of Gaussian densities to deduce

$$\begin{aligned}\tilde{\varphi}_1^{EQ}(x,T) = \hat{\gamma}_T^+ \times \exp\tfrac{\theta}{2}\left(\tilde{\Sigma}_T^+ x.\left(Q_T^+ x + 2m_T^{+,\prime}\right)\right) \\ + \hat{\gamma}_T^- \times \exp\tfrac{\theta}{2}\left(\tilde{\Sigma}_T^- x.\left(Q_T^- x + 2m_T^{-,\prime}\right)\right),\end{aligned} \tag{5.135}$$

where

$$\begin{aligned}\hat{\gamma}_T^+ &= \gamma_T^1 |\tilde{\Sigma}_T^+|^{\frac{1}{2}} \exp\frac{\theta^2}{2}\left(m_T^+ \tilde{\Sigma}_T^+ P_T m_T^{+,\prime} + \eta_T\right), \\ \hat{\gamma}_T^- &= \gamma_T^2 |\tilde{\Sigma}_T^-|^{\frac{1}{2}} \exp\frac{\theta^2}{2}\left(m_T^- \tilde{\Sigma}_T^- P_T m_T^{-,\prime} - \eta_T\right).\end{aligned}$$

Denoting the optimal cost-to-go corresponding to the total cost $J^{\theta}_{\Sigma_{EQ}}(u(\cdot))$ by $S_{EQ}(\cdot)$ and defining

$$I_{t,T}^{EQ} \doteq \exp\frac{\theta}{2}\left\{\int_t^T Tr\left(P_s\left[Q_s - \frac{\Lambda_s}{\theta}\right]\right) ds\right\}$$

for each $u \in \hat{\mathcal{U}}$, the cost-to-go is now given as follows.

Exponential-of-integral cost-to-go.

$$\begin{aligned}S_{EQ}(r,t) = \tfrac{1}{I_{t,T}^{EQ}} \inf_{u\in\hat{\mathcal{U}}} \hat{E}^u \left\{\tilde{\varphi}_1^{EQ}(r_T,T) \times \exp\tfrac{\theta}{2}\left(\int_t^T r_s[Q_s - \tfrac{\Lambda_s}{\theta}].r_s ds\right)\right. \\ \left. \times \exp\tfrac{\theta}{2}\left(\int_t^T \left(R_s u_s.u_s + 2r_s.[m_s - \tfrac{\sigma_s}{\theta}] + 2[n_s u_s - \tfrac{\delta_s}{\theta}]\right) ds\right) |\mathcal{F}_{0,t}^y\right\}\end{aligned} \tag{5.136}$$

subject to observer dynamics given as follows.

Observer dynamics.

$$\begin{aligned}dr_t = \left\{F_t - P_t\left(\Lambda_t - \theta Q_t\right)\right\} r_t dt + \left(f_t - P_t\sigma_t\right) dt \\ + B_t u(t,y) dt + \theta P_t m_t' dt + P_t H_t' N_t^{-1} d\hat{b}_t^u, \quad r(0) = \xi.\end{aligned} \tag{5.137}$$

Formally, the function $S_{EQ}(\cdot)$ satisfies the second-order HJ equations

$$\begin{aligned}\tfrac{\partial}{\partial t} S_{EQ}(r,t) + \tilde{A}^{\theta}(t) S_{EQ}(r,t) \\ + \tfrac{\theta}{2}\left\{r[Q - \tfrac{\Lambda}{\theta}].r + 2r.[m' - \tfrac{\sigma}{\theta}] - 2\tfrac{\delta_s}{\theta}\right\} S_{EQ}(r,t) \\ + \mathcal{H}^{\theta}(r, D_r S_{EQ}(r,t), D_r S_{EQ}(r,t)) = 0, \qquad \Re^n \times [0,T),\end{aligned} \tag{5.138}$$

with terminal condition

$$\begin{aligned}S_{EQ}(r,T) = \tilde{\varphi}_1^{EQ}(r,T) = \hat{\gamma}_T^+ \times \exp\tfrac{\theta}{2}\left(\tilde{\Sigma}_T^+ r.\left(Q_T^+ r + 2m_T^{+,\prime}\right)\right) \\ + \hat{\gamma}_T^- \times \exp\tfrac{\theta}{2}\left(\tilde{\Sigma}_T^- r.\left(Q_T^- r + 2m_T^{-,\prime}\right)\right).\end{aligned} \tag{5.139}$$

Here

$$\tilde{F}_t \doteq F_t - P_t\left(\Lambda_t - \theta Q_t\right), \qquad \tilde{\alpha}_t \doteq P_t H_t' N_t^{-1} H_t P_t,$$

and for $\Phi \in C_x^2(\Re^n)$

$$\tilde{A}^\theta(t)\Phi(x) = \frac{1}{2}Tr\left(\tilde{\alpha}_t D_x^2\Phi(x)\right) + \left(\tilde{F}_t x + \theta P_t m_t' + f_t - P_t\sigma_t\right).D_x\Phi(x),$$
$$\mathcal{H}^\theta(x,p,s) = \inf_{u\in\Re^m}\left\{p.Bu + \frac{\theta}{2}\left(Ru.u + 2nu\right)s\right\}.$$

Hence, the total optimal cost corresponding to control system Σ_{EQ} is obtained from

$$J^\theta_{\Sigma_{EQ}}(u^*(\cdot)) = \inf_{u\in\hat{\mathcal{U}}} J^\theta_{\Sigma_{EQ}}(u(\cdot)) = I^{EQ}_{0,T}S_{EQ}(r(0),0). \tag{5.140}$$

If the terminal condition $S_{EQ}(T,r)$ is an exponential-of-quadratic function of r, the above HJ equations can be solved explicitly to yield optimal controls which are of linear feedback form. Several attempts to solve explicitly the HJ equation (5.138), (5.139) have been unsuccessful. For this reason we shall seek suboptimal control laws.

COROLLARY 5.7. *Consider the HJ equation* (5.138), (5.139) *corresponding to the control system* Σ_{EQ}*, and suppose*

$$\begin{gathered} m = 0, \quad \eta = 0, \quad \zeta = 0, \quad \Delta_T \geq 0, \\ Q_T - \tfrac{1}{\theta}\Delta_T \geq 0, \quad I - \theta P_T(Q_T + \tfrac{1}{\theta}\Delta_T) > 0, \quad \theta > 0. \end{gathered} \tag{5.141}$$

Denoting by $S_{EQ^-}, S_{EQ}, S_{EQ^+}$ *the solutions of* (5.138) *corresponding to the terminal cost functions* $\tilde{\varphi}_1^{EQ^+}, \tilde{\varphi}_1^{EQ}, \tilde{\varphi}_1^{EQ^-}$*, respectively, defined by*

$$\tilde{\varphi}_1^{EQ^+}(x,P_T) \doteq \left(\hat{\gamma}_T^+ + \hat{\gamma}_T^-\right) \times \exp\frac{\theta}{2}\left(x.\tilde{\Sigma}_T^+ Q_T^+ x\right), \tag{5.142}$$

$$\tilde{\varphi}_1^{EQ}(x,P_T) \doteq \hat{\gamma}_T^+ \times \exp\frac{\theta}{2}\left(x.\tilde{\Sigma}_T^+ Q_T^+ x\right) + \hat{\gamma}_T^- \times \exp\frac{\theta}{2}\left(x.\tilde{\Sigma}_T^- Q_T^- x\right), \tag{5.143}$$

$$\tilde{\varphi}_1^{EQ^-}(x,P_T) \doteq \left(\hat{\gamma}_T^+ + \hat{\gamma}_T^-\right) \times \exp\frac{\theta}{2}\left(x.\tilde{\Sigma}_T^- Q_T^- x\right), \tag{5.144}$$

where

$$\hat{\gamma}_T^+ = \gamma_T^1|\tilde{\Sigma}_T^+|^{\frac{1}{2}}, \qquad \hat{\gamma}_T^- = \gamma_T^2|\tilde{\Sigma}_T^-|^{\frac{1}{2}},$$

we have the following bounds:

$$S_{EQ^-}(r,t) \leq S_{EQ}(r,t) \leq S_{EQ^+}(r,t) \qquad \forall (r,t) \in \Re^n \times [0,T]. \tag{5.145}$$

Furthermore,

$$I^{EQ}_{0,T}S_{EQ^-}(r(0),0) \leq J^\theta_{\Sigma_{EQ}}(u^*(\cdot)) \leq I^{EQ}_{0,T}S_{EQ^+}(r(0),0), \tag{5.146}$$

and the suboptimal control laws obtained by solving the HJ equations associated with S_{EQ^-}, S_{EQ^+} *are similar to that of the LEQG tracking problem.*

Proof. This is a direct consequence of Theorem 5.1, part II, which is obtained as follows: from (5.135) and (5.141) we deduce (5.143). Using (5.132)–(5.135) and (5.141) we deduce

$$\begin{gathered} Q_T^+ \geq Q_T^-, \qquad \tilde{\Sigma}_T^+ \geq \tilde{\Sigma}_T^-, \\ \tilde{\varphi}_1^{EQ^-}(r,T) \leq \tilde{\varphi}_1^{EQ}(r,T) = S_{EQ}(r,T) \leq \tilde{\varphi}_1^{EQ_2^+}(r,T). \end{gathered} \tag{5.147}$$

From the theory of dominating solutions of PDEs we derive (5.145). Consequently, we establish (5.146). ∎

Remark 5.8. Lower and upper bounds, such as the ones derived above, can be derived for other nonlinear drift terms $g(\cdot)$ which admit finite-dimensional solutions of the Feynman–Kac information state equation.

6. Conclusion. In general, nonlinear partially observable stochastic optimal control problems have an infinite-dimensional state space. In this paper we have presented an approach for treating systems with nonlinearities which enter the unobservable dynamics as gradients of potential functions, and the observations as quadratic functions, of the unobservable state. When the observations are linear in the unobservable state, sufficient conditions are given to compute optimal control laws explicitly, along the lines of LEQG/LQG tracking problems.

When the cost function is either quadratic or an exponential-of-quadratic function of x and u, we have shown that finite-dimensional sufficient statistics are available, provided the nonlinearities entering the unobservable dynamics are gradients of potential functions and satisfy a generalized version of the Riccati equation. In addition, suboptimal linear feedback control laws are derived for nonlinearities satisfying "sector criteria."

Acknowledgments. The work of this paper arose from the authors' discussion concerning related work pursued independently, at the 33rd IEEE Conference on Decision and Control (1994). The anonymous referees and the associate editor, Professor S. Shreve, have provided several suggestions which improved both the presentation and the technical content of this manuscript.

REFERENCES

[1] W. M. Wonham, *On the separation theorem of stochastic control*, SIAM J. Control, 6 (1968), pp. 312–326.

[2] A. Bensoussan, *Stochastic Control of Partially Observable Systems*, Cambridge University Press, Cambridge, UK, 1992.

[3] A. Bensoussan and J. H. van Schuppen, *Optimal control of partially observable stochastic systems with an exponential-of-integral performance index*, SIAM J. Control Optim., 23 (1985), pp. 599–613.

[4] A. Bensoussan and R. Elliott, *A finite dimensional risk sensitive control problem*, SIAM J. Control Optim., 33 (1996), pp. 1834–1846.

[5] C. Charalambous and J. Hibey, *Minimum principle for partially observable nonlinear risk-sensitive control problems using measure-valued decompositions*, Stochastics Stochastics Rep., 57 (1996), pp. 247–288.

[6] C. Charalambous, *Partially observable nonlinear risk-sensitive control problems: Dynamic programming and verification theorems*, IEEE Trans. Automat. Control, 42 (1997), pp. 1130–1138.

[7] C. Charalambous and R. J. Elliott, *Risk-Sensitive Control Problems and Dynamic Games Featuring the Lur'e-Postnikov Lyapunov Function*, preprint, 1996.

[8] E. Pardoux, *Non-linear filtering, prediction and smoothing*, in Stochastic Systems: The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J. C. Willems, eds., Proceedings of the NATO Advanced Study Institute, D. Reidel, Dordrecht, the Netherlands, 1981.

[9] C. Charalambous, D. Naidu, and K. Moore, *Solvable risk-sensitive control problems with output feedback*, in Proc. 33rd IEEE Conference on Decision and Control, Lake Buena Vista, FL, 1994, pp. 1433–1434.

[10] A. Bensoussan and R. Elliott, *General finite dimensional risk sensitive problems and small noise limits*, IEEE Trans. Automat. Control, 41 (1996), pp. 210–215.

[11] V. Benes, *Exact finite-dimensional filters for certain diffusions with nonlinear drift*, Stochastics, 5 (1981), pp. 65–92.

[12] T. Basar and P. Bernhard, *$\mathcal{H}^\infty$-Optimal Control and Minimax Design Problems*, Birkhäuser, Boston, Basel, Berlin, 1991.

[13] R. S. Liptser and A. N. Shiryayev, *Statistics of Random Processes*, Vol. 1, Springer-Verlag, New York, 1977.

[14] V. E. Benes and I. Karatzas, *On the relation of Zakai's and Mortensen's equations*, SIAM J. Control Optim., 21 (1983), pp. 472–489.

[15] A. Friedman, *Stochastic Differential Equations and Applications*, Academic Press, New York, 1975.

[16] C. Charalambous, *The role of informations state and its adjoint in relating nonlinear output feedback risk-sensitive control and dynamic games*, IEEE Trans. Automat. Control, 42 (1997), pp. 1163–1170.
[17] H. K. Khalil, *Nonlinear Systems*, Macmillan, New York, 1992.
[18] D. Stroock and S. Varadhan, *Multidimensional Diffusion Processes*, Springer-Verlag, Berlin, 1979.
[19] J. S. Baras, G. L. Blankenship, and W. E. Hopkins, *Existence, uniqueness, and asymptotic behavior of solutions to a class of Zakai equations with unbounded coefficients*, IEEE Trans. Automat. Control, 28 (1983), pp. 203–214.
[20] V. Benes and L. Shepp, *Wiener integral associated with diffusion processes*, Teor. Veroyatnost. i Premenen., 3 (1968), pp. 498–501.
[21] U. Haussmann and E. Pardoux, *A conditionally almost linear filtering problem with non-gaussian initial condition*, Stochastics, 23 (1988), pp. 241–275.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 2, pp. 579–589, March 1998

006

A FREE BOUNDARY PROBLEM IN $\mathbb{R}^d$ WITH BOTH SMOOTH AND NONSMOOTH FIT*

J. R. DORROH† AND GUILLERMO FERREYRA†

Abstract. A deterministic infinite-horizon singular control problem with unbounded control set is solved completely. The methods used here are those of dynamic programming and viscosity solutions. The novelty is that the value function is convex, C^1 along a piece of the free boundary and not C^1 along another piece of it.

Key words. singular control, viscosity solution, Bellman equation, smooth fit

AMS subject classifications. 35F30, 49L20, 49L25

PII. S0363012996301440

1. Introduction. We consider the following optimal control problem. Let $d \geq 2$ and let A_d be the space of antisymmetric $d \times d$ matrices a endowed with the norm $|a| = \sqrt{\frac{1}{2}\text{trace}(aa')}$, where a' denotes the transpose of a. The system to be controlled is the bilinear system

$$\dot{x} = a(t)x, \qquad x(0) = x \in \mathbb{R}^d, \tag{1.1}$$

where the control $a(\cdot) \in \mathcal{A}_d$, the space of measurable functions of time valued in A_d. The cost function is defined by

$$v^a(x) = \int_0^\infty e^{-t} \left[\langle x(t), b\rangle + |a(t)|\right] dt, \tag{1.2}$$

where $b \in \mathbb{R}^d \setminus \{0\}$ is fixed throughout and $\langle \cdot, \cdot \rangle$ is the inner product in $\mathbb{R}^d$, and the value function is

$$v(x) = \inf\{v^a(x) \colon a(\cdot) \in \mathcal{A}_d\}. \tag{1.3}$$

We show that v is a convex, Lipschitz viscosity solution of the *free boundary problem*

$$\max(u(x) - \langle b, x\rangle, \lambda(x, \nabla u(x)) - 1) = 0, \qquad x \in \mathbb{R}^d, \tag{1.4}$$

where $\langle \cdot, \cdot \rangle$ denotes the euclidean inner product on $\mathbb{R}^d$ and

$$\lambda(x, p) = |x|\,|p| \sin\theta, \tag{1.5}$$

with θ the angle between x and p. We shall see that the free boundary—the hypersurface where both terms in (1.4) equal zero—consists of two $(d-1)$-dimensional manifolds F_0 and F_1. The manifold F_0 is part of the cylinder in $\mathbb{R}^d$ with axis through b and defining equation

$$\lambda(x, b) = 1. \tag{1.6}$$

*Received by the editors April 3, 1996; accepted for publication (in revised form) January 8, 1997. http://www.siam.org/journals/sicon/36-2/30144.html

†Department of Mathematics, Louisiana State University, Baton Rouge, LA 70808 (dorroh@math.lsu.edu, ferreyra@math.lsu.edu).

The manifold F_1 is harder to describe; that is done in section 3. The value function is C^1 along F_0 but it is not C^1 along F_1. Besides this, the value function is not C^1 along a ray L in the direction of b; see Figure 2. We will also show that the optimal control is either zero or impulsive. In this problem optimal impulsive controls have delicate behavior. In fact, in the region J where the optimal control is impulsive, the optimal trajectories jump along integral curves of (1.1) determined by the feedback control

$$\mathbf{a}(x(t)) = x(t) \wedge \nabla v(x(t)).$$

These control functions turn out to be constant along the trajectories they determine. When we let $x(0)$ vary in J, the optimal impulsive trajectories end up along all of F_0, the only part of the free boundary along which the value function is C^1. The methods used in this paper are those of dynamic programming and viscosity solutions of the Bellman equation. These methods were also used in [8]. The remarkable difference between the behavior of the value function along the free boundary in [8] and that presented here is the existence in our case (1.1)–(1.4) of the piece of the free boundary F_1 along which the value function is not C^1. Moreover, our value function is convex. All stochastic singular control problems with a convex value function found in the literature (see [11, p. 332]) possess the additional property that the value function is C^2 (smooth fit property) along the free boundary. We conjecture that stochastic versions of our problem can provide examples of stochastic singular control problems with a convex value function and nonsmooth fit. Other work dealing with the question of smooth fit can be found in [1], [5], [6], [9], [10], [15], [16], and [17]. The problem (1.1)–(1.4) is somewhat related to the euclidean elastica problem (cf. [13]) if one remembers that the Frenet–Serret formulas for curves in $\mathbb{R}^3$ are a bilinear system with an antisymmetric matrix, as is (1.1) (cf. [2, p. 303]). In such a case, the control system (1.1) describes general smooth curves in $\mathbb{R}^3$ with the matrix $a(t)$ controlling the curvature and the torsion of the curve. The elastica problem is finite horizon with $b = 0$.

2. Derivation of the Bellman equation. If $\xi \in \mathbb{R}^d$, then $|\xi|$ denotes its euclidean length, if $a \in A_d$, then $|a|^2 = \frac{1}{2}\text{trace}(aa') = \sum_{i<j} a_{ij}^2$, and if $a(\cdot) \in \mathcal{A}_d$, then $||a(\cdot)|| = \text{ess sup}\{|a(t)|\colon t \geq 0\}$.

If ξ, η are two vectors in $\mathbb{R}^d \setminus \{0\}$, then the angle between them is given by $\theta = \cos^{-1}(\langle \xi/|\xi|, \eta/|\eta| \rangle) \in [0, \pi]$.

The wedge product $\xi \wedge \eta$ is the antisymmetric $n \times n$ matrix defined by

$$\xi \wedge \eta = \xi\eta' - \eta\xi',$$

and we have

$$\begin{aligned} |\xi \wedge \eta|^2 &= \frac{1}{2}\text{trace}((\xi \wedge \eta)(\xi \wedge \eta)') \\ &= \sum_{i<j} (\xi_i\eta_j - \xi_j\eta_i)^2. \end{aligned}$$

An easy computation shows that $|\xi \wedge \eta|^2 + |\langle \xi, \eta \rangle|^2 = |\xi|^2|\eta|^2$ and hence $|\xi \wedge \eta| = |\xi|\,|\eta| \sin\theta = \lambda(\xi, \eta)$, where θ is the angle between ξ and η.

For $\epsilon > 0$ set

$$v_\epsilon(x) = \inf\{v^a(x)\colon\ \epsilon||a(\cdot)|| \leq 1, a(\cdot) \in \mathcal{A}_d\}.$$

Then an easy approximation argument shows that

$$v(x) = \inf_{\epsilon>0} v_\epsilon(x). \tag{2.1}$$

LEMMA 1. *$v_\epsilon \to v$ uniformly on compact subsets of $\mathbb{R}^d$ as $\epsilon \downarrow 0$ and v, v_ϵ are convex and Lipschitz on $\mathbb{R}^d$ with Lipschitz constant $|b|$ for all $\epsilon > 0$. Moreover, $v(x)$ and $v_\epsilon(x)$ are bounded above by $\langle b, x\rangle$ and below by $-|b|\,|x|$.*

Proof. Let $\Phi(t,s)$ denote the fundamental solution of (1.1) corresponding to a given control $a(\cdot)$. Since $(d/dt)|x(t)|^2 = 2\langle x(t), a(t)x(t)\rangle = 0$, we have $|x(t)| = |x|$, which yields $||\Phi(t,0)|| = 1$ (operator norm). Since

$$\nabla v^a(x) = \int_0^\infty e^{-t}\Phi'(t,0)b\,dt,$$

it follows that $|\nabla v^a(x)| \le |b|$. Since v and v_ϵ are infima of v^a, they are Lipschitz with constant $|b|$.

Let $a(\cdot)$ be a bounded control and suppose $x_\epsilon \to x$. Then

$$v^a(x) = \limsup_{\epsilon\downarrow 0} v^a(x_\epsilon) \ge \limsup_{\epsilon\downarrow 0} v_\epsilon(x_\epsilon),$$

and so $v(x) \ge \limsup_{\epsilon\downarrow 0} v_\epsilon(x_\epsilon)$. By (2.1) we have

$$\liminf_{\epsilon\downarrow 0} v_\epsilon(x_\epsilon) \ge \liminf_{\epsilon\downarrow 0} v(x_\epsilon) \ge v(x).$$

This establishes continuous convergence. Then the local uniform convergence follows; see [4, p. 268]. The last part follows from the fact that $v^0(x) = \langle b, x\rangle$ for all x and from $v(0) = 0$.

Finally, to prove that v is convex, let $\alpha \in [0,1]$ and let $x^\alpha = (1-\alpha)x^0 + \alpha x^1$ be a convex combination of initial states for (1.1). Let $\epsilon > 0$ and let $a_i(\cdot)$ be controls satisfying $v^{a_i}(x^i) \le v(x^i) + \epsilon, i = 0, 1$. Let $a_\alpha = (1-\alpha)a_0 + \alpha a_1$. Then the corresponding solutions of (1.1) satisfy $x^\alpha(t) = (1-\alpha)x^0(t) + \alpha x^1(t)$, and the convexity of (1.2) implies

$$\begin{aligned} v(x^\alpha) &\le v^{a_\alpha}(x^\alpha) \le (1-\alpha)v^{a_0}(x^0) + \alpha v^{a_1}(x^1) \\ &\le (1-\alpha)v(x^0) + \alpha v(x^1) + \epsilon. \end{aligned}$$

Since ϵ was arbitrary, this proves that v is convex. □

We now derive the Bellman equation satisfied by v_ϵ. For the concept of "viscosity solution," see [3].

LEMMA 2. *For all $\epsilon > 0$, v_ϵ is a viscosity solution of*

$$\frac{1}{\epsilon}H(x, \nabla v_\epsilon) + v_\epsilon - \langle x, b\rangle = 0, \qquad x \in \mathbb{R}^d, \tag{2.2}$$

where

$$\begin{aligned} H(x,p) &= \sup\{-\langle ax, p\rangle - |a| \colon\ |a| \le 1, a \in A_d\} \\ &= (|p \wedge x| - 1)^+ = (\lambda(x,p) - 1)^+. \end{aligned} \tag{2.3}$$

Proof. We start with the dynamic programming principle [11], which states that for each $T > 0$

$$v_\epsilon(x) = \inf\left\{\int_0^T e^{-t}[\langle x(t), b\rangle + |a(t)|]dt + e^{-T}v_\epsilon(x(T))\right\}, \tag{2.4}$$

where the infimum is over all controls $a(\cdot) \in \mathcal{A}_d$ satisfying $\epsilon||a(\cdot)|| \le 1$.

Now suppose $\phi \in C^1(\mathbb{R}^d)$ and $x \in \mathbb{R}^d$ with $\phi(x) = v_\epsilon(x)$ and $v_\epsilon - \phi \leq 0$ near x. Then (2.4) yields for all constant $a \in A_d$ satisfying $\epsilon|a| \leq 1$,

$$\phi(x) \leq \int_0^T e^{-t} \left[\langle x(t), b\rangle + |a|\right] dt + e^{-T}\phi(x(T)),$$

which implies by the chain rule and the fundamental theorem of calculus

$$0 \leq \frac{1}{T}\int_0^T e^{-t} \left[\langle x(t), b\rangle + |a| - \phi(x(t)) + \langle \nabla\phi(x(t)), ax(t)\rangle\right] dt;$$

letting $T \downarrow 0$ and taking the supremum over a we obtain

$$\frac{1}{\epsilon}H(x, \nabla\phi(x)) + \phi(x) - \langle b, x\rangle \leq 0.$$

On the other hand, suppose x and $\phi \in C^1(\mathbb{R}^d)$ are such that $\phi(x) = v_\epsilon(x)$ and $v_\epsilon - \phi \geq 0$ near x. Since

$$\sup_{\epsilon||a||\leq 1} \left(\sup_{0\leq t\leq T} |x(t) - x|\right) \to 0 \qquad \text{as } T \downarrow 0, \tag{2.5}$$

(2.4) implies, for $T > 0$ sufficiently small,

$$0 \geq \inf_{\epsilon||a||\leq 1} \left\{\frac{1}{T}\int_0^T e^{-t} \left[\langle x(t), b\rangle + |a(t)| - \phi(x(t)) + \langle \nabla\phi(x(t)), a(t)x(t)\rangle\right] dt\right\}.$$

Now (2.5) allows us to pass to the limit $T \downarrow 0$ and obtain

$$\frac{1}{\epsilon}H(x, \nabla\phi(x)) + \phi(x) - \langle b, x\rangle \geq 0.$$

The maximization in (2.3) is carried out as follows. Let

$$f(a) = -\langle ax, p\rangle - |a| = \sum_{i<j} a_{ij}(x_i p_j - x_j p_i) - |a|.$$

Then by the Cauchy–Schwarz inequality

$$f(a) \leq |a|(|x \wedge p| - 1).$$

Therefore, if $|x \wedge p| \leq 1$, we have $\max f(a) = 0$, attained (not uniquely) at $a = 0$. If $|x \wedge p| > 1$, then $\max f(a) = |x \wedge p| - 1$, attained (uniquely) when

$$a_{ij} = \frac{x_i p_j - x_j p_i}{|x \wedge p|}. \qquad \square$$

THEOREM 1. *v is a Lipschitz viscosity solution of* (1.4).

Proof. Lemma 1 states that v is Lipschitz. Let $x \in \mathbb{R}^d$ and $\phi \in C^1$ be such that $v - \phi$ has a local maximum at x. Then [3, Theorem 1.1, Lemma 1.1] there exists $x_\epsilon \to x$ such that $v_\epsilon - \phi$ has a local maximum at x_ϵ. This implies, by (2.2),

$$\frac{1}{\epsilon}H(x_\epsilon, \nabla\phi(x_\epsilon)) + v_\epsilon(x_\epsilon) - \langle x_\epsilon, b\rangle \leq 0.$$

Since $H \geq 0$ we obtain $v_\epsilon(x_\epsilon) - \langle x_\epsilon, b\rangle \leq 0$; letting $\epsilon \downarrow 0$ yields $v(x) - \langle x, b\rangle \leq 0$. Also multiplying by ϵ and sending $\epsilon \downarrow 0$ yields $H(x, \nabla\phi(x)) \leq 0$. By Lemma 2, we obtain $\lambda(x, \nabla\phi(x)) - 1 \leq 0$. Thus v is a subsolution of (1.4).

Let x and $\phi \in C^1$ be such that $v - \phi$ has a local minimum at x. Choose [3] $x_\epsilon \to x$ such that $v_\epsilon - \phi$ has a local minimum at x_ϵ. Then by (2.2)

$$\frac{1}{\epsilon} H(x_\epsilon, \nabla\phi(x_\epsilon)) + v_\epsilon(x_\epsilon) - \langle x_\epsilon, b\rangle \geq 0.$$

Now if $v(x) - \langle x, b\rangle \geq 0$ then v is a supersolution of (1.4). If not, then it follows that $H(x_\epsilon, \nabla\phi(x_\epsilon)) > 0$, which by Lemma 2 implies $\lambda(x_\epsilon, \nabla\phi(x_\epsilon)) - 1 > 0$, which yields in the limit $\lambda(x, \nabla\phi(x)) - 1 \geq 0$. Thus v is a supersolution of (1.4). □

3. Solution of the free boundary problem. In this section we explicitly construct a candidate value function U that is Lipschitz on $\mathbb{R}^d$ and that solves (1.4) in the classical sense except for lower-dimensional submanifolds of $\mathbb{R}^d$. To gain some intuition on the construction of U, note that from (1.2) we can deduce that if a nonzero control $a(\cdot)$ is optimal along a piece of a trajectory of (1.1), then the inner product $\langle x(\cdot), b\rangle$ must not increase along $x(\cdot)$ and, in the regions where $a = 0$ is optimal, $\langle x(\cdot), b\rangle$ should not be too large. The construction of U is divided into three steps.

Step 1. Let D be the closed region bounded by the half of the cylinder (1.6) in the direction of $-b$:

$$D = \{x \in \mathbb{R}^d \colon \langle x, b\rangle \leq 0, \lambda(x, b) \leq 1\},$$

and let F_0 be the part of the boundary of D where $\lambda(x, b) = 1$. Then F_0 is a half-cylinder. Define

$$U(x) = \langle x, b\rangle, \qquad x \in D. \tag{3.1}$$

Then U is a classical solution of (1.4) in the interior of D. Up to this point there is no justification for the requirement $\langle x, b\rangle \leq 0$ that we put in the definition of D. In fact, below, we will define $U(x) = \langle x, b\rangle$ in a region strictly larger than D.

Step 2. We try to define U outside of D using the method of characteristics [12, Ch. 1, section 7], [7, section 35.1]. The problem is to solve $\lambda(x, \nabla u) = 1$ with boundary condition $u(x) = \langle x, b\rangle$ on F_0.

Since $\lambda \in C^\infty((\mathbb{R}^d \setminus \{0\}) \times (\mathbb{R}^d \setminus \{0\}))$, the flow α_t of the Hamiltonian vector field

$$\begin{aligned} X_\lambda &= \langle \nabla_p \lambda, \nabla_x\rangle - \langle \nabla_x \lambda, \nabla_p\rangle \\ &= \frac{\langle |x|^2 p - \langle x, p\rangle x, \nabla_x\rangle - \langle |p|^2 x - \langle x, p\rangle p, \nabla_p\rangle}{|x \wedge p|} \end{aligned}$$

is well defined.

Let $x_0 \in F_0$, let $\frac{\pi}{2} \leq \theta < \pi$ be the angle between x_0 and b, and let $\Gamma(x_0)$ denote the Hamiltonian trajectory segment $\Gamma(x_0) = \{\alpha_t(x_0, b) : 0 \leq t < \theta\}$. These are curves in phase (x, p)-space whose projections $\Gamma_1(x_0)$ onto position x-space are drawn in Figure 1.

Although the Hamiltonian trajectory segments, being integral curves of the C^∞ vector field X_λ, cannot intersect, their projections onto x-space, the characteristics, can and do in fact intersect. As we shall see below, the locus of points of intersections of the closures of the projections of $\Gamma(x_0)$ is a ray in the direction of b.

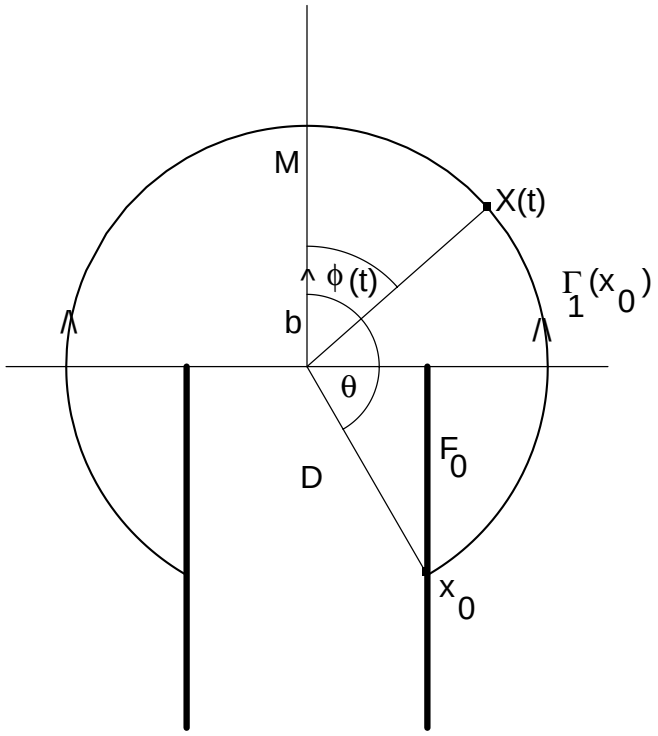


FIG. 1.

Recall that the *Poisson bracket* of λ and $\beta = \beta(x,p)$,

$$\{\lambda,\beta\} = X_\lambda(\beta) = \langle \nabla_p\lambda, \nabla_x\beta\rangle - \langle \nabla_x\lambda, \nabla_p\beta\rangle,$$

vanishes if and only if the function β is a constant of the motion; see [7, section 34.2]. In particular, λ is a constant of the motion and hence $\lambda(x,p) = |x \wedge p| = 1$ on $\Gamma(x_0)$ for all $x_0 \in F_0$. Other constants of the motion are $\langle x,p\rangle$, $|x|$, $|p|$, and each entry $(x_ip_j - x_jp_i)$ of the matrix $x \wedge p$. Thus, $\Gamma_1(x_0)$ is contained on the sphere with center at the origin and radius $|x_0|$.

Next, determine $\Gamma(x_0)$ explicitly for $x_0 \in F_0$. The trajectory $(X(t),P(t)) = \alpha_t(x_0,b)$ starting from (x_0,b) satisfies

$$\begin{aligned} \dot{x}(t) &= -Ax(t) + Bp(t), \qquad x(0) = x_0, \\ \dot{p}(t) &= -Cx(t) + Ap(t), \qquad p(0) = \nabla u(x_0) = b, \end{aligned}$$

where $A = \langle x(t), p(t)\rangle$, $B = |x(t)|^2$, and $C = |p(t)|^2$ are constants.

Then

$$\begin{aligned} X(t) &= (\cos t - A\sin t)\, x_0 + B(\sin t)\, b, \\ P(t) &= -C(\sin t)\, x_0 + (\cos t + A\sin t)\, b. \end{aligned} \tag{3.2}$$

Note that $X(\cdot)$ stays in the plane determined by the vectors b and x_0. Now take the inner product of the first of the pair (3.2) with b and divide throughout by $|X(t)|\,|b|$. Let $\phi(t)$ denote the angle between $X(t)$ and b. Using $|x_0 \wedge b| = 1$ and $\sin\theta = \frac{|x_0 \wedge b|}{|x_0|\,|b|}$

we obtain

$$\begin{aligned}\cos\phi(t) &= \frac{\langle X(t),b\rangle}{|X(t)|\,|b|} = \frac{\cos t\langle x_0,b\rangle + \sin t(-A\langle x_0,b\rangle + B\langle b,b\rangle)}{|x_0|\,|b|}\\ &= \cos t\frac{\langle x_0,b\rangle}{|x_0|\,|b|} + \sin t\frac{(-|\langle x_0,b\rangle|^2 + |x_0|^2\,|b|^2)}{|x_0|\,|b|}\\ &= \cos t\,\cos\theta + \sin t\frac{|x_0\wedge b|^2}{|x_0|\,|b|} = \cos t\,\cos\theta + \sin t\,\sin\theta\\ &= \cos(\theta - t).\end{aligned}$$

Hence $\phi(t) = \theta - t, 0 \le t < \theta$. Thus, space trajectories intersect when $\phi(t) = 0$ or along a ray in the direction of b. Finally, by the method of characteristics, the solution $\hat{U}$ of $\lambda(x,\nabla u) = 1$ satisfies

$$\frac{d}{dt}\hat{U}(X(t)) = \langle \nabla_p\lambda(X(t),P(t)),P(t)\rangle = \lambda^2(X(t),P(t)) = 1.$$

Here we use $\hat{U}$, because to define U, we intend to restrict the domain further. Since $U(x_0) = \langle x_0,b\rangle$ then

$$\hat{U}(X(t)) = t + \langle x_0,b\rangle, \qquad 0 \le t < \theta. \tag{3.3}$$

To express $\hat{U}$ in terms of $X(t)$ note that since $x_0 \in F_0$ then $\theta = \pi - \sin^{-1}(\frac{1}{|b|\,|X(t)|})$, and $|b|^2|X(t)|^2\cos^2(\theta) = |b|^2|X(t)|^2(1-\sin^2(\theta)) = |b|^2|X(t)|^2 - 1$. Therefore, writing x instead of $X(t)$ and ϕ instead of $\phi(t)$

$$\begin{aligned}\hat{U}(x) &= \theta - \phi + |b|\,|x|\cos\theta\\ &= \pi - \sin^{-1}\left(\frac{1}{|b|\,|x|}\right) - \phi - \sqrt{|b|^2|x|^2 - 1}.\end{aligned}$$

Step 3. There is an additional $(d-1)$-dimensional switching manifold F_1 defined by

$$F_1 = \{x \in \mathbb{R}^d : \hat{U}(x) = \langle x,b\rangle \ge 0\}.$$

To analyze F_1, let

$$\begin{aligned}S(r,\phi) &= \hat{U}(x) - \langle x,b\rangle\\ &= \pi - \sin^{-1}\left(\frac{1}{|b|r}\right) - \phi - \sqrt{|b|^2r^2 - 1} - |b|r\cos\phi,\end{aligned} \tag{3.4}$$

where $r = |x|$. Then F_1 is defined implicitly by $S(r,\phi) = 0, 0 \le \phi \le \frac{\pi}{2}$. Now, $S(r,\phi) \ge \pi - \sin^{-1}(\frac{1}{|b|r}) - \phi$ and $r \le \frac{1}{|b|\sin\phi}$ on F_1. Moreover, for $0 \le \phi < \frac{\pi}{2}$ we have $\frac{dr}{d\phi} = -\frac{S_\phi}{S_r}$, where

$$\begin{aligned}S_\phi &= -1 + |b|r\sin\phi \le 0,\\ S_r &= r^{-1}(-\sqrt{|b|^2r^2 - 1} - |b|r\cos\phi).\end{aligned} \tag{3.5}$$

It follows from (3.4), (3.5), and elementary calculations that $S(r,\phi) = 0$ has a unique solution $r = R(\phi), 0 \le \phi \le \frac{\pi}{2}$, that $\frac{1}{|b|} \le R(\phi) \le \frac{1}{|b|\sin\phi}$, and that $R'(\phi) < 0$ on $[0,\frac{\pi}{2})$ so that $R(\cdot)$ is strictly decreasing on $[0,\frac{\pi}{2}]$. Now define the closed regions

$$E = \{x \in \mathbb{R}^d\colon \langle x,b\rangle \ge 0, |x| \le R(\phi)\}, \qquad N = D \cup E, \qquad J = \mathbb{R}^d \setminus \text{interior}(N)$$

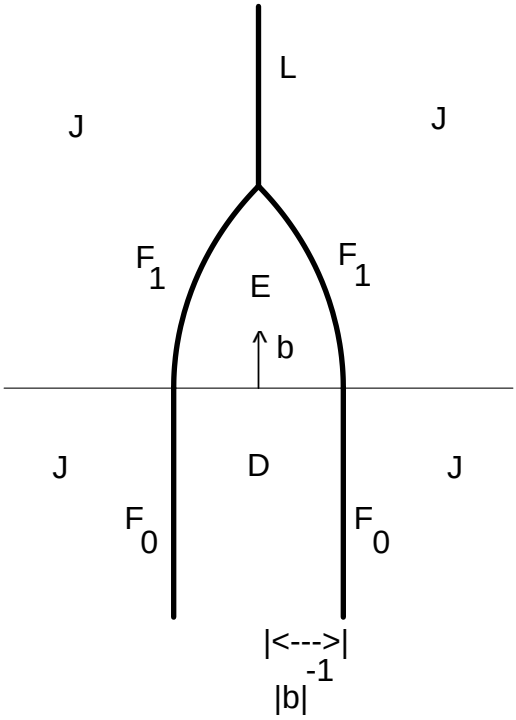


FIG. 2.

and let

$$U(x) = \langle x, b\rangle, \qquad x \in N,$$
$$U(x) = \hat{U}(x), \qquad x \in J. \tag{3.6}$$

Finally, let $L = \{rb/|b| : r \in [r_0, \infty)\}$, with $r_0 = R(0)$, be the ray where the projections Γ_1 intersect. See Figure 2.

Then we have proved the following theorem.

THEOREM 2. *There is a Lipschitz function U on $\mathbb{R}^d$ such that*
(1) *U is $C^{1,1}$ on $\mathbb{R}^d \setminus (F_1 \cup L)$,*
(2) *U is C^∞ on $\mathbb{R}^d \setminus (F_0 \cup F_1 \cup L)$,*
(3) *$\lambda(x, \nabla U) < 1$ in the interior of N,*
(4) *$\lambda(x, \nabla U) = 1$ on the complement of $N \cup L$,*
(5) *$U(x) < \langle x, b\rangle$ on the complement of N,*
(6) *$U(x) = \langle x, b\rangle$ on N;*
in particular, U is a classical solution of (1.4) *on $\mathbb{R}^d \setminus (F_1 \cup L)$.*

4. Equality of U and v.

THEOREM 3. *$U = v$ on $\mathbb{R}^d$.*

Proof. We first show that $v(x) \geq U(x)$ for all $x \in \mathbb{R}^d$. To this end, it is enough to establish $v^a(x) \geq U(x)$ for all bounded controls $a(\cdot)$.

Let $a(\cdot)$ be an arbitrary bounded control and let $x(\cdot)$ denote the solution trajectory of (1.1) starting from x. Since $U(\cdot)$ is Lipschitz and $x(\cdot)$ is absolutely continuous, then

$t \to e^{-t}U(x(t))$ equals the integral of its derivative, and hence

$$e^{-T}U(x(T)) = U(x) + \int_0^T e^{-t}\left(-U(x(t)) + \frac{d}{dt}U(x(t))\right)\,dt \tag{4.1}$$

for all $T > 0$. Let E_0 denote the set of all t such that $x(t) \notin (F_1 \cup L)$, and let E_1 denote the set of all t such that $x(t) \in (F_1 \cup L)$. We have

$$\frac{d}{dt}U(x(t)) = \langle \nabla U(x(t)), a(t)x(t)\rangle$$

almost everywhere (a.e.) for $t \in E_0$, and we will show that $(d/dt)U(x(t)) = 0$ a.e. for $t \in E_1$.

Let E_2 denote the set of all $t \in E_1$ such that $U(x(\cdot))$ is differentiable at t. Since $U(x(\cdot))$ is absolutely continuous, then $E_1 \setminus E_2$ has measure 0. Since each orbit of (1.1) is contained in a sphere centered at the origin (see the proof of Lemma 1), then no orbit intersects both L and $F_1 \setminus L$. ($L \cap F_1$ consists of a single point.)

If the orbit intersects L, then it intersects it in a single point, so that $U(x(\cdot))$ is constant on E_2, and thus has derivative 0 at any $t \in E_2$ that is an accumulation point of E_2. If E_2 has positive measure, then E_2 is uncountable, and by [14, section 23.III, p. 251], all but countably many points of E_2 are condensation points of E_2, and thus accumulation points of E_2.

If the orbit intersects F_1, then for all $t \in E_2$, we have $x(t) \in F_1$, so that

$$U(x(t)) = \langle x(t), b\rangle = |b|\,|x(t)|\cos\phi(t).$$

Since $S(|x(t)|, \phi(t)) = 0$ and $|x(t)|$ is a constant, then $\phi(t)$ is also a constant. This follows since $S(r, \phi) = 0$ defines r as a strictly decreasing function of ϕ on $[0, \frac{\pi}{2}]$, as observed in the proof of Theorem 1 (between (3.5) and (3.6)). Therefore, $U(x(\cdot))$ is constant on E_2, so that $(d/dt)U(x(t)) = 0$ for any $t \in E_2$ that is an accumulation point of E_2. Again, if E_2 has positive measure, then E_2 is uncountable, and all but countably many points of E_2 are accumulation points of E_2.

We know that $-U(x(\cdot)) + [U(x(\cdot))]' \in L^\infty[0, \infty)$, so letting $T \to \infty$ in (4.1), we get

$$0 = U(x) + \int_0^\infty e^{-t}\left(-U(x(t)) + \frac{d}{dt}U(x(t))\right)\,dt.$$

Combining this with (1.2), we get

$$v^a(x) = U(x) + \int_0^\infty e^{-t}\left(\langle x(t), b\rangle - U(x(t)) + \frac{d}{dt}U(x(t)) + |a(t)|\right)\,dt. \tag{4.2}$$

Since $\langle x, b\rangle - U(x) \geq 0$ and $(d/dt)U(x(t)) = 0$ a.e. on E_1, then the integrand in (4.2) is nonnegative a.e. on E_1. On E_0 we have

$$\frac{d}{dt}U(x(t)) + |a(t)| = \langle \nabla U(x(t)), a(t)x(t)\rangle + |a(t)|$$

a.e., and this is nonnegative since by Lemma 2

$$\begin{aligned}
&-\langle \nabla U(x), ax\rangle - |a| = |a|\left(-\left\langle \nabla U(x), \frac{a}{|a|}x\right\rangle - 1\right)\\
&\leq |a|\sup\{-\langle \nabla U(x), ax\rangle - |a|\colon\ |a| \leq 1, a \in A_d\}\\
&= |a|(\lambda(x, \nabla U(x)) - 1)^+ = 0.
\end{aligned}$$

Thus $v^a \geq U$ and $v \geq U$ on $\mathbb{R}^d$.

Next, we show $v \le U$ on $\mathbb{R}^d$. Since $v^0(x) = \langle b, x\rangle$, we have $v = U$ on N. Since L has no interior and since both v and U are Lipschitz, it remains to prove that they are equal on the complement of $N \cup L$. For x in the complement of $N \cup L$ define

$$\mathbf{a}(x) = x \wedge \nabla U(x) = x\,\nabla U(x)' - \nabla U(x)\,x'$$

and check that

$$\begin{aligned}\mathbf{a}(x)x &= (x\,\nabla U(x)')x - (\nabla U(x)\,x')x = \langle x, \nabla U(x)\rangle x - |x|^2 \nabla U(x)\\ &= -\nabla_p \lambda(x, \nabla U(x)).\end{aligned} \tag{4.3}$$

Here we have used (4) of Theorem 2.

Fix x in the complement of $N \cup L$ and let $x_1(t)$, $t \ge 0$, be the integral curve of the vector field $\mathbf{a}(x)x$ starting at x at time zero. Setting $p_1(t) = \nabla U(x_1(t))$, differentiating $\lambda(x, \nabla U(x)) = 1$, and using (4.3) shows

$$\begin{aligned}\dot{x}_1 &= -\nabla_p \lambda(x_1, p_1),\\ \dot{p}_1 &= +\nabla_x \lambda(x_1, p_1).\end{aligned}$$

Thus $(x_1(t), p_1(t))$ is the integral curve of $-X_\lambda$ through $(x, \nabla U(x))$ at $t = 0$ and through (x_0, b), with $x_0 \in F_0$, at some time $t = T$. Hence,

$$(x_1(s), p_1(s)) = \alpha_{T-s}(x_0, b) = (X(T-s), P(T-s)),$$

where $(X(t), P(t))$ is as in (3.2). Now define a sequence of controls $a_\epsilon(\cdot)$ satisfying $\lim_{\epsilon \downarrow 0} v^{a_\epsilon}(x) = U(x)$. Set $a_1(s) = \mathbf{a}(x_1(s))$, $0 \le s < T$, $a_1(s) = 0$, $s \ge T$. It follows that the unique solution of (1.1) corresponding to $a_1(\cdot)$ equals $x_1(t)$, if $0 \le t \le T$, and equals x_0 if $t \ge T$. But we need to spend no time in the complement of N. Accordingly, we define

$$a_\epsilon(t) = \frac{1}{\epsilon} a_1\left(\frac{t}{\epsilon}\right), \qquad t \ge 0. \tag{4.4}$$

Let $x_\epsilon(\cdot)$ and $v^{a_\epsilon}(x)$ be the corresponding trajectory and cost. From (4.2) we obtain

$$v^{a_\epsilon}(x) = U(x) + \int_0^{\epsilon T} e^{-t}[\langle x_\epsilon(t), b\rangle - U(x_\epsilon(t))]dt.$$

Here we have used that (3.3) and (4.3) imply $\langle \nabla U(x_1), a_1\, x_1\rangle = -\lambda^2(x_1, \nabla U(x_1)) = -1 = -|a_1|$ and (3.6). Finally, replacing $x_\epsilon(t) = x_1(t/\epsilon)$ and changing variables, we obtain $v(x) \le \lim_{\epsilon \downarrow 0} v^{a_\epsilon}(x) = U(x)$. □

REFERENCES

[1] V. E. Benes, L. A. Shep, and H. S. Witsenhausen, *Some solvable stochastic control problems*, Stochastics, 4 (1980), pp. 39–83.

[2] W. Boothby, *An Introduction to Differentiable Manifolds and Riemannian Geometry*, Academic Press, New York, 1986.

[3] M. G. Crandall, L. C. Evans, and P. L. Lions, *Some properties of viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 282 (1984), pp. 487–502.

[4] J. Dugundji, *Topology*, Allyn Bacon, Boston, 1966.

[5] J. R. Dorroh and G. Ferreyra, *Optimal advertising in exponentially decaying markets*, J. Optim. Theory Appl. 79 (1993), pp. 219–236.

[6] J. R. Dorroh and G. Ferreyra, *A multi-state, multi-control problem with unbounded controls*, SIAM J. Control Optim. 32 (1994), pp. 1322–1331.
[7] B. A. Dubrovin, A. T. Fomenko, and S. P. Novikov, *Modern Geometry* I, Grad. Texts in Math. 93, Springer-Verlag, New York, 1984,
[8] G. Ferreyra and O. Hijab, *A simple free boundary problem in R^d*, SIAM J. Control Optim., 32 (1994), pp. 501–515.
[9] G. Ferreyra and O. Hijab, *Linear-convex singular control in two dimensions*, in Proc. 33rd Conference on Decision and Control, Orlando, FL, 1994, pp. 2600–2602.
[10] G. Ferreyra and O. Hijab, *Smooth Fit for Some Bellman Equations*, Recent developments in evolution equations (Glasgow, 1994), Pitman Res. Notes Math. Ser. 324, Longman Sci. Tech., Harlow, 1995, pp. 116–122.
[11] W. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.
[12] F. John, *Partial Differential Equations*, Springer-Verlag, New York, 1982.
[13] V. Jurdjevic, *Non-Euclidean elastica*, Amer. J. Math., 117 (1995), pp. 93–124.
[14] K. Kuratowski, *Topology*, Vol. I, Academic Press, New York, 1966.
[15] J. P. Lehoczky and S. E. Shreve, *Absolutely continuous and singular stochastic control*, Stochastics, 17 (1986), pp. 91–109.
[16] S. E. Shreve and H. M. Soner, *Regularity of the value function for a two-dimensional singular stochastic control problem*, SIAM J. Control Optim., 27 (1989), pp. 876–907.
[17] S. E. Shreve and H. M. Soner, *A free boundary problem related to singular stochastic control: The parabolic case*, Comm. Partial Differential Equations, 16 (1991), pp. 373–424.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 2, pp. 590–608, March 1998

007

POSITIVE LINEAR OBSERVERS FOR LINEAR COMPARTMENTAL SYSTEMS*

J. M. VAN DEN HOF†

Abstract. Linear compartmental systems are mathematical systems that are frequently used in biology and mathematics. The inputs, states, and outputs of such systems are positive, because they denote amounts or concentrations of material. For linear dynamic systems the observer problem has been solved. The purpose of the observer problem is to determine a linear observer such that the state can be approximated. The difference between the state and its estimate should converge to zero. The interpretation in terms of a physical system requires that an estimate of the state be positive, like the state itself. In this paper conditions on the system matrices are presented that guarantee that there exists a positive linear observer such that both the error converges to zero and the estimate is positive.

Key words. compartmental systems, positive linear observers, asymptotic stability

AMS subject classifications. 93D20, 15A48

PII. S036301299630611X

1. Introduction.

The purpose of this paper is to derive positive linear observers for linear compartmental systems.

Compartmental systems are mathematical systems that are frequently used in biology and mathematics. In addition, a subclass of the class of chemical processes can be modeled as compartmental systems. A compartmental system consists of several compartments with more or less homogeneous amounts of material. The compartments interact by processes of transportation and diffusion. The dynamics of a compartmental system are derived from mass balance considerations.

In this paper linear compartmental systems consisting of inputs, states, and outputs will be studied. The outputs of these systems are not the real outputs, i.e., material leaving the system, but the observations of the amount or concentrations of material, for example, in one or more compartments. The inputs, states, and outputs are positive, so these systems are called positive linear systems in system theory. As in linear system theory, the purpose is to determine a linear observer such that the state x can be approximated by $\widehat{x}$. The error, $\widehat{x}(t) - x(t)$, should converge to zero. For positive linear systems, the observer provides an approximation of the positive state. Therefore, the observer should be chosen in such a way that the approximation of the state, $\widehat{x}(t)$, is positive, like the state, $x(t)$, itself.

For linear systems the observer problem has been solved by Luenberger [11]. See also [10]. As far as we know, there is no literature on positive observers for positive linear systems, in which the positivity of $\widehat{x}(t)$ is taken into account. It turns out that the existence of a positive linear observer satisfying the above conditions depends largely on the structure of the system matrices, i.e., the zero/nonzero pattern. Some relation can be found in the work of Sontag [13, 14].

The outline of the paper is as follows. In section 2 the problem is posed. In section 3 continuous-time linear compartmental systems are considered, and in section 4 the discrete-time case is treated. Concluding remarks are made in section 5.

*Received by the editors July 8, 1996; accepted for publication (in revised form) January 13, 1997. http://www.siam.org/journals/sicon/36-2/30611.html

†CBS (Statistics Netherlands), P.O. Box 4000, NL 2270 JM Voorburg, the Netherlands (jhof@cbs.nl).

2. Problem formulation. In this section some notation is introduced and the problem is posed.

The set $R_+ = [0, +\infty)$ is called the set of the *positive real numbers*. Let $Z_+ = \{1, 2, \ldots\}$ denote the set of positive integers, $Z_n = \{1, \ldots, n\}$, and $\boldsymbol{N} = \{0, 1, 2, \ldots\}$. Denote by R_+^n the set of n-tuples of the positive real numbers. The set $R_+^{n\times m}$ will be called the set of *positive matrices* of size n by m. Note that R_+^n is not a vector space because it does not admit an inverse with respect to addition. For matrices $A, B \in R^{n\times m}$, we will write $A \geq B$ if $a_{ij} \geq b_{ij}$ for all $i \in Z_n$, $j \in Z_m$, and $A > B$ if $A \geq B$ and $A \neq B$. A matrix $A \in R^{n\times n}$ is said to be a *Metzler matrix* if all its off-diagonal elements are in R_+; see [9]. Metzler matrices can be characterized as follows.

PROPOSITION 2.1. *A matrix $A \in R^{n\times n}$ is a Metzler matrix if and only if there exists an $\alpha \in R_+$ such that $(A + \alpha I) \in R_+^{n\times n}$.*

DEFINITION 2.2. *Consider a continuous-time linear dynamic system*

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), \quad x(t_0) = x_0, \\ y(t) &= Cx(t) + Du(t), \end{aligned} \tag{2.1}$$

with $x(t) \in X \subset R^n$, $u(t) \in U \subset R^m$, $y(t) \in Y \subset R^k$, $t \in T = [t_0, \infty)$. Equation (2.1) *is said to represent a* (continuous-time) positive linear system *if for all $x_0 \in R_+^n$ and for all $u(t) \in R_+^m$, $t \in T$, we have $x(t) \in R_+^n$ and $y(t) \in R_+^k$ for $t \in T$; in other words, $X = R_+^n$, $U = R_+^m$, and $Y = R_+^k$.*

The following proposition provides a characterization of continuous-time positive linear systems.

PROPOSITION 2.3. *A continuous-time linear dynamic system of the form* (2.1) *is a positive linear system if and only if*

$$B \in R_+^{n\times m}, \quad C \in R_+^{k\times n}, \quad D \in R_+^{k\times m}, \quad \text{and} \quad A \text{ is a Metzler matrix.}$$

Proof. Suppose first $u(t) = 0$ for all $t \in T$. For $i \in Z_n$, $x_i(t) \geq 0$ if and only if $\dot{x}_i \geq 0$ whenever $x_i = 0$ and $x_j \geq 0$ for all $j \neq i$. This is equivalent to $a_{ij} \geq 0$ for all $j \neq i$. Moreover, $y(t) = Cx(t) \geq 0$ for $x(t) \geq 0$ if and only if $C \in R_+^{k\times n}$. Now suppose $u(t) \neq 0$. For $i \in Z_n$, $x_i(t) \geq 0$ if and only if $\dot{x}_i \geq 0$ whenever $x_j = 0$ for all $j \in Z_n$. This is equivalent to $b_{ir} \geq 0$ for $r \in Z_m$. Furthermore, if $x(t) = 0$, then $y(t) = Du(t) \geq 0$ if and only if $D \in R_+^{k\times m}$. □

For discrete time, the definition of a positive linear system is presented below.

DEFINITION 2.4. *Consider a discrete-time linear dynamic system*

$$\begin{aligned} x(t+1) &= Ax(t) + Bu(t), \quad x(0) = x_0, \\ y(t) &= Cx(t) + Du(t), \end{aligned} \tag{2.2}$$

with $x \in X \subset R^n$, $u \in U \subset R^m$, $y \in Y \subset R^k$, $t \in T = \boldsymbol{N}$. Equation (2.2) *is said to represent a* (discrete-time) positive linear system *if for all $x_0 \in R_+^n$ and for all $u(t) \in R_+^m$, $t \in T$, we have $x(t) \in R_+^n$ and $y(t) \in R_+^k$ for $t \in T$; in other words, $X = R_+^n$, $U = R_+^m$, and $Y = R_+^k$.*

A characterization of discrete-time positive linear systems is as follows.

PROPOSITION 2.5. *A discrete-time linear system of the form* (2.2) *is a positive linear system if and only if*

$$A \in R_+^{n\times n}, \quad B \in R_+^{n\times m}, \quad C \in R_+^{k\times n}, \quad D \in R_+^{k\times m}.$$

The positive linear observer problem is as follows. A positive linear observer for a positive linear system is a positive linear system described by the equations

$$\dot{\widehat{x}}(t) = H\widehat{x}(t) + Ky(t) + Eu(t), \quad \widehat{x}(t_0) = \widehat{x}_0,$$
$$\widehat{x}(t+1) = H\widehat{x}(t) + Ky(t) + Eu(t), \quad \widehat{x}(t_0) = \widehat{x}_0,$$

for the continuous-time case and the discrete-time case, respectively, which yields an estimate $\widehat{x}(t)$ of the state $x(t)$ at time $t \in T$ of system (2.1), (2.2), respectively. As in linear system theory, the observer has to satisfy the following two conditions:

1. $\widehat{x}(t_0) = x(t_0)$ implies $\widehat{x}(t) = x(t)$ for all $t \geq t_0$ and for all input functions $u(t)$, $t \geq t_0$;
2. $\widehat{x}(t)$ should converge to $x(t)$ for $t \to \infty$, for all input functions $u(t)$, $t \geq t_0$.

For linear systems the problem of finding an observer satisfying 1 and 2 has been completely solved [11]. The solution is

$$\dot{\widehat{x}}(t) = (A - KC)\widehat{x}(t) + Ky(t) + Bu(t),$$
$$\widehat{x}(t+1) = (A - KC)\widehat{x}(t) + Ky(t) + Bu(t),$$

respectively, with $K \in R^{n\times k}$ such that $A - KC$ is *asymptotically stable*; i.e., for the continuous-time case, $\sigma(A - KC) \subseteq \{\lambda \in C \mid \mathrm{Re}(\lambda) < 0\}$, and for the discrete-time case, $\sigma(A - KC) \subseteq \{\lambda \in C \mid |\lambda| < 1\}$. Here $\sigma(A)$ denotes the spectrum of A. The necessary and sufficient conditions for the existence of a matrix $K \in R^{n\times k}$ such that $A - KC$ is asymptotically stable depend on the matrices A and C; i.e., the pair (A, C) should be detectable. Equivalent conditions for detectability can be found in, for example, [3, pp. 259 and 293], respectively. The interpretation in terms of a physical system requires that an estimate $\widehat{x}(t)$ be, like $x(t)$, positive. So a positive linear observer for a positive linear system should also satisfy the following condition:

3. $\widehat{x}(t) \in R^n_+$, for all $t \geq t_0$, if $\widehat{x}(t_0) \in R^n_+$, $y(t) \in R^k_+$, and $u(t) \in R^m_+$ for all $t \geq t_0$.

This third condition is satisfied if and only if $K \in R^{n\times k}_+$ and, for the continuous-time case, $A - KC$ is a Metzler matrix, or for the discrete-time case, $A - KC \in R^{n\times n}_+$. This follows from Propositions 2.3 and 2.5, respectively. Now detectability of (A, C) defined in [3] cannot be used, because then it may be possible that $K \notin R^{n\times k}_+$. Of course, detectability is a necessary condition but is not sufficient. Therefore, new necessary and sufficient conditions on A and C have to be found. The problem considered in this paper is stated below.

Problem 2.6.

Continuous time. Formulate necessary and sufficient conditions on a Metzler matrix $A \in R^{n\times n}$ and a positive matrix $C \in R^{k\times n}_+$ such that there exists a $K \in R^{n\times k}_+$, $K \neq 0$, with

1. $A - KC$ a Metzler matrix;
2. $\sigma(A - KC) \subseteq \{\lambda \in C \mid \mathrm{Re}(\lambda) < 0\}$.

Discrete time. Formulate necessary and sufficient conditions on positive matrices $A \in R^{n\times n}_+$ and $C \in R^{k\times n}_+$ such that there exists a $K \in R^{n\times k}_+$, $K \neq 0$, with

1. $A - KC \in R^{n\times n}_+$;
2. $\sigma(A - KC) \subseteq \{\lambda \in C \mid |\lambda| < 1\}$.

These problems will be solved for linear compartmental systems, which form a subclass of positive linear systems.

3. Continuous time. In this section conditions for the existence of a positive linear observer for continuous-time linear compartmental systems will be derived. First results from the theory on compartmental systems will be presented.

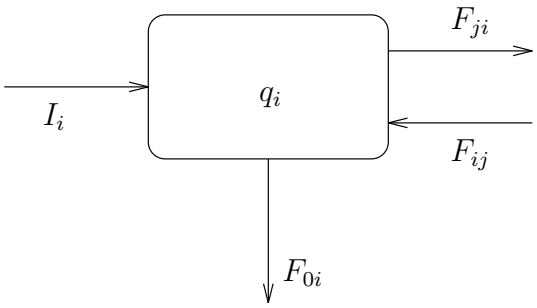


FIG. 3.1. *One compartment with possible flows.*

3.1. Continuous-time compartmental systems. A compartmental system is a system consisting of a finite number of subsystems, which are called compartments. Each compartment is kinetically homogeneous; i.e., any material entering the compartment is instantaneously mixed with the material of the compartment. Compartmental systems are dominated by the law of conservation of mass. They also form natural models for other areas of applications that are subject to conservation laws.

Consider an n-compartmental system. The behavior of the ith compartment can be represented as in Figure 3.1. In this figure, q_i denotes the amount of material considered in compartment i. The arrows represent the flows into and out of the compartment. $I_i \geq 0$ is the flow into compartment i from outside the system, called the *inflow*. $F_{ij} \geq 0$ and $F_{ji} \geq 0$ represent the flow from compartment j into compartment i and the flow from compartment i into compartment j, respectively. Finally, $F_{0i} \geq 0$ is the *outflow* to the environment from compartment i. The mass balance equations for every compartment can be written as

$$\dot{q}_i = \sum_{j \neq i} (-F_{ji} + F_{ij}) + I_i - F_{0i}. \tag{3.1}$$

In this paper the flows F_{ij} will be assumed to be linearly dependent on q_j:

$$F_{ij} = f_{ij} q_j, \quad i = 0, \ldots, n, \quad j = 1, \ldots, n, \quad i \neq j,$$

in which f_{ij} are called the *fractional transfer coefficients*. In general, f_{ij} are functions of q and time t. If f_{ij} is independent of q, the system is a linear system. In this paper it is assumed that f_{ij} is also independent of the time t; i.e., the system is a time-invariant linear system. Using this, (3.1) can be written as

$$\dot{q} = Fq + I,$$

where $q = \begin{pmatrix} q_1 & \cdots & q_n \end{pmatrix}^T \in R_+^n$, $F = (f_{ij}) \in R^{n \times n}$, with $f_{ii} = -(f_{0i} + \sum_{j \neq i} f_{ji})$ and f_{ij} constant for $i \neq j$, and I denotes the inflow from outside the system. Since $q_i \geq 0$ and $I_i \geq 0$, this system is easily seen to be a positive linear system, if the *output* is taken as

$$y = Cq, \quad y \in R^k, \quad C \in R_+^{k \times n},$$

where y denotes the vector of the observations. Note that the output is *not* the outflow of the compartmental system. The outflow, which is sometimes also called *excretion*, represents the flow of material leaving the system. The outputs of an experiment are

measurements and usually differ from the material outflows. On the other hand, the terms *inflow* and *input* can be used interchangeably.

Another property of compartmental systems is that the total flow out of a compartment over any time interval cannot be larger than the amount that was initially present plus the amount that flowed into the compartment during that interval. Together with the constraints on positive linear systems, this comes down to

$$1. \qquad f_{ij} \geq 0 \quad \text{for all } i,j \in Z_n, \quad i \neq j,$$

$$2. \qquad -f_{jj} \geq \sum_{i=1, i\neq j}^{n} f_{ij} \geq 0 \quad \text{for all } j \in Z_n.$$

A matrix F satisfying conditions 1 and 2 above is said to be a *compartmental matrix*. There is an extensive amount of literature on compartmental systems. See, for example, [1, 7, 8, 9]. Condition 2 states that all column sums of F are less than or equal to zero.

Below some properties of compartmental matrices from the literature will be discussed that are needed in this paper. References are [5, 6, 9, 15].

DEFINITION 3.1. *A matrix $A \in R^{n\times n}$ is said to be* reducible *if there exists a permutation matrix $P \in R^{n\times n}$ such that*

$$PAP^T = \begin{pmatrix} U & 0 \\ Q & R \end{pmatrix},$$

with U and R square matrices. A is said to be irreducible *if A is not reducible.*

Let $F \in R^{n\times n}$ be a compartmental matrix. Then it follows from [2, Theorem 6.4.6] that $\sigma(F) \subseteq \{\lambda \in C \mid \mathrm{Re}(\lambda) < 0 \text{ or } \lambda = 0\}$. Since a system $\dot{x} = Fx$ is asymptotically stable if and only if $\sigma(F) \subseteq \{\lambda \in C \mid \mathrm{Re}(\lambda) < 0\}$, a compartmental matrix is asymptotically stable if and only if $0 \notin \sigma(F)$. In the rest of this subsection compartmental matrices with zero eigenvalues are characterized.

PROPOSITION 3.2 (adapted from [15, Theorem III]). *Let $F \in R^{n\times n}$ be an irreducible compartmental matrix. Then $0 \in \sigma(F)$ if and only if $\sum_{i=1}^{n} f_{ij} = 0$ for all $j \in Z_n$*

DEFINITION 3.3. *Consider an n-compartmental system. A* trap *is a compartment or a set of compartments from which there are no transfers or flows to the environment nor to compartments that are not in that set. A trap is said to be* simple *if it does not strictly contain a trap.*

In the physical literature traps are usually referred to as *sinks*.

Let S be a linear compartmental system consisting of the compartments C_1, $C_2, \ldots, C_n$ and let q_j be the amount of material in C_j. Let $T \subseteq S$ be a subsystem of S. Renumbering the compartments, assume T consists of the compartments $C_m, C_{m+1}, \ldots, C_n$, for $m \leq n$. Let $F \in R^{n\times n}$ be the compartmental matrix corresponding to S, consistent with this renumbering. Then T is a trap if and only if

$$(3.2) \quad f_{ij} = 0 \quad \text{for all } (i,j) \text{ such that } j = m, m+1, \ldots, n, \quad i = 0, 1, \ldots, m-1.$$

The following two theorems are due to Fife [5].

THEOREM 3.4. *S contains a trap if and only if one of the following conditions holds:*

1. *for all $j \in Z_n$*

$$\sum_{i=1}^{n} f_{ij} = 0;$$

2. *there exists a permutation matrix* $P \in R^{n\times n}$ *such that*

$$PFP^T = \begin{pmatrix} U & 0 \\ Q & R \end{pmatrix},$$

with U, R *square matrices and the sum of entries of every column of* R *is zero.*

THEOREM 3.5. *S contains a trap if and only if $0 \in \sigma(F)$.*

In response to Fife [5], Foster and Jacquez [6] derived the following result. See also Theorems 1 and 2 together with their proofs in [9].

THEOREM 3.6. *Let S be a compartmental system with system matrix F.*

1. *Zero is an eigenvalue of F of multiplicity $m \in Z_+$ if and only if S contains m simple traps.*

2. *Assume zero is an eigenvalue of F of multiplicity $m \in Z_+$. Then there exists a partition of S into a disjoint union of subsystems*

$$S = S_1 \cup S_2 \cup \cdots \cup S_p$$

such that S_i receives no input from $S_{i+1}, \ldots, S_p$, $i = 1, \ldots, p-1$, and $S_{p-m+1}, \ldots, S_p$ are traps. Relative to this partition the system matrix is given by

$$PFP^T = \widetilde{F} = \begin{pmatrix} F_{11} & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & 0 & & & \\ F_{p-m,1} & & F_{p-m,p-m} & 0 & & \\ F_{p-m+1,1} & & F_{p-m+1,p-m} & F_{p-m+1,p-m+1} & & \\ \vdots & & \vdots & 0 & \ddots & 0 \\ F_{p1} & \cdots & F_{p,p-m} & 0 & 0 & F_{pp} \end{pmatrix},$$

where F_{ii} is irreducible for all $i \in Z_p$ and zero is an eigenvalue of F_{ii} of multiplicity 1 for $i = p-m+1, \ldots, p$, and the sum of entries of every column of F_{ii}, $i = p-m+1, \ldots, p$, is zero.

An additional consequence of this theorem is that if zero is an eigenvalue of a compartmental matrix of (algebraic) multiplicity m, the geometric multiplicity is also m, so there are always m independent eigenvalues for the eigenvalue zero.

3.2. Positive linear observers. In this subsection conditions for the existence of a positive linear observer for continuous-time linear compartmental systems will be derived. Consider a compartmental matrix $F \in R^{n\times n}$. If $F \in R^{n\times n}$ is a compartmental matrix and $K \in R_+^{n\times k}$, $C \in R_+^{k\times n}$ are such that $F - KC$ is a Metzler matrix, then $F - KC$ is also a compartmental matrix, since condition 1 in section 3.1 is satisfied because $F - KC$ is a Metzler matrix and condition 2 becomes

$$\sum_{i=1}^{n} (F - KC)_{ij} = \sum_{i=1}^{n} f_{ij} - (KC)_{ij} \leq \sum_{i=1}^{n} f_{ij} \leq 0.$$

Therefore, for the special class of compartmental matrices, the problem to be solved is the following.

Problem 3.7. Formulate necessary and sufficient conditions on a compartmental matrix $F \in R^{n\times n}$ and a positive matrix $C \in R_+^{k\times n}$ such that there exists a $K \in R_+^{n\times k}$, $K \neq 0$, with

1. $F - KC$ a compartmental matrix;
2. $\sigma(F - KC) \subseteq \{\lambda \in C \mid \mathrm{Re}(\lambda) < 0\}$.

To solve this problem, the notions of positive modifiability and positive detectability will be defined.

DEFINITION 3.8. *Let $F \in R^{n\times n}$ be a compartmental matrix and $C \in R_+^{k\times n}$. The matrix pair (F, C) is said to be* positively modifiable *if there exists a $K \in R_+^{n\times k}$ such that $KC \neq 0$ and $F-KC$ is a compartmental matrix. This implies that $\sigma(F-KC) \subseteq \{\lambda \in C \mid \mathrm{Re}(\lambda) < 0 \text{ or } \lambda = 0\}$. (F, C) is said to be* positively detectable *if there exists a $K \in R_+^{n\times k}$ such that $KC \neq 0$ and $F-KC$ is an asymptotically stable compartmental matrix. This implies that $\sigma(F-KC) \subseteq \{\lambda \in C \mid \mathrm{Re}(\lambda) < 0\}$.*

Note that solving Problem 3.7 is equivalent to checking positive detectability. To solve Problem 3.7, positive modifiability will be used, for which the following characterization can be given.

PROPOSITION 3.9. *Let $F \in R^{n\times n}$ be a compartmental matrix and $C \in R_+^{k\times n}$. The matrix pair (F, C) is positively modifiable if and only if there exists an $i \in Z_n$ and an $r \in Z_k$ such that the rth row in C is nonzero and*

$$\{\text{for all } j \neq i \text{ with } c_{rj} \neq 0, \text{ also } f_{ij} \neq 0\}. \tag{3.3}$$

Remark 3.10. In terms of compartments, Proposition 3.9 can be interpreted as follows: an output can be seen as a strictly positive linear combination of one or more compartments. These compartments contribute to this output. For positive modifiability there should exist an output such that all the compartments contributing to this output have a direct flow to one and the same compartment. A compartmental system with system matrix F can be represented by a unique directed graph; see, for example, [4] or [8, Chapter 3]. Every compartment is represented by a vertex and there is a directed arc from x_i to x_j if and only if $f_{ji} > 0$. If compartment i contributes to output j, i.e., $c_{ji} \neq 0$, this will be represented by a dashed arc. The claim in Proposition 3.9 is equivalent to saying that the graph of F should contain a subgraph of the form given in Figure 3.2(a) if the considered output has one contributing compartment or, for example, Figure 3.2(b) if the considered output has three contributing compartments.

Proof. ($\Rightarrow$) Assume (F, C) is positively modifiable, so a $K \in R_+^{n\times k}$ can be found such that $KC \neq 0$ and $F - KC$ is a compartmental matrix. $KC \in R_+^{n\times n}$, since $K \in R_+^{n\times k}$ and $C \in R_+^{k\times n}$. Therefore there exist $i, s \in Z_n$ such that $0 < (KC)_{is} = \sum_{t=1}^{k} k_{it}c_{ts}$. Hence there exists an $r \in Z_k$ such that $k_{ir}c_{rs} > 0$, which implies $k_{ir} > 0$ and $c_{rs} > 0$. From this it follows that the rth row in C is nonzero. Next, the following holds for $j \in Z_n$, $j \neq i$, since $F - KC$ is a compartmental matrix,

$$0 \leq (F - KC)_{ij} = f_{ij} - \sum_{t=1}^{k} k_{it}c_{tj}. \tag{3.4}$$

Suppose $c_{rj} \neq 0$; i.e., $c_{rj} > 0$. Since $k_{ir} > 0$, this implies $\sum_{t=1}^{k} k_{it}c_{tj} \geq k_{ir}c_{rj} > 0$. Then it follows from (3.4) that $f_{ij} > 0$. So there exist $i \in Z_n$, $r \in Z_k$ such that row r in C is nonzero and for all $j \neq i$ with $c_{rj} \neq 0$, also, $f_{ij} \neq 0$.

($\Leftarrow$) Assume there exist $i \in Z_n$, $r \in Z_k$ such that row r in C is nonzero, and for all $j \neq i$ with $c_{rj} \neq 0$, $f_{ij} \neq 0$ also. Since row r in C is nonzero and $C \in R_+^{k\times n}$, there exists either an $s \in Z_n \setminus \{i\}$ such that $c_{rs} > 0$, or $c_{rs} = 0$ for all $s \in Z_n \setminus \{i\}$ and $c_{ri} > 0$. Assume first that there exists an $s \in Z_n \setminus \{i\}$ such that $c_{rs} > 0$. By assumption, this implies $f_{is} > 0$, and in general, for all $v \in Z_n \setminus \{i\}$, $c_{rv} > 0$ implies $f_{iv} > 0$. Now take

$$0 < k_{ir} < \min_{v\in Z_n\setminus\{i\}, c_{rv}>0} \frac{f_{iv}}{c_{rv}}$$

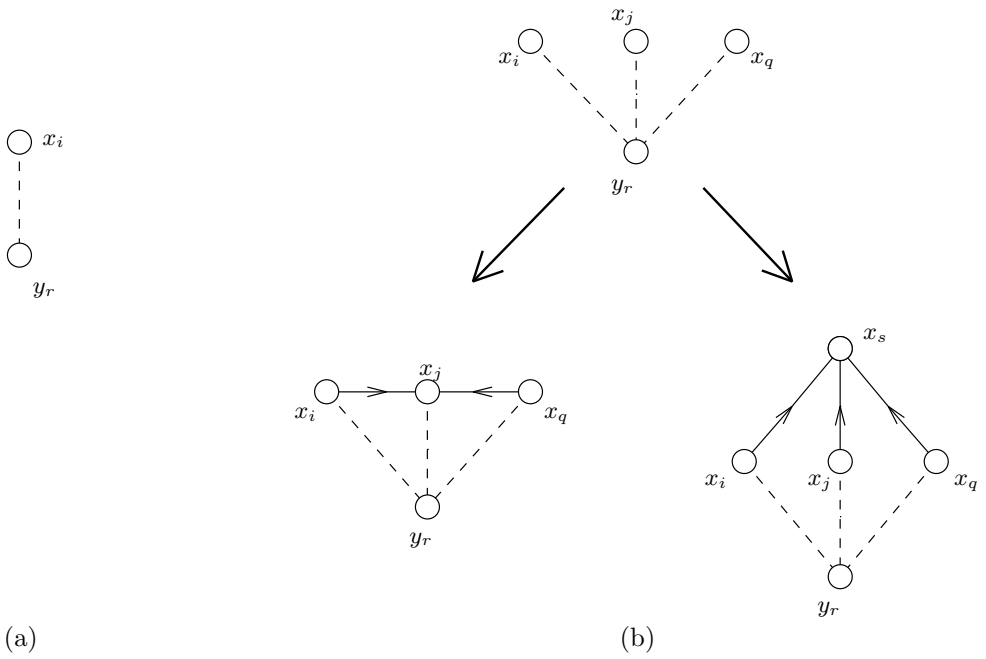


FIG. 3.2. *Subgraphs.*

and all other entries of K equal to zero. Then $K \in R_+^{n\times k}$ and

$$(F-KC)_{iw} = f_{iw} - k_{ir}c_{rw} = \begin{cases} f_{iw} - k_{ir}c_{rw} \in (0, f_{iw}), & \text{if } c_{rw} > 0, \quad w \neq i, \\ f_{iw} \geq 0, & \text{if } c_{rw} = 0, \quad w \neq i; \end{cases}$$

$$(F-KC)_{hw} = f_{hw} \geq 0 \quad \text{for } h \neq i, \quad w \neq h;$$

$$(F-KC)_{ii} = f_{ii} - k_{ir}c_{ri} \leq f_{ii} \leq -\sum_{q=1,q\neq i}^{n} f_{qi} \leq -\sum_{q=1,q\neq i}^{n} (F-KC)_{qi};$$

$$\begin{aligned}(F-KC)_{hh} = f_{hh} &\leq -\sum_{q=1,q\neq h}^{n} f_{qh} \leq -(f_{ih} - k_{ir}c_{rh}) - \sum_{q=1,q\neq h,q\neq i} f_{qh} \\ &= -\sum_{q=1,q\neq h} (F-KC)_{qh} \quad \text{for } h \neq i.\end{aligned}$$

It follows that K given above satisfies $KC \neq 0$, and $F - KC$ satisfies the conditions for a compartmental matrix.

Now assume $c_{rs} = 0$ for all $s \in Z_n \setminus \{i\}$ and $c_{ri} > 0$. Take $k_{ir} > 0$, any positive constant, and all other entries of K equal to zero. Then $K \in R_+^{n\times k}$,

$$(F-KC)_{hw} = f_{hw} \geq 0 \quad \text{for } h \in Z_n, \quad w \neq h;$$

$$(F-KC)_{hh} = f_{hh} \leq -\sum_{q=1,q\neq h}^{n} f_{qh} = -\sum_{q=1,q\neq h} (F-KC)_{qh} \quad \text{for } h \neq i;$$

$$(F-KC)_{ii} = f_{ii} - k_{ir}c_{ri} < f_{ii} \leq -\sum_{q=1,q\neq i}^{n} f_{qi} \leq -\sum_{q=1,q\neq i}^{n} (F-KC)_{qi}.$$

Again it follows that K given above satisfies $KC \neq 0$, and $F - KC$ satisfies the conditions for a compartmental matrix. □

A matrix $K \in R_+^{n\times k}$ such that $KC \neq 0$ and $F - KC$ is a compartmental matrix can be found by the following algorithm.

ALGORITHM 3.11. *Consider $F \in R^{n\times n}$, $C \in R_+^{k\times n}$. Define the sets*

$$R_c = \{(i,r) \in Z_n \times Z_k \mid \text{ row } r \text{ of } C \text{ is nonzero and (3.3) holds for } (i,r)\}$$

and

$$T_{(i,r)} = \{j \in Z_n \setminus \{i\} \mid c_{rj} \neq 0\}.$$

Form the matrix $K \in R_+^{n\times k}$ as follows.

1. *For every pair $(i,r) \in R_c$ with $T_{(i,r)} \neq \emptyset$, take*

$$0 \leq k_{ir} < \min_{j\in T_{(i,r)}} \frac{f_{ij}}{c_{rj}}.$$

2. *For every pair $(i,r) \notin R_c$ with $T_{(i,r)} \neq \emptyset$, take $k_{ir} = 0$.*
3. *For every pair $(i,r) \in Z_n \times Z_k$ with $T_{(i,r)} = \emptyset$, take any positive constant $k_{ir} \geq 0$.*

Of course, for KC to be nonzero, at least one k_{ir}, for a pair $(i,r) \in R_c$, should be strictly positive. It follows from Proposition 3.9 that the set R_c is nonempty if and only if the pair (F,C) is positively modifiable, and if this is the case, K can be chosen in such a way that $KC \neq 0$.

Before presenting the main theorem of this section, the following proposition is stated.

PROPOSITION 3.12. *Let $F \in R^{n\times n}$ be an irreducible compartmental matrix. Assume $C \in R_+^{k\times n}$ and $K \in R_+^{n\times k}$ are such that $KC \neq 0$ and $F-KC$ is a compartmental matrix. Then $F - KC$ is asymptotically stable.*

Proof. Let

$$K = \begin{pmatrix} K_1 \\ \vdots \\ K_n \end{pmatrix}, \quad \text{with } K_i \in R_+^{1\times k}, \quad C = \begin{pmatrix} C_1 & \cdots & C_n \end{pmatrix}, \quad \text{with } C_i \in R_+^{k\times 1}.$$

From $K \in R_+^{n\times k}$ and $C \in R_+^{k\times n}$ it follows that $(F - KC)_{rs} \leq f_{rs}$ for all $r,s \in Z_n$.

First, assume $F - KC$ is irreducible. Since $KC \neq 0$, there exist $i,j \in Z_n$ such that $(KC)_{ij} > 0$, or equivalently, $(F - KC)_{ij} < f_{ij}$. It follows that

$$\sum_{q=1}^{n} (F - KC)_{qj} < \sum_{q=1}^{n} f_{qj} \leq 0.$$

With Proposition 3.2 this implies that zero is not an eigenvalue of $F-KC$, so $F-KC$ is asymptotically stable.

Now, assume $F - KC$ is reducible. Suppose zero is an eigenvalue of $F - KC$. Without loss of generality it may be assumed that

$$F - KC = \begin{pmatrix} U & 0 \\ Q & R \end{pmatrix},$$

where $U \in R^{r\times r}$, with $1 \leq r < n$, $R \in R^{(n-r)\times(n-r)}$, and the sum of entries of every column of R is zero (see Theorem 3.4). For all $j = 1,\ldots,n-r$,

$$\sum_{i=1}^{n-r} R_{ij} = \sum_{i=1}^{n} (F - KC)_{i,r+j} = \sum_{i=1}^{n} (f_{i,r+j} - K_i C_{r+j}).$$

Since F is irreducible, there exists a $t \in Z_r$ such that $f_{t,r+j} > 0$, and because $f_{t,r+j} - K_t C_{r+j} = (F - KC)_{t,r+j} = 0$, $K_t C_{r+j} = f_{t,r+j} > 0$. This implies

$$\sum_{i=1}^{n-r} R_{ij} = \sum_{i=1}^{n} (f_{i,r+j} - K_i C_{r+j}) < \sum_{i=1}^{n} f_{i,r+j} \leq 0.$$

This contradicts the requirements on R. It follows that zero is not an eigenvalue of $F - KC$, so $F - KC$ is asymptotically stable. ∎

From Proposition 3.12 it follows that for an irreducible compartmental matrix $F \in R^{n\times n}$ and a positive matrix $C \in R_+^{k\times n}$, positive modifiability of (F, C) is equivalent to positive detectability of (F, C).

Consider a linear compartmental system S. Assume S contains $m \geq 0$ traps. If $m \geq 1$, then S can be partitioned as in Theorem 3.6, with system matrix

$$F = \begin{pmatrix} F_{11} & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & 0 & & & \\ F_{p-m,1} & & F_{p-m,p-m} & 0 & & \\ F_{p-m+1,1} & & F_{p-m+1,p-m} & F_{p-m+1,p-m+1} & & \\ \vdots & & \vdots & 0 & \ddots & 0 \\ F_{p1} & \cdots & F_{p,p-m} & 0 & 0 & F_{pp} \end{pmatrix} \in R^{n\times n}, \tag{3.5}$$

in which F_{ii} is irreducible for all $i \in Z_p$ and the sum of entries of every column of the square matrices $F_{p-m+1,p-m+1}, \ldots, F_{pp}$ equals zero. Note that

1. $0 \notin \sigma(F_{ii})$ for $i = 1, \ldots, p - m$;
2. $0 \in \sigma(F_{ii})$ with multiplicity 1 for $i = p - m + 1, \ldots, p$.

Consider $C \in R_+^{k\times n}$, $K \in R_+^{n\times k}$, and decompose them to conform to the partition in (3.5):

$$C = \begin{pmatrix} C_1 & \cdots & C_p \end{pmatrix}, \quad K = \begin{pmatrix} K_1 \\ \vdots \\ K_p \end{pmatrix}. \tag{3.6}$$

Now the main theorem of this subsection can be stated. It solves Problem 3.7.

THEOREM 3.13. *Consider a linear compartmental system S, as defined above, with $m \geq 0$ traps.*

1. *If $m = 0$, then $F - KC$ is asymptotically stable for all $K \in R_+^{n\times k}$ with $F - KC$ a compartmental matrix. Moreover, (F, C) is positively detectable if and only if (F, C) is positively modifiable.*

2. *For $m \geq 1$, let F, K, and C be partitioned as in* (3.5) *and* (3.6). *(F, C) is positively detectable if and only if (F_{ii}, C_i) is positively modifiable for all $i = p - m + 1, \ldots, p$.*

Proof. 1. Since S contains no traps, it follows from Theorem 3.5 that $0 \notin \sigma(F)$, so F itself is asymptotically stable. Let $K \in R_+^{n\times k}$ be such that $F - KC$ is a compartmental matrix. Suppose $0 \in \sigma(F - KC)$. Because $0 \notin \sigma(F)$, $KC \neq 0$ and with Theorem 3.4 it follows that either

$$\sum_{q=1}^{n} (F - KC)_{qj} = 0 \quad \text{for all } j \in Z_n \tag{3.7}$$

or there exists a permutation matrix $P \in R^{n\times n}$ such that

$$P(F - KC)P^T = \begin{pmatrix} U & 0 \\ Q & R \end{pmatrix}$$

where $U \in R^{r\times r}$, with $1 \le r < n$, and $R \in R^{(n-r)\times(n-r)}$, with

$$\sum_{q=1}^{n-r} R_{qj} = 0 \quad \text{for all } j = 1,\dots,n-r. \tag{3.8}$$

Since there exist $s,t \in Z_n$ such that $(KC)_{st} > 0$, the equation

$$\sum_{q=1}^{n}(F-KC)_{qt} < \sum_{q=1}^{n} F_{qt} \le 0$$

contradicts (3.7). For the other possibility, assume without loss of generality that $P = I$. Since F contains no traps, the last $n-r$ columns of $F-KC$ cannot be identical to the last $n-r$ columns of F, so there exist $s \in Z_n$ and $t \in Z_{n-r}$ such that $(F-KC)_{s,r+t} < F_{s,r+t}$, from which it follows that

$$\sum_{q=1}^{n-r} R_{qt} = \sum_{q=1}^{n}(F-KC)_{q,r+t} < \sum_{q=1}^{n} F_{q,r+t} \le 0,$$

which contradicts (3.8). It follows that $0 \notin \sigma(F-KC)$ for all $K \in R_+^{n\times k}$ such that $F-KC$ is a compartmental matrix. By the definition of positive modifiability and positive detectability, the second statement in 1 follows.

2. The blocks of $F-KC$ are $F_{ij}-K_iC_j$, for $i,j \in Z_p$. Consider $F_{ii}-K_iC_i$ for $i = p-m+1,\dots,p$. There exists a positive matrix K_i such that $K_iC_i \neq 0$ and $F_{ii}-K_iC_i$ is a compartmental matrix if and only if (F_{ii},C_i) is positively modifiable, by definition.

($\Leftarrow$) Assume (F_{ii},C_i) is positively modifiable for all $i = p-m+1,\dots,p$. Let $i \in \{p-m+1,\dots,p\}$. Since F_{ii} is irreducible, it follows from Proposition 3.12 that if (F_{ii},C_i) is positively modifiable, i.e., $K_iC_i \neq 0$ and $F_{ii}-K_iC_i$ is a compartmental matrix for some positive matrix K_i, then $0 \notin \sigma(F_{ii}-K_iC_i)$. For $j \in \{1,\dots,p-m\}$, $0 \notin \sigma(F_{jj})$, which implies by 1 above that $0 \notin \sigma(F_{jj}-K_jC_j)$ for all positive matrices K_j such that $F_{jj}-K_jC_j$ are compartmental matrices. Hence $0 \notin \sigma(F-KC)$, and it follows that (F,C) is positively detectable.

($\Rightarrow$) Assume (F,C) is positively detectable; i.e., there exists a $K \in R_+^{n\times k}$ such that $KC \neq 0$, $F-KC$ is a compartmental matrix, and $0 \notin \sigma(F-KC)$. Then $0 \notin \sigma(F_{ii}-K_iC_i)$ for all $i \in Z_p$. In particular, $0 \notin \sigma(F_{ii}-K_iC_i)$ for $i = p-m+1,\dots,p$. But $0 \in \sigma(F_{ii})$, which implies $F_{ii}-K_iC_i \neq F_{ii}$; i.e., $K_iC_i \neq 0$. Since $F_{ii}-K_iC_i$ is also a compartmental matrix, it follows that (F_{ii},C_i) is positively modifiable for $i = p-m+1,\dots,p$. □

To construct a positive linear observer, the following algorithm can be used.

ALGORITHM 3.14. *Consider a linear compartmental system S, with system matrix $F \in R^{n\times n}$ and $C \in R_+^{k\times n}$. Assume S contains $m \ge 0$ traps.*

1. *Write F and C in the forms* (3.5) *and* (3.6), *and decompose a matrix $K \in R_+^{n\times k}$ accordingly.*

2. *With Proposition* 3.9 *check positive modifiability of (F_{ii},C_i) for every $i = p-m+1,\dots,p$.*

3. *Execute Algorithm* 3.11 *for every pair (F_{ii},C_i), $i = 1,\dots,p$.*

4. *If (F_{ii},C_i) is positively modifiable for every $i = p-m+1,\dots,p$, step* 3 *provides a positive linear observer. Otherwise (F,C) is not positively detectable.*

Note that in step 3 it is not necessary to have $K_iC_i \neq 0$ for $i = 1,\dots,p-m$, whereas it is necessary for $i = p-m+1,\dots,p$.

To illustrate the theory, this section will be concluded with an example.

Example 3.15. Consider a continuous-time compartmental system with matrices

$$F = \begin{pmatrix} -2 & 0 \\ 1 & 0 \end{pmatrix}, \quad C = \begin{pmatrix} 0 & 1 \end{pmatrix}.$$

The second compartment turns out to be a trap, and F has the form (3.5), with $m = 1$ and $p = 2$. Since $(F_{22}, C_2) = (0, 1)$ satisfies the conditions stated in Proposition 3.9, (F_{22}, C_2) is positively modifiable. Hence with Theorem 3.13, (F, C) is positively detectable, so there exist $k_1, k_2 \in R_+$ such that

$$F - KC = \begin{pmatrix} -2 & -k_1 \\ 1 & -k_2 \end{pmatrix}$$

is an asymptotically stable compartmental matrix. Indeed, this can be achieved by choosing $k_1 = 0$ and $k_2 > 0$. Note that the eigenvalues of $F - KC$ cannot be arbitrarily located in the complex plane, because of the necessary condition $k_1 = 0$. One eigenvalue, -2, cannot be moved. The other eigenvalue can be placed, but only on the real negative axis. The larger k_2, the deeper this latter eigenvalue is placed in the left-half complex plane, but this makes the observer very sensitive to possible observation noise. The problem of choosing a suitable k_2 has not been solved yet. Because of the restriction $k_1 = 0$, the theory for linear optimal observers, as described in, for example, [10], cannot be used.

4. Discrete time. In this section conditions for the existence of a positive linear observer for discrete-time linear compartmental systems will be derived. Most of the results are closely related to the continuous-time case. Again, first, some theory on compartmental systems will be presented.

4.1. Discrete-time compartmental systems. In this subsection discrete-time compartmental systems will be considered. For that purpose it is assumed that transfer of material occurs at discrete times $t_1, t_2, \ldots$, or a continuous-time system is sampled at discrete times, in which case the state at time t_k has been changed into the state at time t_{k+1}. What happens in between will not be considered explicitly. Therefore, this can also be seen as if a transfer has occurred at time t_{k+1}. The discrete times will be assumed to be equally spaced to obtain a time-invariant system. Let this space be the unit time, so $t_{k+1} = t_k + 1$.

Let $q_i(t)$ be the amount of material in the ith compartment at time t. The amount transferred from the jth to the ith compartment between time t and time $t + 1$ is $G_{ij}(t)$. This transferred material will be assumed to be linearly dependent on q_j; i.e., $G_{ij}(t) = g_{ij} q_j(t)$. The state at time $t + 1$ will be given by

$$q_i(t+1) = \sum_{j \neq i} g_{ij} q_j(t) + I_i(t) + g_{ii} q_i(t),$$

where $g_{ii} q_i(t)$ is the amount of material that was in compartment i at time t and is still (or again) in compartment i at time $t + 1$. This amount $g_{ii} q_i(t)$ is equal to $q_i(t)$ minus the amount that left compartment j:

$$g_{ii} q_i(t) = q_i(t) - g_{oi} q_i(t) - \sum_{j=1, j \neq i}^{n} g_{ji} q_i(t) = \left(1 - g_{oi} - \sum_{j=1, j \neq i}^{n} g_{ji} \right) q_i(t).$$

Hence define

$$g_{ii} = 1 - g_{oi} - \sum_{j=1, j\neq i}^{n} g_{ji}.$$

The total outflow of a compartment at time $t+1$ cannot be larger than the amount that was present at time t if the inflow from outside is assumed to be zero. Together with the constraints on positive linear systems (in discrete time), this comes down to

1. $g_{ij} \geq 0$ for all $i, j \in Z_n$;
2. $\sum_{i=1}^{n} g_{ij} \leq 1$ for all $j \in Z_n$.

A matrix G satisfying conditions 1 and 2 above is said to be a *compartmental matrix* (in the discrete-time case). Condition 2 states that all column sums of $G = (g_{ij}) \in R_+^{n\times n}$ are less than or equal to one.

Below properties of compartmental matrices in discrete time will be discussed, analogous to the continuous-time case. In the rest of this subsection, G refers to a discrete-time compartmental matrix, whereas F refers to a continuous-time compartmental matrix.

Let $G \in R_+^{n\times n}$ be a compartmental matrix. Then $\sigma(G) \subseteq \{\lambda \in C \mid |\lambda| \leq 1\}$, since the sum of entries of every column of G is less than or equal to one; see [12, Section 6.2] or [2, Chapter 2]. Because a system $x(t+1) = Gx(t)$ is asymptotically stable if and only if $\sigma(G) \subseteq \{\lambda \in C \mid |\lambda| < 1\}$, it follows from the Perron–Frobenius theorem (see [12]) that a compartmental system is asymptotically stable if and only if the spectral radius $\rho(G) \neq 1$, which is equivalent to $1 \notin \sigma(G)$. Analogously to the continuous-time case, compartmental matrices having spectral radius one are characterized.

PROPOSITION 4.1. *Let $G \in R_+^{n\times n}$ be an irreducible compartmental matrix. Then $\rho(G) = 1$ if and only if $\sum_{i=1}^{n} g_{ij} = 1$ for all $j \in Z_n$.*

Proof. This follows from [2, Theorem 2.2.35]. ∎

A trap in an n-compartmental system is defined in the same way as for continuous-time systems; see Definition 3.3. As in the continuous-time case, let $C_1, C_2, \ldots, C_n$ be the compartments of a linear compartmental system S. After renumbering, let $T \subseteq S$ consist of the compartments $C_m, C_{m+1}, \ldots, C_n$, for $m \leq n$. Then T is a trap if and only if

$$g_{ij} = 0 \quad \text{for all } (i,j) \text{ such that } j = m, m+1, \ldots, n, \quad i = 0, 1, \ldots, m-1, \tag{4.1}$$

where $G = (g_{ij}) \in R_+^{n\times n}$ is the compartmental matrix corresponding to S. Consider $F = G - I$. Since

1. $f_{ij} = g_{ij} \geq 0, \quad \text{for } i, j \in Z_n, \quad i \neq j;$
2. $\sum_{j=1}^{n} f_{ji} = g_{ii} - 1 + \sum_{j=1, j\neq i}^{n} g_{ji} = \sum_{j=1}^{n} g_{ji} - 1 \leq 0.$

The matrix F is a continuous-time compartmental matrix. Assume F is the system matrix for a continuous-time compartmental system S_F and let $T_F \subseteq S_F$ consist of the last $n - m + 1$ compartments $\widetilde{C}_m, \ldots, \widetilde{C}_n$.

PROPOSITION 4.2. *Consider $T \subseteq S$ and $T_F \subseteq S_F$ defined above. Then T is a (simple) trap if and only if T_F is a (simple) trap.*

Proof. T is a trap if and only if (4.1) holds, which is equivalent to

$$\text{(4.2)} \quad \begin{cases} g_{ij} = 0 \quad \text{for all } (i,j) \text{ such that } j = m, m+1, \ldots, n, \quad i = 1, 2, \ldots, m-1, \\ \text{and} \\ g_{jj} = 1 - \sum_{i=1, i \neq j}^{n} g_{ij} \quad \text{for all } j = m, m+1, \ldots, n, \end{cases}$$

since $g_{0j} = 0$. Because $f_{ij} = g_{ij}$ for $i \neq j$ and $f_{jj} = g_{jj} - 1$, (4.2) is equivalent to

$$\begin{cases} f_{ij} = 0 \quad \text{for all } (i,j) \text{ such that } j = m, m+1, \ldots, n, \quad i = 1, 2, \ldots, m-1, \\ \text{and} \\ f_{jj} = -\sum_{i=1, i \neq j}^{n} f_{ij} \quad \text{for all } j = m, m+1, \ldots, n, \end{cases}$$

which is, because $f_{0j} = -\sum_{i=1}^{n} f_{ij}$, equivalent to (3.2); i.e., T_F is a trap. In the same way it can be proved that T is a simple trap if and only if T_F is a simple trap. ∎

Using Proposition 4.2, the following theorems, analogous to Theorems 3.4, 3.5, and 3.6, can be proved.

THEOREM 4.3. *S contains a trap if and only if one of the following conditions holds.*

1. *For all $j \in Z_n$*

$$\sum_{i=1}^{n} g_{ij} = 1;$$

2. *There exists a permutation matrix $P \in R^{n \times n}$ such that*

$$PGP^T = \begin{pmatrix} U_1 & 0 \\ Q_1 & R_1 \end{pmatrix},$$

with U_1, R_1 square matrices and the sum of entries of every column of R_1 being one.

Proof. Since

$$\sum_{i=1}^{n} f_{ij} = \left(\sum_{i=1}^{n} g_{ij} \right) - 1 \quad \text{and}$$

$$PFP^T = P(G - I)P^T = PGP^T - I = \begin{pmatrix} U_1 - I & 0 \\ Q_1 & R_1 - I \end{pmatrix} =: \begin{pmatrix} U & 0 \\ Q & R \end{pmatrix},$$

in which the sum of entries of every column of $R = R_1 - I$ is equal to the sum of entries of every column of R_1 minus one, it follows that the conditions stated in the theorem are equivalent to the conditions stated in Theorem 3.4. The theorem now follows using Proposition 4.2. ∎

THEOREM 4.4. *S contains a trap if and only if $1 \in \sigma(G)$.*

Proof. The following statements are equivalent: (i) $0 \in \sigma(F)$; (ii) $\det(F) = 0$; (iii) $\det(G - I) = 0$; (iv) $1 \in \sigma(G)$; and (v) $\rho(G) = 1$. The last equivalence relation follows from the Perron–Frobenius theorem. With Proposition 4.2, the theorem is proved. ∎

THEOREM 4.5. *Let S be a compartmental system with system matrix G.*

1. *One is an eigenvalue of G of multiplicity $m \in Z_+$ if and only if S contains m simple traps.*

2. *Assume one is an eigenvalue of G of multiplicity $m \in Z_+$. Then there exists a partition of S into a disjoint union of subsystems*

$$S = S_1 \cup S_2 \cup \cdots \cup S_p$$

such that S_i receives no input from $S_{i+1}, \ldots, S_p$, $i = 1, \ldots, p-1$, and $S_{p-m+1}, \ldots, S_p$ are traps. Relative to this partition the system matrix is given by

$$PGP^T = \widetilde{G} = \begin{pmatrix} G_{11} & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & 0 & & & \\ G_{p-m,1} & & G_{p-m,p-m} & 0 & & \\ G_{p-m+1,1} & & G_{p-m+1,p-m} & G_{p-m+1,p-m+1} & & \\ \vdots & & \vdots & 0 & \ddots & 0 \\ G_{p1} & \cdots & G_{p,p-m} & 0 & 0 & G_{pp} \end{pmatrix},$$

where G_{ii} is irreducible for all $i \in Z_p$ and one is an eigenvalue of G_{ii} of multiplicity one for $i = p-m+1, \ldots, p$, and the sum of entries of every column of G_{ii}, $i = p-m+1, \ldots, p$, is one.

Proof. 1. The following statements are equivalent:

(i) one is an eigenvalue of G of multiplicity $m \in Z_+$;

(ii) $\det(G - \lambda I) = (\lambda - 1)^m p(\lambda)$ with $p(1) \neq 0$;

(iii) $\det(F - \lambda I) = \lambda^m p_1(\lambda)$ with $p_1(0) = p(1) \neq 0$;

(iv) zero is an eigenvalue of F of multiplicity $m \in Z_+$.

The equivalence between the second and third statements follows from $\det(F - \lambda I) = \det(G - I - \lambda I) = \det(G - (\lambda + 1)I) = ((\lambda + 1) - 1)^m p(\lambda + 1) = \lambda^m p_1(\lambda)$ with $p_1(\lambda) = p(\lambda + 1)$. Now statement 1 follows from Proposition 4.2 and statement 1 of Theorem 3.6.

2. Consider the following statements.

a. one is an eigenvalue of G of multiplicity $m \in Z_+$;

b. zero is an eigenvalue of F of multiplicity $m \in Z_+$;

c. statement 2 in Theorem 3.6;

d. statement 2 in Theorem 4.5.

From 1 it follows that a $\Leftrightarrow$ b, and Theorem 3.6 provides b $\Rightarrow$ c. Noting that $PGP^T = PFP^T + I$, $G_{ij} = F_{ij}$ for $i \neq j$, and $G_{ii} = F_{ii} + I$, where the sum of entries of every column of G_{ii} is equal to the sum of entries of every column of F_{ii} plus 1; the implication c $\Rightarrow$ d follows from the statements of the proof of part 1 for $m = 1$. This completes the proof of part 2. ∎

4.2. Positive linear observers. In this subsection conditions for the existence of a positive linear observer for discrete-time linear compartmental systems will be derived. Consider a compartmental matrix $G \in R^{n \times n}$. If $G \in R_+^{n \times n}$ is a compartmental matrix and $K \in R_+^{n \times k}$, $C \in R_+^{k \times n}$ are such that $G - KC \in R_+^{n \times n}$, then $G - KC$ is also a compartmental matrix, since condition 1 in section 4.1 is satisfied because $G - KC \in R_+^{n \times n}$ and condition 2 becomes

$$\sum_{i=1}^{n} (G - KC)_{ij} = \sum_{i=1}^{n} g_{ij} - (KC)_{ij} \leq \sum_{i=1}^{n} g_{ij} \leq 1.$$

The problem for the special class of compartmental matrices is stated below.

Problem 4.6. Formulate necessary and sufficient conditions on a compartmental matrix $G \in R_+^{n \times n}$ and a positive matrix $C \in R_+^{k \times n}$ such that there exists a $K \in R_+^{n \times k}$, $K \neq 0$, with

1. $G - KC \in R_+^{n \times n}$ being a compartmental matrix;

2. $\sigma(G - KC) \subseteq \{\lambda \in C \mid |\lambda| < 1\}$.

Note that 2 is equivalent to $\rho(G - KC) < 1$, under the assumption that 1 holds.

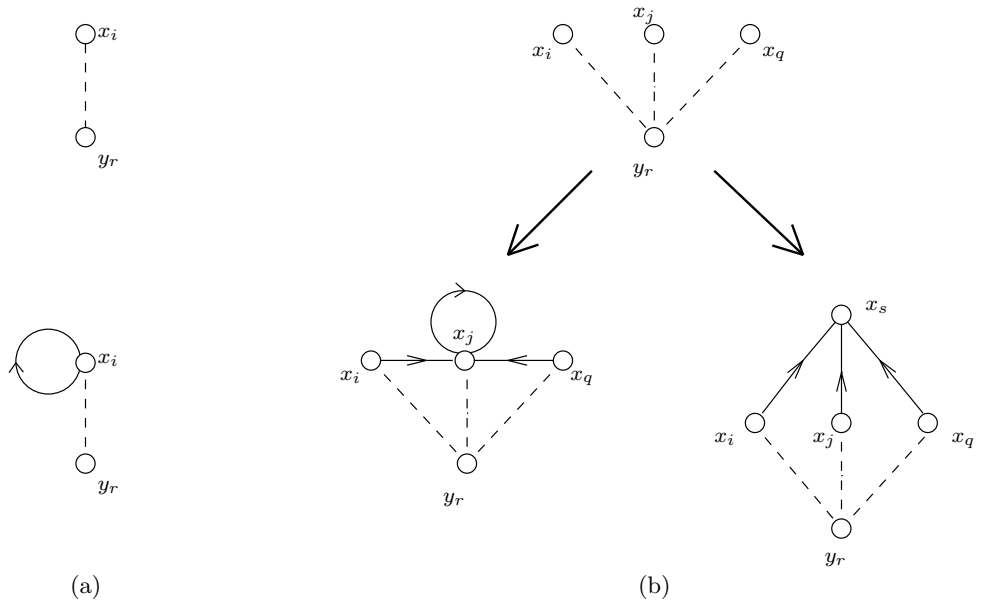


FIG. 4.1. *Subgraphs.*

The notions of positive modifiability and positive detectability are defined analogously to the continuous-time case.

DEFINITION 4.7. *Let $G \in R_+^{n\times n}$ be a compartmental matrix and $C \in R_+^{k\times n}$. The matrix pair (G, C) is said to be* positively modifiable *if there exists a $K \in R_+^{n\times k}$ such that $KC \neq 0$ and $G-KC$ is a compartmental matrix. This implies that $\sigma(G-KC) \subseteq \{\lambda \in C \mid |\lambda| \leq 1\}$. The matrix pair (G, C) is said to be* positively detectable *if there exists a $K \in R_+^{n\times k}$ such that $KC \neq 0$ and $G - KC$ is an asymptotically stable compartmental matrix. This implies that $\sigma(G - KC) \subseteq \{\lambda \in C \mid |\lambda| < 1\}$.*

For the discrete-time case, the condition for positive modifiability is somewhat different, because the diagonal elements of G also play a role.

PROPOSITION 4.8. *Let $G \in R_+^{n\times n}$ be a compartmental matrix and $C \in R_+^{k\times n}$. (G, C) is positively modifiable if and only if there exists an $i \in Z_n$ and an $r \in Z_p$ such that the rth row in C is nonzero and*

$$\{\textit{for all } j \in Z_n \textit{ with } c_{rj} > 0, \textit{ also } g_{ij} > 0\}. \tag{4.3}$$

Remark 4.9. The interpretation of Proposition 4.8 is that for positive modifiability there should exist an output such that all the compartments contributing to this output have a direct flow to one and the same compartment. Note that in contrast to the continuous-time case, if this last mentioned compartment is a compartment that contributes to the output, also is "flow" to itself needed. This means that some of the material in this compartment is still in this compartment one time step ahead. Defining a graph for system matrix G as in Remark 3.10, then the condition in Proposition 4.8 says that the graph of G should contain a subgraph of the form shown in Figure 4.1. In this case, "flow" to itself is represented by a loop, which occurs if $g_{ii} > 0$.

Proof. ($\Rightarrow$) Assume (G, C) is positively modifiable, so a $K \in R_+^{n\times k}$ can be found, such that $KC \neq 0$ and $G - KC$ is a compartmental matrix. $KC \in R_+^{n\times n}$, since

$K \in R_+^{n\times k}$ and $C \in R_+^{k\times n}$. Therefore, there exist $i, s \in Z_n$ such that $0 < (KC)_{is} = \sum_{t=1}^{k} k_{it}c_{ts}$. Hence there exists an $r \in Z_k$ such that $k_{ir}c_{rs} > 0$, which implies $k_{ir} > 0$ and $c_{rs} > 0$. From this it follows that the rth row in C is nonzero. Next,

$$0 \le (G - KC)_{ij} = g_{ij} - \sum_{t=1}^{k} k_{it}c_{tj} \tag{4.4}$$

holds for all $j \in Z_n$, since $G-KC$ is a compartmental matrix. Suppose $c_{rj} > 0$. Since $k_{ir} > 0$, this implies $\sum_{t=1}^{k} k_{it}c_{tj} \ge k_{ir}c_{rj} > 0$. From (4.4) it follows that $g_{ij} > 0$. So there exist $i \in Z_n$, $r \in Z_k$ such that row r in C is nonzero, and for all $j \in Z_n$ with $c_{rj} > 0$, $g_{ij} > 0$ also.

($\Leftarrow$) Assume there exist $i \in Z_n$, $r \in Z_k$ such that row r in C is nonzero, and for all $j \in Z_n$ with $c_{rj} > 0$, $g_{ij} > 0$ also. Since row r in C is nonzero, there exists an $s \in Z_n$ such that $c_{rs} > 0$. This implies $g_{is} > 0$, and in general, for all $v \in Z_n$, $c_{rv} > 0$ implies $g_{iv} > 0$. Now take

$$0 < k_{ir} < \min_{v\in Z_n, c_{rv}>0} \frac{g_{iv}}{c_{rv}}$$

and all other entries of K equal to zero. Then $K \in R_+^{n\times k}$ and

$$(G - KC)_{iw} = g_{iw} - k_{ir}c_{rw} = \begin{cases} g_{iw} - k_{ir}c_{rw} \in (0, g_{iw}) & \text{if } c_{rw} > 0, \\ g_{iw} \ge 0 & \text{if } c_{rw} = 0; \end{cases}$$
$$(G - KC)_{hw} = g_{hw} \ge 0 \quad \text{for } h \ne i.$$

It follows that K given above satisfies $KC \ne 0$, and $G - KC$ satisfies the conditions for a compartmental matrix. □

A matrix $K \in R_+^{n\times k}$ such that $KC \ne 0$ and $G - KC$ is a compartmental matrix can be found by the following algorithm.

ALGORITHM 4.10. *Consider $G \in R_+^{n\times n}$, $C \in R_+^{k\times n}$. Define the sets*

$$R_d = \{(i, r) \in Z_n \times Z_k \mid \text{ row } r \text{ of } C \text{ is nonzero and (4.3) holds for } (i, r)\}$$

and

$$D_{(i,r)} = \{j \in Z_n \mid c_{rj} \ne 0\}.$$

Form the matrix $K \in R_+^{n\times k}$ as follows.

1. *For every pair $(i, r) \in R_d$, take*

$$0 \le k_{ir} < \min_{j\in D_{(i,r)}} \frac{g_{ij}}{c_{rj}}.$$

2. *For every pair $(i, r) \notin R_d$, take $k_{ir} = 0$.*

Of course, for KC to be nonzero, at least one k_{ir}, for a pair $(i, r) \in R_d$, should be strictly positive. It follows from Proposition 4.8 that the set R_d is nonempty if and only if the pair (G, C) is positively modifiable, and if this is the case, K can be chosen in such a way that $KC \ne 0$. Analogous to the continuous-time case, the following results can be stated.

PROPOSITION 4.11. *Let $G \in R_+^{n\times n}$ be an irreducible compartmental matrix. Assume $C \in R_+^{k\times n}$ and $K \in R_+^{n\times k}$ are such that $KC \ne 0$ and $G-KC$ is a compartmental matrix. Then $G - KC$ is asymptotically stable.*

Consider a linear compartmental system S. Assume S contains $m \geq 0$ traps. If $m \geq 1$, then S can be partitioned as in Theorem 4.5, with system matrix

$$G = \begin{pmatrix} G_{11} & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & 0 & & & \\ G_{p-m,1} & & G_{p-m,p-m} & 0 & & \\ G_{p-m+1,1} & & G_{p-m+1,p-m} & G_{p-m+1,p-m+1} & & \\ \vdots & & \vdots & 0 & \ddots & 0 \\ G_{p1} & \cdots & G_{p,p-m} & 0 & 0 & G_{pp} \end{pmatrix} \in R_+^{n\times n}, \tag{4.5}$$

in which G_{ii} is irreducible for all $i \in Z_p$ and the sum of entries of every column of the square matrices $G_{p-m+1,p-m+1}, \ldots, G_{pp}$ equals one. Note that

1. $1 \notin \sigma(G_{ii})$ for $i = 1, \ldots, p-m$;
2. $1 \in \sigma(G_{ii})$ with multiplicity 1 for $i = p-m+1, \ldots, p$.

Consider $C \in R_+^{k\times n}$, $K \in R_+^{n\times k}$, and decompose them to conform to the partition in (4.5):

$$C = \begin{pmatrix} C_1 & \cdots & C_p \end{pmatrix}, \quad K = \begin{pmatrix} K_1 \\ \vdots \\ K_p \end{pmatrix}. \tag{4.6}$$

The main theorem of this subsection, solving Problem 4.6, is stated below.

THEOREM 4.12. *Consider a linear compartmental system S, as defined above, with $m \geq 0$ traps.*

1. *If $m = 0$, then $G - KC$ is asymptotically stable for all $K \in R_+^{n\times k}$ with $G - KC$ a compartmental matrix. Moreover, (G, C) is positively detectable if and only if (G, C) is positively modifiable.*

2. *For $m \geq 1$, let G, K, and C be partitioned as in (4.5) and (4.6). (G, C) is positively detectable if and only if (G_{ii}, C_i) is positively modifiable for all $i = p-m+1, \ldots, p$.*

Proofs of Proposition 4.11 *and Theorem* 4.12. These proofs are analogous to the proofs of Proposition 3.12 and Theorem 3.13, using Proposition 4.1, Theorem 4.3, and Theorem 4.4, respectively, instead of Proposition 3.2, Theorem 3.4, and Theorem 3.5, respectively, and changing F into G and the appropriate zeros into ones. Details are left for the reader. □

To construct a positive linear observer, the following algorithm can be used.

ALGORITHM 4.13. *Consider a linear compartmental system S, with system matrix $G \in R_+^{n\times n}$ and $C \in R_+^{k\times n}$. Assume S contains $m \geq 0$ traps.*

1. *Write G and C in the forms* (4.5) *and* (4.6), *and decompose a matrix $K \in R_+^{n\times k}$ accordingly.*

2. *With Proposition* 4.8 *check positive modifiability of (G_{ii}, C_i) for every $i = p-m+1, \ldots, p$.*

3. *Execute Algorithm* 4.10 *for every pair (G_{ii}, C_i), $i = 1, \ldots, p$.*

4. *If (G_{ii}, C_i) is positively modifiable for every $i = p-m+1, \ldots, p$, step* 3 *provides a positive linear observer. Otherwise (G, C) is not positively detectable.*

Note that in step 3 it is not necessary to have $K_i C_i \neq 0$ for $i = 1, \ldots, p-m$, whereas it is necessary for $i = p-m+1, \ldots, p$.

5. Concluding remarks. Positive linear observers for linear compartmental systems have been considered. Conditions on the system matrices A and C have been

derived for the existence of positive linear observers, i.e., linear observers that provide positive estimates of the state in case the estimate of the initial state and the input is positive. As has been shown in the example in section 3.2, the problem of finding an *optimal* positive linear observer is also worthwhile to be studied. By an *optimal* positive linear observer we mean on the one hand one with a "large" gain K, but on the other hand one that is not too sensitive to possible observation noise. This problem remains to be investigated.

In linear system theory, the dual of the observer problem is the stabilization problem by linear state feedback; see, for example, [3, 10]. Of course, duals of the results in this chapter can be derived. But these results will have no physical meaning since a stabilization problem by linear state feedback would be to design for a positive linear system

$$\dot{x}(t) = Ax(t) + Bu(t)$$

a positive linear control law

$$u(t) = Fx(t) + v(t),$$

with $v(t) \in R_+^m$ a new input, such that the closed loop system

$$\dot{x}(t) = (A + BF)x(t) + Bv(t)$$

is asymptotically stable. For physical reasons, this control law should produce a positive input u, given positive state x and positive input v, so $F \in R_+^{m \times n}$. Therefore $A + BF \geq A$, whereas $A - KC \leq A$. So new results for this problem have to be found, and they are definitely *not* dual to the results in this chapter.

REFERENCES

[1] D. H. Anderson, *Compartmental Modeling and Tracer Kinetics*, Lecture Notes in Biomathematic 50, Springer-Verlag, Berlin, 1983.

[2] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, Computer Science and Applied Mathematics, Academic Press, New York, 1979.

[3] F. M. Callier and C. A. Desoer, *Linear System Theory*, Springer Texts in Electrical Engineering, Springer-Verlag, New York, 1991.

[4] E. J. Davison, *Connectability and structural controllability of composite systems*, Automatica, 13 (1977), pp. 109–123.

[5] D. Fife, *Which compartmental systems contain traps?*, Math. Biosciences, 14 (1972), pp. 311–315.

[6] D. M. Foster and J. A. Jacquez, *Multiple zeros for eigenvalues and the multiplicity of traps of a linear compartmental system*, Math. Biosciences, 26 (1975), pp. 89–97.

[7] K. Godfrey, *Compartmental Models and Their Applications*, Academic Press, London, 1983.

[8] J. A. Jacquez, *Compartmental Analysis in Biology and Medicine*, The University of Michigan Press, Ann Arbor, MI, 1985.

[9] J. A. Jacquez and C. P. Simon, *Qualitative theory of compartmental systems*, SIAM Rev., 35 (1993), pp. 43–79.

[10] H. Kwakernaak and R. Sivan, *Linear Optimal Control Theory*, John Wiley, New York, 1972.

[11] D. G. Luenberger, *An introduction to observers*, IEEE Trans. Automat. Control, 16 (1966), pp. 596–602.

[12] D. G. Luenberger, *Introduction to Dynamic Systems: Theory, Models, and Applications*, John Wiley, New York, 1979.

[13] E. D. Sontag, *Nonlinear regulation: The piecewise linear approach*, IEEE Trans. Automat. Control, 26 (1981), pp. 346–358.

[14] E. D. Sontag, *Remarks on piecwise-linear algebra*, Pacific J. Math., 98 (1982), pp. 183–201.

[15] O. Taussky, *A recurring theorem on determinants*, Amer. Math. Monthly, 56 (1949), pp. 672–676.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 2, pp. 609–653, March 1998

008

EXISTENCE OF MARKOV CONTROLS AND CHARACTERIZATION OF OPTIMAL MARKOV CONTROLS*

THOMAS G. KURTZ† AND RICHARD H. STOCKBRIDGE‡

Abstract. Given a solution of a controlled martingale problem it is shown under general conditions that there exists a solution having Markov controls which has the same cost as the original solution. This result is then used to show that the original stochastic control problem is equivalent to a linear program over a space of measures under a variety of optimality criteria. Existence and characterization of optimal Markov controls then follows. An extension of Echeverria's theorem characterizing stationary distributions for (uncontrolled) Markov processes is obtained as a corollary. In particular, this extension covers diffusion processes with discontinuous drift and diffusion coefficients.

Key words. Markov controls, optimal controls, martingale problems, stationary processes, linear programming, occupation measures

AMS subject classifications. 49A60, 60G35, 60J25, 60J35, 93E20

PII. S0363012995295516

1. Introduction and formulation.

We consider processes whose dynamics are specified through a controlled martingale problem for their generator, that is, by the requirement that

$$f(X(t)) - \int_0^t Af(X(s), u(s))\, ds \tag{1.1}$$

be a martingale for every f in the domain of the generator A. In this expression X is the state process and u is a process which "controls" X. (A detailed formulation of the dynamics is given later in this section.) The controller usually compares control processes by observing their associated "costs" according to a prescribed criterion. In this paper we consider infinite-horizon discounted and long-term average costs, finite-horizon costs, and first passage or first exit costs.

The controller may use any nonanticipating process to control the state. The restriction on the control process is implicit in the martingale requirements; however, it is easy to check that any solution of the controlled martingale problem corresponds to a solution with the same state process and a (relaxed) control adapted to the filtration generated by the state process. Controls may be based on the full history of the state and control processes or they may depend only on the present value of the state. The latter control processes are referred to as Markov or feedback controls.

There are usually two desires which the controller would like to fulfill. The first and most important is to choose the control so as to obtain the minimum cost possible; however, from a practical point of view, the controller would also like to have as simple a control as possible.

*Received by the editors December 4, 1995; accepted for publication (in revised form) January 16, 1997. This research was partially supported by NSF grants DMS 9006674, DMS 9204866, DMS 9404990, and DMS 9504323.

http://www.siam.org/journals/sicon/36-2/29551.html

†Departments of Mathematics and Statistics, University of Wisconsin, Madison, WI 53706 (kurtz@math.wisc.edu).

‡Department of Statistics, University of Kentucky, Lexington, KY 40506–0027 (stockb@ms.uky.edu).

In this paper, we show that it is possible to fulfill both desires simultaneously, at least in one sense, by showing the existence of optimal Markov controls. Furthermore, we develop a method for determining an optimal Markov control.

The question of existence of optimal Markov controls has been studied in [4, 8, 12]. Each paper considers controlled diffusion processes and gives conditions under which existence of optimal Markov controls is assured. The approach of the last two papers is to apply the Krylov selection theorem to obtain a Markov solution to the martingale problem and thus a Markov control. This approach gives existence of an optimal Markov control but does not characterize the control.

The method we use to determine an optimal control is to reformulate the original control problem as a linear program over a collection of measures. This approach has a long history starting with Manne [15] and has been widely applied in a discrete setting (see, for example, [5, 6, 16, 20]). The extension to continuous time processes having continuous state and control processes was given in [19]. However, the optimal controls obtained in [19] were only shown to be progressively measurable and only applied to long-term average control problems. Fleming and Vermes [10] use a similar reformulation in terms of occupation measures. They use convex analysis techniques to show that the solution is given by the upper envelope of the smooth subsolutions of the Hamilton–Jacobi–Bellman equation.

This paper has two main components. The first considers existence of Markov controls for solutions of the controlled martingale problem for the generator A. These existence results typically take two forms. One form starts with a given solution to the martingale problem and establishes existence of another solution having a Markov control whose cost matches the cost of the given solution. The other form begins with an identity in which a measure (or measures) annihilates the generator (or generators) of the process(es) and constructs a solution having a Markov control. The second component develops the LP reformulations of the various control problems.

The organization of this paper is as follows. We give the formulation of the dynamics of the controlled process in section 1.1. Existence of stationary solutions having a Markov control and such that the one-dimensional distributions are prescribed is shown for any prescribed distribution satisfying a stationarity condition (2.1) in section 2. Section 3 extends Echeverria's theorem (see [9, Theorem 4.9.17]) to operators whose ranges include discontinuous functions and considers the forward equation. In section 4, we show that for any solution of the controlled martingale problem there exists another solution having a (time-inhomogeneous) Markov control which has the same one-dimensional state and control distributions as the given solution. Since the long-term average, discounted and finite-horizon cost criteria depend only on the one-dimensional distributions, this result implies that the new solution has the same expected cost for these criteria. The distribution of the Markov control corresponds to the projection of the given control onto the σ-algebra generated by the current state. A corresponding result is also obtained for the first exit criterion. Section 5 provides existence of a solution with a *time-homogeneous* Markov control which matches the costs of a given solution, though the one-dimensional distributions may no longer be the same as the given solution. Finally, in section 6 we determine the equivalent LP formulations of the control problems for each of the cost criteria.

The existence results in section 4 are similar in nature to an observation by Dvoretsky [7] for discrete processes which says that any sequence of random variables has the same one-dimensional distributions as a Markov chain. More recently, for a diffusion having adapted drift and diffusion coefficients, Krylov [13] shows existence of drift

and diffusion coefficients which are functions of the state such that the two diffusions have the same Green measure (see Corollary 5.4 of section 5.1). Gyöngy [11] shows existence of state- and time-dependent coefficients such that the one-dimensional distributions of the diffusions are the same (see Corollary 4.3 of section 4). For controlled diffusions in which the control only affects the drift coefficient, Borkar [3] establishes a similar existence of Markov controls such that the one-dimensional distributions of the diffusions agree. The results in sections 4 and 5 generalize [3, 11, 13].

Following the completion of the first draft of this paper, the authors became aware of work of Bhatt and Borkar [2] that has substantial overlap with the results of this paper. Both papers are based on extensions of Echeverria's theorem to controlled processes given in [18, 19]. Bhatt and Borkar work in the context of complete, separable, metric state spaces and compact control spaces while we treat locally compact (not necessarily compact), separable, metric state and control spaces. Their paper contains analogues of our Theorems 2.2 and 4.1, although the method of proof used in Theorem 2.2 is substantially different. They study the discounted criterion with a fixed discount rate, whereas section 5.1 of our paper allows the discount rate to be state- and control-dependent. This dependence is used to extend Krylov's result on the Green measures. Their characterization of the finite-horizon problem does not include terminal costs since they work with test functions which vanish at the terminal time. Our results (sections 4.2 and 6.3) allow terminal costs. The results on the first-passage criterion (sections 4.1 and 5.2) and the forward equation for Markov processes (section 3) appear only in this paper.

1.1. Formulation. In this paper, we only consider *relaxed* solutions for the martingale problem, so we begin by stating the definition of a relaxed solution.

For a measurable space S, we define $\widehat{C}(S)$ to be the space of continuous functions on S which vanish at ∞, $\overline{C}(S)$ to be the space of bounded, continuous functions on S, $\mathcal{M}(S)$ to be the space of finite Borel measures on S, and $\mathcal{P}(S)$ to be the space of probability measures on S.

Let the state space E and control space U be locally compact, complete, separable metric spaces and let $E^\Delta = E \cup \{\Delta\}$ be the one-point compactification of E. Let $A : \mathcal{D}(A) \subset \widehat{C}(E) \to C(E \times U)$ and $\nu \in \mathcal{P}(E)$. Then an $E \times \mathcal{P}(U)$-valued process (X, Λ) is a relaxed solution of the controlled martingale problem for (A, ν) if there exists a filtration $\{\mathcal{F}_t\}$ such that (X, Λ) is $\{\mathcal{F}_t\}$-progressive, X has initial distribution ν, and for every $f \in \mathcal{D}(A)$,

$$f(X(t)) - \int_0^t \int_U Af(X(s), u)\, \Lambda_s(du)\, ds \tag{1.2}$$

an $\{\mathcal{F}_t\}$-martingale.

We assume that the generator A satisfies the following conditions, which are sufficient to guarantee existence of solutions (at least in E^Δ) to the controlled martingale problem for each $\nu \in \mathcal{P}(E)$:

(i) $\mathcal{D}(A)$ is dense in $\widehat{C}(E)$,

(ii) for each $f \in \mathcal{D}(A)$ and $u \in U$, $Af(\cdot, u) \in \widehat{C}(E)$,

(iii) for each $f \in \mathcal{D}(A)$ and compact $K \subset U$, $\limsup_{x \to \Delta} \sup_{u \in K} |Af(x, u)| = 0$, and

(iv) for each $u \in U$, $A_u f = Af(\cdot, u)$ satisfies the positive maximum principle.

The fundamental existence result (Theorem 2.2) requires the additional condition that

(v) $\mathcal{D}(A)$ is an algebra.

For compactness purposes we also assume that

(vi) there exists $\psi \in C(U)$, $\psi > 0$, such that for each $f \in \mathcal{D}(A)$ there exist constants a_f and b_f satisfying

$$|Af(x,u)| \leq a_f + b_f\psi(u). \tag{1.3}$$

Integrability assumptions of two types will be placed on ψ; the particular assumption will be specified in each theorem. An additional assumption (6.1) relating the cost function c with ψ is imposed in the LP reformulation section.

The results of this paper can be extended to a nonlocally compact state space E as in [1, 2]. We treat the locally compact case for simplicity of exposition.

This paper considers the infinite-horizon discounted cost, the long-term average cost, the finite-horizon cost, and the first passage or first exit cost associated with solutions of the controlled martingale problem. There are two forms to the LP reformulations of each control problem which are based on the existence results contained in sections 4 and 5, respectively. The results in section 4 allow the running cost and terminal or exit cost functions to be time-dependent and lead to LP reformulations in which time is a component. When there is no time-dependence in the cost criteria aside from (possibly) discounting, the LP reformulations are simpler and use the existence of time-homogeneous Markov controls established in section 5.

With this in mind, let $c \in M(E \times U)$ and $g \in M(E)$ be bounded below. ($M(E)$ will denote the Borel measureable functions on E.) The four standard criteria are

- the discounted cost

$$E\left[\int_0^\infty e^{-\alpha t}\int_U c(X(t),u)\Lambda_t(du)\,dt\right]; \tag{1.4}$$

- the finite-horizon cost

$$E\left[\int_0^T \int_U c(X(t),u)\Lambda_t(du)\,dt + g(X(T))\right]; \tag{1.5}$$

- the long-term average cost

$$\limsup_{t\to\infty} E\left[t^{-1}\int_0^t \int_U c(X(s),u)\Lambda_s(du)\,ds\right]; \tag{1.6}$$

- the first passage cost

$$E\left[\int_0^\tau \int_U c(X(s),u)\,\Lambda_s(du)\,ds + g(X(\tau))\right], \tag{1.7}$$

where $E_0 \subset E$ is an open set and $\tau = \inf\{t \geq 0 : X(t) \in E_0^c\}$.

The criteria (1.4), (1.5), and (1.7) are adjusted appropriately when c and g are time-dependent. In addition, for the first passage cost we allow $E_0 \subset \mathbb{R}^+ \times E$ and $\tau = \inf\{t \geq 0 : (t, X(t)) \in E_0^c\}$. We also consider discount rates that depend on the state and control.

2. Stationary solution to the controlled martingale problem. The objective of this section is to establish the existence of a particular form of stationary solution for the controlled martingale problem for a generator A.

Suppose μ is a probability measure on $E \times U$ which satisfies

$$\int_{E\times U} Af(x,u)\,\mu(dx\times du) = 0 \qquad \forall f \in \mathcal{D}(A)\ . \tag{2.1}$$

Denote the state marginal by $\mu_0 = \mu(\cdot \times U)$ and let η be the regular conditional distribution of u given x; that is, η satisfies

$$\mu(\Gamma_1 \times \Gamma_2) = \int_{\Gamma_1} \eta(x,\Gamma_2)\,\mu_0(dx) \quad \forall \Gamma_1 \in \mathcal{B}(E), \quad \Gamma_2 \in \mathcal{B}(U)\ . \tag{2.2}$$

If X is a stationary process with $X(0)$ having distribution μ_0, the pair $(X, \eta(X,\cdot))$ is stationary and the one-dimensional distributions satisfy

$$E[I_{\Gamma_1}(X(t))\,\eta(X(t),\Gamma_2)] = \mu(\Gamma_1\times\Gamma_2), \quad t \geq 0.$$

We show that there exists a stationary process X such that the $E\times\mathcal{P}(U)$-valued process $(X,\eta(X,\cdot))$ is a stationary relaxed solution of the controlled martingale problem for (A,μ_0). The following lemma is essential to this result.

LEMMA 2.1. *Let X_n, X be processes in $D_E[0,\infty)$ with $X_n \Rightarrow X$, and let $D_X = \{t : P\{X(t) \neq X(t-)\} > 0\}$. Suppose, for each $t \geq 0$, that $X_n(t)$ and $X(t)$ have a common distribution $\nu_t \in \mathcal{P}(E)$. Let g be Borel measurable on $[0,\infty)\times E$ and satisfy*

$$\int_0^t \int_E |g(s,x)|\nu_s(dx)ds < \infty$$

for each $t > 0$. Then

$$\int_0^{[n\cdot]/n} g(s,X_n(s))\,ds \Rightarrow \int_0^{\cdot} g(s,X(s))\,ds \tag{2.3}$$

and, in particular, for each $m \geq 1, 0 \leq t_1 \leq \cdots \leq t_m < t_{m+1}$, $t_i \notin D_X$, and $h_1,\ldots,h_m \in \overline{C}(E)$,

$$\begin{aligned}\lim_{n\to\infty} E\left[\int_{[nt_m]/n}^{[nt_{m+1}]/n} g(s,X_n(s))\,ds \prod_{i=1}^m h_i(X_n(t_i))\right]\\ = E\left[\int_{t_m}^{t_{m+1}} g(s,X(s))\,ds \prod_{i=1}^m h_i(X(t_i))\right].\end{aligned} \tag{2.4}$$

Proof. For each $\epsilon > 0$, there exists $g_\epsilon \in \overline{C}([0,\infty)\times E)$ satisfying $\int_0^\infty e^{-t}\int_E |g(s,x) - g_\epsilon(s,x)|\,\nu_s(dx)ds < \epsilon$. Then

$$\begin{aligned}&\left|E\left[\int_{[nt_m]/n}^{[nt_{m+1}]/n} \{g(s,X_n(s)) - g_\epsilon(s,X_n(s))\}\,ds \prod_{i=1}^m h_i(X_n(t_i))\right]\right|\\ \leq\ & \prod_{i=1}^m ||h_i|| E\left[\int_{[nt_m]/n}^{[nt_{m+1}]/n} |g(s,X_n(s)) - g_\epsilon(s,X_n(s))|\,ds\right]\\ \leq\ & \prod_{i=1}^m ||h_i||(t_{m+1}-t_m+1)e^{t_{m+1}}\epsilon.\end{aligned}$$

Similarly,

$$\left| E\left[\int_{t_m}^{t_{m+1}} \{g(s,X(s)) - g_\epsilon(s,X(s))\}\, ds \prod_{i=1}^m h_i(X(t_i)) \right] \right| \le \prod_{i=1}^m ||h_i|| (t_{m+1} - t_m + 1) e^{t_{m+1}} \epsilon.$$

The result now follows since the convergence in (2.4) is immediate with g replaced by g_ϵ. The proof of (2.3) is obtained in a similar manner. □

THEOREM 2.2. *Suppose that E, U, A, and ψ are as in section* 1.1. *Suppose $\mu \in \mathcal{P}(E \times U)$ satisfies* (2.1). *Let η satisfy* (2.2). *Assume that ψ satisfies*

$$\int \psi(u)\, \mu(dx \times du) < \infty. \tag{2.5}$$

Then there exists a stationary process X such that $(X, \eta(X,\cdot))$ is a stationary relaxed solution of the controlled martingale problem for (A, μ_0).

Remark 2.3. It will be clear from the proof that there always exists a modification of X with sample paths in $D_{E^\Delta}[0,\infty)$ (where E^Δ is the one-point compactification of E), but our assumptions do not imply that the process will have sample paths in $D_E[0,\infty)$. For example, let $Af = (1+x^4)(f''(x) + f'(x))$. It is easy to check that $\mu(dx) = c(1+x^4)^{-1}dx$ satisfies $\int_{\mathrm{R}} Af(x)\mu(dx) = 0$, but the corresponding process will repeatedly "go out" at $+\infty$ and "come back in" at $-\infty$.

For clarity of exposition, we break the proof into two parts. The main part of the proof is given in the next theorem, which differs from the more general result in that the range of the generator A consists of *bounded* continuous functions. The second part of the proof consists of applying the theorem to bounded generators which approximate A and showing relative compactness of the solutions.

THEOREM 2.4 (cf. [2, Theorem 2.1 and Corollary 2.1]). *Suppose that E, U are as in section* 1.1. *Let $A : \mathcal{D}(A) \subset \widehat{C}(E) \to \overline{C}(E \times U)$ satisfy conditions* (i)–(v). *Suppose $\mu \in \mathcal{P}(E \times U)$ satisfies* (2.1). *Let η satisfy* (2.2). *Then there exists a stationary process X such that $(X, \eta(X,\cdot))$ is a stationary relaxed solution of the controlled martingale problem for (A, μ_0).*

Proof. As in [18, Theorem 4.1], we may assume E is compact and $A1 = 0$, where 1 denotes the constant function 1, by using E^Δ and extending A to the space $C(E^\Delta)$, if necessary. For $n = 1, 2, 3, \ldots$, define the Yosida approximations A_n by $A_n g = n[(I - n^{-1}A)^{-1} - I]g$ for $g \in \mathcal{R}(I - n^{-1}A)$ and note that for $f \in \mathcal{D}(A)$ and $g = (I - n^{-1}A)f$, $A_n g = Af$.

Let M be the linear subspace of functions of the form

$$\begin{aligned} F(x_1,x_2,u_1,u_2) = &\sum_{i=1}^m \left\{ h_i(x_1) \left[(I - n^{-1}A) f_i(x_2,u_1) + g_i(x_2,u_2) - g_i(x_2,u_1) \right] \right\} \\ &+ h_0(x_2,u_1,u_2), \end{aligned} \tag{2.6}$$

where $h_1, \ldots, h_m \in \overline{C}(E)$, $h_0 \in \overline{C}(E \times U \times U)$, $f_1, \ldots, f_m \in \mathcal{D}(A)$, and $g_1, \ldots, g_m \in \overline{C}(E \times U)$. Define the linear functional Ψ on M by

(2.7)
$$\begin{aligned}\Psi F &= \int_{E\times U}\int_U \sum_{i=1}^m \{h_i(x_2)[f_i(x_2)+g_i(x_2,u_2)-g_i(x_2,u_1)]\}\,\eta(x_2,du_2)\,\mu(dx_2\times du_1)\\ &\quad+\int_{E\times U}\int_U h_0(x_2,u_1,u_2)\,\eta(x_2,du_2)\,\mu(dx_2\times du_1)\\ &=\int_{E\times U}\int_U\left[\sum_{i=1}^m h_i(x_2)f_i(x_2)+h_0(x_2,u_1,u_2)\right]\eta(x_2,du_2)\,\mu(dx_2\times du_1),\end{aligned}$$

in which the second representation follows from the fact that

$$\int_{E\times U}\int_U h(x_2)[g(x_2,u_2)-g(x_2,u_1)]\eta(x_2,du_2)\,\mu(dx_2\times du_1)=0 \tag{2.8}$$

(write $\mu(dx_2\times du_1)=\eta(x_2,du_1)\mu_0(dx_2)$). Also define the linear operator Π: $\mathrm{B}(E\times E\times U\times U)\to\mathrm{B}(E\times E\times U)$ by

$$\Pi F(x_1,x_2,u_1)=\int_U F(x_1,x_2,u_1,u_2)\,\eta(x_2,du_2) \tag{2.9}$$

and the functional p on $\mathrm{B}(E\times E\times U\times U)$ by

$$p(F)=\int_{E\times U}\sup_{x_1}|\Pi F(x_1,x_2,u_1)|\mu(dx_2\times du_1). \tag{2.10}$$

Observe that $\Pi(\Pi F)=\Pi F$, so

$$p(F-\Pi F)=0. \tag{2.11}$$

In order to simplify notation, define the operator B on $\overline{C}(E\times U)$ by

$$Bg(x_2,u_1)=\int_U[g(x_2,u_2)-g(x_2,u_1)]\,\eta(x_2,du_2). \tag{2.12}$$

First, we claim $|\Psi F|\le p(F)$. To see this, fix $F\in M$. Fix $\alpha_i\ge||(I-n^{-1}A)f_i+Bg_i||\vee||f_i||, i=1,\dots,m$, and let ϕ be a polynomial on $\mathbb{R}^m$ which is convex on $\prod_{i=1}^m[-\alpha_i,\alpha_i]$. By the convexity of ϕ and [18, Lemma 3.5],

$$\begin{aligned}&\phi((I-n^{-1}A)f_1+Bg_1,\dots,(I-n^{-1}A)f_m+Bg_m)\\ &\quad\ge\phi(f_1,\dots,f_m)-n^{-1}\nabla\phi(f_1,\dots,f_m)\cdot(Af_1,\dots,Af_m)\\ &\qquad+\nabla\phi(f_1,\dots,f_m)\cdot(Bg_1,\dots,Bg_m)\\ &\quad\ge\phi(f_1,\dots,f_m)-n^{-1}A\phi(f_1,\dots,f_m)+\nabla\phi(f_1,\dots,f_m)\cdot(Bg_1,\dots,Bg_m).\end{aligned}$$

In light of (2.1) and (2.8), integration with respect to μ yields

$$\int\phi((I-n^{-1}A)f_1+Bg_1,\dots,(I-n^{-1}A)f_m+Bg_m)\,d\mu\ge\int\phi(f_1,\dots,f_m)\,d\mu, \tag{2.13}$$

and this inequality can be extended to arbitrary convex functions. Consider, in particular, the convex function $\phi(r_1, \dots, r_m) = \sup_{x_1} \sum_{i=1}^m h_i(x_1) r_i$. It follows that

$$\begin{aligned}
\Psi F &\le \int_{E\times U} \left\{ \sup_{x_1} \sum_{i=1}^m h_i(x_1) f_i(x_2) + \int_U h_0(x_2, u_1, u_2)\, \eta(x_2, du_2) \right\} \mu(dx_2 \times du_1) \\
&= \int_{E\times U} \left\{ \phi(f_1, \dots, f_m)(x_2) + \int_U h_0(x_2, u_1, u_2)\, \eta(x_2, du_2) \right\} \mu(dx_2 \times du_1) \\
&\le \int_{E\times U} \big\{ \phi((I - n^{-1}A) f_1 + Bg_1, \dots, (I - n^{-1}A) f_m + Bg_m)(x_2, u_1) \\
&\qquad\qquad + \int_U h_0(x_2, u_1, u_2)\, \eta(x_2, du_2) \Big\} \mu(dx_2 \times du_1) \\
&= \int_{E\times U} \sup_{x_1} \Pi F(x_1, x_2, u_1)\, \mu(dx_2 \times du_1) \\
&\le p(F).
\end{aligned}$$

Also, $-\Psi F = \Psi(-F) \le p(-F) = p(F)$, so $|\Psi F| \le p(F)$.

Since $\Psi 1 = 1$, observe that for $F \ge 0$,

$$||F|| - \Psi F = \Psi(||F|| - F) \le ||\, ||F|| - F|| \le ||F||,$$

so $\Psi F \ge 0$. As a result, we can apply the Hahn–Banach theorem (cf. [17, p. 187]) to extend Ψ to the entire space $\overline{C}(E \times E \times U \times U)$, still satisfying $|\Psi F| \le p(F)$, and the extension of the Riesz representation theorem in [1, Theorem 2.3] to conclude that there exists a measure $\nu \in \mathcal{P}(E \times E \times U \times U)$ such that

$$\Psi F = \int_{E\times E\times U\times U} F(x_1, x_2, u_1, u_2)\, \nu(dx_1 \times dx_2 \times du_1 \times du_2). \tag{2.14}$$

By considering functions F of particular forms, we observe some of the consequences of this representation of Ψ. First, for F of the form $F(x_1, x_2, u_1, u_2) = h(x_1)(I - n^{-1}A)1(x_2, u_1)$, with 1 being the constant function, it is clear that $\nu(\cdot \times E \times U \times U) = \mu_0(\cdot)$. Second, consider F of the form $F(x_1, x_2, u_1, u_2) = h(x_1)(I - n^{-1}A) f(x_2, u_1)$. Letting $\nu(dx_1 \times dx_2 \times du_1 \times du_2) = \tilde{\eta}(x_1, dx_2 \times du_1 \times du_2)\mu_0(dx_1)$, we have

$$\begin{aligned}
&\int_E h(x_1) f(x_1)\, \mu_0(dx_1) \\
&\qquad = \Psi F \\
&\qquad = \int_{E\times E\times U\times U} h(x_1)(I - n^{-1}A) f(x_2, u_1)\, \nu(dx_1 \times dx_2 \times du_1 \times du_2) \\
&\qquad = \int_E h(x_1) \left[\int_{E\times U} (I - n^{-1}A) f(x_2, u_1)\, \tilde{\eta}(x_1, dx_2 \times du_1 \times U) \right] \mu_0(dx_1),
\end{aligned}$$

and thus letting $\widehat{\eta}(x_1, dx_2 \times du_1) = \tilde{\eta}(x_1, dx_2 \times du_1 \times U)$, it follows that

$$\int_{E\times U} (I - n^{-1}A) f(x_2, u_1) \widehat{\eta}(x_1, dx_2 \times du_1) = f(x_1) \quad \text{a.e. } \mu_0(dx_1). \tag{2.15}$$

Third, observe that $\Psi F = \Psi(\Pi F)$ by (2.11). With $F(x_1, x_2, u_1, u_2) = f(x_1, x_2, u_1) g(u_2)$ and writing $\nu(dx_1 \times dx_2 \times du_1 \times du_2) = \bar{\eta}(x_1, x_2, u_1, du_2)\bar{\nu}(dx_1 \times dx_2 \times du_1)$, we thus

have

$$\begin{aligned}\int_{E\times E\times U} f(x_1,x_2,u_1)\left[\int_U g(u_2)\,\bar\eta(x_1,x_2,u_1,du_2)\right]\bar\nu(dx_1\times dx_2\times du_1)\\ &= \Psi F\\ &= \Psi(\Pi F)\\ &= \int_{E\times E\times U} f(x_1,x_2,u_1)\left[\int_U g(u_2)\eta(x_2,du_2)\right]\bar\nu(dx_1\times dx_2\times du_1).\end{aligned}$$

Therefore

$$\begin{aligned}\nu(dx_1\times dx_2\times du_1\times du_2) &= \eta(x_2,du_2)\bar\nu(dx_1\times dx_2\times du_1)\\ &= \eta(x_2,du_2)\widehat\eta(x_1,dx_2\times du_1)\mu_0(dx_1).\end{aligned}$$

Furthermore, using $F(x_1,x_2,u_1,u_2)=h(x_1)[g(x_2,u_2)-g(x_2,u_1)]$, it follows that

$$\begin{aligned}0 &= \Psi F\\ &= \int_E h(x_1)\left[\int_{E\times U}\int_U \{g(x_2,u_2)-g(x_2,u_1)\}\eta(x_2,du_2)\widehat\eta(x_1,dx_2\times du_1)\right]\mu_0(dx_1)\end{aligned}$$

and so

$$\int_{E\times U}\int_U \{g(x_2,u_2)-g(x_2,u_1)\}\eta(x_2,du_2)\widehat\eta(x_1,dx_2\times du_1)=0 \quad \text{a.e. } \mu_0(dx_1). \tag{2.16}$$

Let $\{(X_k,u_k)\colon k=1,2,\dots\}$ be a Markov chain on $E\times U$ having initial distribution μ and transition function $\widehat\eta$. A straightforward computation shows that the Markov chain is stationary, and by (2.15) and (2.16), for each $f\in\mathcal{D}(A)$ and $g\in\overline{C}(E\times U)$,

$$[(I-n^{-1}A)f](X_k,u_k)-\sum_{i=0}^{k-1} n^{-1}A_n[(I-n^{-1}A)f](X_i,u_i)$$

and

$$\sum_{i=0}^{k} Bg(X_i,u_i)$$

are martingales with respect to the filtration $\mathcal{F}_k=\sigma((X_i,u_i)\colon 0\le i\le k)$.

Define $X_n(\cdot)=X_{[n\cdot]}$, $u_n(\cdot)=u_{[n\cdot]}$, and $\mathcal{F}_t^n=\sigma((X_n(s),u_n(s))\colon 0\le s\le t)$. It immediately follows (recall $A_n(I-n^{-1}A)f=Af$) that

$$[(I-n^{-1}A)f](X_n(t),u_n(t))-\int_0^{[nt]/n} Af(X_n(s),u_n(s))\,ds \tag{2.17}$$

and

$$\int_0^{[nt]/n} Bg(X_n(s),u_n(s))\,ds \tag{2.18}$$

are $\mathcal{F}_t^n$-martingales.

Define the measure-valued random variable Γ_n by

$$\Gamma_n([0,t]\times G) = \int_0^{[nt]/n} I_G(u_n(s))\,ds.$$

Let $\mathcal{L}(U)$ be the space of measures ξ on $[0,\infty)\times U$ such that $\xi([0,t]\times U)<\infty$ for each $t>0$. We take the topology on $\mathcal{L}(U)$ such that $\xi_n\to\xi$ if and only if

$$\int f\,d\xi_n \to \int f\,d\xi$$

for every $f\in \overline{C}([0,\infty)\times U)$ with $\operatorname{supp}(f)\subset[0,t_f]\times U$ for some $t_f<\infty$. Relative compactness of $\{\Gamma_n\}$ follows from the fact that

$$E[\Gamma_n([0,t]\times U)]\le t,$$

and for each $\epsilon>0$, there exists a compact set $K\subset U$ such that

$$E[\Gamma_n([0,t]\times K^c)] = \frac{[nt]}{n}\mu(E\times K^c)\le \epsilon t.$$

See [14, Corollary 1.2]. It follows that (X_n,Γ_n) is relatively compact in $D_E[0,\infty)\times\mathcal{L}(U)$. Along any convergent subsequence with limit point (X,Γ), by [14, Lemma 1.5]

$$\int_{[0,\cdot]\times U} g(X_n(s),u)\Gamma_n(ds\times du) \Rightarrow \int_{[0,\cdot]\times U} g(X(s),u)\Gamma(ds\times du),$$

and as in the proof of [14, Theorem 2.1], there exists a filtration $\{\mathcal{G}_t\}$ such that

$$f(X(t)) - \int_0^t\int_U Af(X(s),u)\,\Gamma(ds\times du) \tag{2.19}$$

is a $\{\mathcal{G}_t\}$-martingale for each $f\in\mathcal{D}(A)$. Since X_n is stationary for time shifts that are multiples of n^{-1}, it follows that X is a stationary process.

Since (2.18) is a martingale, for $h_1,\dots,h_j\in\overline{C}(E), g\in\overline{C}(E\times U)$, and $0\le t_1\le\cdots\le t_j<t_{j+1}$,

$$\begin{aligned}E\left[\int_{[nt_j]/n}^{[nt_{j+1}]/n}\int_U g(X_n(s),u)\,\eta(X_n(s),du)\,ds\prod_{i=1}^j h_i(X_n(t_i))\right]\\ = E\left[\int_{[nt_j]/n}^{[nt_{j+1}]/n}\int_U g(X_n(s),u_n(s))\,\Gamma_n(ds\times du)\prod_{i=1}^j h_i(X_n(t_i))\right].\end{aligned}$$

By the weak convergence of (X_n,Γ_n) to (X,Γ) and Lemma 2.1, letting $n\to\infty$, we obtain

$$\begin{aligned}E\left[\int_{t_j}^{t_{j+1}}\int_U g(X(s),u)\,\eta(X(s),du)\,ds\prod_{i=1}^j h_i(X(t_i))\right]\\ = E\left[\int_{t_j}^{t_{j+1}}\int_U g(X(s),u)\,\Gamma(ds\times du)\prod_{i=1}^j h_i(X(t_i))\right].\end{aligned}$$

In particular, this implies, by taking $g = Af$, that

$$f(X(t)) - \int_0^t \int_U Af(X(s),u)\,\eta(X(s),du)\,ds$$

is an $\{\mathcal{F}_t^X\}$-martingale. □

Proof of Theorem 2.2. For each $n \geq 1$, let $\psi_n = 2^{-n}(2^n \vee \psi)$, $k_n(x) = \int \psi_n(u)\,\eta(x,du)$, and $c_n = \int k_n(x)\,\mu_0(dx) = \int \psi_n(u)\,\mu(dx \times du)$. Observe that $\psi_n \geq 1$ for all n and as $n \to \infty$, $\psi_n(u) \to 1$, $c_n \searrow 1$, and $k_n \searrow 1$. Define the operators A_n on $\mathcal{D}(A)$ by

$$A_n f(x,u) = Af(x,u)/\psi_n(u),$$

and note that $A_n : \mathcal{D}(A) \to \overline{C}(E \times U)$. Also for each n, define the measure $\mu_n \in \mathcal{P}(E \times U)$ by

$$\mu_n(\Gamma) = c_n^{-1} \int_\Gamma \psi_n(u)\mu(dx \times du) \quad \forall \Gamma \in \mathcal{B}(E \times U).$$

Observe that the state marginal of μ_n has density $\frac{k_n(x)}{c_n}$ with respect to μ:

$$\mu_n^0(dx) = \frac{k_n(x)}{c_n}\,\mu_0(dx),$$

and the conditional distribution of u given x under μ_n is given by

$$\eta_n(x,du) = \frac{\psi_n(u)}{k_n(x)}\eta(x,du).$$

The pairs (A_n, μ_n) satisfy the conditions of Theorem 2.4, so there exist stationary processes $\{X_n\}$ such that $(X_n, \eta_n(X_n,\cdot))$ is a solution of the controlled martingale problem for (A_n, μ_n^0).

The relative compactness of $\{X_n\}$ is established by applying [9, Theorem 3.9.1] and [18, Theorem 4.5] exactly as in [18, Theorem 4.7]. Let X be a weak limit of X_n, and without loss of generality, assume the entire sequence converges.

By a monotone class argument, for each $f \in \mathcal{D}(A)$,

$$f(X(t)) - \int_0^t \int_U Af(X(s),u)\,\eta(X(s),du)\,ds$$

is an $\{\mathcal{F}_t^X\}$-martingale if and only if

$$E\left[\left(f(X(t_{m+1})) - f(X(t_m)) - \int_{t_m}^{t_{m+1}} \int_U Af(X(s),u)\,\eta(X(s),du)\,ds\right)\prod_{i=1}^m h_i(X(t_i))\right] = 0 \tag{2.20}$$

for each $m \geq 1$ and $0 \leq t_1 \leq \cdots \leq t_m < t_{m+1}$ and $h_1, \ldots, h_m \in \overline{C}(E)$.

Note that condition (2.21) is satisfied with A_n, η_n, and X_n replacing A, η, and X since $(X_n, \eta_n(X_n,\cdot))$ is a solution of the controlled martingale problem for A_n.

Fix $f \in \mathcal{D}(A)$, $t_1, \ldots, t_{m+1} \in \{t \geq 0 : P(X(t) = X(t-)) = 1\}$, and $h_1, \ldots, h_m \in \overline{C}(E)$. Since $X_n \Rightarrow X$ as $n \to \infty$,

$$E\left[(f(X_n(t_{m+1})) - f(X_n(t_m)))\prod_{i=1}^m h_i(X_n(t_i))\right]$$
$$\to E\left[(f(X(t_{m+1})) - f(X(t_m)))\prod_{i=1}^m h_i(X(t_i))\right].$$

Lemma 2.1 does not apply directly to the integral terms since the distributions of $X_n(t)$ and $X(t)$ are not the same. However, a similar argument can be used by approximating $\int_U Af(x,u)\,\eta(x,du)$ by a continuous function in $\mathcal{L}^1(\mu_0)$ and using the facts that $X_n(t)$ has distribution $\mu_n^0(dx) = k_n(x)/c_n\mu_0(dx)$ and $k_n \searrow 1$ as $n \to \infty$. Therefore, (2.20) is established for $t_1, \ldots, t_{m+1} \in \{t \geq 0 : P(X(t) = X(t-)) = 1\}$. The result is extended to all $t_1, \ldots, t_{m+1}$ by the right continuity of X. Thus

$$f(X(t)) - \int_0^t \int_U Af(X(s),u)\,\eta(X(s),du)\,ds$$

is an $\{\mathcal{F}_t^X\}$-martingale. □

3. Stationary solutions and the forward equation for Markov processes. Theorem 2.2 extends Echeverria's theorem (cf. [9, Theorem 4.9.17]) to include control in the dynamics. This theorem can in turn be used to extend the result in the uncontrolled setting to operators with range in $M(E)$, the (not necessarily bounded) measurable functions on E; that is, we relax both the boundedness and the continuity assumptions of the previous results.

THEOREM 3.1. *Let E be locally compact and separable. Let $\widehat{A} : \mathcal{D}(\widehat{A}) \subset \widehat{C}(E) \to M(E)$, let $\mathcal{D}(\widehat{A})$ be an algebra, and let $\widehat{\mu} \in \mathcal{P}(E)$ satisfy*

$$\int_E \widehat{A}f(x)\,\widehat{\mu}(dx) = 0 \qquad \forall f \in \mathcal{D}(\widehat{A}). \tag{3.1}$$

Suppose that there exists a locally compact, separable, metric space U, an operator $A : \mathcal{D}(A) \equiv \mathcal{D}(\widehat{A}) \subset \widehat{C}(E) \to C(E \times U)$ satisfying conditions (i)–(iv)*, a transition function η from E to U such that*

$$\widehat{A}f(x) = \int_U Af(x,u)\,\eta(x,du) \qquad \forall f \in \mathcal{D}(\widehat{A}),$$

and a $\psi \in C(U)$ satisfying condition (vi) *and*

$$\int_{E\times U} \psi(u)\,\eta(x,du)\widehat{\mu}(dx) < \infty.$$

Then there exists a stationary solution X of the (uncontrolled) martingale problem for $(\widehat{A}, \widehat{\mu})$.

Proof. This theorem follows immediately from Theorem 2.2, defining $\mu \in \mathcal{P}(E \times U)$ by $\mu(dx \times du) = \eta(x,du)\widehat{\mu}(dx)$. □

Theorem 3.1 immediately gives a generalization of Proposition 4.9.19 of [9] regarding solutions of the forward equation

$$\int_E f\,d\nu_t = \int_E f\,d\nu_0 + \int_0^t \int_E \widehat{A}f\,d\nu_s\,ds\,, \quad f \in \mathcal{D}(\widehat{A}), \tag{3.2}$$

for a $\mathcal{P}(E)$-valued function ν. The proof of the next corollary uses the argument of Theorem 4.1 in the next section.

COROLLARY 3.2. *Let $\widehat{A} : \mathcal{D}(\widehat{A}) \subset \widehat{C}(E) \to M(E)$, A, and ψ be as in Theorem* 3.1*, and let ν satisfy* (3.2) *and*

$$\int_0^\infty e^{-\lambda s} \int_{E\times U} \psi(u)\, \eta(x, du)\nu_s(dx)ds < \infty \tag{3.3}$$

for all sufficiently large $\lambda > 0$. Then there exists a solution X of the martingale problem for $(\widehat{A}, \nu_0)$ such that $X(t)$ has distribution ν_t. If uniqueness holds for the martingale problem for $(\widehat{A}, \nu_0)$, then uniqueness holds for (3.2) *among solutions satisfying the integrability condition* (3.3).

Proof. Existence of the process X follows by the proof of Theorem 4.1 when the measure $\pi \in \mathcal{P}(\mathbb{R}^+ \times E \times U)$ is defined to satisfy

$$\int_{\mathbb{R}^+\times E\times U} h(s, x, u)\pi(ds \times dx \times du) = \alpha \int_0^\infty e^{-\alpha s} \int_E \int_U h(s, x, u)\eta(x, du)\nu_s(dx)ds$$

for $h \in \overline{C}(\mathbb{R}^+ \times E \times U)$. The proof of uniqueness is identical to that of Proposition 4.9.19 of [9]. □

Example 3.3 (linear combinations of generators). Suppose

$$\widehat{A}f(x) = \sum_{k=1}^m \beta_k(x)A_k f(x),$$

in which each A_k satisfies conditions (i)–(v), $A_1, \ldots, A_m$ have a common domain, and the coefficients β_k are only assumed to be nonnegative and measurable. Suppose $\widehat{\mu}$ satisfies (3.1) and

$$\int_E \sum_k \beta_k(x)\widehat{\mu}(dx) < \infty.$$

Then there exists a stationary solution of the martingale problem for $(\widehat{A}, \widehat{\mu})$. To see that Theorem 3.1 applies, take $U = [0, \infty)^m$, $Af(x, u) = \sum_{k=1}^m u_k A_k f(x)$, $\eta(x, du) = \prod_{k=1}^m \delta_{\{\beta_k(x)\}}(du_k)$, and $\psi(u) = \sum_{k=1}^m |u_k|$. Similarly, if $\{\nu_t\}$ satisfies (3.2) and

$$\int_0^\infty e^{-\lambda s} \int_E \sum_k \beta_k(x)\nu_s(dx)ds < \infty$$

for all sufficiently large λ, then by Corollary 3.2 there exists a solution X of the martingale problem for $\widehat{A}$ such that ν_t is the distribution of $X(t)$.

Example 3.4 (diffusion operators with discontinuous coefficients). Consider the diffusion generator on $\mathbb{R}^d$,

$$\widehat{A}f(x) = \frac{1}{2}\sum_{i,j=1}^d a_{ij}(x)\frac{\partial^2}{\partial x_i \partial x_j}f(x) + \sum_{i=1}^d b_i(x)\frac{\partial}{\partial x_i}f(x)$$

with domain $\mathcal{D}(\widehat{A}) = C_c^2(\mathbb{R}^d)$, in which the coefficients are only assumed to be measurable. Suppose that $\widehat{\mu} \in \mathcal{P}(E)$ satisfies

$$\int_E \widehat{A}f(x)\, \widehat{\mu}(dx) = 0 \qquad \forall f \in \mathcal{D}(\widehat{A}) \tag{3.4}$$

and

$$\int [\|a(x)\| + |b(x)|]\, \widehat{\mu}(dx) < \infty,$$

where $a(x) = ((a_{ij}(x)))$ and $b(x) = (b_1(x), \ldots, b_d(x))$.

Let $U = \mathbb{M}_+^{d\times d} \times \mathbb{R}^d$, where $\mathbb{M}_+^{d\times d}$ denotes the space of nonnegative definite $d \times d$ matrices, and, with $u = (u^1, u^2)$, define

$$Af(x,u) = \frac{1}{2}\sum_{i,j=1}^{d} u^1_{ij} \frac{\partial^2}{\partial x_i \partial x_j} f(x) + \sum_{i=1}^{d} u^2_i \frac{\partial}{\partial x_i} f(x)\ ,$$

$\eta(x, du) = \delta_{\{a(x)\}}(du^1)\delta_{\{b(x)\}}(du^2)$, and $\psi(u) = \|u^1\| + |u^2|$. Then Theorem 3.1 gives the existence of a stationary solution of the martingale problem for $(\widehat{A}, \widehat{\mu})$. Similarly, if there exists a solution of (3.2) satisfying

$$\int_0^\infty e^{-\lambda s} \int [\|a(x)\| + |b(x)|]\, \nu_s(dx)ds < \infty$$

for all sufficiently large $\lambda > 0$, then there exists a solution X of the martingale problem for $\widehat{A}$ such that for each $t \geq 0$, ν_t is the distribution of $X(t)$.

Example 3.5 (jump processes). Let E be locally compact and

$$\widehat{A}f(x) = \lambda(x) \int_E (f(y) - f(x))\gamma(x, dy),$$

where λ is a nonnegative measurable function on E and γ is a transition function on E. Suppose that $\widehat{\mu} \in \mathcal{P}(E)$ satisfies (3.1) and $\int_E \lambda(x)\widehat{\mu}(dx) < \infty$. Let E^Δ denote the one-point compactification of E and $U = [0, \infty) \times \mathcal{P}(E^\Delta)$. Define

$$Af(x,u) = u_1 \int (f(y) - f(x))u_2(dy)\ ,$$

$\psi(u) = u_1$, and $\eta(x, du) = \delta_{\{\lambda(x)\}}(du_1)\delta_{\{\gamma(x,\cdot)\}}(du_2)$. Then Theorem 3.1 gives the existence of a stationary solution of the martingale problem for $(\widehat{A}, \widehat{\mu})$. Similarly, if there exists a solution of (3.2) satisfying

$$\int_0^\infty e^{-\lambda s}\lambda(x)\, \nu_s(dx)ds < \infty$$

for all sufficiently large $\lambda > 0$, then there exists a solution X of the martingale problem for $\widehat{A}$ such that for each $t \geq 0$, ν_t is the distribution of $X(t)$.

4. Feedback controls. We now use Theorem 2.2 to show that for each solution (X, Λ) of the controlled martingale problem for A there is a process Y and a transition function $\eta(s, y, du)$, $(s, y) \in [0, \infty) \times E$, such that $\{(Y(s), \eta(s, Y(s), \cdot)) : s \geq 0\}$ is a solution of the controlled martingale problem for A and for each $s \geq 0$, $(Y(s), \eta(s, Y(s), \cdot))$ has the same distribution as $(X(s), E[\Lambda_s|X(s)])$. To obtain the desired process, we introduce a time-space generator $\tilde{A}$, identify a stationary distribution for $\tilde{A}$, and apply Theorem 2.2 to obtain a (stationary) solution of the martingale problem for $\tilde{A}$. An absolutely continuous change of measure produces the desired solution of the controlled martingale problem for A.

THEOREM 4.1 (cf. [2, Theorem 2.4]). *Suppose E, U, A, and ψ satisfy the conditions of section* 1.1. *Let (X, Λ) be a relaxed solution of the controlled martingale problem for (A, ν_0), and suppose there exists $\alpha > 0$ such that*

$$\int_0^\infty e^{-\alpha t} E\left[\int_U \psi(u)\Lambda_t(du)\right] dt < \infty. \tag{4.1}$$

Then there exists a process Y and a transition function η from $[0,\infty) \times E$ to U such that $\{(Y(s), \eta(s, Y(s), \cdot)) : s \geq 0\}$ is a relaxed solution of the controlled martingale problem for (A, ν_0) and for each $t \geq 0$, the distribution of $(Y(t), \eta(t, Y(t), \cdot))$ on $E \times \mathcal{P}(U)$ is the same as $(X(t), E[\Lambda_t(\cdot)|X(t)])$.

Proof. Define the time-space generator

$$\tilde{A}(\gamma f)(s,x,u) = \gamma(s)Af(x,u) + \gamma'(s)f(x) + \alpha\left[\gamma(0)\int_E f(y)\nu_0(dy) - \gamma(s)f(x)\right] \tag{4.2}$$

for $f \in \mathcal{D}(A)$ and $\gamma \in \widehat{C}^1(\mathbb{R}^+)$ and observe that $\tilde{A}$ satisfies conditions (i)–(iv).

Define the measure $\pi \in \mathcal{P}(\mathbb{R}^+ \times E \times U)$ by

$$\int_{\mathbb{R}^+\times E\times U} h(s,x,u)\pi(ds \times dx \times du) = \alpha \int_0^\infty e^{-\alpha s} E\left[\int_U h(s, X(s), u)\Lambda_s(du)\right] ds \tag{4.3}$$

for $h \in \overline{C}(\mathbb{R}^+ \times E \times U)$. The following computation verifies that π is a stationary distribution for $\tilde{A}$. Let $f \in \mathcal{D}(A)$ and $\gamma \in \widehat{C}^1(\mathbb{R}^+)$. Then

$$\begin{aligned}
&\int_{\mathbb{R}^+\times E\times U} \tilde{A}(\gamma f)(s,x,u)\,\pi(ds \times dx \times du) \\
&\quad = \alpha\int_0^\infty \gamma(s)e^{-\alpha s}E\left[\int_U Af(X(s),u)\Lambda_s(du)\right] ds + \alpha\int_0^\infty \gamma'(s)e^{-\alpha s}E[f(X(s))]\,ds \\
&\qquad + \alpha^2\int_0^\infty e^{-\alpha s}\,ds\gamma(0)E[f(X(0))] - \alpha^2\int_0^\infty \gamma(s)e^{-\alpha s}E[f(X(s))]\,ds \\
&\quad = \alpha\int_0^\infty \gamma(s)e^{-\alpha s}E\left[\int_U Af(X(s),u)\Lambda_s(du)\right] ds \\
&\qquad + \alpha E\left[\int_0^\infty \frac{d}{ds}(\gamma(s)e^{-\alpha s})f(X(s))\,ds\right] + \alpha\gamma(0)E[f(X(0))] \\
&\quad = \alpha\int_0^\infty \gamma(s)e^{-\alpha s}E\left[\int_U Af(X(s),u)\Lambda_s(du)\right] ds + E\left[\alpha\gamma(s)e^{-\alpha s}f(X(s))\,|_{s=0}^{s=\infty}\right] \\
&\qquad - \alpha\int_0^\infty \gamma(s)e^{-\alpha s}E\left[\int_U Af(X(s),u)\,\Lambda_s(du)\right] ds + \alpha\gamma(0)E[f(X(0))] \\
&\quad = 0.
\end{aligned}$$

Let $\eta(s,x,du)$ denote the regular conditional distribution of u given (s,x) such that

$$\pi(\Gamma_1\times\Gamma_2)=\int_{\Gamma_1}\eta(s,x,\Gamma_2)\,\pi(ds\times dx\times U)\qquad\forall\Gamma_1\in\mathcal{B}(\mathbb{R}^+\times E),\Gamma_2\in\mathcal{B}(U).$$

Since π is a stationary distribution for $\tilde{A}$ and the definition of π together with (4.1) implies (2.5) is satisfied, Theorem 2.2 gives the existence of a stationary time-space process $\{(S(t),Z(t)):t\geq 0\}$ and a filtration $\mathcal{G}_t$ such that $\{(S(t),Z(t),\eta(S(t),Z(t),\cdot)):t\geq 0\}$ is a relaxed solution of the controlled martingale problem for $\tilde{A}$ with distribution π.

Before constructing the desired process Y, we investigate the stationary time process S more carefully. For simplicity, assume that the state space E is compact and $A1=0$; otherwise compactify E as in Theorem 2.2. Then, by choosing $f=1$, we see that

$$\gamma(S(t))-\int_0^t\left(\gamma'(S(r))+\alpha[\gamma(0)-\gamma(S(r))]\right)dr \tag{4.4}$$

is an $\{\mathcal{F}_t^{S,Z}\}$-martingale for every $\gamma\in\widehat{C}^1(\mathbb{R}^+)$. By [9, Theorem 4.4.1], uniqueness holds for the martingale problem (4.4). Now define a process $\tilde{S}$ as follows. Let $\Delta_1,\Delta_2,\Delta_3,\ldots$ be a sequence of independent exponential random variables with parameter α and let $\tilde{S}(0)$ also be an exponential random variable with parameter α which is independent of $\{\Delta_n\}$. Let

$$\tilde{S}(t)=\begin{cases}\tilde{S}(0)+t, & 0\leq t<\Delta_1,\\ \\ t-\Delta_n, & \Delta_n\leq t<\Delta_{n+1}.\end{cases}$$

$\tilde{S}$ is a stationary solution of (4.4) and thus the process S has been identified. Note that we may assume that (S,Z) is defined for all $t\in(-\infty,\infty)$.

Now let $\tau_1=\inf\{t>0:S(t)=0\}$ and, for $k\geq 1$, $\tau_{k+1}=\inf\{t>\tau_k:S(t)=0\}$. Also define $\tau_{-1}=\sup\{t\leq 0:S(t)=0\}$. Define the process Y by $Y(t)=Z(\tau_1+t)$, $t\geq 0$, and the filtration $\mathcal{F}_t=\mathcal{G}_{\tau_1+t}$. Again, for simplicity, compactify the time dimension by taking the one-point compactification $[0,\infty]$ and extend the generator (4.2) by linearity, where for the constant function 1,

$$\tilde{A}(1f)(s,x,u)=Af(x,u)+\alpha\left[\int_E f(y)\nu_0(dy)-f(x)\right]. \tag{4.5}$$

An application of the optional sampling theorem (cf. [9, Theorem 2.2.13]) shows that (with $\gamma=1$)

$$f(Y(t))-\int_0^t\int_U\left[Af(Y(r),u)+\alpha\left(\int_E f(y)\nu_0(dy)-f(Y(r))\right)\right],\\ \eta(S(\tau_1+r),Y(r),du)\,dr$$

is an $\{\mathcal{F}_t\}$-martingale under P.

Define

$$L(t)=[\alpha(\tau_1-\tau_{-1})]^{-1}e^{\alpha t}I_{[0,\tau_2-\tau_1)}(t)$$

and observe that L is an $\{\mathcal{F}_t\}$-martingale with $E[L(t)] = 1$. Let $\widehat{P}$ be a new probability having Radon–Nikodým derivative $L(t)$ on $\mathcal{F}_t$ with respect to the original probability P. Denote expectation with respect to $\widehat{P}$ by $E^{\widehat{P}}[\cdot]$.

Remark 4.2. The Radon–Nikodým derivative $L(t)$ includes the term $[\alpha(\tau_1 - \tau_{-1})]^{-1}$ because we have been unable to show that the $\{\tau_k\}$ are regeneration times for (S, Z). (They are for S alone.) In fact, in general there will be stationary solutions of the controlled martingale problem for $\tilde{A}$ for which the $\{\tau_k\}$ are not regeneration times. When the $\{\tau_k\}$ are regeneration times for (S, Z), the independence between cycles implies

$$E^{\widehat{P}}\left[\int_0^\infty e^{-\alpha t}\int_U h(t, Y(t), u)\,\eta(t, Y(t), du)\,dt\right]$$
$$= E\left[\int_{\tau_1}^{\tau_2}\int_U h(S(t), Z(t), u)\,\eta(S(t), Z(t), du)\,dt\right]$$

and the following argument can be considerably simplified.

We claim that $\{(Y(t), \eta(t, Y(t), \cdot)), t \geq 0\}$ under $\widehat{P}$ is the desired solution. We will first show that this process is a solution to the original martingale problem for A.

A straightforward calculation shows that for $f \in \mathcal{D}(A)$,

$$\lim_{h\searrow 0} h^{-1}E\left[[\alpha(\tau_1 - \tau_{-1})]^{-1}\left\{e^{\alpha(t+h)}I_{[0,\tau_2-\tau_1)}(t+h)f(Y(t+h))\right.\right.$$
$$\left.\left. - e^{\alpha t}I_{[0,\tau_2-\tau_1)}(t)f(Y(t))\right\}|\mathcal{F}_t\right]$$
$$= [\alpha(\tau_1 - \tau_{-1})]^{-1}e^{\alpha t}I_{[0,\tau_2-\tau_1)}(t)\int_U Af(Y(t), u)\,\eta(t, Y(t), du),$$

which implies that

$$[\alpha(\tau_1 - \tau_{-1})]^{-1}e^{\alpha t}I_{[0,\tau_2-\tau_1)}(t)f(Y(t))$$
$$- \int_0^t [\alpha(\tau_1 - \tau_{-1})]^{-1}e^{\alpha s}I_{[0,\tau_2-\tau_1)}(s)\int_U Af(Y(s), u)\,\eta(s, Y(s), du)\,ds$$

is an $\{\mathcal{F}_t\}$-martingale under P. In particular,

$$0 = E\left[[\alpha(\tau_1 - \tau_{-1})]^{-1}\left\{e^{\alpha t_{n+1}}I_{[0,\tau_2-\tau_1)}(t_{n+1})f(Y(t_{n+1})) - e^{\alpha t_n}I_{[0,\tau_2-\tau_1)}(t_n)f(Y(t_n))\right.\right.$$
$$\left.\left. - \int_{t_n}^{t_{n+1}} e^{\alpha s}I_{[0,\tau_2-\tau_1)}(s)\int_U Af(Y(s), u)\,\eta(s, Y(s), du)\,ds\right\}\prod_{i=1}^n h_i(Y(t_i))\right]$$

$$= E\left[[\alpha(\tau_1-\tau_{-1})]^{-1}\left\{e^{\alpha t_{n+1}}I_{[0,\tau_2-\tau_1)}(t_{n+1})f(Y(t_{n+1}))\right.\right.$$
$$- E[e^{\alpha t_{n+1}}I_{[0,\tau_2-\tau_1)}(t_{n+1})|\mathcal{F}_{t_n}]f(Y(t_n))$$
$$- \int_{t_n}^{t_{n+1}} E[e^{\alpha t_{n+1}}I_{[0,\tau_2-\tau_1)}(t_{n+1})|\mathcal{F}_s]$$
$$\left.\left.\times\int_U Af(Y(s),u)\,\eta(s,Y(s),du)\,ds\right\}\prod_{i=1}^n h_i(Y(t_i))\right]$$

$$= E^{\widehat{P}}\left[\left\{f(Y(t_{n+1}))-f(Y(t_n))\right.\right.$$
$$\left.\left.-\int_{t_n}^{t_{n+1}}\int_U Af(Y(s),u)\eta(s,Y(s),du)\,ds\right\}\prod_{i=1}^n h_i(Y(t_i))\right]$$

for each $n\geq 1$, $0\leq t_1\leq\cdots\leq t_n<t_{n+1}$, and $h_1,\ldots,h_n\in\overline{C}(E)$. It follows that

$$f(Y(t))-f(Y(0))-\int_0^t\int_U Af(Y(s),u)\eta(s,Y(s),du)ds$$

is an $\{\mathcal{F}_t^Y\}$-martingale under $\widehat{P}$.

We now derive the one-dimensional distributions of this solution. First, for each $h\in\overline{C}(\mathbb{R}^+\times E\times U)$,

$$E^{\widehat{P}}\left[\alpha\int_0^T e^{-\alpha t}\int_U h(t,Y(t),u)\,\eta(t,Y(t),du)\,dt\right]$$
$$= E\left[[\alpha(\tau_1-\tau_{-1})]^{-1}\alpha e^{\alpha T}I_{[0,\tau_2-\tau_1)}(T)\int_0^T e^{-\alpha t}\int_U h(t,Y(t),u)\,\eta(t,Y(t),du)\,dt\right]$$
$$= E\left[\int_0^T(\tau_1-\tau_{-1})^{-1}E[e^{\alpha T}I_{[0,\tau_2-\tau_1)}(T)|\mathcal{F}_t]e^{-\alpha t}\int_U h(t,Y(t),u)\,\eta(t,Y(t),du)\,dt\right]$$
$$= E\left[(\tau_1-\tau_{-1})^{-1}\int_{\tau_1}^{\tau_2\wedge(\tau_1+T)}\int_U h(S(t),Z(t),u)\,\eta(S(t),Z(t),du)\,dt\right],$$

and letting $T\to\infty$ yields

$$E^{\widehat{P}}\left[\alpha\int_0^\infty e^{-\alpha t}\int_U h(t,Y(t),u)\,\eta(t,Y(t),du)\,dt\right]$$
$$= E\left[(\tau_1-\tau_{-1})^{-1}\int_{\tau_1}^{\tau_2}\int_U h(S(t),Z(t),u)\,\eta(S(t),Z(t),du)\,dt\right]. \tag{4.6}$$

Now, for $t\geq 0$, define $\tau_{-1}^t=\sup\{r\leq t: S(r)=0\}$, $\tau_1^t=\inf\{r>t: S(r)=0\}$, and $\tau_2^t=\inf\{r>\tau_1^t: S(r)=0\}$. Note that $\tau_i^0=\tau_i$ for $i=-1,1$. Observe that

$$(\tau_1^t - \tau_{-1}^t)^{-1} \int_{\tau_1^t}^{\tau_2^t} \int_U h(S(r), Z(r), u)\, \eta(S(r), Z(r), du)\, dr$$

is stationary in t and that for $t \in [\tau_k, \tau_{k+1})$,

$$(\tau_1^t - \tau_{-1}^t)^{-1} \int_{\tau_1^t}^{\tau_2^t} \int_U h(S(r), Z(r), u)\, \eta(S(r), Z(r), du)\, dr$$

$$= (\tau_{k+1} - \tau_k)^{-1} \int_{\tau_{k+1}}^{\tau_{k+2}} \int_U h(S(r), Z(r), u)\, \eta(S(r), Z(r), du)\, dr.$$

Let $N(t)$ denote the number of jumps in the interval $[0, t]$. Then, by stationarity (letting $\tau_0 = \tau_{-1}$),

$$E\left[(\tau_1 - \tau_{-1})^{-1} \int_{\tau_1}^{\tau_2} \int_U h(S(r), Z(r), u)\, \eta(S(r), Z(r), du)\, dr\right]$$

$$= E\left[T^{-1} \int_0^T (\tau_1^t - \tau_{-1}^t)^{-1} \int_{\tau_1^t}^{\tau_2^t} \int_U h(S(r), Z(r), u)\, \eta(S(r), Z(r), du)\, dr\, dt\right]$$

$$= T^{-1} E\left[\sum_{k=1}^{N(T)+1} \frac{T \wedge \tau_k - \tau_{k-1} \vee 0}{\tau_k - \tau_{k-1}} \times \left(\int_{\tau_k}^{\tau_{k+1}} \int_U h(S(r), Z(r), u)\, \eta(S(r), Z(r), du)\, dr\right)\right]$$

$$= T^{-1} E\left[\int_0^T \int_U h(S(r), Z(r), u)\, \eta(S(r), Z(r), du)\, dr\right]$$

$$- T^{-1} E\left[\int_0^{\tau_1 \wedge T} \left(1 - \frac{T \wedge \tau_1}{\tau_1 - \tau_{-1}}\right) \int_U h(S(r), Z(r), u)\, \eta(S(r), Z(r), du)\, dr\right]$$

$$+ T^{-1} E\left[I_{\{N(T)=1\}} \frac{\tau_1}{\tau_1 - \tau_{-1}} \int_T^{\tau_2} \int_U h(S(r), Z(r), u)\, \eta(S(r), Z(r), du)\, dr\right]$$

$$+ T^{-1} E\left[I_{\{N(T)>1\}} \int_T^{\tau_{N(T)+1}} \int_U h(S(r), Z(r), u)\, \eta(S(r), Z(r), du)\, dr\right]$$

$$+ T^{-1} E\left[I_{\{N(T)>0\}} \frac{T - \tau_{N(T)}}{\tau_{N(T)+1} - \tau_{N(T)}} \times \int_{\tau_{N(T)+1}}^{\tau_{N(T)+2}} \int_U h(S(r), Z(r), u)\, \eta(S(r), Z(r), du)\, dr\right].$$

The first term equals $\int h(s, x, u)\, \pi(ds \times dx \times du)$ and the other terms converge to 0

as $T \to \infty$. Thus

$$E\left[(\tau_1 - \tau_{-1})^{-1} \int_{\tau_1}^{\tau_2} \int_U h(S(r), Z(r), u)\, \eta(S(r), Z(r), du)\, dr\right]$$

$$= \int h(s, x, u)\pi(ds \times dx \times du) \tag{4.7}$$

$$= \alpha \int_0^\infty e^{-\alpha s} E\left[\int_U h(s, X(s), u)\, \Lambda_s(du)\right] ds.$$

Combining (4.6) and (4.7) yields

$$E^{\widehat{P}}\left[\alpha \int_0^\infty e^{-\alpha t} \int_U h(t, Y(t), u)\, \eta(t, Y(t), du)\, dt\right]$$

$$= E\left[\alpha \int_0^\infty e^{-\alpha t} \int_U h(t, X(t), u)\, \Lambda_t(du)\, dt\right]. \tag{4.8}$$

Let $\{h_n\} \subset \overline{C}(E)$ be a countable collection which is separating (cf. [9, p. 112]). Taking $h(t, x, u)$ in (4.8) to be of the form $\gamma(t)h_n(x)$, we see that

$$E^{\widehat{P}}[h_n(Y(t))] = E[h_n(X(t))] \quad \text{a.e. } t.$$

Since $\{h_n\}$ is separating, it follows that $Y(t) \stackrel{d}{=} X(t)$, a.e. t, and by right continuity we have

$$Y(t) \stackrel{d}{=} X(t) \quad \forall\ t. \tag{4.9}$$

More generally, modifying $\eta(t, x, du)$ at a set of t of measure zero if necessary, we have that for $h \in \overline{C}(E)$ and $g \in \overline{C}(U)$

$$E^{\widehat{P}}\left[h(Y(t)) \int_U g(u)\, \eta(t, Y(t), du)\right] = E\left[h(X(t)) \int_U g(u)\, \Lambda_t(du)\right] \quad \forall\ t.$$

The relation (4.9) then implies

$$E\left[h(X(t)) \int_U g(u)\, \eta(t, X(t), du)\right] = E\left[h(X(t)) \int_U g(u)\, \Lambda_t(du)\right]$$

for each $h \in \overline{C}(E)$ and $g \in \overline{C}(U)$. Since this is true for each bounded, continuous h, it follows that

$$E\left[\int_U g(u)\Lambda_t(du)|X(t)\right] = \int_U g(u)\eta(t, X(t), du) \quad \text{a.s.}$$

for each bounded continuous g, and hence that the distributions of $\eta(t, Y(t), \cdot)$ and $E[\Lambda_t(\cdot)|X(t)]$ as random measures are the same and also

$$(Y(t), \eta(t, Y(t), \cdot)) \stackrel{d}{=} (X(t), E[\Lambda_t(\cdot)|X(t)]) \quad \forall\ t. \qquad \square \tag{4.10}$$

Theorem 4.1 gives the existence of a relaxed control η which is Markovian in the sense that it only depends on the current state $Y(t)$ and time t and does not depend on

the history of the process. Moreover, (4.10) implies that these solutions have the same cost under any criterion which only depends on the one-dimensional distributions. In particular, this holds for the discounted criterion (1.4), for the finite-horizon criterion (1.5), and for the long-term average criterion (1.6).

The following corollary relaxes the boundedness and uniform ellipticity assumptions in the result of Gyöngy [11].

COROLLARY 4.3. *Suppose that W is an $\mathbb{R}^d$-valued $\{\mathcal{F}_t\}$-Brownian motion; $\widehat{\sigma}$ and $\widehat{b}$ are measurable, $\{\mathcal{F}_t\}$-adapted processes taking values in $\mathbb{M}^{d\times d}$ and $\mathbb{R}^d$, respectively; and $X(0)$ is $\mathbb{R}^d$-valued and $\mathcal{F}_0$-measurable. Let*

$$X(t) = X(0) + \int_0^t \widehat{\sigma}(s)dW(s) + \int_0^t \widehat{b}(s)ds.$$

Then there exist measurable functions $\sigma : [0,\infty)\times\mathbb{R}^d \to \mathbb{M}^{d\times d}$ and $b : \mathbb{R}^d \to \mathbb{R}^d$, an $\mathbb{R}^d$-valued Brownian motion $\tilde{W}$, and a process Y satisfying

$$Y(t) = Y(0) + \int_0^t \sigma(s, Y(s))d\tilde{W}(s) + \int_0^t b(s, Y(s))ds$$

such that for each $t \geq 0$, the distributions of $(X(t), E[\widehat{b}(t)|X(t)], E[\widehat{\sigma}(t)\widehat{\sigma}^T(t)|X(t)])$ and $(Y(t), b(t, Y(t)), \sigma(t, Y(t))\sigma^T(t, Y(t)))$ are the same.

Proof. Let the control space $U = \mathbb{M}_+^{d\times d} \times \mathbb{R}^d$, where $\mathbb{M}_+^{d\times d}$ denotes the space of nonnegative definite $d\times d$ matrices, and, with $u = (u^1, u^2)$, define the time-space generator

$$A(\gamma f)(t,x,u) = \gamma(t)\left(\frac{1}{2}\sum_{i,j=1}^d u^1_{ij} f_{x_i x_j}(x) + \sum_{i=1}^d u^2_i f_{x_i}(x)\right) + \gamma'(t)f(x),$$

where $\gamma \in C_c^1(\mathbb{R}^+)$ and $f \in \mathcal{D}(\widehat{A})$.

Let $\widehat{a}(t) = \widehat{\sigma}(t)\widehat{\sigma}(t)^T$ and $\Lambda_t(du) = \delta_{\{\widehat{a}(t)\}}(du^1)\delta_{\{\widehat{b}(t)\}}(du^2)$. Then (X,Λ) determines a solution to the controlled martingale problem for A. Taking $\psi(u) = ||u^1|| + |u^2|$, the conditions of Theorem 4.1 are satisfied and an application of the theorem gives the existence of a process Y and measurable functions a and b. Let σ be the symmetric square root of a (which will exist even in the degenerate case). Existence of $\tilde{W}$ follows by [9, Theorem 5.3.3], and the result follows. □

4.1. First passage problems. We now turn our attention to the issue of existence of Markov controls for exit problems. In this section, we seek to minimize (1.7), in which c and g are time-dependent and $E_0 \subset \mathbb{R}^+ \times E$, over all solutions (X,Λ) of the controlled martingale problem such that $X(0)$ has distribution ν_0 and where, to avoid degeneracies, we assume $\nu_0\{x : (0,x) \in E_0\} = 1$.

The key to this existence result is the characterization of the occupation measures similar to that given by (2.1). We consider pairs of measures $\mu_0 \in \mathcal{M}(E_0\times U)$ ($\mathcal{M}(E_0\times U)$, the Radon measures on $E_0\times U$) and $\mu_1 \in \mathcal{P}(E_0^c)$, such that

$$\mu_0(E_0\times U) < \infty. \tag{4.11}$$

The condition corresponding to (2.1) is

$$\int_{E_0\times U} [\gamma(s)Af(x,u) + \gamma'(s)f(x)]\,\mu_0(ds\times dx\times du) + \gamma(0)\int f\,d\nu_0 - \int_{E_0^c}\gamma(s)f(x)\,\mu_1(ds\times dx) = 0 \tag{4.12}$$

for all $\gamma \in \widehat{C}^1(\mathbb{R}^+)$ and $f \in \mathcal{D}(A)$.

To motivate this condition on the measures μ_0 and μ_1, consider a process (X, Λ) as above and assume $E[\tau] < \infty$. Then, under appropriate conditions,

$$E\left[\gamma(\tau)f(X(\tau)) - \gamma(0)f(X(0)) \right.$$

$$\left. - \int_0^\tau \int_U [\gamma(s)Af(X(s),u) + \gamma'(s)f(X(s))]\,\Lambda_s(du)\,ds\right] = 0.$$

Let μ_0 be given by

$$\mu_0(B) = E\left[\int_0^\tau \int_U I_B(s, X(s), u)\Lambda_s(du)ds\right],$$

and let μ_1 be the joint distribution of the exit time and exit location $(\tau, X(\tau))$. Then (4.12) is satisfied.

The following lemma will be needed to establish properties of solutions to the controlled martingale problem. Its proof is delayed to the appendix (section 7).

LEMMA 4.4. *Let Q be a nonnegative, $\{\mathcal{F}_t\}$-adapted, cadlag process, let V_1 and V_2 be bounded, nonnegative, measurable, $\{\mathcal{F}_t\}$-adapted processes, and suppose that*

$$g(Q(t)) - \int_0^t (V_1(s)g'(Q(s)) + V_2(s)(g(0) - g(Q(s))))\,ds$$

is an $\{\mathcal{F}_t\}$-martingale for every C^1 function g with g and g' bounded. Let τ be a stopping time, and define $\sigma_0^\tau = \inf\{t > \tau : Q(t) > 0\}$ and $\sigma_1^\tau = \inf\{t > \sigma_0^\tau : Q(t) = 0\}$. Then, for $\tau \le t < \sigma_1^\tau$, $Q(t) - Q(\tau) = \int_\tau^t V_1(s)ds$, and if $\sigma_1^\tau < \infty$ a.s.,

$$P\left\{\int_{\sigma_0^\tau}^{\sigma_1^\tau} V_2(s)\,ds > x \middle| \mathcal{F}_{\sigma_0^\tau}\right\} = e^{-x}, \qquad x \ge 0.$$

The next result demonstrates existence of solutions corresponding to measures μ_0 and μ_1 which satisfy (4.11) and (4.12).

THEOREM 4.5. *Suppose that E, U, A, and ψ satisfy the conditions of section* 1.1 *and that $E_0 \subset \mathbb{R}^+ \times E$ is open. Let μ_0 and μ_1 satisfy* (4.11) *and* (4.12) *and setting $\mu_0^*(\Gamma) = \mu_0(\Gamma \times U)$, $\Gamma \in \mathcal{B}(E_0)$, let η be the transition function satisfying*

$$\mu_0(\Gamma_0 \times \Gamma_1) = \int_{\Gamma_0} \eta(s, x, \Gamma_1)\mu_0^*(ds \times dx).$$

Suppose

$$\int \psi(u)\mu_0(ds \times dx \times du) < \infty. \tag{4.13}$$

Then there exists a process Y with initial distribution ν_0 adapted to a filtration $\{\mathcal{F}_t\}$ and an $\{\mathcal{F}_t\}$-stopping time $\tilde{\tau}$ such that

$$f(Y(t \wedge \tilde{\tau})) - \int_0^{t\wedge\tilde{\tau}} \int_U Af(Y(s), u)\,\eta(s, Y(s), du)\,ds$$

is an $\{\mathcal{F}_t\}$*-martingale for each* $f \in \mathcal{D}(A)$ *and*

$$\begin{aligned} &\text{(i)} \quad P((\tilde{\tau}, Y(\tilde{\tau})) \in E_0^c) = 1 \\ &\textit{and} \\ &\text{(ii)} \quad \{t : (t, Y(t)) \in E_0^c, t < \tilde{\tau}\} \textit{ has Lebesgue measure } 0. \end{aligned} \tag{4.14}$$

Moreover, for each $\Gamma_1 \in \mathcal{B}(E_0 \times U)$,

$$E\left[\int_0^{\tilde{\tau}} \int_U I_{\Gamma_1}(s, Y(s), u)\, \eta(s, Y(s), du)\, ds\right] = \mu_0(\Gamma_1),$$

and for each $\Gamma_2 \in \mathcal{B}(E_0^c)$

$$E[I_{\Gamma_2}(\tilde{\tau}, Y(\tilde{\tau}))] = \mu_1(\Gamma_2).$$

Remark 4.6. Y can be extended to a solution of the controlled martingale problem for A for all $t \geq 0$, with sample paths in $D_{E^\Delta}[0, \infty)$. (See [9, Lemma 4.5.16].)

The hypotheses of the theorem do not rule out the possibility that $\tilde{\tau} > \sigma = \inf\{t : (t, Y(t)) \in E_0^c\}$. However, for many processes one can show that $\inf\{t : (t, Y(t)) \in E_0^c\} = \inf\{t : (t, Y(t)) \in (\overline{E}_0)^c\}$ a.s., which, by (4.14), implies $\tilde{\tau} = \sigma$ a.s.

Proof. The structure of this proof is very similar to that of Theorem 4.1 in that we augment the space, define both a new generator $\overline{A}$ and a corresponding stationary distribution $\overline{\mu}$, invoke Theorem 2.2 to obtain a stationary process, and use the optional sampling theorem in conjunction with an absolutely continuous change of measures to obtain the result. The specifics, however, are substantially different, and we therefore provide complete details.

First we augment the state space with extra time dimensions and augment the control space with a $\{0, 1\}$ component. Thus the state space is $\mathbb{R}^+ \times \mathbb{R}^+ \times E$ and the control space is $U \times \{0, 1\}$. Define the generator $\overline{A}$ by

$$\begin{aligned} \overline{A}(\beta\gamma f)(r, s, x, u, v) &= v\beta(r)[\gamma(s)Af(x, u) + \gamma'(s)f(x)] \\ &\quad + (1 - v)\left[\beta(0)\gamma(0)\int f\, d\nu_0 - \beta(r)\gamma(s)f(x) + \beta'(r)\gamma(s)f(x)\right] \end{aligned}$$

for $\beta, \gamma \in \widehat{C}^1(\mathbb{R}^+)$ and $f \in \mathcal{D}(A)$. It will be shown that the "s" component measures the time the process specified by generator A runs, whereas the "r" component measures the (mean 1) exponential time before all state components are reset. The control "v" determines whether the process governed by A runs or the jump process runs and will be restricted by the stationary measure so that $v = 1$ when $(s, x) \in E_0$ and $v = 0$ when $(s, x) \in E_0^c$.

Let u^* be a fixed point of U and let $K = \mu_0(E_0 \times U) + 1$. Define the measure $\overline{\mu} \in \mathcal{P}(\mathbb{R}^+ \times \mathbb{R}^+ \times E \times U \times \{0, 1\})$ by

$$\begin{aligned} &\int h(r, s, x, u, v)\, \overline{\mu}(dr \times ds \times dx \times du \times dv) \\ &\qquad = K^{-1}\left(\int_{E_0 \times U} h(0, s, x, u, 1)\, \mu_0(ds \times dx \times du)\right. \\ &\qquad\quad \left. + \int_0^\infty \int_{E_0^c} e^{-r} h(r, s, x, u^*, 0)\, \mu_1(ds \times dx)\, dr\right) \end{aligned}$$

for each bounded, continuous h. Observe that the conditional distribution of (u,v) given (r,s,x) under $\overline{\mu}$ is

$$\overline{\eta}(r,s,x,du\times dv)=\begin{cases}\eta(s,x,du)\delta_{\{1\}}(dv), & (s,x)\in E_0,\\ \delta_{\{u^*\}}(du)\delta_{\{0\}}(dv), & (s,x)\in E_0^c.\end{cases} \tag{4.15}$$

Note that we can determine the value of v by observing whether $(s,x)\in E_0$ or E_0^c, so we define $v(s,x)=I_{E_0}(s,x)$. In particular, v is an ordinary feedback control, and we therefore slightly abuse notation and write $\overline{\eta}(r,s,x,du)$ in the sequel. Note also that $I_{E_0}(s,x)=I_{\{0\}}(r)$ a.e. $\bar{\mu}$.

A straightforward computation shows that (4.12) implies that $(\overline{A},\overline{\mu})$ satisfies the stationarity condition (2.1). The conditions of Theorem 2.2 on the state and control spaces and the generator are also satisfied, which therefore implies existence of a stationary $\mathbb{R}^+\times\mathbb{R}^+\times E$-valued process (R,S,X) (which we may assume defined for all $t\in\mathbb{R}$) such that

$$\begin{aligned}&\beta(R(t))\gamma(S(t))f(X(t))\\ &-\int_0^t\int_U \overline{A}(\beta\gamma f)(R(s),S(s),X(s),u,v(S(s),X(s)))\,\overline{\eta}(R(s),S(s),X(s),du)\,ds\end{aligned} \tag{4.16}$$

is an $\{\mathcal{F}_t^{R,S,X}\}$-martingale for all $\beta,\gamma\in\widehat{C}^1(\mathbb{R}^+)$ and $f\in\mathcal{D}(A)$. In addition, we have that $v(S(s),X(s))=I_{E_0}(S(s),X(s))=I_{\{0\}}(R(s))$ a.s. for each $s\geq 0$.

For each $t\geq 0$, define the following random variables (cf. Theorem 4.1): $\sigma_{-1}^t=\sup\{r<t: S(r)=0,R(r)=0\}$, $\sigma_1^t=\inf\{r\geq t: S(r)=0,R(r)=0\}$, $\tau_1^t=\inf\{t>\sigma_1^t: R(t)>0\}$, and $\sigma_2^t=\inf\{r>\sigma_1^t: R(r)=0\}$. For $s\in[\sigma_1^t,\tau_1^t)$, by definition, $R(s)=0$ and by Lemma 4.4 $S(s)=\int_{\sigma_1^t}^s I_{\{0\}}(R(r))dr=(s-\sigma_1^t)$. For $s\in[\tau_1^t,\sigma_2^t)$, by Lemma 4.4, $R(s)=\int_{\tau_1^t}^s I_{(0,\infty)}(R(r))dr=s-\tau_1^t$ a.s. and conditional on $\mathcal{F}_{\tau_1^t}$, $\sigma_2^t-\tau_1^t$ is exponentially distributed with mean 1, and again by Lemma 4.4, $S(s)=S(\tau_1^t)+\int_{\tau_1^t}^s I_{\{0\}}(R(r))dr=S(\tau_1^t)=\tau_1^t-\sigma_1^t$. Starting with $g(r+s)=e^{-\alpha(r+s)}$ and approximating more general g by linear combinations of these exponentials, we see that

$$g(S(t)+R(t))-\int_0^t\left(g'(S(r)+R(r))+(1-v(S(r),X(r)))(g(0)-g(S(r)+R(r)))\right)dr$$

is a martingale for C^1 functions with g and g' bounded. Letting $\tilde{\sigma}_2^t=\inf\{s>\tau_1^t: S(s)+R(s)=0\}$, Lemma 4.4 implies

$$\begin{aligned}P\left\{\int_{\tau_1^t}^{\sigma_2^t}(1-v(S(r),X(r)))dr>x\Big|\mathcal{F}_{\tau_1^t}^{R,S,X}\right\}&=e^{-x}\\ &=P\left\{\int_{\tau_1^t}^{\tilde{\sigma}_2^t}(1-v(S(r),X(r)))dr>x\Big|\mathcal{F}_{\tau_1^t}^{R,S,X}\right\},\end{aligned}$$

and since $\sigma_2^t\leq\tilde{\sigma}_2^t$, we must have $\sigma_2^t=\tilde{\sigma}_2^t$ a.s. In particular, $S(\sigma_2^t)=0$ a.s. Finally, defining $Z(u)=(R(\tau_1^t+u),S(\tau_1^t+u),X(\tau_1^t+u))$ for $u\leq\sigma_2^t-\tau_1^t$, we can extend Z to be a solution of the martingale problem for

$$Cg(r,s,x)=\int g(0,0,y)\nu_0(dy)-g(r,s,x)+\frac{\partial}{\partial r}g(r,s,x).$$

Since any solution of this martingale problem has the property that the final component is constant except for jumps that occur when the first two components jump to zero, it follows that $X(u) = X(\tau_1^t)$ for $\tau_1^t \le u < \sigma_2^t$.

Let h be a fixed, bounded, continuous function, and define

$$H_\epsilon(r) = \int_U e^{-\epsilon(R(r)+S(r))} h(R(r), S(r), X(r), u, v(S(r), X(r))\, \overline{\eta}(R(r), S(r), X(r), du)).$$

Then, as a process in t,

$$(\sigma_1^t - \sigma_{-1}^t)^{-1} \int_{\sigma_1^t}^{\sigma_2^t} H_\epsilon(r)\, dr \tag{4.17}$$

is stationary, and for each t and $s \in [\sigma_{-1}^t, \sigma_1^t)$,

$$(\sigma_1^s - \sigma_{-1}^s)^{-1} \int_{\sigma_1^s}^{\sigma_2^s} H_\epsilon(r)\, dr = (\sigma_1^t - \sigma_{-1}^t)^{-1} \int_{\sigma_1^t}^{\sigma_2^t} H_\epsilon(r)\, dr.$$

These expressions are set equal to 0 whenever $\sigma_{-1}^t = -\infty$ or $\sigma_1^t = +\infty$.

Using stationarity,

$$E\left[(\sigma_1^t - \sigma_{-1}^t)^{-1} \int_{\sigma_1^t}^{\sigma_2^t} H_\epsilon(r)\, dr\right] = T^{-1} \int_0^T E\left[(\sigma_1^t - \sigma_{-1}^t)^{-1} \int_{\sigma_1^t}^{\sigma_2^t} H_\epsilon(r) dr\right] dt, \tag{4.18}$$

in which both sides may be infinite due to the $(\sigma_1^t - \sigma_{-1}^t)^{-1}$ term. The following argument, in fact, shows both terms are finite and identifies their common value.

Let $N(T)$ denote the number of jumps of the process (R, S, X) in the interval $[0, T]$, let $\{\sigma_k : k = 1, \dots, N(T)\}$ denote these jump times, and let $\sigma_{N(T)+1}$ and σ_{-1} ($= \sigma_0$ in the summation) denote the first jump time after time T and the last jump time before time 0, respectively. Then the right-hand side of (4.18) equals

$$\begin{aligned}
&T^{-1} E\left[\sum_{k=1}^{N(T)+1} \frac{T \wedge \sigma_k - \sigma_{k-1} \vee 0}{\sigma_k - \sigma_{k-1}} \int_{\sigma_{k+1}}^{\sigma_{k+2}} H_\epsilon(r)\, dr\right] \\
&\quad = T^{-1} E\left[\int_0^T H_\epsilon(r)\, dr\right] \\
&\qquad - T^{-1} E\left[\int_0^{\sigma_1 \wedge T} \left(1 - \frac{T \wedge \sigma_1}{\sigma_1 - \sigma_{-1}}\right) H_\epsilon(r)\, dr\right] \\
&\qquad + T^{-1} E\left[I_{\{N(T)=1\}} \frac{\sigma_1}{\sigma_1 - \sigma_{-1}} \int_T^{\sigma_2} H_\epsilon(r)\, dr\right] \\
&\qquad + T^{-1} E\left[I_{\{N(T)>1\}} \int_T^{\sigma_{N(T)+1}} H_\epsilon(r)\, dr\right] \\
&\qquad + T^{-1} E\left[I_{\{N(T)>0\}} \frac{T - \sigma_{N(T)}}{\sigma_{N(T)+1} - \sigma_{N(T)}} \int_{\sigma_{N(T)+1}}^{\sigma_{N(T)+2}} H_\epsilon(r)\, dr\right].
\end{aligned}$$

Observe that the first term is

$$\int e^{-\epsilon(r+s)}h(r,s,x,u,v)\,\overline{\mu}(dr\times ds\times dx\times du\times dv) \tag{4.19}$$

and the last four terms are bounded above by $4||h||/(\epsilon T)$. Thus all terms are finite, implying that the terms in (4.18) are finite, and moreover, as $T\to\infty$, these converge to (4.19).

Letting $\epsilon\to 0$ gives, for each bounded, continuous h (and hence for each bounded, measurable h),

$$\begin{aligned}&E\left[(\sigma_1^t-\sigma_{-1}^t)^{-1}\int_{\sigma_1^t}^{\sigma_2^t}\int_U h(R(r),S(r),X(r),u,v(S(r),X(r)))\,\overline{\eta}(R(r),S(r),X(r),du)dr\right]\\ &\qquad=\int h(r,s,x,u,v)\,\overline{\mu}(dr\times ds\times dx\times du\times dv).\end{aligned} \tag{4.20}$$

Then, considering $h(r,s,x,u,v)=I_{\{0\}}(v)$ in (4.20) yields

$$\begin{aligned}K^{-1} &= E[(\sigma_1^t-\sigma_{-1}^t)^{-1}(\sigma_2^t-\tau_1^t)]\\ &= E[(\sigma_1^t-\sigma_{-1}^t)^{-1}],\end{aligned}$$

in which the last equality follows from the fact that, conditional on $\mathcal{F}^{R,S,X}_{\tau_1^t}$, $\sigma_2^t-\tau_1^t$ is exponentially distributed with mean 1.

Now define the process Y by $Y(t)=X(\sigma_1^0+t)$, $\tilde{R}(t)=R(\sigma_1^0+t)$, $\tilde{S}(t)=S(\sigma_1^0+t)$, and the filtration $\{\mathcal{F}_t\}=\{\mathcal{F}^{R,S,X}_{\sigma_1^0+t}\}$. Let $\tilde{\tau}=\inf\{t\ge 0:\tilde{R}(t)>0\}$ and $\sigma=\inf\{t>\tilde{\tau}:\tilde{R}(t)=0\}$, and note that $Y(t)=Y(\tilde{\tau})$ for $\tilde{\tau}\le t<\sigma$. Observe that both σ_1 and σ_{-1} are $\mathcal{F}_0$-measurable. Define a new probability measure $\widehat{P}$ to have Radon–Nikodým derivative $K(\sigma_1-\sigma_{-1})^{-1}$. It then follows from (4.20) that

$$\begin{aligned}&E^{\widehat{P}}\left[\int_0^{\sigma}\int_U h(\tilde{R}(r),\tilde{S}(r),Y(r),u,v(\tilde{S}(r),Y(r)))\,\overline{\eta}(\tilde{R}(r),\tilde{S}(r),Y(r),du)dr\right]\\ &\qquad=\int h(r,s,x,u,v)\,\overline{\mu}(dr\times ds\times dx\times du\times dv)/E[(\sigma_1-\sigma_{-1})^{-1}]\\ &\qquad=\int h(0,s,x,u,1)\mu_0(ds\times dx\times du)\\ &\qquad\quad+\int_0^{\infty}\int e^{-r}h(r,s,x,u^*,0)\mu_1(ds\times dx)dr.\end{aligned} \tag{4.21}$$

Considering $h(r,s,x,u,v)=I_0(r)I_{E_0^c}(s,x)$ and $h(r,s,x,u,v)=I_{(0,\infty)}(r)I_{E_0}(s,x)$ in (4.22) indicates that $(\tilde{S},Y,\tilde{\tau})$ satisfy (4.14).

The optional sampling theorem implies that under $\widehat{P}$,

$$f(Y(t\wedge\tilde{\tau}))-\int_0^{t\wedge\tilde{\tau}}\int_U Af(Y(s),u)\,\eta(s,Y(s),du)\,ds$$

is a martingale with respect to the filtration $\{\mathcal{F}_t\}$.

For $\Gamma_1 \in \mathcal{B}(E_0^c)$ and $h(r,s,x,u,v) = I_{\Gamma_1}(s,x)$, (4.22) implies

$$\begin{aligned}\mu_1(\Gamma_1) &= E\left[\int_{\tilde\tau}^{\sigma} I_{\Gamma_1}(\tilde S(r), Y(r))\,dr\right]\\ &= E[I_{\Gamma_1}(\tilde\tau, Y(\tilde\tau))(\sigma-\tilde\tau)]\\ &= E[I_{\Gamma_1}(\tilde\tau, Y(\tilde\tau))],\end{aligned}$$

where the last equality follows from the fact that $\sigma - \tilde\tau$ is a mean 1 exponential time conditional on $\mathcal{F}_{\tilde\tau}$. Similarly, for $\Gamma_2 \in \mathcal{B}(E_0 \times U)$ and $h(r,s,x,u,v) = I_{\Gamma_2}(s,x,u)$, (4.21) implies

$$E\left[\int_0^{\tilde\tau}\int_U I_{\Gamma_2}(s,Y(s),u)\eta(s,Y(s),du)ds\right] = \mu_0(\Gamma_2). \qquad \square$$

COROLLARY 4.7. *Suppose E, U, A, and ψ satisfy the conditions of section* 1.1, *and $E_0 \subset \mathbb{R}^+ \times E$ is open. Let (X,Λ) be a solution of the controlled martingale problem for A such that $\tau = \inf\{t \ge 0 : (t, X(t)) \in E_0^c\}$ has finite expectation. Suppose*

$$E\left[\int_0^{\tau}\int_U \psi(u)\Lambda_s(du)ds\right] < \infty. \tag{4.22}$$

Then there exists a process Y adapted to a filtration $\{\mathcal{F}_t\}$, a transition function η from $\mathbb{R}^+ \times E$ into U, and an $\{\mathcal{F}_t\}$-stopping time $\tilde\tau$ satisfying (4.14) *such that for $f \in \mathcal{D}(A)$*

$$f(Y(t\wedge\tilde\tau)) - \int_0^{t\wedge\tilde\tau}\int_U Af(Y(s),u)\,\eta(s,Y(s),du)\,ds \tag{4.23}$$

is a martingale with respect to $\{\mathcal{F}_t\}$, and for each $t \ge 0$, $(t\wedge\tilde\tau, Y(t\wedge\tilde\tau), \eta(t\wedge\tilde\tau, Y(t\wedge\tilde\tau),\cdot))$ has the same distribution as $(t\wedge\tau, X(t\wedge\tau), E[\Lambda_{t\wedge\tau}(\cdot)|X(t\wedge\tau)])$.

Proof. Define $\mu_0 \in \mathcal{M}(E_0 \times U)$ by

$$\mu_0(\Gamma) = E\left[\int_0^{\tau}\int_U I_\Gamma(s,X(s),u)\,\Lambda_s(du)\,ds\right] \qquad \forall \Gamma \in \mathcal{B}(E_0\times U),$$

and $\mu_1 \in \mathcal{P}(E_0^c)$ by

$$\mu_1(\Gamma) = E[I_\Gamma(\tau, X(\tau))].$$

The pair (μ_0,μ_1) satisfies (4.12). Theorem 4.5 then implies the existence of a process Y, a transition function η, and random variable $\tilde\tau$ satisfying (4.14) such that (4.23) is a martingale. The fact that

$$(t\wedge\tilde\tau, Y(t\wedge\tilde\tau), \eta(t\wedge\tilde\tau, Y(t\wedge\tilde\tau),\cdot)) \stackrel{d}{=} (t\wedge\tau, X(t\wedge\tau), E[\Lambda_{t\wedge\tau}(\cdot)|X(t\wedge\tau)]) \tag{4.24}$$

follows by essentially the same argument as in Theorem 4.1. $\square$

4.2. Finite-horizon problems. Control problems over a finite horizon can be formulated as first exit problems, and thus the results of the previous section can be applied to finite-horizon problems.

A minor change in the time-space generator of the process from $\gamma(s)Af(x,u) + \gamma'(s)f(x)$ to $\gamma(s)Af(x,u) - \gamma'(s)f(x)$ augments a time component to the original state

component, which decreases linearly at rate 1. Taking $E_0 = (0,T] \times E$, the results of the previous section show that the time-dependence of the feedback control can be taken to only depend on the time remaining. Observe also that the process will exit as soon as it hits the boundary $\{0\} \times E$, so the conclusions are stronger.

The first result establishes existence of a solution for each pair of measures (μ_0, μ_1) which satisfy condition (4.25) given below.

THEOREM 4.8. *Suppose E, U, A, and ψ satisfy the conditions of section* 1.1. *Let $E_0 = (0,T] \times E$ and $\mu_0 \in \mathcal{M}(E_0 \times U)$ and $\mu_1 \in \mathcal{P}(E)$ satisfy*

$$\int_{(0,T]\times E\times U} [\gamma(s)Af(x,u) - \gamma'(s)f(x)]\,\mu_0(ds \times dx \times du) + \gamma(T)\int_E f\,d\nu_0 - \gamma(0)\int_E f(x)\,\mu_1(dx) = 0 \qquad \forall \gamma \in \widehat{C}^1(\mathbb{R}^+), f \in \mathcal{D}(A)\ . \tag{4.25}$$

Let η be the regular conditional distribution of u given (s,x). Suppose ψ satisfies (4.13). *Then there exists a process Y with initial distribution ν_0 such that*

$$\gamma(T-(t\wedge T))f(Y(t\wedge T)) - \int_0^{t\wedge T}\int_U [\gamma(T-s)Af(Y(s),u) - \gamma'(T-s)f(Y(s))]\,\eta(T-s,Y(s),du)\,ds$$

is a martingale with respect to a filtration $\{\mathcal{F}_t\}$. Moreover, for each $\Gamma_1 \in \mathcal{B}(E_0 \times U)$,

$$E\left[\int_0^T \int_U I_{\Gamma_1}(T-s,Y(s),u)\,\eta(T-s,Y(s),du)\,ds\right] = \mu_0(\Gamma_1),$$

and for each $\Gamma_2 \in \mathcal{B}(E)$,

$$E[I_{\Gamma_2}(Y(T))] = \mu_1(\Gamma_2).$$

The next result essentially states that for any given solution (X,Λ) of the finite-horizon problem there exists a process Y and feedback control η whose one-dimensional distributions match the given solution.

COROLLARY 4.9. *Suppose E, U, A, and ψ satisfy the conditions of section* 1.1. *Let (X,Λ) be a solution of the controlled martingale problem for A. Suppose*

$$E\left[\int_0^T \int_U \psi(u)\,\Lambda_s(du)\,ds\right] < \infty.$$

Then there exists a process Y, a transition function η from $(0,T] \times E$ into U such that

$$\gamma(T-(t\wedge T))f(Y(t\wedge T)) - \int_0^{t\wedge T}\int_U [\gamma(T-s)Af(Y(s),u) - \gamma'(T-s)f(Y(s))]\,\eta(T-s,Y(s),du)\,ds$$

is a martingale with respect to a filtration $\{\mathcal{F}_t\}$ and

$$(Y(t), \eta(T-t,Y(t),\cdot)) \stackrel{d}{=} (X(t), E[\Lambda_t|X(t)]) \qquad \forall\ 0 \le t \le T.$$

5. Time-homogeneous Markov controls. When the state space E does not contain any time component, the existence result in Theorem 2.2 indicates that, under the long-term average criterion, to each solution there corresponds a solution having a time-homogeneous Markov control with the same cost. Theorems 4.1 and 4.5 of the previous section require augmenting the state space with a time component, and thus the Markov controls obtained are time-dependent. In fact, time-dependence is necessary in order for (4.10) and (4.24) to hold.

A dynamic programming argument, however, indicates that there should exist optimal controls for the discounted and first exit criteria which are stationary (not time-dependent). The next theorems show that it is possible to have a solution $(Y, \eta(Y,\cdot))$ in which η is a stationary Markov control and for which the discounted or first exit cost matches the corresponding cost of a given solution (X, Λ).

5.1. Discounted problems. We begin by considering the discounted criterion more carefully and allow the discount rate to be a function of the state and control.

THEOREM 5.1. *Suppose that E, U, A, and ψ satisfy conditions* (i)–(vi) *of section* 1.1. *Let (X,Λ) be a relaxed solution of the controlled martingale problem for (A,ν_0), let α be a nonnegative, bounded, and continuous function on $E\times U$, and suppose that*

$$E\left[\int_0^\infty e^{-\int_0^t\int_U \alpha(X(s),u)\Lambda_s(du)ds}\int_U(1+\psi(u))\Lambda_s(du)dt\right]<\infty\,.$$

Then there exists a process Y and a transition function η from E to U such that $(Y,\eta(Y,\cdot))$ is a relaxed solution of the controlled martingale problem for (A,ν_0) and

$$\begin{aligned}&E\left[\int_0^\infty e^{-\int_0^t\int_U \alpha(Y(s),u)\eta(Y(s),du)ds}\int_U c(Y(t),u)\,\eta(Y(t),du)\,dt\right]\\&\quad=E\left[\int_0^\infty e^{-\int_0^t\int_U \alpha(X(s),u)\Lambda_s(du)ds}\int_U c(X(t),u)\,\Lambda_t(du)\,dt\right]\end{aligned}$$

for every $c\in M(E\times U)$ that is bounded below (in the sense that if one side is infinite so is the other).

Remark 5.2. Note that Y and η will, in general, depend on α.

Proof. The proof is very similar to the proof of Theorem 4.1. We therefore only identify the differences. We take the state space to be $\{-1,1\}\times E$ and define the generator

$$A^\alpha(\gamma f)(\theta,x,u)=\gamma(\theta)Af(x,u)+\alpha(x,u)\left[\gamma(-\theta)\int_E f(y)\nu_0(dy)-\gamma(\theta)f(x)\right],\tag{5.1}$$

where $f\in\mathcal{D}(A)$ and $\gamma\in \mathrm{B}(\{-1,1\})$. Let $\Gamma_X(t)=\int_0^t\int_U\alpha(X(s),u)\Lambda_s(du)ds$. Define the measure $\pi\in\mathcal{P}(\{-1,1\}\times E\times U)$ by

$$\begin{aligned}&\int_{\{-1,1\}\times E\times U}h(\theta,x,u)\,\pi(d\theta\times dx\times du)\\&\quad=\frac{E\left[\int_0^\infty e^{-\Gamma_X(s)}\left(\int_U h(-1,X(s),u)\,\Lambda_s(du)+\int_U h(1,X(s),u)\,\Lambda_s(du)\right)ds\right]}{2E\left[\int_0^\infty e^{-\Gamma_X(s)}ds\right]}\end{aligned}\tag{5.2}$$

for $h \in \overline{C}(\{-1,1\} \times E \times U)$. The fact that π is a stationary measure for A^α follows from the fact that

$$e^{-\Gamma_X(t)} f(X(t)) - f(X(0)) - \int_0^t e^{-\Gamma_X(s)} \int_U (Af(X(s),u) - \alpha(X(s),u) f(X(s))) \, \Lambda_s(du) ds$$

is a martingale which, taking expectations and letting $t \to \infty$, implies

$$E\left[\int_0^\infty e^{-\Gamma_X(s)} \int_U \left(Af(X(s),u) + \alpha(X(s),u) \left[\int f \, d\nu_0 - f(X(s)) \right] \right) \Lambda_s(du) ds \right] = 0.$$

Observe that π can be written as

$$\begin{aligned} \pi(d\theta \times dx \times du) &= \frac{1}{2}(\delta_{\{-1\}}(d\theta) + \delta_{\{1\}}(d\theta)) \widehat{\pi}(dx \times du) \\ &= \frac{1}{2}(\delta_{\{-1\}}(d\theta) + \delta_{\{1\}}(d\theta)) \pi_0(dx) \eta(x, du) \, , \end{aligned}$$

where $\widehat{\pi}$ and π_0 denote the marginals of π on $E \times U$ and E, respectively, and η does not depend on θ. Theorem 2.2 therefore gives the existence of a stationary process (Θ, Z) with marginal distribution $\frac{1}{2}(\delta_{\{-1\}}(d\theta) + \delta_{\{1\}}(d\theta))\pi_0(dx)$ such that $(\Theta, Z, \eta(Z, du))$ is a solution of the controlled martingale problem for A^α.

Let $\tau_1 = \inf\{t > 0 : \Theta(t) \neq \Theta(0)\}$, $\tau_{-1} = \sup\{t < 0 : \Theta(t) \neq \Theta(0)\}$, and $\tau_{k+1} = \inf\{t > \tau_k : \Theta(t) \neq \Theta(\tau_k)\}$. As in the proof of Theorem 4.1, define $Y(t) = Z(\tau_1 + t)$. It is no longer the case that the τ_k are the jump times of a Markov chain; however, taking $C = E[(\tau_1 - \tau_{-1})^{-1}]$,

$$L(t) = [C(\tau_1 - \tau_{-1})]^{-1} e^{\int_0^t \int_U \alpha(Y(s),u)\eta(Y(s),du)ds} I_{[0,\tau_2 - \tau_1)}(t)$$

is still a mean 1 martingale. (Note that C is finite by the boundedness of α.) The remainder of the proof is the same as before, with (4.8) replaced by

$$\begin{aligned} &E^{\widehat{P}}\left[\int_0^\infty e^{-\int_0^t \int_U \alpha(Y(s),u)\eta(Y(s),du)ds} \int_U c(Y(t),u) \, \eta(Y(t),du) \, dt\right] \\ &\quad = E\left[\int_0^\infty e^{-\Gamma_X(t)} \int_U c(X(t),u) \, \Lambda_t(du) \, dt\right], \end{aligned}$$

which gives the desired result. □

The following result is a consequence of the construction in the proof of Theorem 5.1 and will be used in our discussion of the linear programing approach to the solution of optimal discounted control problems.

COROLLARY 5.3. *Suppose that E, U, A, and ψ satisfy conditions* (i)–(vi) *of section* 1.1. *Suppose that $\widehat{\pi} \in \mathcal{P}(E \times U)$ satisfies*

$$\int_{E \times U} \left[Af(x,u) + \alpha(x,u) \left(\int f(y)\nu_0(dy) - f(x) \right) \right] \widehat{\pi}(dx \times du) = 0 \quad \forall f \in \mathcal{D}(A)$$

and

$$\int \psi(u) \, \widehat{\pi}(dx \times du) < \infty.$$

Let η satisfy $\widehat{\pi}(dx \times du) = \widehat{\pi}_0(dx)\eta(x, du)$. Then there exists a process Y such that $(Y, \eta(Y, \cdot))$ is a relaxed solution of the controlled martingale problem for (A, ν_0) and

$$\frac{E\left[\int_0^\infty e^{-\int_0^t \int_U \alpha(Y(s),u)\eta(Y(s),du)ds} \int_U c(Y(t),u)\,\eta(Y(t),du)\,dt\right]}{E\left[\int_0^\infty e^{-\int_0^t \int_U \alpha(Y(s),u)\eta(Y(s),du)ds}\,dt\right]} \tag{5.3}$$

$$= \int_{E\times U} c(x,u)\widehat{\pi}(dx \times du)$$

for every $c \in B(E \times U)$ and for every nonnegative $c \in M(E \times U)$ (in the sense that if one side is infinite so is the other).

Proof. Define $\pi(d\theta \times dx \times du) = \frac{1}{2}(\delta_{\{-1\}}(d\theta) + \delta_{\{1\}}(d\theta))\widehat{\pi}(dx \times du)$. Then π is a stationary distribution for A^α given by (5.1), and the construction in the proof of Theorem 5.1 gives the desired process. □

Since the discount rate is allowed to be state and control dependent, Theorem 5.1 implies an extension of Theorem 1 of Krylov [13] in which it is shown that to each diffusion $\xi(t)$ having nonanticipating drift and diffusion coefficients and nonanticipating killing rate $\gamma(t)$, there exists a diffusion $x(t)$ having drift and diffusion coefficients and a killing rate g which are all functions of the state alone such that they have the same Green measure: for each $\Gamma \in \mathcal{B}(\mathbb{R}^d)$,

$$\begin{aligned}\mu(\Gamma) &= E\left[\int_0^\infty I_\Gamma(\xi(t))e^{-\int_0^t \gamma(s)\,ds}\,dt\right] \\ &= E\left[\int_0^\infty I_\Gamma(x(t))e^{-\int_0^t g(x(s))\,ds}\,dt\right].\end{aligned}$$

Theorem 1 of [13] requires a uniform ellipticity assumption on the diffusion ξ which the following corollary relaxes.

COROLLARY 5.4. *Suppose that W is an $\mathbb{R}^d$-valued, $\{\mathcal{F}_t\}$-Brownian motion; $\widehat{\sigma}$ and $\widehat{b}$ are measurable, $\{\mathcal{F}_t\}$-adapted processes taking values in $\mathbb{M}^{d\times d}$ and $\mathbb{R}^d$, respectively; $X(0)$ is $\mathbb{R}^d$-valued and $\mathcal{F}_0$-measurable; and γ is a nonnegative, bounded, $\{\mathcal{F}_t\}$-adapted process such that*

$$E\left[\int_0^\infty e^{-\int_0^t \gamma(s)ds}(1 + \|\widehat{\sigma}(t)\widehat{\sigma}^T(t)\| + |\widehat{b}(t)| + \gamma(t))dt\right] < \infty.$$

Let

$$X(t) = X(0) + \int_0^t \widehat{\sigma}(s)dW(s) + \int_0^t \widehat{b}(s)ds.$$

Then there exist measurable functions $\sigma : \mathbb{R}^d \to \mathbb{M}^{d\times d}$, $b : \mathbb{R}^d \to \mathbb{R}^d$, and $g : \mathbb{R}^d \to [0, \infty)$, an $\mathbb{R}^d$-valued Brownian motion $\tilde{W}$, and a process Y satisfying

$$Y(t) = Y(0) + \int_0^t \sigma(Y(s))d\tilde{W}(s) + \int_0^t b(Y(s))ds$$

such that for each $\Gamma \in \mathcal{B}(\mathbb{R}^d)$

$$E\left[\int_0^\infty I_\Gamma(X(t))e^{-\int_0^t \gamma(s)ds}dt\right] = E\left[\int_0^\infty I_\Gamma(Y(t))e^{-\int_0^t g(Y(s))ds}dt\right].$$

Proof. Let $E = \mathbb{R}^d$ and $U = \mathbb{M}_+^{d\times d} \times \mathbb{R}^d \times \mathbb{R}$. Define the generator for $f \in C_c^2(\mathbb{R}^d)$ (and $u = (u^1, u^2, u^3)$) by

$$Af(x,u) = \frac{1}{2}\sum_{i,j=1}^{d} u_{ij}^1 f_{x_i x_j}(x) + \sum_{i=1}^{d} u_i^2 f_{x_i}(x)$$

and let $\alpha(x,u) = u^3$ and $\psi(u) = ||u^1|| + |u^2| + |u^3|$. Letting $\widehat{a}(t) = \widehat{\sigma}(t)\widehat{\sigma}(t)^T$ and $\Lambda_t(du) = \delta_{\{\widehat{a}(t)\}}(du^1)\delta_{\{\widehat{b}(t)\}}(du^2)\delta_{\{\gamma(t)\}}(du^3)$, (X,Λ) is a solution of the controlled martingale problem for A. Using the cost function $c(x,u) = I_\Gamma(x)$, Theorem 5.1 yields the result where

$$a(x) = \int_U u^1 \eta(x,du), \qquad b(x) = \int_U u^2 \eta(x,du), \qquad g(x) = \int_U \alpha(x,u)\eta(x,du),$$

σ is the symmetric square root of a (which exists even in the degenerate case), and $\tilde{W}$ is obtained from [9, Theorem 5.3.3]. □

5.2. First passage problems. The previous formulation of the exit problem includes time as a component of the state of the process and allows the running cost c and exit cost g to depend on the time component. The result naturally has time-dependence in the Markov control.

We now consider cost structures *in which the only time-dependence is through discounting at a rate* $\alpha \geq 0$ that can depend on the state and the control. We show existence of solutions with a time-homogeneous Markov control. In this model, the cost is defined using the state process up to the time it leaves an open region $\tilde{E}_0 \subset E$ rather than the time-state process leaving an open region $E_0 \subset \mathbb{R}^+ \times E$.

THEOREM 5.5. *Suppose E, U, A, and ψ satisfy conditions* (i)–(vi) *of section* 1.1. *Let $\tilde{E}_0 \subset E$ be open. Let (X,Λ) be a solution of the controlled martingale problem for A, let α be a nonnegative, bounded, and continuous function on $E \times U$, define $\tau = \inf\{t \geq 0 : X(t) \in \tilde{E}_0^c\}$, and assume that*

$$E\left[\int_0^\tau e^{-\int_0^t \int_U \alpha(X(s),u)\Lambda_s(du)ds}\left(1 + \int_U \psi(u)\Lambda_t(du)\right)dt\right] < \infty.$$

Then there exists a process Y adapted to a filtration $\{\mathcal{F}_t\}$, a transition function η from E into U, and an $\{\mathcal{F}_t\}$-stopping time $\tilde{\tau}$ such that

$$f(Y(t\wedge\tilde{\tau})) - \int_0^{t\wedge\tilde{\tau}} \int_U Af(Y(s),u)\,\eta(Y(s),du)\,ds$$

is an $\{\mathcal{F}_t\}$-martingale for all $f \in \mathcal{D}(A)$, $P\{Y(\tilde{\tau}) \notin \tilde{E}_0^c, \tilde{\tau} < \infty\} = 0$, the Lebesgue measure of $\{t : Y(t) \in \tilde{E}_0^c, t < \tilde{\tau}\}$ is zero a.s., and, setting

$$\Gamma_Y(t) = \int_0^t \int_U \alpha(Y(s),u)\eta(Y(s),du)ds$$

and

$$\Gamma_X(t) = \int_0^t \int_U \alpha(X(s),u)\Lambda_s(du)ds,$$

$$E\left[\int_0^{\tilde\tau}\int_U e^{-\Gamma_Y(s)}c(Y(s),u)\,\eta(Y(s),du)\,ds + e^{-\Gamma_Y(\tilde\tau)}g(Y(\tilde\tau))\right]$$

$$= E\left[\int_0^{\tau}\int_U e^{-\Gamma_X(s)}c(X(s),u)\,\Lambda_s(du)\,ds + e^{-\Gamma_X(\tau)}g(X(\tau))\right] \tag{5.4}$$

for every $c \in M(E\times U)$ and $g \in M(E)$ that are bounded below (cf. Theorem 5.1*).*

Proof. Augment the state with a nonnegative component to form a new state space $\mathbb{R}^+ \times E$. Add a new point u^* to U giving $\overline{U} = U \cup \{u^*\}$ and augment the control space with a $\{0,1\}$ component to form a new control space $\overline{U}\times\{0,1\}$. Define $c(x,u^*) = 0$, $\alpha(x,u^*) = 1$ and $\psi(u^*) = 1$. Define the generator $\overline{A}$ by

$$\overline{A}(\gamma f)(s,x,u,v) = v\gamma(s)Af(x,u) + (1-v)\gamma'(s)f(x)$$

for all $f \in \mathcal{D}(A)$ and $\gamma \in \widehat{C}^1(\mathbb{R}^+)$.

Define

$$(\overline{S}(t),\overline{X}(t)) = (t - t\wedge\tau, X(t\wedge\tau))$$

and the relaxed control on $\overline{U}\times\{0,1\}$,

$$\overline{\Lambda}_t(du\times dv) = [\Lambda_t(du)\times\delta_{\{1\}}(dv)]\cdot I_{[0,\tau)}(t) + [\delta_{\{u^*\}}(du)\times\delta_{\{0\}}(dv)]\cdot I_{[\tau,\infty)}(t).$$

Then $(\overline{S},\overline{X},\overline{\Lambda})$ is a relaxed solution to the controlled martingale problem for $\overline{A}$ with $\overline{S}(0) = 0$ and $\overline{X}(0)$ having distribution ν_0, where ν_0 is the initial distribution of X. Let $\overline{\nu}_0 = \delta_{\{0\}}\times\nu_0$. As in the proof of Theorem 4.5, we can define $v(x) = I_{\tilde E_0}(x)$, and we will have $v(\overline{X}(t)) = I_{\{0\}}(\overline{S}(t))$ a.s.

Observe that with

$$\overline{c}(x,u,v) = \begin{cases} c(x,u) & \text{for} \quad x\in\tilde E_0, u\in\overline{U},\\ \alpha(x,u)g(x) & \text{for} \quad x\in\tilde E_0^c, u\in\overline{U},\end{cases}$$

the infinite-horizon discounted cost of $(\overline{S},\overline{X},\overline{\Lambda})$ satisfies

$$E\left[\int_0^\infty\int_{\overline{U}\times\{0,1\}} e^{-\int_0^t\int_{\overline{U}\times\{0,1\}}\alpha(\overline{X}(s),u)\overline{\Lambda}_s(du\times dv)ds}\overline{c}(\overline{X}(t),u,v)\overline{\Lambda}_t(du\times dv)\,dt\right]$$

$$= E\left[\int_0^\tau\int_U e^{-\int_0^t\int_U\alpha(X(s),u)\Lambda_s(du)}c(X(s),u)\Lambda_s(du)\,ds\right.$$

$$\left. +\ e^{-\int_0^\tau\int_U\alpha(X(s),u)\Lambda_s(du)}g(X(\tau))\right].$$

By Theorem 5.1, there exists a process (S,Y) and a transition function η from $\mathbb{R}^+\times E$ to $\overline{U}$ such that $(S,Y,\eta(S,Y,\cdot))$ is a relaxed solution of the controlled martingale problem for $(\overline{A},\overline{\nu}_0)$ and

$$E\left[\int_0^\infty\int_{\overline{U}} e^{-\Gamma_{\bar X}(t)}h(\overline{S}(t),\overline{X}(t),u,v)\,\overline{\Lambda}_t(du\times dv)\,dt\right]$$

$$= E\left[\int_0^\infty\int_{\overline{U}} e^{-\Gamma_Y(t)}h(S(t),Y(t),u,v(Y(t)))\eta(S(t),Y(t),du)\,dt\right] \tag{5.5}$$

for every measurable function h which is bounded below. In particular, taking $h(s,y,u,v) = |I_{\{0\}}(s) - v|$, we see that $I_{\{0\}}(S(t)) = v(Y(t))$ a.s.

Define $\tilde{\tau} = \inf\{r > 0 : S(r) > 0\}$ and $\sigma = \inf\{t > \tilde{\tau} : S(t) \leq 0\}$. Then for $f \in \mathcal{D}(A)$ and $\gamma \in \widehat{C}^1(\mathbb{R}^+)$,

(5.6)
$$\gamma(S((\tilde{\tau}+t)\wedge\sigma))f(Y((\tilde{\tau}+t)\wedge\sigma)) - \gamma(S(\tilde{\tau}))f(Y(\tilde{\tau})) - \int_{\tilde{\tau}}^{(\tilde{\tau}+t)\wedge\sigma} \gamma'(S(s))f(Y(s))ds$$

(taking (5.6) to be zero if $\tilde{\tau} = \infty$) is an $\{\mathcal{F}^{S,Y}_{\tilde{\tau}+t}\}$-martingale, and it follows from uniqueness for the martingale problem for $B\gamma f(s,x) = \gamma'(s)f(x)$ that $S(\tilde{\tau}+t) \equiv t$ and $Y(\tilde{\tau}+t) \equiv Y(\tilde{\tau})$, when $\tilde{\tau} < \infty$. In particular, if h does not depend on s, (5.5) becomes

$$E\left[\int_0^\tau \int_U e^{-\Gamma_X(t)}h(X(t),u,1)\,\Lambda_t(du)\,dt + e^{-\Gamma_X(\tau)}h(X(\tau),u^*,0)\right]$$
$$= E\left[\int_0^{\tilde{\tau}} \int_U e^{-\Gamma_Y(t)}h(Y(t),u,1)\eta(0,Y(t),du)\,dt + e^{-\Gamma_Y(\tilde{\tau})}h(Y(\tilde{\tau}),u^*,0)\right]$$

and (5.4) follows. □

COROLLARY 5.6. *Suppose that E, U, A, and ψ satisfy conditions* (i)–(vi) *of section* 1.1 *and that $\tilde{E}_0 \subset E$ is open. Let $E_0 = \mathbb{R}^+ \times \tilde{E}_0$. Let $\nu_0 \in \mathcal{P}(\tilde{E}_0)$, $\mu_0 \in \mathcal{M}(E_0 \times U)$, and $\mu_1 \in \mathcal{P}(E_0^c)$ satisfy* (4.11), (4.12), *and* (4.13). *Let $\alpha \in [0,\infty)$ and η be the transition function from E to U satisfying*

$$\int_{[0,\infty)\times E\times U} e^{-\alpha s}h(x,u)\mu_0(ds\times dx\times du) = \int_{E\times U} h(x,u)\eta(x,du)\mu_0^*(dx),$$

where $\mu_0^*(\Gamma) = \mu_0(\Gamma \times U)$, $\Gamma \in \mathcal{B}(E_0)$. Then there exists a process Y adapted to a filtration $\{\mathcal{F}_t\}$, and an $\{\mathcal{F}_t\}$-stopping time $\tilde{\tau}$ such that

$$f(Y(t\wedge\tilde{\tau})) - \int_0^{t\wedge\tilde{\tau}} \int_U Af(Y(s),u)\,\eta(Y(s),du)\,ds$$

is an $\{\mathcal{F}_t\}$-martingale for all $f \in \mathcal{D}(A)$, $P\{Y(\tilde{\tau}) \in \tilde{E}_0^c\} = 1$, the Lebesgue measure of $\{t : Y(t) \in \tilde{E}_0^c, t < \tilde{\tau}\}$ is zero a.s., and

$$E\left[\int_0^{\tilde{\tau}} \int_U e^{-\alpha s}c(Y(s),u)\,\eta(Y(s),du)\,ds + e^{-\alpha\tilde{\tau}}g(Y(\tilde{\tau}))\right]$$
$$= \int e^{-\alpha s}c(x,u)\,\mu_0(ds\times dx\times du) + \int e^{-\alpha s}g(x)\mu_1(ds\times dx)$$

for every $c \in M(E\times U)$ and $g \in M(E)$ that are bounded below (cf. Theorem 5.1*).*

Proof. The result is an immediate consequence of Theorems 4.5 and 5.5. □

For the undiscounted exit problem, we can simplify the conditions on μ_0 and μ_1.

COROLLARY 5.7. *Suppose that E, U, A, and ψ satisfy conditions* (i)–(vi) *of section* 1.1 *and that $\tilde{E}_0 \subset E$ is open. Suppose that $\nu_0 \in \mathcal{P}(\tilde{E}_0)$, $\mu_0 \in \mathcal{M}(\tilde{E}_0 \times U)$, and $\mu_1 \in \mathcal{P}(\tilde{E}_0^c)$ satisfy*

$$\int_{\tilde{E}_0\times U} Af(x,u)\mu_0(dx\times du) + \int_{\tilde{E}_0} f(x)\nu_0(dx) - \int_{\tilde{E}_0^c} f(x)\mu_1(dx) = 0, \quad f \in \mathcal{D}(A)$$

(*cf.* (4.11) *and* (4.12)) *and that*

$$\int_{\tilde{E}_0\times U}(1+\psi(u))\mu_0(dx\times du)<\infty.$$

Let η *satisfy* $\mu_0(dx\times du)=\eta(x,du)\mu_0^*(dx)$. *Then there exists a process* Y *adapted to a filtration* $\{\mathcal{F}_t\}$, *and an* $\{\mathcal{F}_t\}$*-stopping time* $\tilde{\tau}$ *such that*

$$f(Y(t\wedge\tilde{\tau}))-\int_0^{t\wedge\tilde{\tau}}\int_U Af(Y(s),u)\,\eta(Y(s),du)\,ds$$

is an $\{\mathcal{F}_t\}$*-martingale for all* $f\in\mathcal{D}(A)$, $P\{Y(\tilde{\tau})\in\tilde{E}_0^c\}=1$, *the Lebesgue measure of* $\{t:Y(t)\in\tilde{E}_0^c, t<\tilde{\tau}\}$ *is zero a.s., and*

$$E\left[\int_0^{\tilde{\tau}}\int_U c(Y(s),u)\,\eta(Y(s),du)\,ds+g(Y(\tilde{\tau}))\right]$$
$$=\int_{\tilde{E}_0\times U}c(x,u)\,\mu_0(dx\times du)+\int_{\tilde{E}_0^c}g(x)\mu_1(dx)$$

for every $c\in M(E\times U)$ *and* $g\in M(E)$ *that are bounded below (cf. Theorem* 5.1*).*

Proof. Consider the generator $\tilde{A}$ for a process in $[0,\infty)\times E$ with control space $U\times\{0,1\}$ defined by

$$\tilde{A}(\zeta f)(s,x,u,v)=v\zeta(s)Af(x,u)+(1-v)\zeta'(s)f(x)+(1-v)\left(\zeta(0)\int f\,d\nu_0-\zeta(s)f(x)\right),$$

and for some fixed $u^*\in U$, define $\pi\in\mathcal{P}([0,\infty)\times E\times U\times\{0,1\})$ by

$$\pi(ds\times dx\times du\times dv)=C(\delta_{\{0\}}(ds)\delta_{\{1\}}(dv)\mu_0(dx\times du)+e^{-s}ds\delta_{\{0\}}(dv)\delta_{\{u^*\}}(du)\mu_1(dx))\,,$$

where C is a constant normalizing π to be a probability measure. Then π is a stationary measure for $\tilde{A}$, and defining $\alpha(s,x,u,v)=(1-v)$, we see that the conditions of Corollary 5.3 are satisfied with A replaced by

$$A_0(\zeta f)(s,x,u,v)=v\zeta(s)Af(x,u)+(1-v)\zeta'(s)f(x).$$

Consequently, noting that

$$\eta_0(s,x,du\times dv)=I_{\{0\}}(s)\eta(x,du)\delta_{\{1\}}(dv)+I_{(0,\infty)}(s)\delta_{\{0\}}(dv)\delta_{\{u^*\}}(du),$$

there exists a process (S,Y) such that $((S,Y),\eta_0(S,Y,\cdot))$ is a relaxed solution of the controlled martingale problem for $(A_0,\nu_0\times\delta_{\{0\}})$. Using the cost function $c(s,x,u,v)=c(x,u)v+g(x)(1-v)$, (5.3) implies

$$\frac{E\left[\int_0^\infty e^{-\int_0^t I_{(0,\infty)}(S(s))ds}\left(\int_U c(Y(t),u)I_{\{0\}}(S(t))\eta(Y(t),du)+g(Y(t))I_{(0,\infty)}(S(t))\right)dt\right]}{E\left[\int_0^\infty e^{-\int_0^t I_{(0,\infty)}(S(s))ds}\,dt\right]}$$
$$=C\left(\int_{\tilde{E}_0\times U}c(x,u)\mu_0(dx\times du)+\int_{\tilde{E}_0^c}g(x)\mu_1(dx)\right).$$

Finally, since for any solution of the martingale problem for A_0, $S(t)>0$ implies that $S(r)>0$ for all $r>t$, if we define $\tilde{\tau}=\inf\{t:S(t)>0\}$, we have the desired result. □

6. LP reformulations. In this section we reformulate the original control problems as linear programs over appropriate spaces of measures. The importance of these reformulations is that they provide a way to *characterize* an optimal Markov control. Compute the measure(s) μ, $\widehat{\pi}$, π, or μ_0 and μ_1 satisfying the appropriate constraints which minimizes the cost criterion. The relaxed control is then the conditional distribution η on the control space U given the state.

Throughout this section we assume that E, U, A, and ψ satisfy the conditions of section 1.1. In addition, we assume that there are constants a and b such that

$$\psi(u) \le a + bc(x,u) \quad \text{or} \quad \psi(u) \le a + bc(s,x,u) \tag{6.1}$$

and

$$\{(x,u) : c(x,u) \le a\} \quad \text{or} \quad \{(s,x,u) : c(s,x,u) \le a\} \quad \text{is compact for each } a > 0. \tag{6.2}$$

6.1. Long-term average problems. The LP reformulation is especially straightforward for long-term average control problems provided the minimization is over all solutions of the controlled martingale problem for A without any restrictions on the initial distribution.

THEOREM 6.1. *Suppose that E, U, A, and ψ satisfy conditions* (i)–(vi) *of section* 1.1. *Let $c : E \times U \to \mathbb{R}$ be lower semicontinuous, bounded below, and satisfy* (6.1) *and* (6.2). *Then the long-term average control problem of minimizing* (1.6) *over all solutions of the controlled martingale problem for A is equivalent to the LP of minimizing*

$$\int_{E\times U} c(x,u)\,\mu(dx \times du) \tag{6.3}$$

over all distributions $\mu \in \mathcal{P}(E \times U)$ satisfying

$$\int_{E\times U} Af(x,u)\,\mu(dx \times du) = 0 \qquad \forall f \in \mathcal{D}(A). \tag{6.4}$$

Remark 6.2. This result is a combination of Theorems 3.2 and 3.3 in [19]. However, the result stated here is stronger in that the proof of the equivalence uses the existence of a solution with Markov controls given in Theorem 2.2 in place of Theorem 4.1 in [18].

Proof. Let (X, Λ) be a solution of the controlled martingale problem for A, and define

$$\mu_t(\Gamma_1 \times \Gamma_2) = E\left[t^{-1}\int_0^t \int_U I_{\Gamma_1\times\Gamma_2}(X(s),u)\Lambda_s(du)ds\right].$$

If (1.6) is finite, then the conditions on c imply that $\{\mu_t\}$ is relatively compact. The lower semicontinuity of c implies that for any limit point μ of $\{\mu_t\}$, $\int c d\mu$ is smaller than (1.6). Furthermore, $\int Af d\mu = 0$ for all $f \in \mathcal{D}(A)$. It follows that the minimum cost for the LP is a lower bound for (1.6) for all solutions of the controlled martingale problem.

The conditions on c imply that if there exists at least one μ satisfying (6.4) for which (6.3) is finite, then there exists a solution μ^* for the LP. But by Theorem 2.2 there exists a stationary solution of the controlled martingale problem with marginals given by μ^*, and hence with long-run average cost given by $\int c\,d\mu^*$, that is, the minimal value of the LP. □

6.2. Discounted problems. We first reformulate the control problem in which the *only* time-dependence is through discounting.

THEOREM 6.3. *Suppose that E, U, A, and ψ satisfy conditions* (i)–(vi) *of section* 1.1. *Let c be a measurable function which is bounded below, and which satisfies* (6.1) *and* (6.2). *Then, for each $\alpha > 0$, the discounted control problem of minimizing* (1.4) *over all solutions of the controlled martingale problem for (A, ν_0) is equivalent to the linear programming problem of minimizing*

$$\alpha^{-1} \int_{E\times U} c(x,u)\widehat{\pi}(dx \times du) \tag{6.5}$$

over all distributions $\widehat{\pi} \in \mathcal{P}(E \times U)$ satisfying

$$\int_{E\times U} A^{\alpha} f(x,u)\widehat{\pi}(dx \times du) = 0 \quad \forall f \in \mathcal{D}(A), \tag{6.6}$$

where

$$A^{\alpha} f(x,u) = Af(x,u) + \alpha \left[\int f(y)\nu_0(dy) - f(x) \right].$$

Proof. Let (X, Λ) be a solution of the controlled martingale problem for A, and define $\widehat{\pi}$ by

$$\int_{E\times U} h(x,u)\widehat{\pi}(dx \times du) = \alpha \int_0^{\infty} e^{-\alpha s} E\left[\int_U h(X(s),u)\Lambda_s(du) \right] ds.$$

If the α-discounted cost is finite, that is,

$$E\left[\int_0^{\infty} e^{-\alpha s} \int_U c(X(s),u)\, \Lambda_s(du)\, ds \right] < \infty,$$

then by (6.1), (4.1) holds, and as in the proof of Theorem 5.1, $\widehat{\pi}$ satisfies (6.6). Moreover, the definition of $\widehat{\pi}$ implies that the α-discounted cost satisfies

$$E\left[\int_0^{\infty} e^{-\alpha s} \int_U c(X(s),u)\, \Lambda_s(du)\, ds \right] = \alpha^{-1} \int_{E\times U} c(x,u)\widehat{\pi}(dx \times du)\ .$$

Conversely, if $\widehat{\pi}$ satisfies (6.6) and (6.5) is finite, then by (6.1), the conditions of Corollary 5.3 are satisfied and hence there exists a solution $(Y, \eta(Y,\cdot))$ of the controlled martingale problem for the original generator A whose α-discounted cost is also given by (6.5). □

When the running cost function is allowed to be time-dependent, the existence result in Theorem 4.1 can be used in the preceding argument in place of Theorem 5.1, resulting in a different LP formulation. The resulting Markov control is now time-dependent. This observation is summarized in the next theorem. It is interesting to note that both LP formulations are equivalent to the discounted problem when c does not depend on time, and thus the two LPs have the same value.

THEOREM 6.4. *Suppose that E, U, A, and ψ satisfy conditions* (i)–(vi) *of section* 1.1. *Let c be a measurable function which is bounded below and satisfies* (6.1) *and* (6.2). *Then the discounted control problem of minimizing* (1.4) *over all solutions of*

the controlled martingale problem for (A, ν_0) is equivalent to the linear programming problem of minimizing

$$\alpha^{-1} \int_{E\times U} c(s,x,u)\pi(ds \times dx \times du)$$

over all distributions $\pi \in \mathcal{P}(\mathbb{R}^+ \times E \times U)$ satisfying

$$\int_{E\times U} \tilde{A}(\gamma f)(s,x,u)\, \pi(ds \times dx \times du) = 0 \qquad \forall f \in \mathcal{D}(A)\ , \gamma \in \widehat{C}^1(\mathbb{R}^+),$$

where $\tilde{A}$ is defined by

$$\tilde{A}(\gamma f)(s,x,u) = \gamma(s)Af(x,u) + \gamma'(s)f(x) + \alpha\left[\gamma(0)\int f(y)\nu_0(dy) - \gamma(s)f(x)\right].$$

6.3. Finite-horizon problems. Since the optimal control depends on the time remaining, the equivalent LP formulation for a finite-horizon control problem is the same regardless of whether or not the running cost c is time-dependent.

THEOREM 6.5. *Suppose that E, U, A, and ψ satisfy conditions* (i)–(vi) *of section* 1.1. *Let c and g be measurable functions which are bounded below and suppose c satisfies* (6.1) *and* (6.2). *Then the finite-horizon control problem of minimizing* (1.5) *over all solutions of the controlled martingale problem for (A, ν_0) is equivalent to the linear programming problem of minimizing*

$$\int c(s,x,u)\, \mu_0(ds \times dx \times du) + \int g(x)\, \mu_1(dx)$$

over all measures $\mu_0 \in \mathcal{M}((0,T] \times E \times U)$ and $\mu_1 \in \mathcal{P}(E)$ satisfying

$$\int \overline{A}(\gamma f)(s,x,u)\, \mu_0(ds \times dx \times du) + \int \overline{B}(\gamma f)(x)\, \mu_1(dx) = 0$$
$$\forall\, f \in \mathcal{D}(A)\ , \gamma \in \widehat{C}^1(\mathbb{R}^+),$$

where $\overline{A}$ is defined by

$$\overline{A}(\gamma f)(s,x,u) = \gamma(s)Af(x,u) - \gamma'(s)f(x)$$

and $\overline{B}$ is

$$\overline{B}(\gamma f)(x) = \gamma(T)\int f(y)\, \nu_0(dy) - \gamma(0)f(x).$$

Proof. The proof is essentially the same as for Theorem 6.3 but uses Theorem 4.8 and Corollary 4.9 in place of Theorem 5.1 and Corollary 5.3. □

6.4. First passage problems. We consider three slightly different cases for the first passage criterion and develop equivalent LP formulations. We first consider the general case in which the running cost c and exit cost g are time-dependent.

THEOREM 6.6. *Suppose that E, U, A, and ψ satisfy conditions* (i)–(vi) *of section* 1.1. *Let c and g be measurable functions which are bounded below and suppose c*

satisfies (6.1) *and* (6.2). *Then the first passage control problem of minimizing* (1.7) *over all solutions of the controlled martingale problem for* (A, ν_0) *is equivalent to the linear programming problem of minimizing*

$$\int_{E_0} c(s,x,u)\,\mu_0(ds \times dx \times du) + \int_{E_0^c} g(s,x)\mu_1(ds \times dx)$$

over all measures $\mu_0 \in \mathcal{M}(E_0 \times U)$ *and* $\mu_1 \in P(E_0^c)$ *satisfying*

$$\int \widehat{A}(\gamma f)(s,x,u)\,\mu_0(ds \times dx \times du) + \int \widehat{B}(\gamma f)(s,x)\,\mu_1(ds \times dx) = 0$$

$$\forall\, f \in \mathcal{D}(A)\ , \gamma \in \widehat{C}^1(\mathbb{R}^+),$$

where $\widehat{A}$ *is defined by*

$$\widehat{A}(\gamma f)(s,x,u) = \gamma(s)Af(x,u) + \gamma'(s)f(x)$$

and $\widehat{B}$ *is*

$$\widehat{B}(\gamma f)(s,x) = \gamma(0)\int f(y)\,\nu_0(dy) - \gamma(s)f(x).$$

Proof. The proof is essentially the same as for Theorem 6.3 but uses Theorem 4.5 and Corollary 4.7 in place of Theorem 5.1 and Corollary 5.3. □

We now consider the special case in which the only time-dependence is through discounting at a rate $\alpha > 0$; i.e., $c(s,x,u) = e^{-\alpha s}\tilde{c}(x,u)$ and $g(s,x) = e^{-\alpha s}\tilde{g}(x)$.

THEOREM 6.7. *Let* c *and* g *be measurable functions which are bounded below and suppose* c *satisfies* (6.1) *and* (6.2). *Then the first passage control problem of minimizing* (1.7) *over all solutions of the controlled martingale problem for* (A, ν_0) *is equivalent to the linear programming problem of minimizing*

$$\int_{E_0} e^{-\alpha s}c(x,u)\,\mu_0(ds \times dx \times du) + \int_{E_0^c} e^{-\alpha s}g(x)\mu_1(ds \times dx)$$

over all measures $\mu_0 \in \mathcal{M}(E_0 \times U)$ *and* $\mu_1 \in P(E_0^c)$ *satisfying*

$$\int \widehat{A}(\gamma f)(s,x,u)\,\mu_0(ds \times dx \times du) + \int \widehat{B}(\gamma f)(s,x)\,\mu_1(ds \times dx) = 0$$

$$\forall\, f \in \mathcal{D}(A)\ , \gamma \in \widehat{C}^1(\mathbb{R}^+),$$

where $\widehat{A}$ *is defined by*

$$\widehat{A}(\gamma f)(s,x,u) = \gamma(s)Af(x,u) + \gamma'(s)f(x)$$

and $\widehat{B}$ *is*

$$\widehat{B}(\gamma f)(s,x) = \gamma(0)\int f(y)\,\nu_0(dy) - \gamma(s)f(x).$$

Proof. The proof is essentially the same as for Theorem 6.3 but uses Theorem 5.5 and Corollary 5.6 in place of Theorem 5.1 and Corollary 5.3. □

Note that, in general, the optimal control η^* obtained from Theorem 6.6 will be time-dependent, whereas that from Theorem 6.7 will be time-homogeneous.

Finally, we consider the undiscounted criterion in which the running cost and exit cost functions do not depend on the time.

THEOREM 6.8. *Let c and g be measurable functions which are bounded below and suppose c satisfies* (6.1) *and* (6.2). *Then the first passage control problem of minimizing* (1.7) *over all solutions of the controlled martingale problem for (A, ν_0) is equivalent to the linear programming problem of minimizing*

$$\int_{\tilde{E}_0} c(x,u)\,\mu_0(dx \times du) + \int_{\tilde{E}_0^c} g(x)\mu_1(dx)$$

over all measures $\mu_0 \in \mathcal{M}(\tilde{E}_0 \times U)$ and $\mu_1 \in P(\tilde{E}_0^c)$ satisfying

$$\int_{\tilde{E}_0 \times U} Af(x,u)\mu_0(dx \times du) + \int_{\tilde{E}_0^c} Bf(x)\mu_1(dx) = 0, \quad f \in \mathcal{D}(A),$$

where B is defined by

$$Bf(x) = \int f(y)\,\nu_0(dy) - f(x).$$

Proof. The proof is essentially the same as for Theorem 6.3 but uses Theorem 5.5 and Corollary 5.7 in place of Theorem 5.1 and Corollary 5.3. □

7. Appendix. The proof of Lemma 4.4 relies on the following result.

LEMMA 7.1. *Let $S(t)$ be a nonnegative, real-valued, cadlag stochastic process adapted to a filtration $\{\mathcal{F}_t\}$ and let ν be a random measure on $(0,\infty)\times(0,\infty)$ adapted to $\{\mathcal{F}_t\}$ in the sense that $\nu((0,t]\times A)$ is $\mathcal{F}_t$-measurable for each $A \in \mathcal{B}((0,\infty))$ and each $t \geq 0$. Suppose that for each continuously differentiable γ with γ and γ' bounded*

$$M_\gamma(t) = \gamma(S(t)) - \int_0^t \gamma'(S(u))du - \int_{(0,t]\times(0,\infty)} (\gamma(0) - \gamma(s))\nu(du \times ds) \tag{7.1}$$

is an $\{\mathcal{F}_t\}$-local martingale and that there exists a sequence of stopping times with $\alpha_n \to \infty$ a.s. such that for each $t > 0$,

$$E[\nu((0, t\wedge\alpha_n]\times(0,\infty))] < \infty. \tag{7.2}$$

Then, except for a discrete set of time points at which S jumps to zero, S increases linearly at rate 1. *Defining $\mathcal{Z} = \{u \geq 0 : S(u) = 0\}$ and letting $N(t)$ be the cardinality of $\mathcal{Z} \cap [0,t]$, $N(t)$ is a counting process such that*

$$N(t) - \nu((0,t]\times(0,\infty)) \tag{7.3}$$

is an $\{\mathcal{F}_t\}$-local martingale.

Proof. Note that if we define

$$S_n(t) = \begin{cases} S(t), & t < \alpha_n, \\ S(\alpha_n) + t - \alpha_n, & t \geq \alpha_n, \end{cases}$$

and

$$\nu_n((0,t]\times(a,b]) = \nu((0, t\wedge\alpha_n]\times(a,b])$$

and replace (S,ν) in the statement of the lemma by (S_n,ν_n), then M_γ will be a martingale (rather than a local martingale), and if the conclusion of the lemma holds for (S_n,ν_n) for all n, then the conclusion of the lemma holds for (S,N). With that observation, we assume that M_γ is a martingale and that $E[\nu((0,t]\times(0,\infty))]<\infty$. Under these assumptions, we will show that (7.3) is a martingale.

We approximate the set $\mathcal{Z}$ using the following level crossing times. Let $0<\delta<\epsilon$ and define $\tau_0^{\delta,\epsilon}=0$:

$$\begin{aligned}\sigma_1^{\delta,\epsilon}&=\inf\{t>0:S(t)>\epsilon\},\\ \tau_k^{\delta,\epsilon}&=\inf\{t>\sigma_k^{\delta,\epsilon}:S(t)\le\delta\},\\ \sigma_{k+1}^{\delta,\epsilon}&=\inf\{t>\tau_k^{\delta,\epsilon}:S(t)>\epsilon\},\end{aligned}$$

and $\Gamma_t=\cup_{k=1}^\infty(t\wedge\sigma_k^{\delta,\epsilon},t\wedge\tau_k^{\delta,\epsilon}]$. Note that

$$Z\cap[0,t]\subset[0,\sigma_1^{\delta,\epsilon})\cup\cup_{k=1}^\infty[t\wedge\tau_k^{\delta,\epsilon},t\wedge\sigma_{k+1}^{\delta,\epsilon})\ .$$

Let $\gamma\ge0$ with $\gamma(s)=0$ for $s\le\epsilon$ and $\gamma(s)>0$ for $s>\epsilon$. Then for $k\ge1$, the optional sampling theorem implies

$$\begin{aligned}0&=E[M_\gamma(t\wedge\sigma_{k+1}^{\delta,\epsilon})-M_\gamma(t\wedge\tau_k^{\delta,\epsilon})]\\ &=E\left[\gamma(S(t\wedge\sigma_{k+1}^{\delta,\epsilon}))+\int_{(t\wedge\tau_k^{\delta,\epsilon},t\wedge\sigma_{k+1}^{\delta,\epsilon}]\times(0,\infty)}\gamma(s)\nu(du\times ds)\right].\end{aligned}\tag{7.4}$$

Since $\gamma(s)>0$ on (ϵ,∞), it follows that

$$\nu((\tau_k^{\delta,\epsilon},\sigma_{k+1}^{\delta,\epsilon}]\times(\epsilon,\infty))=0\tag{7.5}$$

a.s. for each $k\ge1$ such that $\tau_k^{\delta,\epsilon}<\infty$, and $S(\sigma_k^{\delta,\epsilon})=\epsilon$ for each $k\ge2$ such that $\sigma_k^{\delta,\epsilon}<\infty$.

Now let γ satisfy $\gamma(0)=1$, $0\le\gamma(s)<1, s>0$, and $\gamma(s)=0, s\ge\delta$. Then, since $M(t\wedge\tau_k^{\delta,\epsilon})-M(t\wedge\sigma_k^{\delta,\epsilon})$ is a martingale for each k by the optional sampling theorem, (7.2) and the monotone convergence theorem imply that

$$\begin{aligned}\sum_{k=1}^\infty(M(t\wedge\tau_k^{\delta,\epsilon})-M(t\wedge\sigma_k^{\delta,\epsilon}))&=\sum_{\tau_k^{\delta,\epsilon}\le t}\gamma(S(\tau_k^{\delta,\epsilon}))-\gamma(0)\nu(\Gamma_t\times(0,\infty))\\ &\quad+\int_{\Gamma_t\times(0,\delta)}\gamma(s)\nu(du\times ds)\end{aligned}\tag{7.6}$$

is a martingale. It follows that

$$\gamma(0)E[\nu(\Gamma_t\times(0,\infty))]=E\left[\sum_{\tau_k^{\delta,\epsilon}\le t}\gamma(S(\tau_k^{\delta,\epsilon}))+\int_{\Gamma_t\times(0,\delta)}\gamma(s)\nu(du\times ds)\right].\tag{7.7}$$

Since the left side of (7.7) depends on γ only through $\gamma(0)$, which we fix as 1, and both terms on the right side are monotone increasing in γ, it follows that $\nu(\Gamma_t\times(0,\delta))=0$ a.s. and that $S(\tau_k^{\delta,\epsilon})=0$ for each $k\ge1$ such that $\tau_k^{\delta,\epsilon}<\infty$ (otherwise changing γ would change the value of the right side). Since δ and ϵ are arbitrary, it in turn

follows that S is monotone increasing except for jumps to 0. Since $\gamma(0) = 1$, $N^\epsilon(t) = \sum_{\tau_k^{\delta,\epsilon} \le t} \gamma(S(\tau_k^{\delta,\epsilon}))$ (which does not depend on δ) simply counts the number of jumps of S from above ϵ to 0. The martingale in (7.10) can be written

$$N^\epsilon(t) - \nu(\Gamma_t \times [0,\infty)),$$

keeping in mind that Γ_t does depend on ϵ. Since

$$E[N^\epsilon(t)] = E[\nu(\Gamma_t \times [0,\infty))] \le E[\nu((0,t] \times [0,\infty))] < \infty,$$

$N(t) = \lim_{\epsilon \to 0} N^\epsilon(t)$ exists, and letting $\delta, \epsilon \to 0$, (7.5) implies that (7.6) converges to

$$N(t) - \nu((0,t] \times (0,\infty)), \tag{7.8}$$

which, consequently, is a martingale. Note that if we show that S is strictly increasing except for the jumps to zero we will have that $N(t)$ is the cardinality of $\mathcal{Z} \cap [0,t]$.

Now let $\tau_x^r = \inf\{t > r : S(t) \ge x\}$ and assume that $\gamma \ge 0$ and $\gamma(z) = 0$ for $z \le x$ and $\gamma(z) > 0$ for $z > x$. Then

$$E[I_{\{S(r)<x\}}\gamma(S(\tau_x^r \wedge t))] = -E\left[I_{\{S(r)<x\}} \int_{(r,\tau_x^r \wedge t] \times (0,\infty)} \gamma(z)\nu(du \times dz)\right], \tag{7.9}$$

and it follows that both sides must be zero so that $S(\tau_x^r) = x$, if $S(r) < x$ and $\tau_x^r < \infty$. Note that this conclusion implies that S has no upward jumps and hence is continuous except for jumps to zero. Consequently,

$$\gamma(S(t)) - \int_0^t (\gamma(0) - \gamma(S(u-)))dN(u)$$

is continuous. The fact that (7.8) is a martingale implies that

$$\int_0^t (\gamma(0) - \gamma(S(u-)))dN(u) - \int_{(0,t]\times(0,\infty)} (\gamma(0) - \gamma(S(u-)))\nu(du \times ds) \tag{7.10}$$

is a martingale, and adding (7.10) to (7.1), we see that

$$\begin{aligned} \gamma(S(t)) - \int_0^t \gamma'(S(u))du - \int (\gamma(0) - \gamma(S(u-)))dN(u) \\ + \int_{(0,t]\times(0,\infty)} (\gamma(s) - \gamma(S(u-)))\nu(du \times ds) \end{aligned} \tag{7.11}$$

is a martingale. Equation (7.9) also implies that $\nu((r, \tau_x^r] \times (x,\infty)) = 0$, and hence for $0 = r_0 < r_1 < \cdots$ and $0 = x_0 < x_1 < \cdots$, we have

$$\sum_{i,j} I_{(S(r_i-),\infty)}(x_j)\nu((r_i, \tau_{x_j}^{r_i} \wedge r_{i+1}] \times (x_j, x_{j+1}]) = 0.$$

Letting $\max_i(r_{i+1} - r_i) \to 0$ and observing that if $x_j > S(u-)$, then $\tau_{x_j}^r > u$,

$$\liminf \sum_i I_{(S(r_i-),\infty)}(x_j) I_{(r_i, \tau_{x_j}^{r_i} \wedge r_{i+1}]}(u) \ge I_{(S(u-),\infty)}(x_j),$$

and hence

$$\sum_j \int_{(0,\infty)} I_{(S(u-),\infty)}(x_j)\nu(du \times (x_j, x_{j+1}]) = 0.$$

Now letting $\max_j(x_{j+1} - x_j) \to 0$, we have

$$\lim \sum_j I_{(S(u-),\infty)}(x_j) I_{(x_j,x_{j+1}]}(s) = I_{(S(u-),\infty)}(s),$$

and hence

$$\int_{(0,\infty)\times(0,\infty)} I_{(S(u-),\infty)}(s)\nu(du \times ds) = 0. \tag{7.12}$$

Let $\tau_r = \inf\{t > r : S(t) = 0\}$, and let $\gamma(0) = 1$, $\gamma(x) < 1$ for $0 < x < a$ and $\gamma(x) = 1$ for $x \geq a$. Then

$$E\left[I_{\{S(r)\geq a\}} \int_{(r,\tau_r\wedge t]\times(0,\infty)} (1-\gamma(s))\nu(du \times ds)\right] = E[I_{\{S(r)\geq a\}}(M_\gamma(\tau_r \wedge t) - M_\gamma(r))] = 0,$$

and it follows that

$$I_{\{S(r)\geq a\}}\nu((r, \tau_r] \times (0, a)) = 0 \qquad \text{a.s.} \tag{7.13}$$

Since a is arbitrary, it follows that (7.13) holds for all rational a a.s. and hence, by taking limits, for all a a.s. Consequently, approximating as in the proof of (7.12), (7.13) implies

$$\int_{(0,\infty)\times(0,\infty)} I_{(0,S(u-))}(s)\nu(du \times ds) = 0 \tag{7.14}$$

and (7.12) and (7.14) imply that the support of ν is contained in $\{(u, S(u-)) : u > 0\}$, and hence the last term in (7.11) is zero. But that observation implies that (7.11) is a continuous martingale and hence is constant. Taking $\gamma(s) = s$ gives

$$S(t) - S(0) + \int_0^t S(u-)dN(u) = t,$$

and it follows that S increases linearly at rate 1 except for jumps to zero. □

Proof of Lemma 4.4. For $\lambda > 0$, let $Z(t) = e^{\lambda t}Q(t)$. Then for g in C^1 with g and g' bounded,

$$g(Z(t)) - \int_0^t \left((V_1(s)e^{\lambda s} + \lambda Z(s))g'(Z(s)) + V_2(s)(g(0) - g(Z(s)))\right) ds$$

is an $\{\mathcal{F}_t\}$-local martingale. For simplicity, assume that $\int_0^\infty V_1(u)du = \infty$ a.s. (If not, we can modify Q so that the integral is infinite and the conclusions of the lemma for the modified process imply the conclusions for the original.) Let $\zeta(t) = \inf\{r > \tau : \int_\tau^r (V_1(u)e^{\lambda s} + \lambda Z(s))\, du > t\}$, $\tilde{Z}(t) = Z \circ \zeta(t)$, and

$$\nu((0,t] \times (0,a]) = \int_\tau^{\zeta(t)} V_2(u) I_{(0,a]}(Z(u))du = \int_{\sigma_0^\tau}^{\zeta(t)} V_2(u) I_{(0,a]}(Z(u))du.$$

Then, by the optional sampling theorem,

$$g(\tilde{Z}(t)) - \int_0^t g'(\tilde{Z}(s))\,ds - \int_{(0,t]\times(0,\infty)} (g(0)-g(s))\nu(du\times ds) \tag{7.15}$$

is an $\{\mathcal{F}_{\zeta(t)}\}$-local martingale. By Lemma 7.1, $\tilde{Z}$ increases linearly at rate one except for jumps to zero. Consequently, for $\zeta(t) < \sigma_1^\tau$

$$Z(\zeta(t)) - Z(\zeta(0)) = \tilde{Z}(t) - \tilde{Z}(0) = t = \int_\tau^{\zeta(t)} (V_1(u)e^{\lambda s} + \lambda Z(s))\,du\,,$$

where the first equality is the definition of $\tilde{Z}$, the second is the consequence of Lemma 7.1, and the third is the definition of $\zeta(t)$. Note that if $Z(\tau) > 0$, then $\zeta(0) = \tau$, and if $Z(\tau) = 0$, then $Z(\zeta(0)) = 0$. In either case, since λ is arbitrary, it follows that $Q(t) - Q(\tau) = \int_\tau^t V_1(u)du$ for $\tau \le t < \sigma_1^\tau$ and that $\zeta(0) = \sigma_0^\tau$.

Now assume that $\sigma_1^\tau < \infty$ a.s. Let N be the counting process of jumps to zero by $\tilde{Z}$, and let

$$\Lambda(t) = \int_{\zeta(0)}^{\zeta(t)} V_2(s) I_{(0,\infty)}(Z(s))ds.$$

Then, again by Lemma 7.1,

$$N(t) - \Lambda(t) \tag{7.16}$$

is an $\{\mathcal{F}_{\zeta(t)}\}$-local martingale, and since ζ is continuous on any interval on which $\tilde{Z}$ is positive, Λ is continuous. Let τ_1 be the time of the first jump of N. Then $\zeta(\tau_1) = \sigma_1^\tau$. Note that

$$\Lambda(\tau_1) = \int_{\zeta(0)}^{\zeta(\tau_1)} V_2(s) I_{(0,\infty)}(Z(s))ds = \int_{\sigma_0^\tau}^{\sigma_1^\tau} V_2(s)ds.$$

Let $\eta(x) = \inf\{u : \Lambda(u) > x\}$, and for bounded positive f, define

$$L_f(t) = f(N(t)) \exp\left\{ - \int_0^t \frac{f(N(s)+1) - f(N(s))}{f(N(s))} d\Lambda(s) \right\}.$$

Then, by Itô's formula,

$$L_f(t) = f(0) + \int_0^t L_f(s-)d(N(s) - \Lambda(s)),$$

and hence L_f is a local martingale. In particular, it follows that

$$\begin{aligned} 1 &= E[L_f(\tau_1 \wedge \eta(x))|\mathcal{F}_{\zeta(0)}] \\ &= E[f(1)I_{\{\tau_1 < \eta(x)\}} e^{-\frac{f(1)-f(0)}{f(0)}\Lambda(\tau_1)\wedge x} |\mathcal{F}_{\zeta(0)}] + E[f(0)I_{\{\tau_1 \ge \eta(x)\}} e^x |\mathcal{F}_{\zeta(0)}]. \end{aligned}$$

Letting $f(0) = 1$ and then taking a limit as $f(1) \to 0$, we have

$$P\{\Lambda(\tau_1) \ge x|\mathcal{F}_{\sigma_0^\tau}\} = P\{\Lambda(\tau_1) \ge x|\mathcal{F}_{\zeta(0)}\} = P\{\tau_1 \ge \eta(x)\} = e^{-x}. \qquad \square$$

REFERENCES

[1] A. G. Bhatt and R. L. Karandikar, *Invariant measures and evolution equations for Markov processes*, Ann. Probab., 21 (1993), pp. 2246–2268.

[2] A. G. Bhatt and V. S. Borkar, *Occupation measures for controlled Markov processes: Characterization and optimality*, Ann. Probab., 24 (1996), pp. 1531–1562.

[3] V. S. Borkar, *A remark on the attainable distributions of controlled diffusions*, Stochastics, 18 (1986), pp. 17–23.

[4] V. S. Borkar and M. K. Ghosh, *Ergodic control of multidimensional diffusions* I: *The existence results*, SIAM J. Control Optim., 26 (1988), pp. 112–126.

[5] C. Derman, *On sequential decisions and Markov chains*, Management Sci., 9 (1962), pp. 16–24.

[6] E. V. Denardo, *On linear programming in a Markov decision problem*, Management Sci., 16 (1970), pp. 281–288.

[7] A. Dvoretsky, *Asymptotic normality for sums of dependent random variables*, in Proc. 6th Berkeley Symposium on Mathematical Statistics and Probability, L. M. Le Cam, J. Neyman, and E. L. Scott, eds., University of California Press, Berkeley, CA, 1972, pp. 513–535.

[8] N. El Karoui, D. Nguyen, and M. Jeanblanc-Picqué, *Compactification methods in the control of degenerate diffusions: Existence of an optimal control*, Stochastics, 20 (1987), pp. 169–219.

[9] S. N. Ethier and T. G. Kurtz, *Markov Processes: Characterization and Convergence*, Wiley, New York, 1986.

[10] W. H. Fleming and D. Vermes, *Convex duality approach to the optimal control of diffusions*, SIAM J. Control Optim., 27 (1989), pp. 1136–1155.

[11] I. Gyöngy, *Mimicking the one-dimensional marginal distributions of processes having an Ito differential*, Probab. Theory Related Fields, 71 (1986), pp. 501–516.

[12] U. G. Haussmann and J. P. Lepeltier, *On the existence of optimal controls*, SIAM J. Control Optim., 28 (1990), pp. 851–902.

[13] N. V. Krylov, *Once more about the connection between elliptic operators and Itô's stochastic equations*, in Statistics and Control of Stochastic Processes, Steklov Seminar, 1984, Optimization Software Inc., New York, 1985, pp. 214–229.

[14] T. G. Kurtz, *Averaging for martingale problems and stochastic approximation*, in Proc. Joint U.S.–French Workshop on Applied Stochastic Analysis, Lecture Notes in Control and Inform. Sci. 177, Springer-Verlag, New York, 1992, pp. 186–209.

[15] A. S. Manne, *Linear programming and sequential decisions*, Management Sci., 6 (1960), pp. 259–267.

[16] B. G. Pittel, *A linear programming problem connected with optimal stationary control in a dynamic decision problem*, Theory Probab. Appl., 16 (1971), pp. 724–728.

[17] H. L. Royden, *Real Analysis*, 2nd ed., Macmillan, New York, 1968.

[18] R. H. Stockbridge, *Time-average control of martingale problems: Existence of a stationary solution*, Ann. Probab., 18 (1990), pp. 190–205.

[19] R. H. Stockbridge, *Time-average control of martingale problems: A linear programming formulation*, Ann. Probab., 18 (1990), pp. 206–217.

[20] P. Wolfe and G. B. Dantzig, *Linear programming in a Markov chain*, Oper. Res., 10 (1962), pp. 702–710.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 2, pp. 654–679, March 1998

009

A CONSTRUCTION OF RATIONAL WAVELETS AND FRAMES IN HARDY–SOBOLEV SPACES WITH APPLICATIONS TO SYSTEM MODELING*

NICHOLAS F. DUDLEY WARD† AND JONATHAN R. PARTINGTON†

Abstract. Using the Daubechies wavelet theory we establish rational wavelet decompositions of the Hardy–Sobolev classes on the half-plane. The decay of wavelet coefficients is analyzed and error bounds for approximation are given. We give applications to the modeling of linear systems and to the model reduction of infinite-dimensional systems.

Key words. wavelets, frames, atomic decompositions, matching pursuits, infinite-dimensional systems, Hardy–Sobolev spaces

AMS subject classifications. 41, 93

PII. S0363012996297339

1. Introduction.

1.1. Notation and conventions.

$\mathbb{C}_+ = \{s = x + iy : x > 0\}$ right half-plane,
$\mathbb{I} = \{iy : y \in \mathbb{R}\}$ imaginary axis.

For f belonging to $L^2(\mathbb{R})$ the Fourier transform $\hat{f}$ is defined using the following convention:

$$\hat{f}(\xi) = \int_{-\infty}^{\infty} f(t)e^{-i\xi t}dt.$$

For g belonging to $L^2((0,\infty))$ we write $G = (\mathcal{L}g)(s)$ for the Laplace transform of g:

$$G(s) = (\mathcal{L}g)(s) = \int_0^{\infty} g(t)e^{-st}dt.$$

$H^2(\mathbb{C}_+)$ denotes the Hardy space of functions $F(s)$ analytic in the right half-plane and such that

$$\|F\|_2 = \sup_{x>0} \int_{\infty}^{\infty} |F(x+iy)|^2 dy < \infty.$$

By the Paley–Wiener theorem every $F(s)$ belonging to $H^2(\mathbb{C}_+)$ is the Laplace transform of some $f(t) \in L^2((0,\infty))$ such that $\|F\|_2 = (2\pi)^{1/2}\|f\|_2$. Upper-case letters will be used to denote the Laplace transform of the corresponding lower-case letter, e.g., $\Psi(s) = (\mathcal{L}\psi)(s)$. (For further details of Hardy spaces defined on a half-plane see [28] or [36].)

Lettered results, e.g., Theorem A, will denote a known result in the literature. Numbered results will be proved.

*Received by the editors January 17, 1996; accepted for publication (in revised form) January 17, 1997. The research of the first author was supported by the EPSRC.
http://www.siam.org/journals/sicon/36-2/29733.html
†School of Mathematics, University of Leeds, Leeds LS2 9JT, UK (pmt6ndw@amsta.leeds.ac.uk, pmt6jrp@leeds.ac.uk).

1.2. General background. The approximation and identification of transfer functions of stable linear time-invariant infinite-dimensional systems by those of finite-dimensional systems is an important part of modern systems theory, and there are two norms which are most commonly considered. The H^∞ norm has found applications in robust control [20], whereas the H^2 norm is the basis of linear quadratic Gaussian control. However, it is widely regarded as desirable to be able to consider both H^2 and H^∞ criteria simultaneously: see, for example, the articles of Foias, Frazho, and Tannenbaum [19] and Bernstein and Haddad [5]. One approach to this is by means of Hardy–Sobolev classes.

Hardy–Sobolev classes for the disc were considered in [15] and [2]. For the upper half-plane the Hardy–Sobolev class $H^{2,m}$, $m \in \mathbb{R}$, is defined as follows: let $H^{2,m}$, $m \in \mathbb{R}$, be the class of functions $F(s)$ analytic in the right half-plane such that $F(s) = (\mathcal{L}f)(s)$ and

$$\|F\|_{2,m} = \left(\int_0^\infty |f(t)|^2 (1+t)^{2m} dt \right)^{1/2} < \infty. \tag{1.1}$$

If $m = 0$, $H^{2,0}$ corresponds to the classical Hardy space H^2 for the half-plane, and if m is a positive integer, $F(s)$ belonging to $H^{2,m}$ is equivalent to the first m derivatives of $F(s)$ belonging to H^2. An equivalent norm can then be defined on $H^{2,m}$ by the formula

$$\|F\|_{2,m}^2 = \|F\|_2^2 + \|F'\|_2^2 + \cdots + \|F^m\|_2^2.$$

In this investigation we shall be concerned mainly with the range $|m| \leq 1$. The range $m > 1/2$ is of particular interest since for $F(s)$ belonging to $H^{2,m}$ we have $\|F\|_\infty \leq C_m \|F\|_{2,m}$, where C_m is a constant, and so approximation in the Hardy–Sobolev norm gives simultaneous approximations in both the uniform and the L^2 norms. Indeed, we can say more: since we obtain, using the Cauchy–Schwarz inequality,

$$\begin{aligned} \|f\|_{L^1} &= \int_0^\infty |f(t)|\, dt \\ &\leq \|f(t)(1+t)\|_{L^2} \|1/(1+t)\|_{L^2} \\ &= \|F\|_{2,1}, \end{aligned}$$

we obtain simultaneous approximation in the L^1 norm as well, itself important in bounded input, bounded output (BIBO) stable control theory (cf. [10]).

The approach to approximation that we shall take will involve the theory of wavelets, and we now outline this.

1.3. Wavelet classes for H^2. We give a definition of wavelet which is perhaps somewhat more geometric than is normal, and which includes the more usual analytic definition and is appropriate for both the disc and half-plane. By a lattice $\mathcal{L}$ we mean a discrete set of points in the right half-plane which is both sufficiently dense and separated with respect to the pseudohyperbolic metric: there exists $0 < \delta_1 \leq \delta_2 < 1$ such that the union of balls $B(S, \delta_2)$, $S \in \mathcal{L}$, covers $\mathbb{C}_+$ (i.e., δ_2 dense) and S, $T \in \mathcal{L}, S \neq T \Longrightarrow \rho(S,T) > \delta_1$ (i.e., δ_1 separated), where

$$\rho(S,T) = \left| \frac{S-T}{S+\overline{T}} \right|.$$

(See [42] for details.)

In effect, to say that a set is separated in the above sense amounts to the fact that the ratio of the distance between neighboring points in the right half-plane to the distance from the imaginary axis is approximately constant. In this paper we are interested in lattices of the following form:

$$\mathcal{L} = \left\{ S_{j,k} : S_{j,k} = \frac{1}{2^j} + i\frac{k}{2^j} b_0; j, k \in \mathbb{Z} \right\},$$

where b_0 is some fixed positive constant. Clearly, $\mathcal{L}$ is separated and $C_1 \operatorname{dist}(S_{j,k}, \mathbb{I}) \leq \operatorname{dist}(S_{j,k}, S_{j',k'}) \leq C_2 \operatorname{dist}(S_{j,k}, \mathbb{I})$. Associated with a given lattice $\mathcal{L}$ will be a wavelet class $\mathcal{W} = \{F_W : W \in \mathcal{L}\}$. Notice that any point in $\mathcal{L}$ can be transformed onto another by a dilation followed by a translation. We construct $\mathcal{W}$ so that each F_W is obtained from some fixed *mother wavelet* $\Psi(y) \in H^2$ in a similar fashion. Thus $\mathcal{W}$ actually consists of the system of functions $\Psi_{j,k}(y) = 2^{j/2}\Psi(2^j y - kb_0)$ where we have normalized them in L^2.

In this investigation we are interested in wavelet classes for H^2 with mother wavelets of the form $\Psi(y) = (1 + iy)^p$, where p is a fixed positive integer. Suppose that $W = X + iY = 2^{-j} + ikb_0 2^{-j} \in \mathcal{L}$. Then the corresponding function $\Psi_{j,k}$ can be written

$$\begin{aligned} \Psi_W(y) &= \frac{X^{-1/2}}{(1 + i(y - Y)/X)^p} \\ &= \frac{X^{p-1/2}}{((X - iY) + iy)^p} \\ &= \frac{X^{p-1/2}}{(iy + \overline{W})^p}. \end{aligned}$$

For $p = 1$ Ψ_W is just the normalized Cauchy kernel with pole at $-\overline{W}$, and for $p = 2$ Ψ_W is the Bergman kernel evaluated on the imaginary axis. From the practical point of view the usefulness of the wavelet classes just given is based on the observation that one has a system of rational functions whose poles cluster on the left of the imaginary axis. Much of the present investigation centers about the general problem of obtaining decompositions by using wavelets as elementary building blocks in a sense which will be made precise below.

Using a fundamental result of Daubechies we obtain decompositions of H^2 where the mother wavelets consist of powers of the Cauchy kernel for the right half-plane. Consider for a moment the space $H^{2,1}$ of the disc. It is easy to see that $f(z) = \sum_{n=0}^{\infty} a_n z^n \in H^{2,1}$ is equivalent to $\|f\|_{2,1}^2 = \sum_{n=0}^{\infty}(1 + n^2)|a_n|^2 < \infty$. That is, one can determine whether a function $f(z)$ in H^2 also has L^2 bounded derivative by examining decay of the Taylor coefficients $a_n = \langle f, z^n \rangle$. We obtain similar conditions for $H^{2,1}(\mathbb{C}_+)$ where the building blocks are replaced by a nonorthogonal system of rational wavelets. We also consider in section 6 the half-plane algebra $A(\mathbb{C}_+)$ for the right half-plane and use a result of Hayman and Lyons to show that suitable sets of Cauchy kernels are fundamental for $A(\mathbb{C}_+)$. In section 7 we consider error estimates, and in section 8 we deduce algorithms which may be applied to the model reduction of a system.

The following results will be established.

THEOREM 1.1. *Let $\Psi(y) = (1 + iy)^{-3}$, let $\Psi_{j,k}(y) = 2^{j/2}\Psi(2^j y - b_0 k)$ where $j, k \in \mathbb{Z}$, $b_0 > 0$, and let $\langle \cdot, \cdot \rangle$ be the usual L^2-inner product.*

If $b_0 > 0$ is sufficiently small, then for each m with $-1 \leq m \leq 1$ there exist positive constants A_m and B_m such that for $F(s) \in H^{2,m}$ the wavelet coefficients

$\langle F, \Psi_{j,k}\rangle$ satisfy the following pair of inequalities:

$$A_m\|F\|_{2,m}^2 \le \sum_{j,k} |\langle F, \Psi_{j,k}\rangle|^2 (1+2^j)^{2m} \le B_m\|F\|_{2,m}^2. \tag{1.2}$$

Our next result is an atomic decomposition of $H^{2,m}$. Let $\ell^2((1+2^{2j})^m)$ denote the weighted ℓ^2-space

$$\ell^2((1+2^j)^{2m}) = \left\{\boldsymbol{\lambda} = (\lambda_{j,k}) : \|\boldsymbol{\lambda}\|_{2,m}^2 = \sum |\lambda_{j,k}|^2 (1+2^j)^{2m} < \infty\right\}.$$

A sequence $\Psi_{j,k}$ in $H^{2,m}$ will be called a *set of atoms (with respect to* $\ell^2((1+2^j)^{2m})$*)* for $H^{2,m}$ if the mapping $S : \ell^2((1+2^j)^{2m}) \to H^{2,m}$ defined by

$$S : \boldsymbol{\lambda} \mapsto \sum_{j,k} \lambda_{j,k}\Psi_{j,k}$$

is a surjection. Let $\Psi_{j,k} = 2^{j/2}\Psi(2^j y - b_0 k)$ where $\Psi(t) = (1+iy)^{-3}$.

THEOREM 1.2. *The system $(\Psi_{j,k})$, $j,k \in \mathbb{Z}$, is a set of atoms for $H^{2,m}$ for sufficiently small b_0 in the sense defined above. Furthermore, for each F belonging to $H^{2,m}$ we have the inequalities*

$$\frac{1}{B_{-m}}\|F\|_{2,m}^2 \le \inf\left\{\|(\lambda_{j,k})\|_{2,m}^2 : F = \sum_{j,k}\lambda_{j,k}\Psi_{j,k}\right\} \le \frac{1}{A_{-m}}\|F\|_{2,m}^2. \tag{1.3}$$

THEOREM 1.3. *Let T be the operator defined on $H^{2,m}$, where $m = 1$ or $m = -1$ by the formula*

$$TF = \sum \langle F, \Psi_{j,k}\rangle \Psi_{j,k}, \quad F \in H^{2,m}.$$

Then, for sufficiently small b_0, T is a bounded map $H^{2,m} \to H^{2,m}$. Furthermore, T is a surjection and invertible.

Finally, we have the following frame decomposition of $H^{2,m}$ for both $m = 1$ and $m = -1$.

COROLLARY 1.4. *Every $F \in H^{2,m}$, $m = \pm 1$, has the wavelet series representation*

$$F = TT^{-1}F = \sum_{j,k} \left\langle T^{-1}F, \Psi_{j,k}\right\rangle \Psi_{j,k}.$$

We shall prove Theorem 1.1 for the cases $m = \pm 1$ in section 3, and for the general case at the end of section 5. Theorems 1.2 and 1.3 will be proved in sections 4 and 5, respectively. We also obtain estimates for the various constants for a range of b_0 which appear in the given results by means of computer, and for these we refer to the tables.

2. Frames and wavelet decompositions of H^2.

2.1. Brief exposition of the theory of frames. Frames were introduced by Duffin and Schaeffer in [17] in the context of nonharmonic Fourier series, and give a technique for expanding vectors in a Hilbert space by means of systems of non-orthogonal vectors.

We give below a brief exposition of the theory of frames. For further details and fuller proofs we refer to [17], [13], and [34]. Let H be a Hilbert space with inner

product $\langle \cdot, \cdot \rangle$. A system (ϕ_j) of vectors in H will be called a *frame* for H if there exist positive constants A and B so that the following pair of inequalities obtain:

$$A\|f\|^2 \leq \sum_j |\langle f, \phi_j \rangle|^2 \leq B\|f\|^2, \quad f \in H. \tag{2.1}$$

We define the operator T on H by means of the formula

$$Tf = \sum_j \langle f, \phi_j \rangle \phi_j, \quad f \in H.$$

From (2.1) it follows that T is a positive operator on H such that $\langle Tf, f \rangle \geq A\|f\|^2$. Next one can establish that T is a surjection and is invertible by means of the following piggyback closed range theorem.

LEMMA A. *If U is a positive operator on H such that $\langle Uf, f \rangle \geq \alpha\|f\|^2$ for all $f \in H$ then U is invertible on H and its inverse U^{-1} satisfies $\|U^{-1}\| \leq 1/\alpha$.*

Proof. We first note that range(U) is a closed subspace of H. Let f_n be a Cauchy sequence in range(U). Then $f_n = U(g_n)$, and by hypothesis,

$$\|g_n - g_m\|^2 \leq \alpha^{-1} \langle U(g_n - g_m), g_n - g_m \rangle \leq \alpha^{-1} \|U(g_n - g_m)\| \cdot \|g_n - g_m\|.$$

Thus g_n is also a Cauchy sequence in H with limit g. It follows from the continuity of U that U has closed range.

Next we show that range(U) is dense in H: if $g \in H$ is such that $\langle Uf, g \rangle = 0$ for all $f \in H$, it follows in particular that $\langle Ug, g \rangle = 0$. We have $\alpha\|g\|^2 \leq \langle Ug, g \rangle$, and hence we deduce that $g = 0$. Hence, by the first part, range$(U) = H$.

Finally, since any $f \in H$ can be written as $f = Ug$ we may define $U^{-1}f = g$ and

$$\alpha\|U^{-1}f\|^2 \leq \langle UU^{-1}f, U^{-1}f \rangle = \langle f, U^{-1}f \rangle \leq \|f\| \cdot \|U^{-1}f\|,$$

whence $\|U^{-1}f\| \leq \alpha^{-1}\|f\|$. This completes the proof. □

In fact, one can show using (2.1) that $I - 2(A + B)^{-1}T$ is a contraction and so compute T^{-1} by a Neumann series: suppose that $f \in H$, and define a *residual vector* $R(f)$ by the equation

$$R(f) = f - \frac{2}{A + B} \sum_j \langle f, \phi_j \rangle \phi_j.$$

Then $-\frac{B-A}{B+A} I \leq R \leq \frac{B-A}{B+A} I$, and so, since R is symmetric, $\|R\| \leq \frac{B-A}{B+A}$. If B is close to A we obtain a good reconstruction of f. Otherwise we can write down an algorithm, viz., a Neumann series for the reconstruction of f with exponential convergence. (See [13] for details.)

Since T^{-1} is symmetric it follows that $\langle T^{-1}f, \phi_j \rangle = \langle f, T^{-1}\phi_j \rangle$ and so $f \in H$ has the expansions

$$f = TT^{-1}f = \sum \langle T^{-1}f, \phi_j \rangle \phi_j = \sum \langle f, T^{-1}\phi_j \rangle \phi_j.$$

The system defined by $\tilde{\phi}_j = T^{-1}\phi_j$ is called the *dual frame* associated with ϕ_j. It can also be shown that $\tilde{\phi}_j$ is itself a frame.

2.2. Fundamental theorem of Daubechies. Given a function $\Psi \in L^2(\mathbb{R})$, and $a_0 > 1$ and $b_0 > 0$, we define the system of functions

$$\Psi_{j,k}(t) = {a_0}^{j/2}\Psi(a_0^j t - kb_0), \quad j,k \in \mathbb{Z}. \tag{2.2}$$

We ask for conditions on Ψ and b_0 such that the system given by (2.2) is a frame for $L^2(\mathbb{R})$. That is, there exist positive constants A, B such that

$$A\|f\|_2^2 \le \sum_{j,k} |\langle f, \Psi_{j,k}\rangle|^2 \le B\|f\|_2^2, \quad f \in L^2(\mathbb{R}).$$

Daubechies [12], [13], proves the following fundamental result.

THEOREM B. *Suppose that $\Psi \in L^2(\mathbb{R})$ and $a_0 > 1$ are such that*

$$\inf_{1\le|\xi|\le a_0} \sum_{-\infty}^{\infty} |\hat\Psi(a_0^j\xi)|^2 > 0,$$

$$\sup_{1\le|\xi|\le a_0} \sum_{-\infty}^{\infty} |\hat\Psi(a_0^j\xi)|^2 < \infty,$$

and if $\beta(s) = \sup_{1\le|\xi|\le a_0} \sum |\hat\Psi(a_0^j\xi)||\hat\Psi(a_0^j\xi+s)|$ decays at least as fast as $(1+|s|)^{-(1+\epsilon)}$ with $\epsilon > 0$, then there exists a $\tilde b_0 > 0$ such that the $\Psi_{j,k}$ constitute a frame for all $b_0 < \tilde b_0$. For such b_0 the following equalities are frame bounds for the $\Psi_{j,k}$:

$$A = \frac{2\pi}{b_0}\left\{\inf_{1\le|\xi|\le a_0} \sum_{-\infty}^{\infty} |\hat\Psi(a_0^j\xi)|^2 - \sum_{\substack{k=-\infty\\k\ne0}}^{\infty} \left[\beta\left(\frac{2\pi}{b_0}k\right)\beta\left(\frac{-2\pi}{b_0}k\right)\right]^{1/2}\right\},$$

$$B = \frac{2\pi}{b_0}\left\{\sup_{1\le|\xi|\le a_0} \sum_{-\infty}^{\infty} |\hat\Psi(a_0^j\xi)|^2 + \sum_{\substack{k=-\infty\\k\ne0}}^{\infty} \left[\beta\left(\frac{2\pi}{b_0}k\right)\beta\left(\frac{-2\pi}{b_0}k\right)\right]^{1/2}\right\}.$$

Since the Fourier transforms of H^2 functions of the lower half-plane are functions in $L^2(\mathbb{R})$ supported on the positive real axis, we may apply Theorem B to $H^2(\mathbb{P})$ where $\mathbb{P} = \{x + iy : y < 0\}$. By rotating, we see that Theorem B is applicable to establishing frames for $H^2(\mathbb{C}_+)$. In particular, it is routine to check that the estimates given in the hypotheses of Theorem B are satisfied by mother wavelets of the form $\Psi(y) = (1 + iy)^{-p}$ where p is a positive integer not less than 2 (since $\Psi(y)$ is the Laplace transform of $\psi(t) = t^{p-1}e^{-t}/(p-1)!$, $t > 0$).

In the main we will be interested in the system $\Psi_{j,k}(y) = 2^{j/2}\Psi(2^j y - b_0 k)$ where Ψ is the mother wavelet $\Psi(y) = (1+iy)^{-3}$. However, in order to establish our results, we also consider some auxiliary systems of wavelets—in particular, ones generated by $\Phi(y) = (1 + iy)^{-2}$. With $a_0 = 2$ and a given mother wavelet Ψ, one can then easily establish by means of a computer a range of b_0 for which the corresponding system $\Psi_{j,k}(y)$ is a frame for H^2. See Tables 2.1 and 2.2 for the wavelets $\Psi(y) = (1 + iy)^{-3}$ and $\Phi(y) = (1 + iy)^{-2}$.

3. Proof of Theorem 1.1 for $m = \pm 1$.

TABLE 2.1
Estimates of frame bounds for the mother wavelet $\Psi(y) = (1+iy)^{-3}$.

b_0	A	B	B/A
0.25	13.579	13.615	1.003
0.5	6.786	6.811	1.004
1.0	3.114	3.685	1.183
2	0.460	2.939	6.387
3	−0.591	2.857	−4.832

TABLE 2.2
Estimates of frame bounds for the mother wavelet $\Psi(y) = (1+iy)^{-2}$.

b_0	A	B	B/A
0.25	9.064	9.066	1.000
0.5	4.532	4.533	1.000
1.0	2.208	2.325	1.053
2	0.738	1.528	2.069
3	0.118	1.392	11.777

3.1. Proof of Theorem 1.1: $m = 1$. The proof of Theorem 1.1 for $m = 1$ follows from the considerations of the previous section. Since $\Psi_{j,k}(y)$ is a frame for H^2 it follows that for F belonging to H^2,

$$A_\Psi \|F\|_2^2 \le \sum_{j,k} |\langle F, \Psi_{j,k}\rangle|^2 \le B_\Psi \|F\|_2^2. \tag{3.1}$$

If, in addition, F' also belongs to H^2 (so that $F \in H^{2,1}$) then since $\Phi_{j,k}(y)$ with $\Phi(y) = 2(1+iy)^{-2}$ is also a frame for H^2 we have

$$A_\Phi \|F'\|_2^2 \le \sum_{j,k} |\langle F', \Phi_{j,k}\rangle|^2 \le B_\Phi \|F'\|_2^2. \tag{3.2}$$

Up to a constant, Ψ is attained from Φ by differentiation with respect to y. Therefore we may integrate by parts and obtain

$$\begin{aligned}\langle F', \Phi_{j,k}\rangle &= \int_{-\infty}^{\infty} F'(it)\overline{\Phi}_{j,k}(t)dt \\ &= 2^j \int_{-\infty}^{\infty} F(it)\overline{\Psi}_{j,k}(t)dt \\ &= 2^j \langle F, \Psi_{j,k}\rangle .\end{aligned}$$

Therefore, (3.2) may be replaced by

$$A_\phi \|F'\|_2^2 \le \sum_{j,k} |\langle F, \Psi_{j,k}\rangle|^2 2^{2j} \le B_\phi \|F'\|_2^2. \tag{3.3}$$

Adding (3.1) and (3.3) and noting that

$$\frac{1}{2}(1+x)^2 \le (1+x^2) \le (1+x)^2,$$

we obtain the conclusion of Theorem 1.1 with $A_1 = 1/4 \min\{A_\Psi, A_\phi\}$ and $B_1 = \max\{B_\Psi, B_\phi\}$.

Remark. One could prove Theorem 1.1 directly using the same technique as the proof of Theorem 3.1 below. However, the proof given has the merit of being elementary, although it will not give the best estimates for A_1 and B_1.

3.2. Proof of Theorem 1.1: $m = -1$. We wish to establish the existence of positive constants A_{-1} and B_{-1} such that for $G = \mathcal{L}g$ belonging to $H^{2,-1}$, the following inequalities obtain:

$$A_{-1}\|G\|_{2,-1}^2 \le \sum_{j,k} |\langle G, \Psi_{j,k}\rangle|^2 (1+2^j)^{-2m} \le B_{-1}\|G\|_{2,-1}^2.$$

Since $G = \mathcal{L}(g) \in H^{2,-1}$ it follows that $\tilde{g}(t) = g(t)(1+t)^{-1} \in L^2((0,\infty))$. If $\Phi_{j,k} = \mathcal{L}(\phi_{j,k})$ is a frame for H^2 there exist constants A_ϕ, B_ϕ such that

$$A_\phi\|\tilde{g}\|_2^2 \le \sum_{j,k} |\langle \tilde{g}, \phi_{j,k}\rangle|^2 \le B_\phi\|\tilde{g}\|_2^2,$$

i.e.,

$$A_\phi\|g\|_{2,-1}^2 \le \sum_{j,k} |\langle \tilde{g}, \phi_{j,k}\rangle|^2 \le B_\phi\|g\|_{2,-1}^2.$$

Now

$$\langle \tilde{g}, \phi_{j,k}\rangle = \left\langle g, (1+t)^{-1}\phi_{j,k}\right\rangle.$$

We would like to choose a frame $\Phi_{j,k}$ for H^2 so that

$$\phi_{j,k} = (1+t)(1+2^j)^{-1}\psi_{j,k},$$

where Ψ is the mother wavelet $\Psi(y) = \mathcal{L}(t^2e^{-t}/2) = (1+iy)^{-3}$, $\Psi_{j,k} = 2^{j/2}\Psi(2^j y - b_0 k)$, and $\Psi_{j,k} = \mathcal{L}(2^{-j/2}(2^{-j}t)^2 e^{-2^{-j}(t-ib_0k)})$.

Our next result is an analogue of Theorem B.

THEOREM 3.1. *Suppose that $\Psi(y) \in H^2(\mathbb{C}_+)$ and that $\Psi_{j,k}(y) = 2^{j/2}\Psi(2^j y - kb_0)$, $j,k \in \mathbb{Z}$. Define the system $\Phi_{j,k}$ by $\phi_{j,k}(t) = (1+t)(1+2^j)^{-1}\psi_{j,k}(t)$. Suppose further that*

$$\inf_{0<t<\infty}(1+t)^2\sum_{-\infty}^{\infty}(1+2^j)^{-2}|\psi(2^{-j}t)|^2 > 0,$$

$$\sup_{0<t<\infty}(1+t)^2\sum_{-\infty}^{\infty}(1+2^j)^{-2}|\psi(2^{-j}t)|^2 < \infty,$$

and if $\gamma(s) = \sup_{0<t<\infty}(1+t)^2\sum(1+2^j)^{-2}|\psi(2^{-j}t)||\psi(2^{-j}t+s)|$ decays at least as fast as $(1+|s|)^{-(1+\epsilon)}$ with $\epsilon > 0$, then there exists a $\tilde{b}_0 > 0$ such that the $\Phi_{j,k}$ constitute a frame for H^2 for all $b_0 < \tilde{b}_0$. For such b_0 the following equalities are frame bounds for the $\Phi_{j,k}$:

$$A_{-1} = \frac{2\pi}{b_0}\left\{\inf_{0<t<\infty}(1+t)^2\sum_{-\infty}^{\infty}(1+2^j)^{-2}|\psi(2^{-j}t)|^2 - \sum_{\substack{k=-\infty\\k\neq 0}}^{\infty}\left[\gamma\left(\frac{2\pi}{b_0}k\right)\gamma\left(\frac{-2\pi}{b_0}k\right)\right]^{1/2}\right\},$$

$$B_{-1} = \frac{2\pi}{b_0}\left\{\sup_{0<t<\infty}(1+t)^2\sum_{-\infty}^{\infty}(1+2^j)^{-2}|\psi(2^{-j}t)|^2 + \sum_{\substack{k=-\infty\\k\neq 0}}^{\infty}\left[\gamma\left(\frac{2\pi}{b_0}k\right)\gamma\left(\frac{-2\pi}{b_0}k\right)\right]^{1/2}\right\}.$$

Suppose that $G = \mathcal{L}g \in H^2$. We proceed to analyze $\sum_{j,k} |\langle g, \phi_{j,k}\rangle|^2$ using the method of Daubechies:

$$\begin{aligned}
\sum_{j,k} |\langle g, \phi_{j,k}\rangle|^2 &= \sum_{j,k} \int_0^\infty g(t)\overline{\phi_{j,k}}(t)dt \int_0^\infty \overline{g}(t')\phi_{j,k}(t')dt' \\
&= \sum_{j,k} (1+2^j)^{-2} \int_0^\infty g(t)(1+t)\overline{\psi_{j,k}}(t)dt \\
&\qquad \times \int_0^\infty \overline{g}(t')(1+t')\psi_{j,k}(t')dt' \\
&= \sum_{j,k} (1+2^j)^{-2} 2^{-j} \int_0^\infty g(t)(1+t)\overline{\psi}(2^{-j}t)e^{-i2^{-j}b_0kt}dt \\
&\qquad \times \int_0^\infty \overline{g}(t')(1+t')\psi(2^{-j}t')e^{i2^{-j}b_0kt'}dt' \\
&= \frac{2\pi}{b_0} \sum_{j,k} (1+2^j)^{-2} \int_0^\infty \int_0^\infty g(t)\overline{g}(t')(1+t)\overline{\psi}(2^{-j}t)(1+t')\psi(2^{-j}t') \\
&\qquad \times \delta(t'-t-2\pi k 2^j b_0^{-1})\,dt\,dt' \\
&= \frac{2\pi}{b_0} \sum_{j,k} (1+2^j)^{-2} \int_0^\infty g(t)\overline{g}(t-2\pi 2^j b_0^{-1}k)(1+t)\overline{\psi}(2^{-j}t) \\
&\qquad \times (1+t-2\pi k 2^j b_0^{-1})\psi(2^{-j}t-2\pi k b_0^{-1})dt \\
&= \frac{2\pi}{b_0} \sum_{j} (1+2^j)^{-2} \int_0^\infty |g(t)|^2(1+t)^2|\psi(2^{-j}t)|^2dt + \mathrm{rest}(g),
\end{aligned}$$

where in the fourth line we have used the Poisson summation formula $\sum_{l\in\mathbb{Z}} \exp(ilax) = 2\pi a^{-1}\sum_{k\in\mathbb{Z}} \delta(x - 2\pi k a^{-1})$. This is justified by supposing first that g is smooth and has compact support and then proceeding by a standard density argument. If

$$\begin{aligned}
m &= \inf_{0<t<\infty} (1+t)^2 \sum_{-\infty}^{\infty} (1+2^j)^{-2}|\psi(2^{-j}t)|^2, \\
M &= \sup_{0<t<\infty} (1+t)^2 \sum_{-\infty}^{\infty} (1+2^j)^{-2}|\psi(2^{-j}t)|^2
\end{aligned}$$

we have

$$m\|g\|_2^2 \le \sum_j (1+2^j)^{-2} \int_0^\infty |g(t)|^2(1+t)^2|\psi(2^{-j}t)|^2dt \le M\|g\|_2^2. \tag{3.4}$$

3.3. Analysis of rest(g). We have

$$\begin{aligned}
\mathrm{rest}(g) = \frac{2\pi}{b_0} \sum_j \sum_{k\neq 0} (1+2^j)^{-2} \int_0^\infty g(t)\overline{g}(t-2\pi k 2^j b_0^{-1})(1+t)\overline{\psi}(2^{-j}t) \\
\times(1+t-2\pi k 2^j b_0^{-1})\psi(2^{-j}t-2\pi k b_0^{-1})dt.
\end{aligned}$$

TABLE 3.1
Estimates of A_{-1} and B_{-1} in Theorem 1.1.

b_0	A_{-1}	B_{-1}
0.25	13.699	65.524
0.5	6.797	32.814
1.0	0.775	19.031
1.1	−0.324	18.330

By the Cauchy–Schwarz inequality for integrals we have

$$|\operatorname{rest}(g)| \le \frac{2\pi}{b_0}\sum_j\sum_{k\neq 0}(1+2^j)^{-2}\left\{\int_0^\infty |g(t)|^2(1+t)^2|\psi(2^{-j}t)||\psi(2^{-j}t-2\pi kb_0^{-1})|dt\right\}^{1/2}$$
$$\times\left\{\int_0^\infty |g(t-2\pi k2^jb_0^{-1})|^2(1+t-2\pi k2^jb_0^{-1})^2|\psi(2^{-j}t)||\psi(2^{-j}t-2\pi kb_0^{-1})|dt\right\}^{1/2}.$$

We make the substitution $t-2\pi k2^jb_0^{-1}\mapsto t$ in the second integral and deduce that

$$|\operatorname{rest}(g)| \le \frac{2\pi}{b_0}\sum_j\sum_{k\neq 0}(1+2^j)^{-2}\left\{\int_0^\infty |g(t)|^2(1+t)^2|\psi(2^{-j}t)||\psi(2^{-j}t-2\pi kb_0^{-1})|dt\right\}^{1/2}$$
$$\times\left\{\int_0^\infty |g(t)|^2(1+t)^2|\psi(2^{-j}t)||\psi(2^{-j}t+2\pi kb_0^{-1})|dt\right\}^{1/2}.$$

Next we apply the Cauchy–Schwarz inequality for sums to the sum indexed by j and obtain

$$|\operatorname{rest}(g)| \le \frac{2\pi}{b_0}\sum_{k\neq 0}\left\{\sum_j(1+2^j)^{-2}\int_0^\infty |g(t)|^2(1+t)^2|\psi(2^{-j}t)||\psi(2^{-j}t-2\pi kb_0^{-1})|dt\right\}^{1/2}$$
$$\times\left\{\sum_j(1+2^j)^{-2}\int_0^\infty |g(t)|^2(1+t)^2|\psi(2^{-j}t)||\psi(2^{-j}t+2\pi kb_0^{-1})|dt\right\}^{1/2}.$$

Therefore, if we write

$$\gamma(s)=\sup_{0<t<\infty}(1+t)^2\sum_j(1+2^j)^{-2}|\psi(2^{-j}t)||\psi(2^{-j}t+s)|,$$

we obtain

$$|\operatorname{rest}(g)| \le \frac{2\pi}{b_0}\sum_{\substack{k=-\infty\\k\neq 0}}^{\infty}\left[\gamma\left(\frac{2\pi}{b_0}k\right)\gamma\left(\frac{-2\pi}{b_0}k\right)\right]^{1/2}. \tag{3.5}$$

Combining (3.4) with (3.5), we obtain the conclusion of Theorem 3.1.

We apply Theorem 3.1 to the considerations above with $\Psi(y)=(1+iy)^{-3}$. Since $\psi(t)=2^{-1}t^2e^{-t}$ it is easy to see that the hypotheses of the theorem are satisfied. We refer to Table 3.1 for estimates of frame bounds for different b_0. This completes the proof of Theorem 1.1 in the case $m=-1$.

4. Proof of Theorem 1.2. Although $H^{2,m}$ is a Hilbert space and so, by the Riesz–Fréchet theorem, is its own dual, we identify $(H^{2,m})^*$ with $H^{2,-m}$ by means of the following pairing:

$$\langle F, G\rangle = \int F\overline{G}, \quad F \in H^{2,m}, \quad G \in H^{2,-m}.$$

In fact, the map $M_{(1+t)^{2m}} : H^{2,m} \to H^{2,-m}$, $F(s) \mapsto (\mathcal{L}f(t)(1+t)^{2m})(s)$ is an isometric isomorphism. (For further details, see [21].) Note also that $|\langle F, G\rangle| \le \|F\|_{2,m}\|G\|_{2,-m}$ and that $\sup\{|\langle F, G\rangle| : \|G\|_{2,-m} = 1\} = \|F\|_{2,m}$.

4.1. Atomic decompositions of $H^{2,m}$. A useful tool for establishing atomic decompositions for a Banach space is Banach's closed range theorem, which we state below for the convenience of the reader. We refer to the articles of Bonsall [6], [7] (see also [14]) and the references contained therein for further information.

For a subset E of a normed space X let $E^\perp = \{\phi \in X^* : \phi(x) = 0 \quad \text{for} \quad x \in X\}$; for a subset F of X^* let $F_\perp = \{x \in X : \phi(x) = 0 \quad \text{for} \quad \phi \in X^*\}$.

THEOREM C (Banach's closed range theorem). *Let T be a bounded linear mapping of a Banach space X into a Banach space Y.*

(i) *If T^*Y^* is closed in X^*, then TX is closed in Y and $TX = (\ker T^*)_\perp$.*

(ii) *If TX is closed in Y, then T^*Y^* is closed in X^* and $T^*Y^* = (\ker T)^\perp$.*

LEMMA 4.1. *The adjoint S^* from $(H^{2,m})^* = H^{2,-m}$ to $(\ell^2((1+2^j)^{2m}))^* = \ell^2((1+2^j)^{-2m})$ is given by the formula*

$$S^* : G \mapsto \{\langle G, \Psi_{j,k}\rangle\}, \quad G \in H^{2,-m}.$$

Proof. Suppose that $G \in H^{2,-m}$ and let ϕ_G be the functional corresponding to G. Then, for any sequence $\boldsymbol{\lambda} = (\lambda_{j,k})$ belonging to $\ell^2((1+2^j)^{2m})$, it follows from the definition of S^* that

$$\begin{aligned}(S^*\phi_G)(\boldsymbol{\lambda}) &= \langle G, S\boldsymbol{\lambda}\rangle \\ &= \sum_{j,k} \lambda_{j,k} \langle G, \Psi_{j,k}\rangle .\end{aligned}$$

Thus S^* maps $G \mapsto \langle G, \Psi_{j,k}\rangle$. □

LEMMA 4.2. *S^* has zero kernel.*

Proof. By definition, if $G \in H^{2,-m}$,

$$\begin{aligned}\|S^*G\|_{2,-m}^2 &= \sup_{\boldsymbol{\lambda}\in\ell^2((1+2^j)^{2m})} \frac{|(S^*G)(\boldsymbol{\lambda})|^2}{\|\boldsymbol{\lambda}\|_{2,m}^2} \\ &= \sup_{\boldsymbol{\lambda}\in\ell^2((1+2^j)^{2m})} \frac{\left|\sum_{j,k} \lambda_{j,k} \langle G, \Psi_{j,k}\rangle\right|^2}{\|\boldsymbol{\lambda}\|_{2,m}^2}. \qquad (4.1)\end{aligned}$$

We now define the sequence $(\lambda_{j,k})$ by

$$\lambda_{j,k} = \overline{\langle G, \Psi_{j,k}\rangle}(1+2^j)^{-2m}.$$

Then

$$\begin{aligned}\|(\lambda_{j,k})\|_{2,m}^2 &= \sum_{j,k} |\langle G, \Psi_{j,k}\rangle|^2 (1+2^j)^{-4m}(1+2^j)^{2m} \\ &= \sum |\langle G, \Psi_{j,k}\rangle|^2 (1+2^j)^{-2m} \\ &\le B_{-m}\|G\|_{2,-m}^2\end{aligned}$$

by Theorem 1.1, and so is a candidate for the supremum in equation (4.1). Therefore

$$\begin{aligned}\|S^*G\|_{2,-m}^2 &\geq \frac{\left(\sum_{j,k} |\langle G, \Psi_{j,k}\rangle|^2 (1+2^j)^{-2m}\right)^2}{\sum_{j,k} |\langle G, \Psi_{j,k}\rangle|^2 (1+2^j)^{-2m}} \\ &= \sum_{j,k} |\langle G, \Psi_{j,k}\rangle|^2 (1+2^j)^{-2m} \\ &\geq A_{-m}\|G\|_{2,-m}^2, \end{aligned} \tag{4.2}$$

by Theorem 1.1. This completes the proof. □

LEMMA 4.3. *S^* has closed range.*

Proof. Let $((\lambda_{j,k}^n))_n$ be a Cauchy sequence in $\ell^2(1+2^j)^{-2m}$ in the range of S^*, so that $((\lambda_{j,k}^n))_n \to (\lambda_{j,k})$. Then each sequence $(\lambda_{j,k}^n) = S^*G_n$, where G_n belongs to $H^{2,m}$. By the inequality derived in (4.2) it follows that (G_n) is a Cauchy sequence in $H^{2,-m}$ so that $G_n \to G$. Therefore, it follows from the continuity of S^* that $(\lambda_{j,k}) = S^*(G)$ and we have the desired result. □

The first part of Theorem 1.2 follows immediately from Lemmas 4.1, 4.2, 4.3, and Banach's closed range theorem.

The lower bound in (1.3) follows immediately from an elementary duality argument. To obtain the upper bound let $N = \ker S$ and $X = \ell^2((1+2^j)^{2m})/N$ and define $U : X \to H^{2,m}$ by $U[\boldsymbol{\lambda}] = S\boldsymbol{\mu}$, $(\boldsymbol{\mu} \in [\boldsymbol{\lambda}] \in X)$. Plainly, U is a bounded linear bijection of X onto $H^{2,m}$ and so, by Banach's isomorphism theorem, has a bounded inverse. Therefore U^* is a bounded linear bijection from $H^{2,-m}$ onto X^*. From the proof of Lemma 4.2 we know that

$$\begin{aligned} A_{-m}\|G\|_{2,-m}^2 &\leq \sup\{\|(S^*G)(\boldsymbol{\lambda})\|_{2,-m}^2 : \boldsymbol{\lambda} \in \ell^2((1+2^j)^{2m})\} \\ &= \sup\{\|\langle F, U[\boldsymbol{\lambda}]\rangle : [\boldsymbol{\lambda}] \in X, \|[\boldsymbol{\lambda}]\| \leq 1\} \\ &= \|U^*G\|. \end{aligned}$$

Hence $\|U^{-1}\| = \|(U^*)^{-1}\| \leq A_{-m}^{-1}$. This completes the proof of Theorem 1.2.

Remark. The usefulness of Banach's closed range theorem in establishing atomic decompositions is clear from papers of D. H. Luecking (see, for example, [29] and [30]), who used it to prove decompositions of Bergman and Hardy spaces. In particular, a duality proof of the Coifman–Rochberg decomposition of L_a^p, $p > 1$, was given [9]. In [6] Bonsall gave an abstract formulation of this method giving various applications, and in [7] he gave an elementary proof of a general atomic decomposition theorem.

5. Proof of Theorem 1.3. We consider the operator

$$TF = \sum \langle F, \Psi_{j,k}\rangle \Psi_{j,k}, \quad F \in H^{2,m}, \tag{5.1}$$

where $\langle \cdot, \cdot \rangle$ is the L^2-inner product. We proceed to show that T is bounded on both $H^{2,1}$ and $H^{2,-1}$ and in fact is a bijection. Note that T is symmetric with respect to $\langle \cdot, \cdot \rangle$ and bounded on H^2.

LEMMA 5.1. *T is a bounded operator $H^{2,m} \to H^{2,m}$ for both $m = 1$ and $m = -1$.*

Proof. Let T_N be the partial sum operator

$$T_N = \sum_{|j|,|k| \leq N} \langle \cdot, \Psi_{j,k}\rangle \Psi_{j,k}.$$

Let $F \in H^{2,m}$ and $G \in H^{2,-m}$. Then, by the Cauchy–Schwarz inequality,

$$\begin{aligned}
|\langle T_N F, G\rangle|^2 &= \left|\left\langle \sum \langle F, \Psi_{j,k}\rangle \Psi_{j,k}, G\right\rangle\right|^2 \\
&= \left|\sum \langle F, \Psi_{j,k}\rangle \overline{\langle G, \Psi_{j,k}\rangle}\right|^2 \\
&\le \sum |\langle F, \Psi_{j,k}\rangle|^2 (1+2^j)^{2m} \sum |\langle G, \Psi_{j,k}\rangle|^2 (1+2^j)^{-2m} \\
&\le B_m \|F\|_{2,m}^2 \times B_{-m}\|G\|_{2,-m}^2,
\end{aligned}$$

by Theorem 1.1. Now, by duality,

$$\|T_N F\|_{2,m}^2 = \sup_{G\in H^{2,-m}} \frac{|\langle T_N F, G\rangle|^2}{\|G\|_{2,-m}^2} \le B_m B_{-m} \|F\|_{2,m}^2,$$

so that T_N is bounded (independently of N) on $H^{2,m}$ and so must T be. By the symmetry of T (with respect to the duality pairing given by $\langle\cdot,\cdot\rangle$) we conclude also that T is bounded on $H^{2,-m}$. □

Next we wish to establish that T is a bijection $H^{2,m} \to H^{2,m}$, for $m = 1$ and $m = -1$. To do this we estimate $\|TF\|_{2,m}^2$ from below. It follows from duality that

$$\begin{aligned}
\|TF\|_{2,m} &= \sup_{G\in H^{2,-m}} \frac{|\langle TF, G\rangle|}{\|G\|_{2,-m}} \\
&= \sup_{G\in H^{2,-m}} \frac{|\sum_{j,k} \langle F, \Psi_{j,k}\rangle \overline{\langle G, \Psi_{j,k}\rangle}|}{\|G\|_{2,-m}}.
\end{aligned}$$

We need to find a suitable candidate for G so that we may show that $\|TF\|_{2,m}^2 \ge K_m \|F\|_{2,m}^2$. We proceed to analyze $\sum_{j,k} \langle F, \Psi_{j,k}\rangle \overline{\langle G, \Psi_{j,k}\rangle} = \sum_{j,k} \langle f, \psi_{j,k}\rangle \overline{\langle g, \psi_{j,k}\rangle}$ using the same technique as in the proof of Theorem 3.1. We obtain

(5.2)
$$\begin{aligned}
\sum_{j,k} \langle F, \Psi_{j,k}\rangle \overline{\langle G, \Psi_{j,k}\rangle} &= \frac{2\pi}{b_0} \sum_j \int_0^\infty f(t)\overline{g}(t)|\psi(2^{-j}t)|^2 dt \\
&\quad + \frac{2\pi}{b_0} \sum_j \sum_{k\ne 0} \int_0^\infty f(t)\overline{g}(t - 2\pi k 2^j b_0^{-1}) \overline{\psi}(2^{-j}t)\psi(2^{-j}t - 2\pi k b_0^{-1}) dt.
\end{aligned}$$

It is clear in view of (5.2) that an appropriate candidate for G is given by $g(t) = f(t)(1+t)^{2m}$ (so that $\|g\|_{2,-m}^2 = \|f\|_{2,m}^2$). Next we analyze the second sum on the right of (5.2) with this choice of g, which we denote by $\mathrm{rest}(f)$. We deduce

$$|\,\mathrm{rest}(f)| \le \frac{2\pi}{b_0} \sum_j \sum_{k\ne 0} I_1 \times I_2,$$

where

$$I_1 = \left\{ \int_0^\infty |f(t)|^2 (1+t)^{2m} (1+t-2\pi k 2^j b_0^{-1})^{2m} |\psi(2^{-j}t)||\psi(2^{-j}t - 2\pi k b_0^{-1})| dt \right\}^{1/2}$$

and

$$\begin{aligned}
I_2 = \Bigg\{ \int_0^\infty |f(t - 2\pi k 2^j b_0^{-1})|^2 (1+t-2\pi k 2^j b_0^{-1})^{2m} (1+t)^{-2m} |\psi(2^{-j}t)| \\
\times |\psi(2^{-j}t - 2\pi k b_0^{-1})| dt \Bigg\}^{1/2}.
\end{aligned}$$

TABLE 5.1
Estimates of the lower bound for $\|TF\|_{2,1}$.

b_0	0.1	0.2	0.25	0.3	0.5
K	33.947	16.974	13.578	11.297	−4.222

We define

$$\mu = \inf_{0<t<\infty} \sum_j |\psi(2^{-j}t)|^2,$$
$$\delta_1^m(s) = \sup_{0<t<\infty} \sum_j (1+t+2^j s)^{2m} |\psi(2^{-j}t)||\psi(2^{-j}t+s)|,$$
$$\delta_2^m(s) = \sup_{0<t<\infty} \sum_j (1+t+2^j s)^{-2m} |\psi(2^{-j}t)||\psi(2^{-j}t+s)|.$$

If $\sum_{k\neq 0}\{\delta_1^m(\frac{2\pi k}{b_0})\delta_2^m(\frac{2\pi k}{b_0})\}^{1/2} = \sum_{k\neq 0}\{\delta_1^{-m}(\frac{2\pi k}{b_0})\delta_2^{-m}(\frac{2\pi k}{b_0})\}^{1/2}$ converges and diminishes to 0 with b_0, there exists $\tilde{b_0}$ such that for $b < \tilde{b_0}$, the quantity $K = \frac{2\pi}{b_0}(\mu - \sum_{k\neq 0}\{\delta_1^m(\frac{2\pi k}{b_0})\delta_2^m(\frac{2\pi k}{b_0})\}^{1/2}) > 0$, and we obtain $\|TF\|_{2,m} \geq K\|F\|_{2,m}$. Plainly, we can apply this argument to the mother wavelet $\Psi(y) = (1+iy)^{-3}$. We refer to Table 5.1 for estimates of the constant K for $\Psi(y)$ and for different b_0. Note that for $b_0 = 0.5$ the value of K is negative, and so we cannot deduce that T is bounded below.

As in section 4 it is now plain that $T : H^{2,m} \to H^{2,m}$ has closed range. Next we observe that $\{TF : F \in H^{2,m}\}$ is dense in $H^{2,m}$. Suppose that $G \in H^{2,-m}$ and that

$$\langle TF, G\rangle = \langle F, TG\rangle = 0, \quad F \in H^{2,m}.$$

If we take $F = \mathcal{L}f$ defined by $f(t) = g(t)(1+t)^{-2m}$, we deduce that

$$0 = \frac{\sup |\langle TF, G\rangle|}{\|G\|_{2,-m}} \geq K\|G\|_{2,-m},$$

and so $G = 0$. It follows from the Hahn–Banach theorem that the set $\{TF : F \in H^{2,m}\}$ is dense in $H^{2,m}$. Hence, since T has closed range in $H^{2,m}$, we deduce that range$(T) = H^{2,m}$. This completes the proof of Theorem 1.3.

Remark. The proof given here should be compared with the proof of Proposition 2.12 in [12], which unfortunately contains an error. However, we achieve similarly symmetric lower and upper bounds for $\|TF\|_{2,m}$, and as Daubechies points out in Remark 2, it follows by interpolation that T is bounded above and below on $H^{2,m'}$ for $-1 \leq m' \leq 1$. (See section IX.4, particularly Example 3 (rigged Hilbert spaces), in [41].)

5.1. Proof of Theorem 1.1: General case. To obtain Theorem 1.1 for $-1 \leq m \leq 1$ it is simplest to proceed (as a referee has observed) by interpolation. The upper bound

$$\sum_{j,k} |\langle F, \Psi_{j,k}\rangle|^2 (1+2^j)^{2m} \leq B_m \|F\|_{2,1}^2 \tag{5.3}$$

is obtained by applying a theorem of Stein (Theorem 3.6 in Chapter 4 of [4]) to the operator R defined by

$$RF = (\langle F, \Psi_{j,k}\rangle),$$

which, as we have seen, can be regarded as a bounded operator between weighted Hilbert spaces as follows:

$$R : L^2(0,\infty;\,(1+t)^{(1-2\theta)}dt) \to \ell^2((1+2^j)^{(1-2\theta)})$$

for $\theta = 0$ and 1. Hence, by Stein's theorem, it is bounded for all intermediate values of θ, and we can take $B_{(1-2\theta)} \le B_1^{1-\theta} B_{-1}^{\theta}$.

The lower bound is now most simply obtained by duality between $H^{2,m}$ and $H^{2,-m}$. Namely, by Theorem 1.3 we can find a constant $C_m > 0$ such that $\|TF\| \ge C_m\|F\|$ for each $F \in H^{2,m}$. Then there exists a normalized vector $G \in H^{2,-m}$ such that

$$|\,\langle TF, G\rangle\,| \ge C_m \|F\|_{2,m}.$$

Hence, by the Cauchy–Schwarz inequality,

$$\sum |\,\langle F, \Psi_{j,k}\rangle\,|^2 (1+2^j)^{2m} \sum |\,\langle G, \Psi_{j,k}\rangle\,|^2 (1+2^j)^{-2m} \ge C_m^2 \|F\|_{2,m}^2$$

and so

$$\sum |\,\langle F, \Psi_{j,k}\rangle\,|^2 (1+2^j)^{2m} \ge C_m^2 B_{-m}^{-2} \|F\|_{2,m}^2,$$

as required.

6. Wavelets for the half-plane algebra. An interesting class of wavelets is obtained when $p = 1$ in section 1.3. That is, $\Psi(y) = (1+iy)^{-1}$ is the Cauchy kernel. This wavelet does not fall under the Daubechies theory since is does not have vanishing mean value, but the system $\Psi_{j,k}$ does constitute a fundamental set for the half-plane algebra $A(\mathbb{C}_+)$, and so one can use the Cauchy kernel to obtain approximations in the uniform norm. In [16] it was shown using the Hayman–Lyons theory that certain sets of Cauchy kernels formed complete model sets in the disc algebra. By using inequalities due to Borwein and Erdélyi [8] it was shown how these model sets could be used for worst-case identification.

In this section we consider the half-plane analogue of the main results in [16]. Since they are deduced from [16] by conformal mapping the proofs will be abbreviated.

Let $\mathbb{D}$ be the unit disc and $A(\mathbb{D}) = \{f(z) : f(z)$ is analytic in $\mathbb{D}$ and continuous in $\overline{\mathbb{D}}\}$ be the disc algebra. Next let $Q_{j,k}$ be the Whitney cube partition of $\mathbb{D}$ defined for $j = 1, 2, \ldots$ and $k = 1, 2, \ldots, 2^j$ by

$$Q_{j,k} = \left\{ z : 1 - \frac{1}{2^j} \le |z| \le 1 - \frac{1}{2^{j+1}}, \frac{2k\pi}{2^j} \le \arg z \le \frac{2(k+1)\pi}{2^j} \right\}.$$

If $A \subset D$ we set $A_{j,k} = A \cap Q_{j,k}$ and $z_{j,k} = \left(1 - \frac{1}{2^j}\right) \exp \frac{2\pi i k}{2^j}$. Finally, we define $s(\theta)$ by

$$s(\theta) = s(\theta, A) = \sum_{A_{j,k} \ne \emptyset} \left(\frac{1 - |z_{j,k}|}{|z_{j,k} - e^{i\theta}|} \right)^2. \tag{6.1}$$

We say that A satisfies the *Hayman–Lyons condition* if, and only if, $s(\theta) = +\infty$ for all $\theta \in [0, 2\pi]$. In [16] the following result was established.

THEOREM D. *Suppose that $A \subset \mathbb{D}$ and that A satisfies the Hayman–Lyons condition. Then if $f \in A(\mathbb{D})$ and $\epsilon > 0$, there exists $\lambda_1, \dots, \lambda_n \in \mathbb{C}$ and $a_1, \dots, a_N \in A$ such that*

$$\left\| f(z) - \sum_{k=1}^{N} \lambda_k \frac{1}{1 - \overline{a}_k z} \right\|_\infty < \epsilon.$$

That is, the set $\mathcal{W} = \{C_w(z) : w \in A\}$, where $C_w(z) = (1 - \overline{w}z)^{-1}$ is a fundamental set for $A(\mathbb{D})$.

Our next result is a half-plane version of Theorem D. Let $A(\mathbb{C}_+)$ be the half-plane algebra for $\mathbb{C}_+$, that is, the set of $F(s)$ analytic in $\mathbb{C}_+$, continuous up to the imaginary axis and such that $\lim_{y \to \pm\infty} F(iy)$ exists. Plainly, the map $M_z = (1 - z)(1 + z)^{-1}$ gives an isometric isomorphism between $A(\mathbb{C}_+)$ and $A(\mathbb{D})$.

THEOREM 6.1. *Let $C'_w(s) = (\overline{w} + s)^{-1}$ be the Cauchy kernel for $\mathbb{C}_+$ and $\mathcal{L} \subset \mathbb{C}_+$ be the lattice*

$$\mathcal{L} = \{Z_{j,k} : \Re(Z_{j,k}) = 2^{-j}, \Im(Z_{j,k}) = k2^{-j}, j, k \in \mathbb{Z}\}.$$

Then the set $\mathcal{W} = \{1\} \cup \{C'_w(iy) : w \in \mathcal{L}\}$ is a fundamental set for $A(\mathbb{C})$. That is, given $F(s) \in A(\mathbb{C}_+)$ and $\epsilon > 0$ there exists $\lambda_1, \dots, \lambda_N, K \in \mathbb{C}$ and $w_1, \dots, w_N \in \mathcal{L}$ such that

$$\left\| F(s) - \sum_{k=1}^{N} \lambda_k \frac{1}{\overline{w}_k + s} - K \right\|_\infty < \epsilon.$$

We sketch the proof. We need to reformulate the Hayman–Lyons condition. Let $\tilde{Q}_{j,k}$ be the Whitney cube division of $\mathbb{C}_+$ defined by

$$\tilde{Q}_{j,k} = \left\{ s = x + iy : \frac{1}{2^j} \le x \le \frac{1}{2^{j-1}}, \frac{k}{2^j} \le y \le \frac{(k+1)}{2^j} \right\}.$$

For a set A in $\mathbb{C}_+$ let $A_{j,k} = A \cap Q_{j,k}$ and $Z_{j,k} = X_{j,k} + iY_{j,k} = 2^{-j} + ik2^{-j}$. It is easy to see by conformal transformation that the series in the Hayman–Lyons condition (6.1) takes the form

$$S(iy) = S(iy, A) = \sum_{A_{j,k} \neq \emptyset} \left(\frac{X_{j,k}}{|Z_{j,k} - iy|} \right)^2.$$

If iy is infinite $S(\infty)$ is defined by

$$S(\infty) = S(\infty, A) = \sum_{A_{j,k} \neq \emptyset} \left(\frac{X_{j,k}}{|Z_{j,k}|} \right)^2.$$

It is then straightforward to show that if $S(iy) = \infty$ for all y in $\overline{\mathbb{R}}$ where for each y we sum over all those cubes $Q_{j,k}$ which meet a disc with center iy, that $\{C'_w(iy) : w \in A\} \cup \{1\}$ is a fundamental set for $A(\mathbb{C}_+)$. Suppose that $G(s) \in A(\mathbb{C}_+)$. Then $G((1-z)(1+z)^{-1}) \in A(\mathbb{D})$. Let $\mathcal{A}$ be the inverse image of A by $(1-s)(1+s)^{-1}$. Then, since the hyperbolic metric is conformally invariant it follows that $s(e^{i\theta}, \mathcal{A}) = +\infty$

for $\theta \in [0, 2\pi]$, so that given $\epsilon > 0$, there exists $\lambda_1, \ldots, \lambda_N \in \mathbb{C}$ and $a_1, \ldots, a_N \in \mathcal{A}$ such that

$$\left\| G\left(\frac{1-z}{1+z}\right) - \sum_{k=1}^{N} \lambda_k \frac{1}{1-\overline{a}_k z} \right\|_\infty < \epsilon. \tag{6.2}$$

Suppose that $\overline{a}_k = (1-\overline{w}_k)(1+\overline{w}_k)^{-1}$ and $z = (1-s)(1+s)^{-1}$. Then

$$\begin{aligned} \frac{1}{1-\overline{a}_k z} &= \frac{1+s}{(1-\overline{a}_k) + s(1+\overline{a}_k)} \\ &= \frac{1+\overline{w}_k}{2}\left(1 + (1-\overline{w}_k) C'_{w_k}(s)\right). \end{aligned}$$

Hence (6.2) takes the form

$$\left\| G(s) - \sum_{k=1}^{N} \lambda'_k C'_{w_k}(s) - K \right\|_\infty < \epsilon.$$

Plainly, the lattice $\mathcal{L}$ satisfies the Hayman–Lyons condition for the right half-plane, and so for A in the above argument we may take the lattice $\mathcal{L}$ and deduce that the corresponding set of wavelets $\mathcal{W}$ is a fundamental set for $A(\mathbb{C}_+)$

Remark. The Hayman–Lyons condition can be formulated in a potential theoretic way: let $A \subset \mathbb{C}_+$ and define the sequence $A'_{j,k}$ to be $A_{j,k}$ if $A_{j,k} \neq \emptyset$ and $\emptyset$ otherwise. Let $Z_{j,k}$ be a point in $A'_{j,k}$ and $K(Z_{j,k}, \delta)$ be a hyperbolic ball with center $Z_{j,k}$. Then $S(iy_0, A) = \infty$ if the union of hyperbolic balls $K(Z_{j,k}, \delta)$ is not minimally thin at iy_0 (see [18] for more on this topic).

An interesting application of Theorem 6.1 is given to a problem in robust identification. Here one is given corrupted frequency response measurements $a_k = F(iy_k) + \eta_k$, $k = 1, \ldots, n$, on the imaginary axis of an unknown function $F(s)$ in the half-plane algebra $A(\mathbb{C}_+)$, where η_k is the measurement error (sometimes called *noise*, although it could for example be due to deterministic effects such as nonlinearities), which is assumed to be small (say, $|\eta_k| \leq \epsilon$) for each k. Then the intention is to produce an approximate model $\tilde{F} \in A(\mathbb{C}_+)$ for F in such a way that the following *robust convergence* condition is satisfied:

$$\lim_{\substack{n\to\infty \\ \epsilon\to 0}} \sup_{\|\eta\| \leq \epsilon} \|\tilde{F} - F\|_\infty = 0 \quad \text{for all } F \in A(\mathbb{C}_+). \tag{6.3}$$

An algorithm for reconstructing $F(s)$ is given by applying the following result [38].

THEOREM E. *Let X be a separable infinite-dimensional normed space over a field $\mathbb{F}$, and let $\{\phi_k\}_{k=1}^{\infty}$ be a uniformly bounded sequence of elements in the dual space X^*. Then there exist maps $T_n : \mathbb{F} \to X$ such that*

$$\lim_{\substack{n\to\infty \\ \epsilon\to 0}} \sup_{\|\eta\| \leq \epsilon} \|x_n - x\| = 0 \quad \textit{for all } x \in X,$$

with $x_n = T_n((\phi_k(x) + \eta_k)_{k=1}^n)$, if and only if there exists $\delta > 0$ with

$$\sup |\phi_k(x)| \geq \delta \|x\| \quad \textit{for all } x \in X.$$

In order to implement Theorem E we first note that for ϕ_k we can take evaluation functionals $\phi_k(F) = F(iy_k)$ and then we choose a complete model set $\{X_p\}_{p\geq 1}$ for $A(\mathbb{C}_+)$; that is, a sequence of finite-dimensional subspaces $\{X_p\}_{p\geq 1}$ of $A(\mathbb{C}_+)$ which satisfies the following two conditions:

(i) $X_1 \subset X_2 \subset X_3 \subset \cdots$,
(ii) $\cup_{p=1}^{\infty} X_p$ is dense in $A(\mathbb{C}_+)$.

The vectors $T_n(a_1, \ldots, a_n)$ are then constructed to be solutions $G_p \in X_p$ of the minimax problem:

$$s = \min_{G \in X_p} \max_{1 \le k \le n} |\phi_k(G) - a_k|,$$

where $a_k = F_k(x) + \eta_k$. Here p is chosen, depending on n, to be as large as possible such that

$$\max_{1 \le k \le n} |\phi_k(G)| \ge (\delta/2)\|G\|_\infty \quad \text{for all } G \in X_p.$$

Since this optimization problem can be solved by linear programming, the computational burden is not excessive.

We have just shown in Theorem 6.1 that $\{C_w(iy) : w \in \mathcal{L}\} \cup \{1\}$ is a complete model set for $A(\mathbb{C}_+)$, and we may for instance take X_p to be the finite-dimensional space spanned by the set

$$\{1\} \cup \{C'_w(iy) : s = 2^{-j} + ik2^{-j}, -p \le k \le p, 0 \le j \le p\}.$$

On transforming to the disc, this subspace corresponds to a subspace $\tilde{X}_p$ of the disc algebra spanned by Cauchy kernels $C_a(z) = 1/(1 - \bar{a}z)$, where the points a lie on a finite lattice $\mathcal{L}_p$, say. Now if $w = u + iv$ and $a = (1 - w)/(1 + w)$, then the following identity holds:

$$1 - \left|\frac{1-w}{1+w}\right|^2 = \frac{4|u|}{|1+w|^2},$$

and this implies that the points of $\mathcal{L}_p$ are not too close to the unit circle. Specifically, we easily obtain the uniform bound $1 - |a|^2 \ge (2/3)2^{-p}$, which implies that the poles of the functions in $\tilde{X}_p$ lie in the region $\{|z| \ge R\}$, where $(R+1)/(R-1) \le 6.2^p$.

We may now use the Borwein–Erdélyi extension of Bernstein's inequality which applies to rational functions with poles in a region $\{|z| \ge R\}$, namely,

$$\|g'\|_\infty \le N \frac{R+1}{R-1} \|g\|_\infty$$

where $\dim \tilde{X}_p = N$ [8], and perform a calculation similar to that of [16] to show that, provided that the interpolation points (iy_k), $k = 1, \ldots, n(p)$, on the imaginary axis are chosen so that the maximum gap Δ_n between their images $(1 - iy_k)/(1 + iy_k)$ on the unit circle is sufficiently small, one does indeed have

$$\max |G(iy_k)| \ge (1/2)\|G\|_\infty \quad \text{for all } G \in X_p.$$

In the example we are considering this condition is satisfied if $\Delta_n \le 1/(6N2^p)$ and $N = 2p^2 + 3p + 3$.

This means that the Cauchy kernels can be used as a complete model set for identification purposes along the lines of [38, 16].

7. Error estimates. There is extensive literature on the model reduction of infinite-dimensional linear systems, both in the H^∞ norm [22, 23, 24, 26] and the H^2 norm [25, 1]. Recent research has focused on two approaches which can be applied

to a wide class of systems. First, there is the use of orthogonal basis functions such as Laguerre and Kautz models [26, 31, 37, 27, 43] where the poles normally lie in a fixed finite set. Second, there is the approach based on rational wavelets [39, 40, 15] where the poles lie on some appropriate infinite lattice, and this is the approach we have adopted in this paper.

In order to illustrate the techniques involved, we shall consider the class of retarded delay systems in the sense of Bellman and Cooke [3]. These have perhaps the simplest irrational transfer functions that occur frequently and, unlike the transfer functions obtained from partial differential equations, tend to be difficult to approximate efficiently. Other examples, the discrete-time fractional filters, were considered by analogous methods in [15], and delay systems were considered by transforming them to the disc.

In this section we shall specialize to delay systems of the form $R(s)e^{-sT}$, $R(s)$ rational, because the results on approximation of such systems have a particularly simple form. However, general retarded delay systems (of the form $N(s)/D(s)$ where N and D are sums of terms of the form $p(s)e^{-Ts}$ with p a polynomial and $T > 0$) can be analyzed by similar means (the techniques for decomposing such systems can be found in [23, 25, 37]).

For a delay system of the form $G(s) = R(s)e^{-sT}$, where $R(s)$ is asymptotic to Cs^{-k} at infinity, the following approximation results are known [22, 23, 24, 25].

There exist constants A, $B > 0$ such that the minimal achievable error in H^∞ and H^2 rational approximation by degree-n systems satisfies

$$An^{-k} \le E_n^\infty(G) \le Bn^{-k}$$

and

$$An^{-k+1/2} \le E_n^2(G) \le Bn^{-k+1/2}.$$

This implies immediately that the $H^{2,1}$ error must satisfy

$$E_n^{2,1}(G) \ge An^{-k+1/2},$$

but it seems that a tight upper estimate is not known.

Now if we consider the integral operator Ω on $L^2(0,\infty)$ defined by

$$(\Omega u)(t) = \int_0^\infty w(t)h(t+\tau)w(\tau)u(\tau)\,d\tau, \tag{7.1}$$

where w is a suitable weight function, then its rank is the same as the rank of the usual Hankel operator and its Hilbert–Schmidt norm squared is

$$\int_{t=0}^\infty \int_{\tau=0}^\infty |w(t)|^2|h(t+\tau)|^2|w(\tau)|^2\,dt\,d\tau,$$

which equals

$$\int_{r=0}^\infty \int_{t=0}^r |w(t)|^2|h(r)|^2|w(r-t)|^2\,dt\,dr.$$

Choosing $w(t) = \sqrt{(2\sqrt{t} + b/\sqrt{t})}$ gives a Hilbert–Schmidt norm squared of

$$\int_0^\infty (r^2 + 4br + 2b^2)|h(r)|^2\pi/2\,dr,$$

since one has

$$\int_0^r (2\sqrt{t} + b/\sqrt{t})(2\sqrt{r-t} + b/\sqrt{r-t})\,dt = (r^2 + 4br + 2b^2)\pi/2.$$

We do not know a weight w such that

$$\int_0^r |w(t)|^2\,|w(r-t)|^2\,dt = (r+1)^2,$$

or even r^2+1, but by choosing $b = 1/\sqrt{2}$ we have the bounds

$$(r+1)^2 \le r^2 + 4br + 2b^2 \le (1/2 + 1/\sqrt{2})(r+1)^2$$

for $r \ge 0$, and so we have the fairly tight inequality

$$(\pi/2)\|h\|_{2,1}^2 \le \|\Omega\|_{HS}^2 \le ((\sqrt{2}+1)\pi/4)\|h\|_{2,1}^2,$$

with corresponding bounds for rational approximation errors in the $H^{2,1}$ norm. Namely, if the singular values of Ω are $(\omega_r)_{r\ge1}$, then we have the following error bound for $\|h - h_n\|_{2,1}$ when h_n is the impulse response of any degree-n system and corresponds to an operator Ω_n of rank n:

$$\begin{aligned}\|h - h_n\|_{2,1}^2 &\ge C\|\Omega - \Omega_n\|_{HS}^2 \\ &\ge C\sum_{r=n+1}^{\infty} \omega_r^2,\end{aligned}$$

where $C = 4(\sqrt{2}-1)/\pi$. Similar weighted Hankel integral operators were considered in [25] for the purposes of estimating error bounds in L^2 approximation, and in [11] for the purposes of estimating the singular values of Hankel operators on Bergman spaces. Here we have the additional complication that it does not seem possible to choose w in closed form so that the Hilbert–Schmidt norm of Ω and the $H^{2,1}$ norm of h are equal, although we can obtain equivalence of norms.

The following result generalizes both Theorem 3.2 of [25] and Lemma 14 of [11]. It applies directly to transfer functions of the form $(e^{-sT}-1)/s$ with $T > 0$, and, by standard decomposition techniques [37], to many other delay systems.

THEOREM 7.1. *Let $w(t)$ be a smooth positive function in $L^2(0,T)$ and let $h(t) = \chi_{[0,T]}(t)$. Then the singular values of the scaled Hankel operator Ω defined by* (7.1) *satisfy $r\omega_r \to J(w)$, where*

$$J(w) = \frac{1}{\pi}\int_0^T w(t)w(T-t)\,dt.$$

Proof. The proof here is a simple modification of the proof given in [25], so we shall omit many of the details. The operator Ω is self-adjoint and solving for its eigenvalues leads directly to the Sturm–Liouville differential equation

$$\frac{d}{dt}\left(\frac{v'(t)}{w(T-t)^2}\right) + \frac{v(t)w(t)^2}{\lambda^2} = 0,$$

with boundary conditions $v(T) = 0 = v'(0)$. Analysis of this equation by the Liouville–Green method [33, 35] shows that its eigenvalues are asymptotic to those

of the equation $z'' + z/\lambda^2$ over the interval $[0, X]$ with boundary conditions $z(0) = z(X) = 0$, where

$$X = \int_0^T w(t)w(T-t)\,dt,$$

and this gives the result. ∎

Unfortunately, for the weights w considered above, the integral $J(w)$ cannot be obtained in closed form, although it can be estimated numerically.

We conclude this section with a further result which applies particularly to delay systems, and can be regarded as an appendix to Theorem 1.1 above. There we considered the wavelet coefficients $\langle F, \Psi_{j,k}\rangle$ for $F \in H^{2,1}$ regarded as functions of j. It would clearly be useful to be able to estimate them as functions of k as well (so that an infinite series could be truncated to a finite sum by discarding the less significant terms) and this is what we now do.

LEMMA 7.2. *Suppose that $f(y)$, $g(y) \in L^2(\mathbb{R})$ satisfy $|f(y)| \le A|y|^{-m}$ and $|g(y)| \le B|y-C|^{-n}$, where m, $n > 1/2$ and $C > 0$. Then the following inequality holds:*

$$\langle f, g\rangle \le \frac{B\|f\|_2}{\sqrt{2n-1}(C/2)^{n-1/2}} + \frac{A\|g\|_2}{\sqrt{2m-1}(C/2)^{m-1/2}}.$$

Proof. Using the Cauchy–Schwarz inequality we have, for any r with $0 < r < C$,

$$\langle f, g\rangle \le \|f\|_2 \left(\int_{-\infty}^{r} |g(y)|^2\,dy\right)^{1/2} + \|g\|_2 \left(\int_{r}^{\infty} |f(y)|^2\,dy\right)^{1/2},$$

and this easily gives the result when we take $r = C/2$. ∎

The conditions of the following theorem will apply to any strictly proper delay system, and one can always take $m \ge 1$. Clearly, the result is of greatest use when $j < 0$, because then we obtain a good approximation with fewer translates $\Psi_{j,k}$.

THEOREM 7.3. *Let $\Psi(y) = (1+iy)^{-3}$ and let $\Psi_{j,k}(y) = 2^{j/2}\Psi(2^j y - kb_0)$ where $j, k \in \mathbb{Z}$, and let $\langle\cdot,\cdot\rangle$ be the usual L^2-inner product.*

Suppose that $|F(iy)| \le A|y|^{-m}$, where $m > 1/2$. Then for $k \ne 0$ the wavelet coefficients $\langle F, \Psi_{j,k}\rangle$ satisfy

$$\langle F, \Psi_{j,k}\rangle \le \frac{2^{5/2}\|f\|_2}{\sqrt{5}|kb_0|^{5/2}} + \frac{A\|\Psi\|_2}{\sqrt{2m-1}(|kb_0|2^{-j-1})^{m-1/2}}.$$

Proof. Clearly, we may assume without loss of generality that $k > 0$. We now apply Lemma 7.2, with $g = \Psi_{j,k}$, noting that $\|\Psi_{j,k}\|_2 = \|\Psi\|_2$ and that

$$|\Psi_{j,k}(y)| \le 2^{-5j/2}|y - kb_0 2^{-j}|^{-3}. \qquad ∎$$

8. Applications to system modeling. It is desired to approximate an infinite-dimensional model by a finite-dimensional model in the Hardy–Sobolev norm. That is, given a transfer function in $H^{2,1}$, we desire to approximate $F(s)$ by a degree n rational function. The framework developed in this paper gives a method by which we can approximate $F(s)$ by a sum of degree 3 rational functions.

In order to implement the techniques of Theorems 1.2 or 1.3 in the model reduction of a transfer function in the Hardy–Sobolev space $H^{2,1}$, we present three algorithms.

ALGORITHM 1. This is based on Theorem 1.2 and a matching pursuit (MP) algorithm of Mallat and Zhang [32]. With $\Psi(y)$ as in Theorem 1.2, it follows that the linear span of the collection $\mathcal{D} = \{\Psi_{j,k}(y) : j,k \in \mathbb{Z}\}$ is dense in $H^{2,1}$. Let $\langle \cdot, \cdot \rangle_{2,1}$ be the $H^{2,1}$-inner product. Take $0 < \alpha \leq 1$ and $F_1 \in H^{2,1}$. Choose $\Psi_{j_1,k_1}(y)$ such that

$$|\langle F_1, \Psi_{j_1,k_1} \rangle_{2,1}| \geq \alpha \sup_{\Psi_{j,k} \in \mathcal{D}} |\langle F_1, \Psi_{j,k} \rangle_{2,1}|.$$

There exists F_2 in the orthogonal complement of the subspace spanned by Ψ_{j_1,k_1} with $F_1 = \langle F_1, \Psi_{j_1,k_1} \rangle_{2,1} \Psi_{j_1,k_1} + F_2$. Next we choose $\Psi_{j_2,k_2} \in \mathcal{D}$ such that

$$|\langle F_2, \Psi_{j_2,k_2} \rangle_{2,1}| \geq \alpha \sup_{\Psi_{j,k} \in \mathcal{D}} |\langle F_2, \Psi_{j,k} \rangle_{2,1}|.$$

There exists F_3 in the orthogonal complement of the subspace spanned by Ψ_{j_2,k_2} with $F_2 = \langle F_2, \Psi_{j_2,k_2} \rangle_{2,1} \Psi_{j_2,k_2} + F_3$. Proceeding in this manner we obtain a sequence $\{\Psi_{j_n,k_n}\}_{n=1}^{\infty}$ and a sequence of *residual vectors* $\{F_n\}_{n=1}^{\infty}$ such that

$$F_n = \langle F_n, \Psi_{j_n,k_n} \rangle_{2,1} \Psi_{j_n,k_n} + F_{n+1}$$

and

$$\|F_n\|_{2,1}^2 = |\langle F_n, \Psi_{j_n,k_n} \rangle_{2,1}|^2 + \|F_{n+1}\|_{2,1}^2.$$

Therefore,

$$\begin{aligned} F_1 &= \sum_{n=1}^{m-1} (F_n - F_{n+1}) + F_m \\ &= \sum_{n=1}^{m-1} \langle F_n, \Psi_{j_n,k_n} \rangle_{2,1} \Psi_{j_n,k_n} + F_m, \end{aligned}$$

and

$$\begin{aligned} \|F_1\|_{2,1}^2 &= \sum_{n=1}^{m-1} (\|F_n\|_{2,1}^2 - \|F_{n+1}\|_{2,1}^2) + \|F_m\|_{2,1}^2 \\ &= \sum_{n=1}^{m-1} |\langle F_n, \Psi_{j_n,k_n} \rangle_{2,1}|^2 + \|F_m\|_{2,1}^2. \end{aligned}$$

Mallat and Zhang (Theorem 1, [32]) prove that $\|F_m\|_{2,1} \to 0$, and so

$$F_1 = \sum_{n=1}^{\infty} \langle F_m, \Psi_{j_n,k_n} \rangle_{2,1} \Psi_{j_n,k_n}$$

and $\|F_1\|_{2,1}^2 = \sum_{n=1}^{\infty} |\langle F_n, \Psi_{j_n,k_n} \rangle_{2,1}|^2$.

ALGORITHM 2. This is an adaptation of Algorithm 1. MP may converge somewhat slowly (see [15, section 4.2]), but it is possible to adapt it so that we obtain geometric convergence in the $H^{2,1}$ norm by projecting onto larger subspaces. For example, let $V_n = \text{span}\{\Psi_{j,k} : -n \leq j \leq n, -2^n \leq k \leq 2^n\}$ and let P_{V_n} be the projection onto V_n in the L^2 norm. Let $\langle \cdot, \cdot \rangle$ denote the L^2-inner product.

To proceed with the algorithm fix a positive ϵ and $F_1 \in H^{2,1}$. By Theorem 1.1 we may choose n_1 so large that

$$(A_1 - \epsilon)\|F_1\|_{2,1}^2 \leq \sum_{|k|,|j|\leq n_1} |\langle F_1, \Psi_{j,k}\rangle|^2(1+2^j)^2. \tag{8.1}$$

The vector $F_2 = F_1 - P_{V_{n_1}}F_1$ is orthogonal to V_{n_1} in L^2 and so $\langle F_2, \Psi_{j,k}\rangle = 0$. Therefore $\langle F_1, \Psi_{j,k}\rangle = \langle P_{V_{n_1}}F_1, \Psi_{j,k}\rangle$, $|j|, |k| \leq n_1$. Therefore, by (8.1) and Theorem 1.1, we have

$$(A_1 - \epsilon)\|F_1\|_{2,1}^2 \leq \sum_{|k|,|j|\leq n_1} |\langle P_{V_{n_1}}F_1, \Psi_{j,k}\rangle|^2(1+2^j)^2 \leq B_1\|P_{V_{n_1}}F_1\|_{2,1}^2.$$

Thus $B_1\|P_{V_{n_1}}F_1\|_{2,1}^2 \geq (A_1-\epsilon)\|F_1\|_{2,1}^2$. Increasing n_1 if necessary, we obtain similarly that $B_1\|P_{V_{n_1}}F_1\|_{2,1}^2 \geq (A_1 - \epsilon)\|P_{V_{n_1}}F_1 - F_2\|_{2,1}^2$. By the parallelogram law we have

$$\|F_2\|_{2,1}^2 + \|P_{V_{n_1}}F_1\|_{2,1}^2 = \frac{1}{2}(\|F_1\|_{2,1}^2 + \|P_{V_{n_1}}F_1 - F_2\|_{2,1}^2).$$

Therefore

$$\begin{aligned}\|F_2\|_{2,1}^2 &\leq \left(\frac{1}{2} - \left(1 - \frac{B_1}{2(A_1-\epsilon)}\right)\frac{(A_1-\epsilon)}{B_1}\right)\|F_1\|_{2,1}^2 \\ &= \left(1 - \frac{(A_1-\epsilon)}{B_1}\right)\|F_1\|_{2,1}^2. \end{aligned} \tag{8.2}$$

With the same ϵ and F_1 replaced by F_2 we repeat the procedure. Continuing in this manner we obtain a sequence $\{F_k\}_{k\geq 1}$ such that $F_k = P_{V_{n_k}}F_k + F_{k+1}$, where the $P_{V_{n_k}}F_k$ and F_{k+1} are orthogonal in the L^2 norm and such that

$$\|F_{k+1}\|_{2,1}^2 \leq \left(1 - \frac{A_1-\epsilon}{B_1}\right)^k \|F_1\|_{2,1}^2.$$

In (8.2) we have assumed b_0 is chosen so that $B_1 < 2A_1$. We may take, for example, $b_0 = 0.5$ and estimate by computer $A_1 = 4.503$ and $B_1 = 6.789$.

Remark. It is perhaps surprising that the algorithm given is based on taking projections in the L^2 norm and obtaining an expansion which converges in the Hardy–Sobolev norm.

ALGORITHM 3. This is based on ideas in Theorem 1.3. Since the operator TF is not a positive operator on $H^{2,1}$ it does not follow immediately that we can choose some constant K such that $\|I - KT\|_{2,1} < 1$. However, using a simple manipulation of Daubechies' techniques one may establish the following elementary result.

PROPOSITION 8.1. *Let $\Psi_{j,k}$ be as in Theorem* 1.3*, and*

$$RF = F - K\sum_{j,k}\langle F, \Psi_{j,k}\rangle\,\Psi_{j,k}, \qquad F \in H^{2,1}.$$

Define

$$\begin{aligned} m &= \inf_{0<t<\infty}\sum_j |\psi(2^{-j}t)|^2, \quad M = \sup_{0<t<\infty}\sum_j |\psi(2^{-j}t)|^2, \\ \beta_1(s) &= \sup_{0<t<\infty}(1+t)^{-2}\sum_j |\psi(2^{-j}t)||\psi(2^{-j}t - s)|, \\ \beta_2(s) &= \sup_{0<t<\infty}(1+t)^{2}\sum_j |\psi(2^{-j}t)||\psi(2^{-j}t + s)|. \end{aligned}$$

If $K = b_0/\pi(m+M)$, *then*

$$\|RF\|_{2,1} \le \frac{M - m + 2\sum_{k\neq 0}\left\{\beta_1\left(\frac{2\pi}{b_0}k\right)\beta_2\left(\frac{2\pi}{b_0}k\right)\right\}^{1/2}}{m+M}\|F\|_{2,1}.$$

Proof. The proof follows the lines of the proof of Theorem 1.3, and so the details are only sketched. We let G belong to $H^{2,-1}$ and estimate the modulus of

$$\begin{aligned}\langle RF, G\rangle &= \langle Rf, g\rangle \\ &= \langle f, g\rangle - K\sum_{j,k}\langle f, \psi_{j,k}\rangle\,\overline{\langle g, \psi_{j,k}\rangle} \\ &= \int_0^\infty f(t)\overline{g}(t)dt - \frac{2\pi K}{b_0}\left\{\sum_j\int_0^\infty f(t)\overline{g}(t)|\psi(2^{-j}t)|^2dt + \operatorname{rest}(f,g)\right\} \\ &= \int_0^\infty f(t)\overline{g}(t)\left\{1 - \frac{2\pi K}{b_0}\sum_j|\psi(2^{-j}t)|^2\right\}dt + \operatorname{rest}(f,g). \qquad (8.3)\end{aligned}$$

Analyzing the rest term we obtain with $\beta_1(s)$ and $\beta_2(s)$ defined above that

$$|\operatorname{rest}(f,g)| \le \frac{2\pi}{b_0}\sum_{k\neq 0}\left\{\beta_1\left(\frac{2\pi}{b_0}k\right)\beta_2\left(\frac{2\pi}{b_0}k\right)\right\}^{1/2}\|f\|_{2,1}\|g\|_{2,-1}. \qquad (8.4)$$

Therefore, substituting (8.4) into (8.3), we deduce

$$\|RF\|_{2,1} \le \left\{1 - \frac{2\pi K}{b_0}\left(m - \sum_{k\neq 0}\left\{\beta_1\left(\frac{2\pi}{b_0}k\right)\beta_2\left(\frac{2\pi}{b_0}k\right)\right\}^{1/2}\right)\right\}\|F\|_{2,1}.$$

With $K = b_0/2\pi(m+M)$ we obtain the conclusion of the proposition.

In a forthcoming paper it is intended to present some numerical results on the model reduction of several systems using the methods which appear in this study.

Acknowledgments. In conclusion, we would like to express our gratitude to the following friends and colleagues who have helped us in conversations and correspondence to understand frames and wavelets: Professors Ingrid Daubechies, Maurice Dodson, Walter Hayman F.R.S., and E. Christopher Lance. It will be plain to the informed reader the debt we owe to the work of Daubechies. We are also very grateful to one of the referees for reading a manuscript very carefully and making some pertinent comments. Finally, we would like to acknowledge a special debt to Ingrid Dudley Ward for writing some programs in C to estimate the frame constants given in this paper.

REFERENCES

[1] L. Baratchart, M. Olivi, and F. Wielonsky, *Asymptotic properties of rational L_2 approximation*, in Lecture Notes in Control and Inform. Sci. 144, Springer-Verlag, New York, 1990, pp. 477–486.

[2] L. Baratchart and M. Zerner, *On the recovery of functions from pointwise boundary values in a Hardy-Sobolev class of the disc*, J. Comput. Appl. Math., 46 (1993), pp. 255–269.

[3] R. Bellman and K. L. Cooke, *Differential-Difference Equations*, Academic Press, New York, 1963.
[4] C. Bennett and R. Sharpley, *Interpolation of Operators*, Academic Press, New York, 1988.
[5] D. S. Bernstein and W. M. Haddad, *LQG control with an H_∞ performance bound*, IEEE Trans. Automat. Control, 34 (1989), pp. 293–305.
[6] F. F. Bonsall, *Decompositions of functions as sums of elementary functions*, Quart. J. Math. Oxford Ser. (2), 37 (1986), pp. 129–136.
[7] F. F. Bonsall, *A general atomic decomposition theorem and Banach's closed range theorem*, Quart. J. Math. Oxford Ser. (2), 42 (1991), pp. 9–14.
[8] P. Borwein and T. Erdélyi, *Polynomials and Polynomial Inequalities*, Springer-Verlag, New York, 1995.
[9] R. R. Coifman and R. Rochberg, *Representation theorems for holomorphic and harmonic functions in L^p*, Asterisque, 77 (1980), pp. 12–66.
[10] M. A. Dahleh and J. B. Pearson, *Optimal rejection of persistent disturbances, robust stability, and mixed sensitivity minimization*, IEEE Trans. Automat. Control, 33 (1988), pp. 722–731.
[11] N. Das and J. R. Partington, *Little Hankel operators on the half plane*, Integral Equations Operator Theory, 20 (1994), pp. 306–324.
[12] I. Daubechies, *The wavelet transform, time-frequency localisation and signal analysis*, IEEE Trans. Inform. Theory, 36 (1990), pp. 961–1005.
[13] I. Daubechies, *Ten Lectures on Wavelets*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 61, SIAM, Philadelphia, PA, 1992.
[14] N. F. Dudley Ward, *Atomic Decompositions of Integrable or Continuous Functions*, Ph.D. thesis, University of York, UK, 1991.
[15] N. F. Dudley Ward and J. R. Partington, *Rational wavelet decompositions of transfer functions in Hardy-Sobolev classes*, Math. Controls Signals Systems, 8 (1996), pp. 257–278.
[16] N. F. Dudley Ward and J. R. Partington, *Robust identification in the disc algebra using rational wavelets and orthonormal basis functions*, Internat. J. Control, 64 (1996), pp. 409–423.
[17] R. J. Duffin and A. C. Schaeffer, *A class of nonharmonic Fourier series*, Trans. Amer. Math. Soc., 72 (1952), pp. 341–366.
[18] M. Éssen, H. L. Jackson, and P. J. Rippon, *On minimally thin and rarefied sets in R^p, $p \geq 2$*, Hiroshima Math J., 15 (1985), pp. 393–410.
[19] C. Foias, A. Frazho, and A. Tannenbaum, *On combined H^∞-H^2 suboptimal interpolants*, Linear Algebra Appl., 204 (1994), pp. 443–469.
[20] B. A. Francis, *A Course in H_∞ Control Theory*, Springer-Verlag, New York, 1987.
[21] F. G. Friedlander, *Introduction to the Theory of Distributions*, Cambridge University Press, Cambridge, UK, 1982.
[22] K. Glover, R. F. Curtain, and J. R. Partington, *Realization and approximation of linear infinite-dimensional systems with error bounds*, SIAM J. Control Optim., 26 (1988), pp. 863–898.
[23] K. Glover, J. Lam, and J. R. Partington, *Rational approximation of a class of infinite-dimensional systems* I*: Singular values of Hankel operators*, Math. Control Signals Systems, 3 (1990), pp. 325–344.
[24] K. Glover, J. Lam, and J. R. Partington, *Rational approximation of a class of infinite-dimensional systems* II*: Optimal convergence rates for L_∞ approximants*, Math. Control Signals Systems, 4 (1991), pp. 233–246.
[25] K. Glover, J. Lam, and J. R. Partington, *Rational approximation of a class of infinite-dimensional systems: The L_2 case*, in Progress in Approximation Theory, P. Nevai and A. Pinkus, eds., Academic Press, New York, 1991.
[26] G. Gu, P. P. Khargonekar, and E. B. Lee, *Approximation of infinite dimensional systems*, IEEE Trans. Automat. Control, 34 (1989), pp. 610–618.
[27] P. Heuberger, P. Van den Hof, and O. Bosgra, *Modelling linear dynamical systems through generalized orthonormal basis functions*, in Proc. 12th IFAC World Congress, Vol. 5, Sydney, 1993, pp. 283–286.
[28] K. Hoffman, *Banach Spaces of Analytic Functions*, Prentice-Hall, Englewood Cliffs, NJ, 1962.
[29] D. H. Luecking, *Forward and reverse Carleson inequalities for functions in Bergman spaces and their derivatives*, Amer. J. Math., 107 (1985), pp. 85–111.
[30] D. H. Luecking, *Representation and duality in weighted spaces of analytic functions*, Indiana Univ. Math. J., 34 (1985), pp. 319–336.
[31] P. M. Mäkilä, *Laguerre series approximation of infinite dimensional systems*, Automatica, 26 (1990), pp. 985–995.

[32] S. Mallat and Z. Zhang, *Matching pursuit with time-frequency dictionaries*, IEEE Trans. Signal Process., 41 (1993), pp. 3397–3415.
[33] H. Margenau and G. M. Murphy, *The Mathematics of Physics and Chemistry*, Van Nostrand, Princeton, NJ, 1964.
[34] Y. Meyer, *Wavelets and Operators*, Cambridge Stud. Adv. Math. 37, Cambridge University Press, Cambridge, UK, 1992.
[35] F. W. J. Olver, *Asymptotics and Special Functions*, Academic Press, New York, 1974.
[36] J. R. Partington, *An Introduction to Hankel Operators*, Cambridge University Press, Cambridge, UK, 1989.
[37] J. R. Partington, *Approximation of delay systems by Fourier-Laguerre series*, Automatica, 27 (1991), pp. 569–572.
[38] J. R. Partington, *Interpolation in normed spaces from the values of linear functionals*, Bull. Lond. Math. Soc., 26 (1994), pp. 165–170.
[39] Y. C. Pati, *Wavelets and Time-Frequency Methods in Linear Systems and Neural Networks*, Ph.D. thesis, University of Maryland, College Park, MD, 1992.
[40] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, *Orthogonal matching pursuit; recursive function approximation with applications to wavelet decompositions*, in Proc. 27th Annual Asilomar Conference on Signals Systems and Computers, 1993.
[41] M. Reed and B. Simon, *Methods of Modern Mathematical Physics* II*: Fourier Analysis, Self-Adjointness*, Academic Press, New York, 1975.
[42] R. Rochberg, *Decomposition Theorems for Bergman Spaces and Their Applications*, in Operators and Function Theory, S. Power, ed., D. Reidel, Dordrecht, the Netherlands, 1985.
[43] B. Wahlberg and P. M. Mäkilä, *On approximations of stable linear dynamical systems using Laguerre and Kautz functions*, Automatica, 32 (1996), pp. 693–708.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 2, pp. 680–697, March 1998

010

SLIDING MODES IN SOLVING CONVEX PROGRAMMING PROBLEMS*

MICHAEL P. GLAZOS†, STEFEN HUI‡, AND STANISLAW H. ŻAK§

Abstract. Sliding modes are used to analyze a class of dynamical systems that solve convex programming problems. The analysis is carried out using concepts from the theory of differential equations with discontinuous right-hand sides and Lyapunov stability theory. It is shown that the equilibrium points of the system coincide with the minimizers of the convex programming problem, and that irrespective of the initial state of the system the state trajectory converges to the solution set of the problem. The dynamic behavior of the systems is illustrated by two numerical examples.

Key words. sliding modes, differential inclusions, convex programming, stability, continuous algorithms, gradient system

AMS subject classifications. 93C15, 93D05, 90C25, 34D30

PII. S0363012993255880

1. Introduction.

Most of the traditional methods for solving constrained optimization problems are iterative algorithms [17]. However, over the past two decades considerable effort has been given to developing continuous-time methods for solving constrained optimization problems. The impetus for much of the early development in this area was a desire to solve constrained optimization problems using an electronic analog computer. Perhaps the first to develop a continuous-time algorithm was Pyne [18], who in 1956 proposed a method for solving linear programming problems using an electronic analog computer. Soon after, other methods [12, 21, 22] were proposed for solving various mathematical programming problems on an electronic analog computer. More recently, a class of analog systems known as artificial neural networks have been used to solve certain constrained optimization problems; see, for example, [3, 4, 13, 15, 16, 19, 23]. Many of these networks are suitable for monolithic implementation and are thus well suited for applications that require on-line optimization.

An approach commonly used in developing analog optimizers is to first convert the constrained optimization problem into an associated unconstrained optimization problem, and then design an analog network that solves the unconstrained problem. Such a network is typically an implementation of the dynamic gradient system for minimizing the objective function of the unconstrained problem. (See [6, 9, 8, 11, 14, 5] and references therein for other continuous-time methods for solving constrained optimization problems.)

One method for converting a constrained optimization problem into an unconstrained optimization problem is the penalty function method [17]. The idea behind the penalty function method is to replace the constrained optimization problem

$$\begin{array}{ll} \text{minimize} & f(\boldsymbol{x}) \\ \text{subject to} & \boldsymbol{x} \in \Omega \end{array}$$

*Received by the editors September 20, 1993; accepted for publication (in revised form) January 18, 1997.
http://www.siam.org/journals/sicon/36-2/25588.html
†Department of Electrical Engineering, Rochester Institute of Technology, Rochester, NY 14623 (mpgeee@rit.edu).
‡Department of Mathematical Sciences, San Diego State University, San Diego, CA 92182.
§School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907.

with an unconstrained problem of the form

$$\text{minimize} \quad f(\boldsymbol{x}) + \zeta p(\boldsymbol{x}),$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is continuous on $\mathbb{R}^n$, $\Omega \subset \mathbb{R}^n$, ζ is a positive constant, and $p : \mathbb{R}^n \to \mathbb{R}$ is continuous on $\mathbb{R}^n$ and satisfies

$$\text{(i)} \quad p(\boldsymbol{x}) \geq 0 \quad \text{for all} \quad \boldsymbol{x} \in \mathbb{R}^n$$

and

$$\text{(ii)} \quad p(\boldsymbol{x}) = 0 \quad \text{if and only if} \quad \boldsymbol{x} \in \Omega.$$

If for a finite value of the penalty parameter ζ the solution of the unconstrained problem coincides with the solution of the constrained problem, then the penalty function p is said to be exact. Bertsekas [2] has shown that except for trivial cases an exact penalty function must not be everywhere differentiable. Quite often we can find an exact penalty function that results in the function $f(\boldsymbol{x}) + \zeta p(\boldsymbol{x})$ being piecewise smooth. However, the dynamic gradient system for such a function will have a discontinuous right-hand side. As such, an analysis of its behavior cannot be carried out using only methods derived from the classical theory of differential equations. Rather, the analysis of the system must be carried out using other methods, for example, those reported in [7, 1]. Such an approach was taken by Karpinskaya [12], and more recently by Utkin [25] and Żak et al. [27], among others.

Rodríguez-Vázquez et al. [19] proposed a class of neural networks for optimization problems whose design is based on concepts from the penalty function method. This class of optimizers is particularly attractive for two reasons. First, their design does not require the calculation of a penalty parameter. Second, these networks can be realized using switched-capacitor technology, and thus are suitable for monolithic implementation. In light of these properties we feel that a rigorous analysis of the dynamics of these systems is in order. Since such an analysis is not provided in [19], we do so here.

In this paper we perform an analysis of the dynamics of the analog networks proposed in [19] when applied to a broad class of convex programming problems. The paper is organized as follows. The statement of the problem is given in section 2. The main results of the paper are presented in section 3. In section 4 we illustrate the dynamic behavior of the analyzed networks by presenting the results of two computer simulations. Concluding remarks are offered in section 5.

2. Problem statement. We consider the constrained optimization problem

$$\begin{array}{ll} \text{minimize} & f(\boldsymbol{x}) \\ \text{subject to} & g_i(\boldsymbol{x}) \leq 0, \quad i = 1, 2, \ldots, m, \end{array} \tag{1}$$

where $\boldsymbol{x} \in \mathbb{R}^n$, $f : \mathbb{R}^n \to \mathbb{R}$, and $g_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, 2, \ldots, m$. Before proceeding further we first introduce some notation. We let Ω denote the feasible region for problem (1); that is,

$$\Omega = \bigcap_{i=1}^{m} \{\boldsymbol{x} : g_i(\boldsymbol{x}) \leq 0\}.$$

The collection of all interior points of Ω is denoted by Ω^o, and the boundary points of Ω are denoted by $\partial\Omega$. Also, we define

$$\Delta_i = \{\boldsymbol{x} : g_i(\boldsymbol{x}) = 0\}, \quad i = 1, 2, \ldots, m,$$

and

$$\Delta = \bigcup_{i=1}^{m} \Delta_i.$$

We let Γ denote the collection of all relative minimizers of problem (1). Lastly, we introduce the index sets

$$I(\boldsymbol{x}) = \{i : g_i(\boldsymbol{x}) = 0\}$$

and

$$J(\boldsymbol{x}) = \{i : g_i(\boldsymbol{x}) > 0\}.$$

We assume the following.

A1. The sets $\{\boldsymbol{x} : g_i(\boldsymbol{x}) > 0\}$, $i = 1, 2, \ldots, m$, are all nonempty.

A2. The functions f and g_i, $i = 1, 2, \ldots, m$, are convex over $\mathbb{R}^n$ and have continuous first partial derivatives on $\mathbb{R}^n$; that is, $f \in \mathrm{C}^1$ and $g_i \in \mathrm{C}^1$, $i = 1, 2, \ldots, m$.

A3. The set Ω is nonempty and bounded.

A4. The constraints are everywhere regular; that is, the vectors $\nabla g_i(\boldsymbol{x})$, $i \in I(\boldsymbol{x})$, are linearly independent for any $\boldsymbol{x} \in \mathbb{R}^n$.

Remark 1. Note that there is no loss of generality by assuming A1. Indeed, when treating a constrained optimization problem we can simply ignore any constraints that are satisfied everywhere.

Remark 2. It follows from A2 that problem (1) is a convex programming problem. Specifically, by A3, we are asked to minimize a convex function f over a compact convex set Ω.

Remark 3. It follows from A1–A4 that

1. the sets $\{\boldsymbol{x} : g_i(\boldsymbol{x}) < 0\}$, $i = 1, 2, \ldots, m$, are all nonempty;
2. the points in Δ_i define a smooth $(n-1)$-dimensional surface in $\mathbb{R}^n$ that separates the regions $\{\boldsymbol{x} : g_i(\boldsymbol{x}) > 0\}$ and $\{\boldsymbol{x} : g_i(\boldsymbol{x}) < 0\}$, $i = 1, 2, \ldots, m$;
3. $\Omega^o = \bigcap_{i=1}^{m} \{\boldsymbol{x} : g_i(\boldsymbol{x}) < 0\}$, $\partial\Omega = \Omega \cap \Delta$, and both Ω^o and $\partial\Omega$ are nonempty.

The class of analog networks proposed by Rodríguez-Vázquez et al. [19] for solving problem (1) is modeled by

$$\begin{aligned} \tau\dot{\boldsymbol{x}}(t) &= \boldsymbol{h}(\boldsymbol{x}(t)) \\ &= \begin{cases} -\mu \sum_{i \in J(\boldsymbol{x}(t))} \nabla g_i(\boldsymbol{x}(t)) & \text{if} \quad \boldsymbol{x}(t) \notin \Omega, \\ -\nabla f(\boldsymbol{x}(t)) & \text{if} \quad \boldsymbol{x}(t) \in \Omega, \end{cases} \\ \boldsymbol{x}(t_0) &= \boldsymbol{x}_0, \end{aligned} \tag{2}$$

where τ and μ are positive design constants. Observe that the function $\boldsymbol{h}$ is piecewise continuous over $\mathbb{R}^n$ and, in general, discontinuous on the surfaces $\Delta_1, \Delta_2, \ldots, \Delta_m$. As such, our analysis must use methods from the theory of differential equations with discontinuous right-hand sides. Here, as in [7], we take a solution of (2) to be a solution of the differential inclusion

$$\tau\dot{\boldsymbol{x}}(t) \in H(\boldsymbol{x}(t)), \tag{3}$$

where for each $\boldsymbol{x}$, $H(\boldsymbol{x})$ is the smallest closed convex set containing the cluster values of the function $\boldsymbol{h}(\boldsymbol{y})$ as $\boldsymbol{y} \to \boldsymbol{x}$, $\boldsymbol{y} \notin \Delta$. That is, a solution of (2) is an absolutely continuous function $\boldsymbol{x}(t)$ defined on an interval or segment L for which $\tau\dot{\boldsymbol{x}}(t) \in H(\boldsymbol{x}(t))$

almost everywhere on L [7]. Observe that if the function $\boldsymbol{h}$ is continuous at a point $\boldsymbol{x}$, that is, $\boldsymbol{x} \notin \Delta$, then $H(\boldsymbol{x})$ consists of a single point, namely $\boldsymbol{h}(\boldsymbol{x})$, and the solution satisfies (2) in the usual sense. However, if $\boldsymbol{x} \in \Delta$, then $\boldsymbol{x}$ lies on one or more of the surfaces $\Delta_1, \Delta_2, \ldots, \Delta_m$. Denoting the elements of $I(\boldsymbol{x})$ by $i_1, i_2, \ldots, i_k$ it follows from A4 that for sufficiently small $\delta > 0$ the surfaces $\Delta_{i_1}, \Delta_{i_2}, \ldots, \Delta_{i_k}$ partition the δ-neighborhood of $\boldsymbol{x}$ into 2^k regions, $\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_{2^k}$, in each of which the function $\boldsymbol{h}$ is continuous. Let $\boldsymbol{h}_j$ denote the function $\boldsymbol{h}$ restricted to the region $\mathcal{R}_j$. Then, $H(\boldsymbol{x})$ is the smallest closed convex set containing the set

$$\bigcup_{j=1}^{2^k} \bigcap_{n=1}^{\infty} \boldsymbol{h}_j \left(\left\{ \boldsymbol{y} \in \mathcal{R}_j : \|\boldsymbol{y} - \boldsymbol{x}\| \leq \frac{1}{n} \right\} \right).$$

We note that a rigorous justification for using (3) to analyze the behavior of system (2), as well as theorems guaranteeing the existence of solutions of system (2), is provided in [7, Chapter 2]. We also point out that differential inclusions are also used in continuous-time algorithms for solving nonsmooth convex programming problems; see, for example, [9, 8].

Our goal is to show that the solution set, Γ, of the optimization problem (1) is precisely the set of equilibrium points of system (2), and that all trajectories of system (2) converge to Γ, where convergence is understood in the sense of the following definition.

DEFINITION 1. *A trajectory $\boldsymbol{x} : [t_0, \infty) \to \mathbb{R}^n$ is said to converge to the solution set, Γ, if*

$$\lim_{t \to \infty} d(\boldsymbol{x}(t), \Gamma) = 0,$$

where

$$d(\boldsymbol{x}, \Gamma) = \inf_{\boldsymbol{y} \in \Gamma} \|\boldsymbol{x} - \boldsymbol{y}\|_2.$$

Recalling the well-known result that any relative minimizer of a convex programming problem is a global minimizer, it then follows from A2 and A3 that Γ is both closed and bounded. Therefore, the infimum in Definition 1 is achieved; that is,

$$d(\boldsymbol{x}, \Gamma) = \min_{\boldsymbol{y} \in \Gamma} \|\boldsymbol{x} - \boldsymbol{y}\|_2.$$

3. Main results. In this section we show that the equilibrium points of system (2) coincide with the minimizers of problem (1), and that all trajectories of system (2) converge to the solution set, Γ, of problem (1). Before proceeding further, we make a remark concerning the analysis of system (2). A phenomenon commonly occurring in systems such as (2) is the so-called sliding mode, where the motion of the system is confined to one or more of the discontinuity surfaces—see [14, 24, 25] for accounts of sliding modes and their applications in control and optimization. Consequently, we must consider two cases when analyzing the dynamic behavior of system (2). The first case is when the motion of the system is not confined to any of the surfaces $\Delta_1, \Delta_2, \ldots, \Delta_m$. The second case is when the system is in a sliding mode on one or more of the surfaces $\Delta_1, \Delta_2, \ldots, \Delta_m$. We only need to consider these two cases, for every trajectory of system (2) is composed of these two types of motion. Namely, every trajectory $\boldsymbol{x}(t)$ of system (2) can be broken up over a countable number of intervals, $(t_0, t_1), (t_1, t_2), (t_2, t_3), \ldots$, on each of which both index sets $I(\boldsymbol{x}(t))$ and

$J(\boldsymbol{x}(t))$ are constant. If $I(\boldsymbol{x}(t))$ is not empty, then the system is in a sliding mode on that interval, and if $I(\boldsymbol{x}(t))$ is empty, then the system is not in a sliding mode.

We begin our analysis by introducing the function $V : \mathbb{R}^n \to \mathbb{R}$ as

$$V(\boldsymbol{x}) = \sum_{i=1}^{m} \max\{g_i(\boldsymbol{x}), 0\},$$

or equivalently,

$$V(\boldsymbol{x}) = \begin{cases} \sum_{i \in J(\boldsymbol{x})} g_i(\boldsymbol{x}) & \text{if} \quad \boldsymbol{x} \notin \Omega, \\ 0 & \text{if} \quad \boldsymbol{x} \in \Omega. \end{cases} \tag{4}$$

Remark 4. It follows from A2 that the function V is continuous and convex on $\mathbb{R}^n$. Also, by definition, $V(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \notin \Omega$, and $V(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \in \Omega$.

We first show that every trajectory of system (2) reaches the feasible region Ω in finite time, and is thereafter confined to Ω. To prove the claim we need the following technical result.

LEMMA 1. *V is a decreasing function of time when evaluated on any trajectory of system* (2).

Proof. Let $\boldsymbol{x} : [t_0, \infty) \to \mathbb{R}^n$ be any particular trajectory of system (2). It follows directly from Remark 4 and the fact that $\boldsymbol{x}(t)$ is absolutely continuous that to prove the lemma it is enough to show that $\frac{d}{dt}V(\boldsymbol{x}(t)) \leq 0$ almost everywhere on $\{t : \boldsymbol{x}(t) \notin \Omega\}$. As noted earlier, we must consider two cases when analyzing the dynamic behavior of system (2).

Case 1. Suppose that on the interval (t_{l-1}, t_l) the trajectory $\boldsymbol{x}(t)$ does not intersect Ω or any of the surfaces $\Delta_1, \Delta_2, \ldots, \Delta_m$; that is,

$$\boldsymbol{x}(t) \notin \Omega \cup \Delta \quad \text{for all} \quad t \in (t_{l-1}, t_l). \tag{5}$$

Given $t' \in (t_{l-1}, t_l)$, let $\tilde{V} : \mathbb{R}^n \to \mathbb{R}$ be the function defined as

$$\tilde{V}(\boldsymbol{x}) = \sum_{i \in J(\boldsymbol{x}(t'))} g_i(\boldsymbol{x}).$$

It follows from (5) and the fact that $\boldsymbol{x}(t)$ is absolutely continuous that $I(\boldsymbol{x}(t)) = \emptyset$ and $J(\boldsymbol{x}(t)) = J(\boldsymbol{x}(t')) \neq \emptyset$ for all $t \in (t_{l-1}, t_l)$. Thus, on the interval (t_{l-1}, t_l), the trajectory $\boldsymbol{x}(t)$ must satisfy (2) in the usual sense; that is,

$$\tau \dot{\boldsymbol{x}}(t) = -\mu \nabla \tilde{V}(\boldsymbol{x}(t))$$

almost everywhere on the interval (t_{l-1}, t_l). Also, it follows from (4) that $V(\boldsymbol{x}(t)) = \tilde{V}(\boldsymbol{x}(t))$ for all $t \in (t_{l-1}, t_l)$. Applying the chain rule yields

$$\frac{d}{dt}V(\boldsymbol{x}(t)) = -\frac{\mu}{\tau} \left\| \nabla \tilde{V}(\boldsymbol{x}(t)) \right\|_2^2 \tag{6}$$

almost everywhere on the interval (t_{l-1}, t_l). This concludes the analysis for the first case.

Case 2. Suppose that on the interval (t_{l-1}, t_l) the trajectory $\boldsymbol{x}(t)$ is confined to one or more of the surfaces $\Delta_1, \Delta_2, \ldots, \Delta_m$ and does not intersect Ω. Specifically, suppose there exist nonempty index sets $\tilde{I}$ and $\tilde{J}$ such that

(i) $I(\boldsymbol{x}(t)) = \tilde{I}$ for all $t \in (t_{l-1}, t_l)$, and
(ii) $J(\boldsymbol{x}(t)) = \tilde{J}$ for all $t \in (t_{l-1}, t_l)$.

Let $i_1, i_2, \ldots, i_k$ denote the elements of $\tilde{I}$, and let S denote the surface to which the trajectory $\boldsymbol{x}(t)$ is confined on the interval (t_{l-1}, t_l); that is,

$$S = \bigcap_{i \in \tilde{I}} \Delta_i.$$

As in Case 1, let

$$\tilde{V}(\boldsymbol{x}) = \sum_{i \in \tilde{J}} g_i(\boldsymbol{x}).$$

We next define the function $\tilde{\boldsymbol{G}} : \mathbb{R}^n \to \mathbb{R}^{n \times k}$ as

$$\tilde{\boldsymbol{G}}(\boldsymbol{x}) = \left[\begin{array}{cccc} \nabla g_{i_1}(\boldsymbol{x}) & \nabla g_{i_2}(\boldsymbol{x}) & \cdots & \nabla g_{i_k}(\boldsymbol{x}) \end{array} \right].$$

We note that it follows from A4 that $\tilde{\boldsymbol{G}}(\boldsymbol{x})$ is of full rank for any $\boldsymbol{x} \in S$. Given $t' \in (t_{l-1}, t_l)$, it follows from A4 and (i) above that for sufficiently small $\delta > 0$ the surfaces $\Delta_{i_1}, \Delta_{i_2}, \ldots, \Delta_{i_k}$ partition the δ-neighborhood of $\boldsymbol{x}(t')$ into 2^k regions, in each of which the function $\boldsymbol{h}$ is continuous. One can show that $H(\boldsymbol{x}(t'))$ is the set of all vectors $\boldsymbol{w}$ having the form

$$\boldsymbol{w} = -\mu \nabla \tilde{V}(\boldsymbol{x}(t')) - \mu \sum_{j=0}^{2^k - 1} \alpha_j \tilde{\boldsymbol{G}}(\boldsymbol{x}(t')) \boldsymbol{u}_j, \tag{7}$$

where $\alpha_j \geq 0$, $j = 0, 1, \ldots, 2^k - 1$,

$$\sum_{j=0}^{2^k - 1} \alpha_j = 1,$$

and $\boldsymbol{u}_j \in \mathbb{R}^k$, $j = 0, 1, \ldots, 2^k - 1$, are defined as

$$\begin{aligned}
\boldsymbol{u}_0 &= \left[\begin{array}{ccccccc} 0 & 0 & 0 & \cdots & 0 & 0 & 0 \end{array} \right]^T, \\
\boldsymbol{u}_1 &= \left[\begin{array}{ccccccc} 0 & 0 & 0 & \cdots & 0 & 0 & 1 \end{array} \right]^T, \\
\boldsymbol{u}_2 &= \left[\begin{array}{ccccccc} 0 & 0 & 0 & \cdots & 0 & 1 & 0 \end{array} \right]^T, \\
\boldsymbol{u}_3 &= \left[\begin{array}{ccccccc} 0 & 0 & 0 & \cdots & 0 & 1 & 1 \end{array} \right]^T, \\
&\vdots \\
\boldsymbol{u}_{2^k-3} &= \left[\begin{array}{ccccccc} 1 & 1 & 1 & \cdots & 1 & 0 & 1 \end{array} \right]^T, \\
\boldsymbol{u}_{2^k-2} &= \left[\begin{array}{ccccccc} 1 & 1 & 1 & \cdots & 1 & 1 & 0 \end{array} \right]^T, \\
\boldsymbol{u}_{2^k-1} &= \left[\begin{array}{ccccccc} 1 & 1 & 1 & \cdots & 1 & 1 & 1 \end{array} \right]^T.
\end{aligned}$$

Let $T(\boldsymbol{x})$ denote the tangent plane to the surface S at the point $\boldsymbol{x}$. Observe that $\dot{\boldsymbol{x}}(t) \in T(\boldsymbol{x}(t))$ almost everywhere on the interval (t_{l-1}, t_l) because the trajectory $\boldsymbol{x}(t)$ is confined to the surface S on the interval (t_{l-1}, t_l). Thus, $\boldsymbol{x}(t)$ is an absolutely continuous function that satisfies

$$\tau \dot{\boldsymbol{x}}(t) \in H(\boldsymbol{x}(t)) \cap T(\boldsymbol{x}(t))$$

almost everywhere on the interval (t_{l-1}, t_l). We note that since the trajectory $\boldsymbol{x}(t)$ is confined to the surface S on the interval (t_{l-1}, t_l), the set $H(\boldsymbol{x}(t)) \cap T(\boldsymbol{x}(t))$ is

by assumption nonempty for all $t \in (t_{l-1}, t_l)$. In particular, we see that $H(\boldsymbol{x}(t')) \cap T(\boldsymbol{x}(t'))$ is the set of all vectors $\boldsymbol{w}$ that have the form (7) and also lie on the tangent plane $T(\boldsymbol{x}(t'))$. Observe, however, that

$$\sum_{j=0}^{2^k-1} \alpha_j \tilde{\boldsymbol{G}}(\boldsymbol{x}(t'))\boldsymbol{u}_j \in \operatorname{span}\left\{\nabla g_i(\boldsymbol{x}(t')) : i \in \tilde{I}\right\}$$

and thus is orthogonal to $T(\boldsymbol{x}(t'))$. Therefore, the set $H(\boldsymbol{x}(t')) \cap T(\boldsymbol{x}(t'))$ contains exactly one element, namely, the orthogonal projection of the vector $-\mu\nabla\tilde{V}(\boldsymbol{x}(t'))$ on the tangent plane $T(\boldsymbol{x}(t'))$. Hence, for each $t \in (t_{l-1}, t_l)$ the set $H(\boldsymbol{x}(t)) \cap T(\boldsymbol{x}(t))$ contains exactly one element, and thus $\dot{\boldsymbol{x}}(t)$ is uniquely determined almost everywhere on the interval (t_{l-1}, t_l). Specifically,

$$\tau\dot{\boldsymbol{x}}(t) = -\mu \boldsymbol{P}_{\boldsymbol{x}(t)} \nabla\tilde{V}(\boldsymbol{x}(t))$$

almost everywhere on the interval (t_{l-1}, t_l), where $\boldsymbol{P}_{\boldsymbol{x}}$ is the orthogonal projector onto the tangent plane $T(\boldsymbol{x})$; that is,

$$\boldsymbol{P}_{\boldsymbol{x}} = \boldsymbol{I}_n - \tilde{\boldsymbol{G}}(\boldsymbol{x})\left(\tilde{\boldsymbol{G}}^T(\boldsymbol{x})\tilde{\boldsymbol{G}}(\boldsymbol{x})\right)^{-1}\tilde{\boldsymbol{G}}^T(\boldsymbol{x}),$$

where $\boldsymbol{I}_n$ denotes the $n \times n$ identity matrix. Note that from (4) and (ii) above, $V(\boldsymbol{x}(t)) = \tilde{V}(\boldsymbol{x}(t))$ for all $t \in (t_{l-1}, t_l)$. Applying the chain rule yields

$$\frac{d}{dt}V(\boldsymbol{x}(t)) = -\frac{\mu}{\tau}\nabla^T\tilde{V}(\boldsymbol{x}(t))\boldsymbol{P}_{\boldsymbol{x}(t)}\nabla\tilde{V}(\boldsymbol{x}(t))$$

almost everywhere on the interval (t_{l-1}, t_l). Observing that

$$\boldsymbol{P}_{\boldsymbol{x}} = \boldsymbol{P}_{\boldsymbol{x}}^T = \boldsymbol{P}_{\boldsymbol{x}}^2,$$

we conclude that

$$\frac{d}{dt}V(\boldsymbol{x}(t)) = -\frac{\mu}{\tau}\left\|\boldsymbol{P}_{\boldsymbol{x}(t)}\nabla\tilde{V}(\boldsymbol{x}(t))\right\|_2^2 \tag{8}$$

almost everywhere on the interval (t_{l-1}, t_l). This completes the analysis for the second case.

It follows directly from (6) and (8) that $\frac{d}{dt}V(\boldsymbol{x}(t)) \leq 0$ almost everywhere on $\{t : \boldsymbol{x}(t) \notin \Omega\}$. This completes the proof of the lemma. □

Before stating the next result we introduce the following notation. For $\boldsymbol{y} \in \mathbb{R}^n$ and $r > 0$, we let $\mathcal{B}(\boldsymbol{y}, r)$ denote the open ball with center $\boldsymbol{y}$ and radius r; that is,

$$\mathcal{B}(\boldsymbol{y}, r) = \{\boldsymbol{x} : \|\boldsymbol{x} - \boldsymbol{y}\|_2 < r\}.$$

THEOREM 1. *Every trajectory of system* (2) *reaches the feasible set,* Ω*, in finite time and is thereafter confined to* Ω*.*

Proof. Let $\boldsymbol{x} : [t_0, \infty) \to \mathbb{R}^n$ be any particular trajectory of system (2). To prove the theorem, we must show that there exists a number $T_\Omega(\boldsymbol{x}_0) \geq t_0$ such that $\boldsymbol{x}(t) \in \Omega$ for all $t \geq T_\Omega(\boldsymbol{x}_0)$. It follows from Remark 4 and Lemma 1 that $T_\Omega(\boldsymbol{x}_0) = t_0$ if $\boldsymbol{x}_0 \in \Omega$. We now consider the case when $\boldsymbol{x}_0 \notin \Omega$. Once again, we need to consider two cases when analyzing the dynamic behavior of system (2). In the proof, we use the fact that the sublevel set

$$\{\boldsymbol{x} : V(\boldsymbol{x}) \leq a\}$$

is bounded for any $a \in \mathbb{R}$ (see, for example, [25, p. 229]).

Case 1. Consider again the analysis presented in Case 1 of the proof of Lemma 1. By the boundedness of the sublevel sets of V, there exist $\boldsymbol{y} \in \Omega^o$ and $r(\boldsymbol{x}_0) > 0$ such that

$$\mathcal{B}(\boldsymbol{y}, r(\boldsymbol{x}_0)) \supset \{\boldsymbol{x} : V(\boldsymbol{x}) \leq V(\boldsymbol{x}_0)\}.$$

Observe that by A2, the function $\tilde{V}$ is convex over $\mathbb{R}^n$ and has continuous first partial derivatives on $\mathbb{R}^n$. Consequently, the inequality

$$\nabla^T \tilde{V}(\boldsymbol{x}(t))\,(\boldsymbol{x}(t) - \boldsymbol{y}) \geq \tilde{V}(\boldsymbol{x}(t)) - \tilde{V}(\boldsymbol{y}) \tag{9}$$

is satisfied for all $t \in (t_{l-1}, t_l)$. Applying the Cauchy–Schwarz inequality to (9) yields

$$\left\| \nabla \tilde{V}(\boldsymbol{x}(t)) \right\|_2 \|\boldsymbol{x}(t) - \boldsymbol{y}\|_2 \geq \tilde{V}(\boldsymbol{x}(t)) - \tilde{V}(\boldsymbol{y}) \tag{10}$$

for all $t \in (t_{l-1}, t_l)$. We see from Remark 3 that $\tilde{V}(\boldsymbol{y}) < 0$. Also, by definition, $\tilde{V}(\boldsymbol{x}(t)) > 0$ for all $t \in (t_{l-1}, t_l)$. Combining these two facts together with (10), we conclude that

$$\left\| \nabla \tilde{V}(\boldsymbol{x}(t)) \right\|_2 \geq -\frac{\tilde{V}(\boldsymbol{y})}{\|\boldsymbol{x}(t) - \boldsymbol{y}\|_2} \tag{11}$$

for all $t \in (t_{l-1}, t_l)$. It follows from the fact that $\mathcal{B}(\boldsymbol{y}, r(\boldsymbol{x}_0)) \supset \{\boldsymbol{x} : V(\boldsymbol{x}) \leq V(\boldsymbol{x}_0)\}$ and from Lemma 1 that $\boldsymbol{x}(t) \in \mathcal{B}(\boldsymbol{y}, r(\boldsymbol{x}_0))$ for all $t \in (t_{l-1}, t_l)$. Hence, $\|\boldsymbol{x}(t) - \boldsymbol{y}\|_2 \leq r(\boldsymbol{x}_0)$ for all $t \in (t_{l-1}, t_l)$. This fact, together with (11), implies that

$$\left\| \nabla \tilde{V}(\boldsymbol{x}(t)) \right\|_2 \geq -\frac{\tilde{V}(\boldsymbol{y})}{r(\boldsymbol{x}_0)} \tag{12}$$

for all $t \in (t_{l-1}, t_l)$. Now let $\tilde{\eta}(\boldsymbol{x}_0)$ be the positive constant defined as

$$\tilde{\eta}(\boldsymbol{x}_0) = \frac{\mu}{\tau} \left(\frac{\tilde{V}(\boldsymbol{y})}{r(\boldsymbol{x}_0)} \right)^2. \tag{13}$$

Then, it follows from (6), (12), and (13) that

$$\frac{d}{dt} V(\boldsymbol{x}(t)) \leq -\tilde{\eta}(\boldsymbol{x}_0) \tag{14}$$

almost everywhere on the interval (t_{l-1}, t_l). However, (14) combined with the fact that there is a finite number of constraints, that is, m is finite, implies the existence of a number $\eta_1(\boldsymbol{x}_0) > 0$ such that

$$\frac{d}{dt} V(\boldsymbol{x}(t)) \leq -\eta_1(\boldsymbol{x}_0) \tag{15}$$

almost everywhere on $\{t : \boldsymbol{x}(t) \notin \Omega \cup \Delta\}$. This concludes the analysis for the first case.

Case 2. Consider again the analysis presented in Case 2 of the proof of Lemma 1. By the boundedness of the sublevel sets of V, there exist $\boldsymbol{y} \in \Omega^o$ and $r(\boldsymbol{x}_0) > 0$ such that

$$\mathcal{B}(\boldsymbol{y}, r(\boldsymbol{x}_0)) \supset \{\boldsymbol{x} : V(\boldsymbol{x}) \leq V(\boldsymbol{x}_0)\}.$$

Following the argument of our analysis for Case 2 of the proof of Lemma 1, we conclude that there exist nonnegative constants $\beta_{i_1}, \beta_{i_2}, \ldots, \beta_{i_k}$, such that

$$\boldsymbol{P}_{\boldsymbol{x}(t')} \nabla \tilde{V}(\boldsymbol{x}(t')) = \nabla \tilde{V}(\boldsymbol{x}(t')) + \sum_{j \in \tilde{I}} \beta_j \nabla g_j(\boldsymbol{x}(t')). \tag{16}$$

Let $\hat{V} : \mathbb{R}^n \to \mathbb{R}$ be the function defined as

$$\hat{V}(\boldsymbol{x}) = \tilde{V}(\boldsymbol{x}) + \sum_{j \in \tilde{I}} \beta_j g_j(\boldsymbol{x}).$$

By virtue of A2, the function $\hat{V}$ is convex over $\mathbb{R}^n$ and has continuous first partial derivatives on $\mathbb{R}^n$. Using arguments similar to those used in Case 1, we have

$$\left\| \nabla \hat{V}(\boldsymbol{x}(t')) \right\|_2 \geq -\frac{\tilde{V}(\boldsymbol{y})}{r(\boldsymbol{x}_0)}.$$

It then follows from (16) and the definition of $\hat{V}$ that

$$\left\| \boldsymbol{P}_{\boldsymbol{x}(t)} \nabla \tilde{V}(\boldsymbol{x}(t)) \right\|_2 = \left\| \nabla \hat{V}(\boldsymbol{x}(t)) \right\|_2 \geq -\frac{\tilde{V}(\boldsymbol{y})}{r(\boldsymbol{x}_0)}$$

for all $t \in (t_{l-1}, t_l)$. As in Case 1, we conclude that there exists a number $\eta_2(\boldsymbol{x}_0) > 0$, such that

$$\frac{d}{dt} V(\boldsymbol{x}(t)) \leq -\eta_2(\boldsymbol{x}_0) \tag{17}$$

almost everywhere on $\{t : \boldsymbol{x}(t) \in \Delta \setminus \Omega\}$. This concludes the analysis for the second case.

Let $\eta(\boldsymbol{x}_0) = \min\{\eta_1(\boldsymbol{x}_0), \eta_2(\boldsymbol{x}_0)\}$. Then, combining (15) and (17) we obtain

$$\frac{d}{dt} V(\boldsymbol{x}(t)) \leq -\eta(\boldsymbol{x}_0) \tag{18}$$

almost everywhere on $\{t : \boldsymbol{x}(t) \notin \Omega\}$. Let

$$T_\Omega(\boldsymbol{x}_0) = t_0 + \frac{V(\boldsymbol{x}_0)}{\eta(\boldsymbol{x}_0)}. \tag{19}$$

By Lemma 1, Remark 4, (18), and (19), the trajectory $\boldsymbol{x}(t) \in \Omega$ for all $t \geq T_\Omega(\boldsymbol{x}_0)$. This completes the proof. □

Having established this last result, we now turn our attention to analyzing the dynamic behavior of system (2) when its motion is confined to the feasible set Ω. We begin by introducing the function $F : \mathbb{R}^n \to \mathbb{R}$ as

$$F(\boldsymbol{x}) = f(\boldsymbol{x}) - f^*,$$

where f^* is the optimal value of f for problem (1).

Remark 5. It follows from A2 that the function F is convex over $\mathbb{R}^n$ and has continuous first partial derivatives on $\mathbb{R}^n$. Also, by definition, $F(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \in \Gamma$, and $F(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in \Omega \setminus \Gamma$.

We will show that F tends to zero with time when evaluated on any trajectory of system (2). We need the following technical result.

LEMMA 2. *F is a decreasing function of time when evaluated on any trajectory of system* (2) *while it is confined to the feasible set* Ω.

Proof. Let $\boldsymbol{x} : [t_0, \infty) \to \mathbb{R}^n$ be any particular trajectory of system (2). It follows from Remark 5 and the fact that $\boldsymbol{x}(t)$ is absolutely continuous that to prove the lemma it is enough to show that $\frac{d}{dt}F(\boldsymbol{x}(t)) \leq 0$ almost everywhere on $\{t : \boldsymbol{x}(t) \in \Omega\}$. As before, we consider two cases when analyzing the dynamic behavior of system (2).

Case 1. Suppose that $\boldsymbol{x}(t) \in \Omega^o$ for all $t \in (t_{l-1}, t_l)$. Using arguments similar to those used in Case 1 of the proof of Lemma 1, we conclude that

$$\tau \dot{\boldsymbol{x}}(t) = -\nabla f(\boldsymbol{x}(t)) \tag{20}$$

almost everywhere on the interval (t_{l-1}, t_l). Applying the chain rule, we obtain

$$\frac{d}{dt}F(\boldsymbol{x}(t)) = -\frac{1}{\tau}\left\|\nabla f(\boldsymbol{x}(t))\right\|_2^2 \tag{21}$$

almost everywhere on the interval (t_{l-1}, t_l). This concludes the analysis for the first case.

Case 2. Suppose that $\boldsymbol{x}(t) \in \partial\Omega$ and $I(\boldsymbol{x}(t))$ is constant for all $t \in (t_{l-1}, t_l)$. We use the same notation as in Case 2 of the proof of Lemma 1. The first part of the proof is almost identical to that of Case 2 of the proof of Lemma 1, and we omit the details. We only observe that now we have the following.

(i) $H(\boldsymbol{x}(t'))$ is the set of all vectors $\boldsymbol{w}$ having the form

$$\boldsymbol{w} = -\alpha_0 \nabla f(\boldsymbol{x}(t')) - \mu \sum_{j=1}^{2^k-1} \alpha_j \tilde{\boldsymbol{G}}(\boldsymbol{x}(t'))\boldsymbol{u}_j,$$

where $\alpha_j \geq 0$, $j = 0, 1, \ldots, 2^k - 1$, $\sum_{j=0}^{2^k-1} \alpha_j = 1$, and $\boldsymbol{u}_j \in \mathbb{R}^k$, $j = 1, 2, \ldots, 2^k - 1$, are as defined in the proof of Lemma 1.

(ii) Suppose $\hat{\boldsymbol{w}} \in H(\boldsymbol{x}(t')) \cap T(\boldsymbol{x}(t'))$, with

$$\hat{\boldsymbol{w}} = -\hat{\alpha}_0 \nabla f(\boldsymbol{x}(t')) - \mu \sum_{j=1}^{2^k-1} \hat{\alpha}_j \tilde{\boldsymbol{G}}(\boldsymbol{x}(t'))\boldsymbol{u}_j, \tag{22}$$

where $\hat{\alpha}_j \geq 0$, $j = 0, 1, \ldots, 2^k - 1$, and $\sum_{j=0}^{2^k-1} \hat{\alpha}_j = 1$. Then, the vector $\hat{\boldsymbol{w}}$ must satisfy the equation

$$\hat{\boldsymbol{w}} = -\hat{\alpha}_0 \boldsymbol{P}_{\boldsymbol{x}(t')} \nabla f(\boldsymbol{x}(t')). \tag{23}$$

It follows from (22) and (23) that

$$-\hat{\alpha}_0 \left(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{x}(t')}\right) \nabla f(\boldsymbol{x}(t')) = \mu \sum_{j=1}^{2^k-1} \hat{\alpha}_j \tilde{\boldsymbol{G}}(\boldsymbol{x}(t'))\boldsymbol{u}_j. \tag{24}$$

Now, let $\boldsymbol{U} \in \mathbb{R}^{k \times (2^k-1)}$ be the matrix with columns $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_{2^k-1}$, and let $\hat{\boldsymbol{\alpha}} \in \mathbb{R}^{2^k-1}$ be the column vector with components $\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_{2^k-1}$; that is,

$$\boldsymbol{U} = \begin{bmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 & \cdots & \boldsymbol{u}_{2^k-1} \end{bmatrix}$$

and

$$\hat{\boldsymbol{\alpha}} = \begin{bmatrix} \hat{\alpha}_1 & \hat{\alpha}_2 & \cdots & \hat{\alpha}_{2^k-1} \end{bmatrix}^T.$$

Using the above notation, we can rewrite (24) as

$$-\frac{\hat{\alpha}_0}{\mu}\left(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{x}(t')}\right)\nabla f(\boldsymbol{x}(t')) = \tilde{\boldsymbol{G}}(\boldsymbol{x}(t'))\boldsymbol{U}\hat{\boldsymbol{\alpha}}. \tag{25}$$

Premultiplying both sides of (25) by $\left(\tilde{\boldsymbol{G}}^T(\boldsymbol{x}(t'))\tilde{\boldsymbol{G}}(\boldsymbol{x}(t'))\right)^{-1}\tilde{\boldsymbol{G}}^T(\boldsymbol{x}(t'))$, we obtain

$$-\frac{\hat{\alpha}_0}{\mu}\left(\tilde{\boldsymbol{G}}^T(\boldsymbol{x}(t'))\tilde{\boldsymbol{G}}(\boldsymbol{x}(t'))\right)^{-1}\tilde{\boldsymbol{G}}^T(\boldsymbol{x}(t'))\nabla f(\boldsymbol{x}(t')) = \boldsymbol{U}\hat{\boldsymbol{\alpha}}.$$

Taking the 1-norm yields

$$\frac{\hat{\alpha}_0}{\mu}\left\|\left(\tilde{\boldsymbol{G}}^T(\boldsymbol{x}(t'))\tilde{\boldsymbol{G}}(\boldsymbol{x}(t'))\right)^{-1}\tilde{\boldsymbol{G}}^T(\boldsymbol{x}(t'))\nabla f(\boldsymbol{x}(t'))\right\|_1 = \left\|\boldsymbol{U}\hat{\boldsymbol{\alpha}}\right\|_1.$$

Hence,

$$\frac{\hat{\alpha}_0}{\mu}\left\|\left(\tilde{\boldsymbol{G}}^T(\boldsymbol{x}(t'))\tilde{\boldsymbol{G}}(\boldsymbol{x}(t'))\right)^{-1}\tilde{\boldsymbol{G}}^T(\boldsymbol{x}(t'))\right\|_1 \left\|\nabla f(\boldsymbol{x}(t'))\right\|_1 \geq \left\|\boldsymbol{U}\hat{\boldsymbol{\alpha}}\right\|_1. \tag{26}$$

However,

$$\left\|\boldsymbol{U}\hat{\boldsymbol{\alpha}}\right\|_1 \geq \sum_{j=1}^{2^k-1}\hat{\alpha}_j = 1 - \hat{\alpha}_0. \tag{27}$$

Moreover, by A2 and the fact that $\Omega \cap S$ is by definition compact, there exist nonnegative constants M_1 and M_2 such that

$$M_1 \geq \left\|\left(\tilde{\boldsymbol{G}}^T(\boldsymbol{x})\tilde{\boldsymbol{G}}(\boldsymbol{x})\right)^{-1}\tilde{\boldsymbol{G}}^T(\boldsymbol{x})\right\|_1 \quad \text{for all} \quad \boldsymbol{x} \in \Omega \cap S, \tag{28}$$

and

$$M_2 \geq \left\|\nabla f(\boldsymbol{x})\right\|_1 \quad \text{for all} \quad \boldsymbol{x} \in \Omega \cap S. \tag{29}$$

Using (26)–(29) we obtain

$$\frac{\hat{\alpha}_0 M_1 M_2}{\mu} \geq 1 - \hat{\alpha}_0,$$

which implies that $\hat{\alpha}_0 > 0$. Therefore

$$\frac{M_1 M_2}{\mu} \geq \frac{1-\hat{\alpha}_0}{\hat{\alpha}_0} = \frac{1}{\hat{\alpha}_0} - 1,$$

and hence

$$\hat{\alpha}_0 \geq \frac{\mu}{\mu + M_1 M_2}. \tag{30}$$

Let $\sigma = \frac{\mu}{\mu + M_1 M_2}$; then (23) and (30) imply that

$$H(\boldsymbol{x}(t)) \cap T(\boldsymbol{x}(t)) \subset \left\{-\alpha \boldsymbol{P}_{\boldsymbol{x}(t)} \nabla f(\boldsymbol{x}(t)) : \sigma \leq \alpha \leq 1\right\} \tag{31}$$

for all $t \in (t_{l-1}, t_l)$, and therefore

$$\tau \dot{\boldsymbol{x}}(t) \in \left\{-\alpha \boldsymbol{P}_{\boldsymbol{x}(t)} \nabla f(\boldsymbol{x}(t)) : \sigma \leq \alpha \leq 1\right\} \tag{32}$$

almost everywhere on the interval (t_{l-1}, t_l). Applying the chain rule and observing that $\boldsymbol{P}_{\boldsymbol{x}} = \boldsymbol{P}_{\boldsymbol{x}}^T = \boldsymbol{P}_{\boldsymbol{x}}^2$, we obtain

$$\frac{d}{dt} F(\boldsymbol{x}(t)) \in \left\{-\frac{\alpha}{\tau} \left\| \boldsymbol{P}_{\boldsymbol{x}(t)} \nabla f(\boldsymbol{x}(t)) \right\|_2^2 : \sigma \leq \alpha \leq 1\right\}$$

almost everywhere on the interval (t_{l-1}, t_l). Hence,

$$\frac{d}{dt} F(\boldsymbol{x}(t)) \leq -\frac{\sigma}{\tau} \left\| \boldsymbol{P}_{\boldsymbol{x}(t)} \nabla f(\boldsymbol{x}(t)) \right\|_2^2 \tag{33}$$

almost everywhere on the interval (t_{l-1}, t_l). This concludes the analysis for the second case.

It now follows from (21) and (33) that $\frac{d}{dt} F(\boldsymbol{x}(t)) \leq 0$ almost everywhere on $\{t : \boldsymbol{x}(t) \in \Omega\}$. The proof is complete. □

Before presenting the next lemma we introduce the following notation. For each $\varepsilon > 0$, let

$$\Phi_\varepsilon = \Omega \cap \{\boldsymbol{x} : F(\boldsymbol{x}) < \varepsilon\}.$$

LEMMA 3. *F tends to zero with time when evaluated on any trajectory of system (2).*

Proof. Let $\boldsymbol{x} : [t_0, \infty) \to \mathbb{R}^n$ be any particular trajectory of system (2). To prove the lemma, it is enough to show that given any $\varepsilon > 0$, there exists a number $T(\boldsymbol{x}_0, \varepsilon) \geq t_0$ such that $\boldsymbol{x}(t) \in \Phi_\varepsilon$ for all $t \geq T(\boldsymbol{x}_0, \varepsilon)$. By Theorem 1, there is a number $T_\Omega(\boldsymbol{x}_0) \geq t_0$ such that $\boldsymbol{x}(t) \in \Omega$ for all $t \geq T_\Omega(\boldsymbol{x}_0)$. It follows from Theorem 1 and Lemma 2 that

$$T(\boldsymbol{x}_0, \varepsilon) = T_\Omega(\boldsymbol{x}_0) \quad \text{if} \quad \boldsymbol{x}\left(T_\Omega(\boldsymbol{x}_0)\right) \in \Phi_\varepsilon.$$

We now consider the case when $\boldsymbol{x}\left(T_\Omega(\boldsymbol{x}_0)\right) \notin \Phi_\varepsilon$. Once again, we must consider two cases when analyzing the dynamic behavior of system (2).

Case 1. Consider again the analysis presented in Case 1 of the proof of Lemma 2, and suppose that $\boldsymbol{x}(t) \in \Omega \setminus \Phi_{\varepsilon/2}$ for all $t \in (t_{l-1}, t_l)$. Let $\boldsymbol{y} \in \Gamma$ and $r > 0$ such that $\mathcal{B}(\boldsymbol{y}, r) \supset \Omega$. The existence of an open ball with this property is a consequence of A3. By Remark 5

$$\nabla^T F(\boldsymbol{x}(t))(\boldsymbol{x}(t) - \boldsymbol{y}) \geq F(\boldsymbol{x}(t)) - F(\boldsymbol{y}) \tag{34}$$

for all $t \in (t_{l-1}, t_l)$. Applying the Cauchy–Schwarz inequality to (34) yields

$$\|\nabla F(\boldsymbol{x}(t))\|_2 \, \|\boldsymbol{x}(t) - \boldsymbol{y}\|_2 \geq F(\boldsymbol{x}(t)) - F(\boldsymbol{y}) \tag{35}$$

for all $t \in (t_{l-1}, t_l)$. Observe from Remark 5 that $F(\boldsymbol{y}) = 0$. Also, by definition, $F(\boldsymbol{x}(t)) \geq \varepsilon/2$ for all $t \in (t_{l-1}, t_l)$. Combining these two facts together with (35), we obtain

$$\|\nabla F(\boldsymbol{x}(t))\|_2 \geq \frac{\varepsilon}{2\|\boldsymbol{x}(t) - \boldsymbol{y}\|_2}$$

for all $t \in (t_{l-1}, t_l)$. However, $\|\boldsymbol{x}(t) - \boldsymbol{y}\|_2 \leq r$ for all $t \in (t_{l-1}, t_l)$ since $\mathcal{B}(\boldsymbol{y}, r) \supset \Omega$ and $\boldsymbol{x}(t) \in \Omega$ for all $t \in (t_{l-1}, t_l)$. Hence,

$$\|\nabla F(\boldsymbol{x}(t))\|_2 \geq \frac{\varepsilon}{2r}$$

for all $t \in (t_{l-1}, t_l)$. Let $\eta_1(\varepsilon)$ be the positive constant defined as

$$\eta_1(\varepsilon) = \frac{\varepsilon^2}{4\tau r^2}.$$

It now follows from (21) and the definition of F that

$$\frac{d}{dt} F(\boldsymbol{x}(t)) \leq -\eta_1(\varepsilon) \tag{36}$$

almost everywhere on $\{t : \boldsymbol{x}(t) \in \Omega^o \setminus \Phi_{\varepsilon/2}\}$. This concludes the analysis for the first case.

Case 2. Consider again the analysis presented in Case 2 of the proof of Lemma 2, and suppose that $\boldsymbol{x}(t) \in \Omega \setminus \Phi_{\varepsilon/2}$ for all $t \in (t_{l-1}, t_l)$. Let $\boldsymbol{y} \in \Gamma$ and $r > 0$ such that $\mathcal{B}(\boldsymbol{y}, r) \supset \Omega$. It follows from (22), (23), and (30) that there exist nonnegative constants $\beta_{i_1}, \beta_{i_2}, \ldots, \beta_{i_k}$ such that

$$\boldsymbol{P}_{\boldsymbol{x}(t')} \nabla f(\boldsymbol{x}(t')) = \nabla f(\boldsymbol{x}(t')) + \sum_{j \in \tilde{I}} \beta_j \nabla g_j(\boldsymbol{x}(t')). \tag{37}$$

Let $\tilde{F} : \mathbb{R}^n \to \mathbb{R}$ be the function defined as

$$\tilde{F}(\boldsymbol{x}) = F(\boldsymbol{x}) + \sum_{j \in \tilde{I}} \beta_j g_j(\boldsymbol{x}).$$

Observe that by virtue of A2 and Remark 5, the function $\tilde{F}$ is convex over $\mathbb{R}^n$ and has continuous first partial derivatives on $\mathbb{R}^n$. Therefore, the inequality

$$\nabla^T \tilde{F}(\boldsymbol{x}(t'))(\boldsymbol{x}(t') - \boldsymbol{y}) \geq \tilde{F}(\boldsymbol{x}(t')) - \tilde{F}(\boldsymbol{y})$$

holds. Applying the Cauchy–Schwarz inequality and observing that $\tilde{F}(\boldsymbol{y}) \leq 0$ and $\tilde{F}(\boldsymbol{x}(t')) \geq \varepsilon/2$, we obtain

$$\left\| \nabla \tilde{F}(\boldsymbol{x}(t')) \right\|_2 \geq \frac{\varepsilon}{2\|\boldsymbol{x}(t') - \boldsymbol{y}\|_2}.$$

However, $\|\boldsymbol{x}(t') - \boldsymbol{y}\|_2 \leq r$ since $\boldsymbol{x}(t') \in \Omega$ and $\mathcal{B}(\boldsymbol{y}, r) \supset \Omega$. Therefore,

$$\left\| \nabla \tilde{F}(\boldsymbol{x}(t')) \right\|_2 \geq \frac{\varepsilon}{2r}.$$

Taking into account (37) and the definition of $\tilde{F}$ it follows that

$$\left\| \boldsymbol{P}_{\boldsymbol{x}(t)} \nabla f(\boldsymbol{x}(t)) \right\|_2 \geq \frac{\varepsilon}{2r} \tag{38}$$

for all $t \in (t_{l-1}, t_l)$. Now let $\tilde{\eta}(\varepsilon)$ be the positive constant defined as

$$\tilde{\eta}(\varepsilon) = \frac{\sigma \varepsilon^2}{4\tau r^2}.$$

It then follows from (33) that

$$\frac{d}{dt}F(\boldsymbol{x}(t)) \leq -\tilde{\eta}(\varepsilon) \tag{39}$$

almost everywhere on the interval (t_{l-1}, t_l). However, (39) combined with the fact that there is a finite number of constraints, that is, m is finite, implies the existence of a constant $\eta_2(\varepsilon) > 0$ such that

$$\frac{d}{dt}F(\boldsymbol{x}(t)) \leq -\eta_2(\varepsilon) \tag{40}$$

almost everywhere on $\left\{t : \boldsymbol{x}(t) \in \partial\Omega \setminus \Phi_{\varepsilon/2}\right\}$. This concludes the analysis for the second case.

Let $\eta(\varepsilon) = \min\{\eta_1(\varepsilon), \eta_2(\varepsilon)\}$. It follows directly from (36) and (40) that

$$\frac{d}{dt}F(\boldsymbol{x}(t)) \leq -\eta(\varepsilon) \tag{41}$$

almost everywhere on $\left\{t : \boldsymbol{x}(t) \in \Omega \setminus \Phi_{\varepsilon/2}\right\}$. Now let

$$T(\boldsymbol{x}_0, \varepsilon) = T_\Omega(\boldsymbol{x}_0) + \frac{F\left(\boldsymbol{x}(T_\Omega(\boldsymbol{x}_0))\right) - \frac{\varepsilon}{2}}{\eta(\varepsilon)}. \tag{42}$$

Then, by Theorem 1, Lemma 2, (41), and (42), we have $\boldsymbol{x}(t) \in \Phi_\varepsilon$ for all $t \geq T(\boldsymbol{x}_0, \varepsilon)$, which completes the proof of the lemma. □

We are now ready to present the main results of this paper. Before doing so we introduce the following notation. For each $\delta > 0$, let

$$\Gamma_\delta = \Omega \cap \{\boldsymbol{x} : d(\boldsymbol{x}, \Gamma) < \delta\}.$$

THEOREM 2. *Every trajectory of system* (2) *converges to the solution set,* Γ*, of problem* (1)*.*

Proof. Let $\boldsymbol{x} : [t_0, \infty) \to \mathbb{R}^n$ be any particular trajectory of system (2). To prove the theorem, it is enough to show that given any $\delta > 0$, there exists a number $T_\delta(\boldsymbol{x}_0) \geq t_0$ such that $\boldsymbol{x}(t) \in \Gamma_\delta$ for all $t \geq T_\delta(\boldsymbol{x}_0)$. Let

$$\begin{aligned}\varepsilon &= \varepsilon(\delta)\\ &= \min\{F(\boldsymbol{x}) : \boldsymbol{x} \in \Omega \cap \{\boldsymbol{x} : d(\boldsymbol{x}, \Gamma) = \delta\}\}.\end{aligned}$$

Note that ε is well defined since, by definition, the function F is continuous and the set $\Omega \cap \{\boldsymbol{x} : d(\boldsymbol{x}, \Gamma) = \delta\}$ is compact. Also, observe that, by definition, $\varepsilon > 0$. Now it is a direct consequence of the fact that F is a convex function and Ω is a convex set that $\Phi_{\varepsilon/2} \subset \Gamma_\delta$. By Lemma 3, there exists a number $T(\boldsymbol{x}_0, \varepsilon/2) \geq t_0$ such that $\boldsymbol{x}(t) \in \Phi_{\varepsilon/2}$ for all $t \geq T(\boldsymbol{x}_0, \varepsilon/2)$. Let $T_\delta(\boldsymbol{x}_0) = T(\boldsymbol{x}_0, \varepsilon/2)$. Then, we have $\boldsymbol{x}(t) \in \Gamma_\delta$ for all $t \geq T_\delta(\boldsymbol{x}_0)$. This completes the proof. □

We now show that the equilibrium points of system (2) coincide with the minimizers of the optimization problem (1).

THEOREM 3. *A point* $\boldsymbol{x}^* \in \mathbb{R}^n$ *is an equilibrium point of system* (2) *if and only if it is a minimizer of problem* (1)*.*

Proof. It follows from Theorem 2 that any equilibrium point of system (2) must be contained in Γ. Thus, it remains to show that every point in Γ is an equilibrium point of system (2). To this end, let $\boldsymbol{x} : [t_0, \infty) \to \mathbb{R}^n$ be any particular trajectory

of system (2) with $\boldsymbol{x}_0 \in \Gamma$. Note that by Remarks 4 and 5, and Lemmas 1 and 2, $\boldsymbol{x}(t) \in \Gamma$ for all $t \geq t_0$. When analyzing the behavior of the trajectory $\boldsymbol{x}(t)$ we need only consider the two cases considered in the proof of Lemma 2.

Case 1. Consider again the analysis presented in Case 1 of the proof of Lemma 2. It follows from the first-order necessary conditions for problem (1), and the fact that $\boldsymbol{x}(t) \in \Gamma$ for all $t \geq t_0$, that $\nabla f(\boldsymbol{x}(t)) = \boldsymbol{0}$ for all $t \in (t_{l-1}, t_l)$. It then follows from (20) that $\dot{\boldsymbol{x}}(t) = \boldsymbol{0}$ almost everywhere on the interval (t_{l-1}, t_l). This concludes the analysis for the first case.

Case 2. Consider again the analysis presented in Case 2 of the proof of Lemma 2. It follows from the first-order necessary conditions for problem (1), and the fact that $\boldsymbol{x}(t) \in \Gamma$ for all $t \geq t_0$, that there exist nonnegative constants, $\lambda_{i_1}, \lambda_{i_2}, \ldots, \lambda_{i_k}$, such that

$$\nabla f(\boldsymbol{x}(t')) + \sum_{j \in \tilde{I}} \lambda_j \nabla g_j(\boldsymbol{x}(t')) = \boldsymbol{0}. \tag{43}$$

It now follows from (43) and (22) that $\boldsymbol{0} \in H(\boldsymbol{x}(t')) \cap T(\boldsymbol{x}(t'))$. However, we see from (31) that all of the elements of $H(\boldsymbol{x}(t')) \cap T(\boldsymbol{x}(t'))$ differ by a positive multiplicative constant. Therefore, $H(\boldsymbol{x}(t')) \cap T(\boldsymbol{x}(t')) = \{\boldsymbol{0}\}$, and hence, $H(\boldsymbol{x}(t)) \cap T(\boldsymbol{x}(t)) = \{\boldsymbol{0}\}$ for all $t \in (t_{l-1}, t_l)$. Using (32), we conclude that $\dot{\boldsymbol{x}}(t) = \boldsymbol{0}$ almost everywhere on the interval (t_{l-1}, t_l). This concludes the analysis for the second case.

It now follows from the above arguments that $\dot{\boldsymbol{x}}(t) = \boldsymbol{0}$ almost everywhere on the interval $[t_0, \infty)$, and hence, $\boldsymbol{x}(t) = \boldsymbol{x}_0$ for all $t \geq t_0$. This completes the proof. ∎

We see from (20) and (32) that while confined to Ω, (2) can be viewed as a continuous-time gradient projection method. We note that a discrete-time gradient projection method for nonlinear programming was first proposed by Rosen [20] and that other continuous-time methods using gradient projections are reported in [5] and references therein.

Similarities between the dynamic system approach to solving optimization problems and the so-called interior point methods are discussed in [26]. For a review of interior point methods, as well as path-following methods, we refer the reader to Gonzaga [10].

4. Examples. In this section, we illustrate the dynamic behavior of system (2) by presenting the results of two computer simulations.

Example 1. In this example, we consider the quadratic programming problem

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} + \boldsymbol{c}^T\boldsymbol{x}, \quad \boldsymbol{x} \in \mathbb{R}^2, \\ \text{subject to} & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}, \end{array}$$

where

$$\boldsymbol{Q} = \begin{bmatrix} 4 & 1 \\ 1 & 2 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} -12 \\ -10 \end{bmatrix}, \quad \boldsymbol{A} = \begin{bmatrix} 5 & 1 \\ -5 & 1 \\ -1 & -2 \end{bmatrix}, \quad \text{and} \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 10 \\ 10 \end{bmatrix}.$$

The above optimization problem clearly satisfies A1–A4, and one can verify that the point $\boldsymbol{x}^* = \begin{bmatrix} -19/22 & 95/22 \end{bmatrix}^T$ is a unique minimizer for the problem; that is, $\Gamma = \{\boldsymbol{x}^*\}$. Note that $\boldsymbol{x}^* \in \Delta_1$. A phase-plane portrait for system (2) that solves the above optimization problem is shown in Figure 1. Observe that each of the trajectories in the phase-plane portrait converges to the point $\boldsymbol{x}^*$ while sliding along the surface Δ_1.

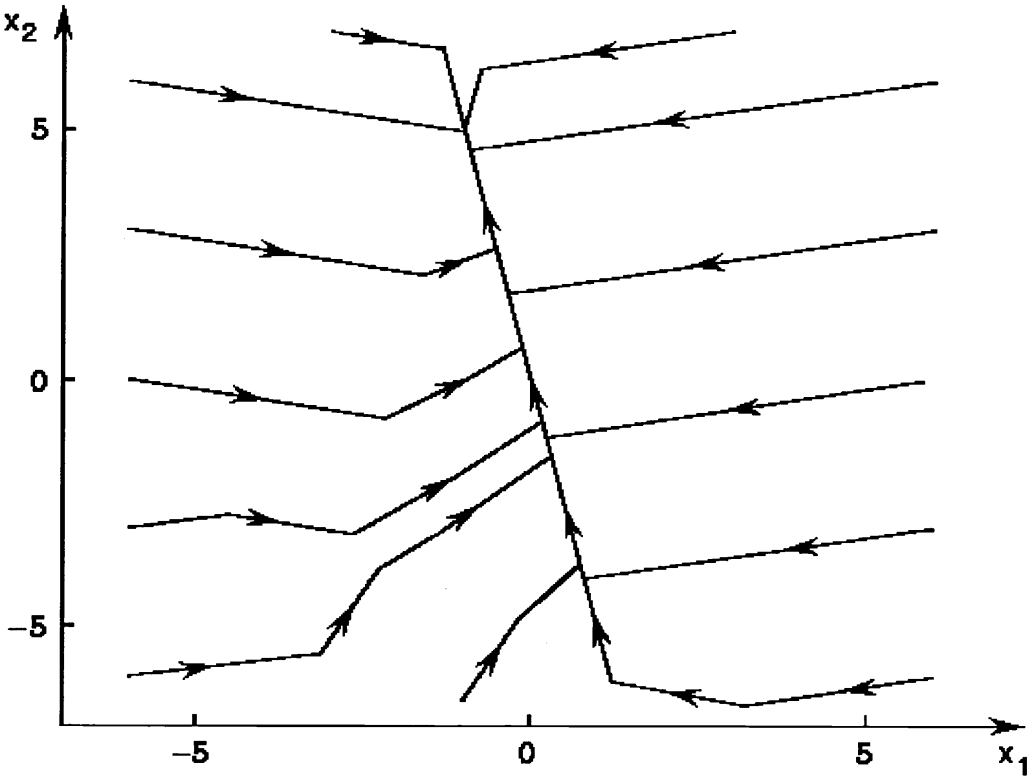


FIG. 1. *Phase-plane portrait for the system in Example* 1.

Example 2. In this example, we consider the convex programming problem

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} + \boldsymbol{c}^T\boldsymbol{x}, \quad \boldsymbol{x} \in \mathbb{R}^2, \\ \text{subject to} & g_i(\boldsymbol{x}) \leq 0, \quad i = 1, 2, 3, \end{array}$$

where

$$\begin{aligned} g_1(\boldsymbol{x}) &= x_1^2 + (x_2 - 4)^2 - 64, \\ g_2(\boldsymbol{x}) &= (x_1 + 3)^2 + x_2^2 - 36, \\ g_3(\boldsymbol{x}) &= (x_1 - 3)^2 + x_2^2 - 36, \end{aligned}$$

and $\boldsymbol{Q}$ and $\boldsymbol{c}$ are as defined in Example 1. One can easily verify that the above optimization problem satisfies A1–A4 and that the point $\boldsymbol{x}^* = \begin{bmatrix} 1.776 & 3.629 \end{bmatrix}^T$ is a unique minimizer for the problem; that is, $\Gamma = \{\boldsymbol{x}^*\}$. Note that $\boldsymbol{x}^* \in \Delta_2$. A phase-plane portrait for system (2) solving the above optimization problem is shown in Figure 2. Note that each of the trajectories in the phase-plane portrait converges to the point $\boldsymbol{x}^*$ while sliding along the surface Δ_2.

We close this section by noting that both of the simulations were performed on a Northgate 486 personal computer using the SIMNON software package.

5. Conclusions. We analyzed a class of dynamic systems proposed by Rodríguez-Vázquez et al. [19] for solving convex programming problems. We showed that the equilibrium points of the system coincide with the minimizers of the convex programming problem, and that all trajectories of the system converge to the solution set of

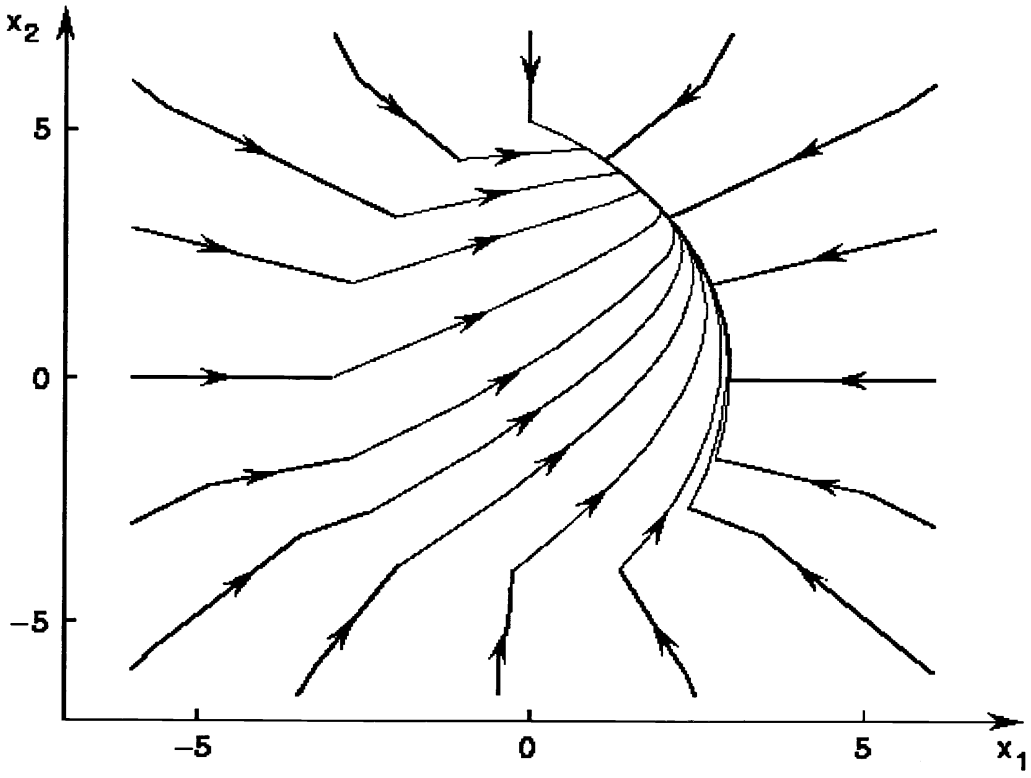


FIG. 2. *Phase-plane portrait for the system in Example* 2.

the problem. In carrying out the analysis we used concepts from the theory of differential equations with discontinuous right-hand sides and Lyapunov stability theory. Our analysis method can also be applied to other classes of analog dynamic optimizers whose designs are based on exact penalty functions. An open problem is to extend the results obtained herein to a more general class of mathematical programming problems.

Acknowledgment. We gratefully acknowledge the constructive remarks of the SICON editor, J. Burke, and the reviewers.

REFERENCES

[1] J.-P. AUBIN AND A. CELLINA, *Differential Inclusions: Set-Valued Maps and Viability Theory*, Springer-Verlag, Berlin, 1984.

[2] D. P. BERTSEKAS, *Necessary and sufficient conditions for a penalty method to be exact*, Mathematical Programming, 9 (1975), pp. 87–99.

[3] L. O. CHUA AND G.-N. LIN, *Nonlinear programming without computation*, IEEE Trans. Circuits Systems, CAS-31 (1984), pp. 182–188.

[4] A. CICHOCKI AND R. UNBEHAUEN, *Neural Networks for Optimization and Signal Processing*, John Wiley, Chichester, UK, 1993.

[5] Y. G. EVTUSHENKO, *Numerical Optimization Techniques*, Translations Series in Mathematics and Engineering, Optimization Software, Inc., New York, 1985.

[6] L. FAYBUSOVICH, *Dynamical systems that solve linear programming problems*, in Proc. 31st Conf. on Decision and Control, Tucson, AZ, 1992, pp. 1626–1631.

[7] A. F. FILIPPOV, *Differential Equations with Discontinuous Righthand Sides*, Math. Appl. (Soviet Ser.), Kluwer, Dordrecht, the Netherlands, 1988.

[8] S. Flåm and J. Zowe, *A primal-dual differential method for convex programming*, in Optimization and Nonlinear Analysis, A. Ioffe, M. Marcus, and S. Reich, eds., Longman Scientific and Technical, Essex, UK, 1992, pp. 119–129.

[9] S. D. Flåm, *Solving convex programs by means of ordinary differential equations*, Math. Oper. Res., 17 (1992), pp. 290–302.

[10] C. C. Gonzaga, *Path-following methods for linear programming*, SIAM Rev., 34 (1992), pp. 167–224.

[11] U. Helmke and J. B. Moore, *Optimization and Dynamical Systems*, Comm. Control Engrg. Ser., Springer-Verlag, London, 1994.

[12] N. N. Karpinskaya, *Method of "penalty" functions and the foundations of Pyne's method*, Automat. Remote Control, 28 (1967), pp. 124–129.

[13] M. P. Kennedy and L. O. Chua, *Neural networks for nonlinear programming*, IEEE Trans. Ciruits Systems, 35 (1988), pp. 554–562.

[14] S. K. Korovin and V. I. Utkin, *Using sliding modes in static optimization and nonlinear programming*, Automatica, 10 (1974), pp. 525–532.

[15] W. E. Lillo, S. Hui, and S. H. Żak, *Neural networks for constrained optimization problems*, Internat. J. Circuit Theory Appl., 21 (1993), pp. 385–399.

[16] W. E. Lillo, M. H. Loh, S. Hui, and S. H. Żak, *On solving constrained optimization problems with neural networks: A penalty method approach*, IEEE Trans. Neural Networks, 4 (1993), pp. 931–940.

[17] D. G. Luenberger, *Linear and Nonlinear Programming*, 2nd ed., Addison-Wesley, Reading, MA, 1984.

[18] I. B. Pyne, *Linear programming on an electronic analogue computer*, Trans. Amer. Inst. Electrical Engineers, 75 (1956), pp. 139–143.

[19] A. Rodríguez-Vázquez, R. Domínguez-Castro, A. Rueda, J. L. Huertas, and E. Sánchez-Sinencio, *Nonlinear switched-capacitor "neural" networks for optimization problems*, IEEE Trans. Circuits Systems, 37 (1990), pp. 384–398.

[20] J. B. Rosen, *The gradient projection method for nonlinear programming, Part* II*: Nonlinear constraints*, J. Soc. Indust. Appl. Math., 9 (1961), pp. 514–532.

[21] M. V. Rybashov, *The gradient method of solving convex programming problems on electronic analog computers*, Automat. Remote Control, 26 (1965), pp. 1886–1898.

[22] M. V. Rybashov, *Gradient method of solving linear and quadratic programming problems on electronic analog computers*, Automat. Remote Control, 26 (1965), pp. 2079–2089.

[23] D. W. Tank and J. J. Hopfield, *Simple 'neural' optimization networks: An A/D converter, signal decision circuit, and a linear programming circuit*, IEEE Trans. Circuits Systems, CAS-36 (1986), pp. 533–541.

[24] V. I. Utkin, *Sliding Modes and Their Application in Variable Structure Systems*, Mir Publishers, Moscow, 1978.

[25] V. I. Utkin, *Sliding Modes in Control and Optimization*, Springer-Verlag, Berlin, Heidelberg, 1992.

[26] M. Vidyasagar, *Minimum-seeking properties of analog neural networks with multilinear objective functions*, IEEE Trans. Automat. Control, 40 (1995), pp. 1359–1375.

[27] S. H. Żak, V. Upatising, W. E. Lillo, and S. Hui, *A dynamical systems approach to solving linear programming problems*, in Differential Equations, Dynamical Systems, and Control Science: A Festschrift in Honor of Lawrence Markus, Chap. 54, Lecture Notes in Pure and Appl. Math. 152, K. D. Elworthy, W. N. Everitt, and E. B. Lee, eds., Marcel Dekker, New York, 1994, pp. 913–946.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 2, pp. 698–718, March 1998

011

LIPSCHITZIAN STABILITY FOR STATE CONSTRAINED NONLINEAR OPTIMAL CONTROL*

A. L. DONTCHEV† AND W. W. HAGER‡

Abstract. For a nonlinear optimal control problem with state constraints, we give conditions under which the optimal control depends Lipschitz continuously in the L^2 norm on a parameter. These conditions involve smoothness of the problem data, uniform independence of active constraint gradients, and a coercivity condition for the integral functional. Under these same conditions, we obtain a new nonoptimal stability result for the optimal control in the L^∞ norm. And under an additional assumption concerning the regularity of the state constraints, a new tight L^∞ estimate is obtained. Our approach is based on an abstract implicit function theorem in nonlinear spaces.

Key words. optimal control, state constraints, Lipschitzian stability, implicit function theorem

AMS subject classifications. 49K40, 90C31

PII. S0363012996299314

1. Introduction.

We consider the following optimal control problem involving a parameter:

$$\text{minimize} \int_0^1 h_p(x(t), u(t))dt \tag{1}$$

subject to

$$\dot{x}(t) = f_p(x(t), u(t)) \text{ a.e. } t \in [0,1], \quad x(0) = x^0,$$
$$g_p(x(t)) \le 0 \text{ for all } t \in [0,1], \ \ u \in L^\infty, \ \ x \in W^{1,\infty},$$

where the state $x(t) \in \mathbf{R}^n$, $\dot{x} \equiv \frac{d}{dt}x$, the control $u(t) \in \mathbf{R}^m$, the parameter p lies in a metric space, the functions $h_p : \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}$, $f_p : \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}^n$, and $g_p : \mathbf{R}^n \to \mathbf{R}^k$. Throughout the paper, $L^\alpha(J; \mathbf{R}^m)$ denotes the usual Lebesgue space of functions $u : J \to \mathbf{R}^m$ with $|u(\cdot)|^\alpha$ integrable, equipped with its standard norm

$$\| u \|_{L^\alpha} = \left(\int_J |u(t)|^\alpha dt \right)^{1/\alpha},$$

where $|\cdot|$ is the Euclidean norm. Of course, $\alpha = \infty$ corresponds to the space of essentially bounded functions. Let $W^{m,\alpha}(J; \mathbf{R}^n)$ be the usual Sobolev space consisting of vector-valued functions whose jth derivative lies in L^α for all $0 \le j \le m$; its norm is

$$\| u \|_{W^{m,\alpha}} = \sum_{j=0}^{m} \| u^{(j)} \|_{L^\alpha} .$$

When either the domain J or the range $\mathbf{R}^n$ is clear from context, it is omitted. We let H^m denote the space $W^{m,2}$, and Lip denote $W^{1,\infty}$, the space of Lipschitz continuous

*Received by the editors February 26, 1996; accepted for publication (in revised form) January 21, 1997. This research was supported by National Science Foundation grant DMS 9404431.
http://www.siam.org/journals/sicon/36-2/29931.html
†Mathematical Reviews, University of Michigan, Ann Arbor, MI 48107 (ald@math.ams.org).
‡Department of Mathematics, University of Florida, Gainesville, FL 32611-8105 (hager@math.ufl.edu).

functions. Subscripts on spaces are used to indicate bounds on norms; in particular, $W^{m,\alpha}_{\kappa}$ denotes the set of functions in $W^{m,\alpha}$ with the property that the L^α norm of the mth derivative is bounded by κ, and Lip_κ denotes the space of Lipschitz continuous functions with Lipschitz constant κ. Throughout, c is a generic constant, independent of the parameter p and time t, and $B_a(x)$ is the closed ball centered at x with radius a. The L^2 inner product is denoted $\langle\cdot,\cdot\rangle$, the complement of a set $\mathcal{A}$ is $\mathcal{A}^c$, and the transpose of a matrix B is B^T. Given a vector $y \in \mathbf{R}^m$ and a set $\mathcal{A} \subset \{1, 2, \ldots, m\}$, $y_\mathcal{A}$ denotes the subvector consisting of components associated with indices in $\mathcal{A}$. And if $Y \in \mathbf{R}^{m\times n}$, then $Y_\mathcal{A}$ is the submatrix consisting of rows associated with indices in $\mathcal{A}$.

We wish to study how a solution to either (1) or the associated variational system representing the first-order necessary condition depends on the parameter p. We assume that the problem (1) has a local minimizer $(x, u) = (x_*, u_*)$ corresponding to a reference value $p = p_*$ of the parameter, and the following smoothness condition holds.

Smoothness. The local minimizer (x_*, u_*) of (1) lies in $W^{2,\infty}\times$ Lip. There exists a closed set $\Delta \subset \mathbf{R}^n \times \mathbf{R}^m$ and a $\delta > 0$ such that $B_\delta(x_*(t), u_*(t)) \subset \Delta$ for every $t \in [0, 1]$. The function values and first two derivatives of $f_p(x, u)$, $g_p(x, u)$, and $h_p(x, u)$, and the third derivatives of $g_p(x)$, with respect to x and u, are uniformly continuous relative to p near p_* and $(x, u) \in \Delta$. And when either the first two derivatives of $f_p(x, u)$ and $h_p(x, u)$ or the first three derivatives of $g_p(x)$, with respect to x and u, are evaluated at (x_*, u_*), the resulting expression is differentiable in t, and the L^∞ norm of the time derivative is uniformly bounded relative to p near p_*.

Let A, B, and K be the matrices defined by

$$A = \nabla_x f_*(x_*, u_*), \quad B = \nabla_u f_*(x_*, u_*), \quad \text{and } K = \nabla_x g_*(x_*).$$

Here and elsewhere the * subscript is always associated with p_*. Let $\mathcal{A}(t)$ be the set of indices of the active constraints at $(x_*(t), p_*)$; that is,

$$\mathcal{A}(t) = \{i \in \{1, 2, \cdots, k\} : g_*(x_*(t))_i = 0\}.$$

We introduce the following assumption.

Uniform independence at $\mathcal{A}$. The set $\mathcal{A}(0)$ is empty and there exists a scalar $\alpha > 0$ such that

$$\left| \sum_{i\in\mathcal{A}(t)} v_i K_i(t) B(t) \right| \geq \alpha |v_{\mathcal{A}(t)}|$$

for each $t \in [0, 1]$ where $\mathcal{A}(t) \neq \emptyset$ and for each choice of v.

Uniform independence implies that the state constraints are first-order (see [12] for the definition of the order of a state constraint). This condition can be generalized to higher order state constraints (see Maurer [17]), however, the generalization of the stability results in this paper to higher order state constraints is not immediate.

It is known (see, for instance, Theorem 7.1 of the recent survey [12] and the regularity analysis in [8]) that under appropriate assumptions, the first-order necessary conditions (Pontryagin's minimum principle) associated with a solution (x_*, u_*) of (1) can be written in the following way. There exist $\psi_* \in W^{2,\infty}$ and $\nu_* \in$ Lip such that $x = x_*$, $\psi = \psi_*$, $u = u_*$, and $\nu = \nu_*$ are a solution at $p = p_*$ of the variational system:

$$\dot{x} = f_p(x, u), \; x(0) = x^0, \tag{2}$$

$$\dot{\psi} = -\nabla_x H_p(x, \psi, u, \nu), \ \psi(1) = 0, \tag{3}$$

$$0 = \nabla_u H_p(x, \psi, u, \nu), \tag{4}$$

$$g_p(x) \in N(\nu), \ \nu(1) \le 0, \ \dot{\nu} \ge 0 \ \text{ a.e.} \tag{5}$$

Here H_p is the Hamiltonian defined by

$$H_p(x, \psi, u, \nu) = h_p(x, u) + \psi^{\mathsf{T}} f_p(x, u) - \nu^{\mathsf{T}} \nabla g_p(x) f_p(x, u),$$

and the set-valued map N is defined in the following way: given a nondecreasing Lipschitz continuous function ν, a continuous function y lies in $N(\nu)$ if and only if

$$y(t) \le 0, \ \dot{\nu}(t)^{\mathsf{T}} y(t) = 0 \text{ for a.e. } t \in [0, 1], \ \text{ and } \ \nu(1)^{\mathsf{T}} y(1) = 0.$$

Defining

$$Q = \nabla_{xx} H_*(w_*), \ \ M = \nabla_{xu} H_*(w_*), \text{ and } R = \nabla_{uu} H_*(w_*),$$

where $w_* = (x_*, \psi_*, u_*, \nu_*)$, let $\mathcal{B}$ be the quadratic form

$$\mathcal{B}(x, u) = \frac{1}{2} \int_0^1 x(t)^{\mathsf{T}} Q(t) x(t) + u(t)^{\mathsf{T}} R(t) u(t) + 2x(t)^{\mathsf{T}} M(t) u(t) dt,$$

and let L be the linear and continuous operator from $H^1 \times L^2$ to L^2 defined by $L(x, u) = \dot{x} - Ax - Bu$. We introduce the following growth assumption for the quadratic form.

Coercivity. There exists a constant $\alpha > 0$ such that

$$\mathcal{B}(x, u) \ \ge \ \alpha \langle u, u \rangle \quad \text{for all } \ (x, u) \in \mathcal{M},$$

where

$$\mathcal{M} = \{(x, u) : x \in H^1, \ u \in L^2, L(x, u) = 0, \ x(0) = 0\}. \tag{6}$$

In the terminology of [12], the form of the minimum principle we employ is the "indirect adjoining approach with continuous adjoint function." A different approach, found in [13], for example, involves a different choice for the multipliers and for the Hamiltonian. The multipliers in these two approaches are related in a linear fashion as shown in [11]. Normally, the multiplier ν, associated with the state constraint, and the derivative of ψ have bounded variation. In our statement of the minimum principle above, we are implicitly assuming some additional regularity so that ν and $\dot{\psi}$ are not only of bounded variation, but Lipschitz continuous. This regularity can be proved under the uniform independence and coercivity conditions (see [8]).

In section 3 we establish the following result.

THEOREM 1.1. *Suppose that the problem* (1) *with* $p = p_*$ *has a local minimizer* (x_*, u_*) *and that the smoothness and the uniform independence conditions hold. Let* ψ_* *and* ν_* *be the associated multipliers satisfying the variational system* (2)–(5) *with* $\psi_* \in W^{2,\infty}$ *and* $\nu_* \in$ *Lip. If the coercivity condition holds, then there exist a constant* μ *and neighborhoods* V *of* p_* *and* U *of* $w_* = (x_*, \psi_*, u_*, \nu_*)$ *in* $W^{1,\infty} \times W^{1,\infty} \times L^\infty \times L^\infty$, *such that for every* $p \in V$, *there is a unique solution* $w = (x, \psi, u, \nu) \in U$ *to the first-order necessary conditions* (2)–(5) *with the property that* $(\dot{x}, \dot{\psi}, u, \nu) \in \mathrm{Lip}_\mu$ *and* (x, u) *is a local minimizer of the problem* (1) *associated with* p. *Moreover, for every* $p_i \in V, i = 1, 2$, *if* $w_i = (x_i, \psi_i, u_i, \nu_i)$ *is the corresponding solution of* (2)–(5), *the following estimate holds:*

$$\|x_1 - x_2\|_{H^1} + \|\psi_1 - \psi_2\|_{H^1} + \|u_1 - u_2\|_{L^2} + \|\nu_1 - \nu_2\|_{L^2} \le c E_2, \tag{7}$$

where

$$E_\alpha = \|f_{p_1}(x_1,u_1) - f_{p_2}(x_1,u_1)\|_{L^\alpha} + \|\nabla_x H_{p_1}(w_1) - \nabla_x H_{p_2}(w_1)\|_{L^\alpha} + \|\nabla_u H_{p_1}(w_1) - \nabla_u H_{p_2}(w_1)\|_{L^\alpha} + \|g_{p_1}(x_1) - g_{p_2}(x_1)\|_{W^{1,\alpha}}.$$

In addition, we have

$$\|x_1 - x_2\|_{W^{1,\infty}} + \|\psi_1 - \psi_2\|_{W^{1,\infty}} + \|u_1 - u_2\|_{L^\infty} + \|\nu_1 - \nu_2\|_{L^\infty} \le cE_2^{2/3}.$$

The proof of Theorem 1.1 is based on an abstract implicit function theorem appearing in section 2. In section 4 we show that the L^∞ estimate of Theorem 1.1 can be sharpened if the points where the state constraints change between active and inactive are separated. In section 5 we comment briefly on related work.

2. An implicit function theorem in nonlinear spaces. The following lemma provides a generalization of the implicit function theorem that can be applied to nonlinear spaces. To simplify the notation, we let $\|x - y\|_X$ denote the distance between the elements x and y of the metric space X.

LEMMA 2.1. *Let X and Π be metric spaces with X complete, let Y be a subset of Π, and let P be a set. Given $w_* \in X$ and $r > 0$, let W denote the ball $B_r(w_*)$ in X and suppose that $T : W \times P \to Y$ and $F : X \to 2^\Pi$ (the subsets of Π) have the following properties.*

(P1) *$T(w_*, p_*) \in F(w_*)$ for some $p_* \in P$.*

(P2) *For some $\beta > 0$, $\|T(w_*, p_*) - T(w_*, p)\|_\Pi \le \beta$ for all $p \in P$.*

(P3) *For some $\epsilon > 0$, $\|T(w_1, p) - T(w_2, p)\|_\Pi \le \epsilon \|w_1 - w_2\|_X$ for all $w_1, w_2 \in W$ and $p \in P$.*

(P4) *F^{-1} restricted to Y is single-valued and Lipschitz continuous, with Lipschitz constant λ.*

If $\epsilon\lambda < 1$ and $r \ge \lambda\beta/(1 - \epsilon\lambda)$, then for each $p \in P$, there exists a unique $w \in W$ such that $T(w, p) \in F(w)$. Moreover, for every $p_i \in P, i = 1, 2$, if w_i denotes the w associated with p_i, then we have

$$\|w_1 - w_2\|_X \le \frac{\lambda}{1 - \lambda\epsilon} \|T(w_1, p_1) - T(w_1, p_2)\|_\Pi. \tag{8}$$

Proof. Fix $p \in P$ and define $\Phi(w) = F^{-1}(T(w, p))$ for $w \in W$. Observe that

$$\begin{aligned}\|\Phi(w_1) - \Phi(w_2)\|_X &= \|F^{-1}(T(w_1, p)) - F^{-1}(T(w_2, p))\|_X \\ &\le \lambda \|T(w_1, p) - T(w_2, p)\|_\Pi \le \lambda\epsilon \|w_1 - w_2\|_X\end{aligned}$$

for each $w_1, w_2 \in W$. Since $\lambda\epsilon < 1$, Φ is a contraction on W with contraction constant $\lambda\epsilon$. Let $w \in W$. Since $w_* = F^{-1}(T(w_*, p_*))$ and $r \ge \lambda\beta/(1 - \epsilon\lambda)$, we have

$$\begin{aligned}\|w_* - \Phi(w)\|_X &= \|F^{-1}(T(w_*, p_*)) - F^{-1}(T(w, p))\|_X \\ &\le \lambda(\|T(w, p) - T(w_*, p)\|_\Pi + \|T(w_*, p) - T(w_*, p_*)\|_\Pi) \\ &\le \lambda(\epsilon r + \beta) \le r.\end{aligned}$$

Thus Φ maps W into itself. By the Banach contraction mapping principle, there exists a unique $w \in W$ such that $w = \Phi(w)$. Since $w = \Phi(w)$ is equivalent to $T(w, p) \in F(w)$ for $w \in W$, we conclude that for each $p \in P$, there is a unique $w(p) \in W$ such that

$T(w(p),p) \in F(w(p))$. Defining $w_1 = w(p_1)$ and $w_2 = w(p_2)$, we have

$$\begin{aligned}\|w_1 - w_2\|_X &= \|F^{-1}(T(w_1,p_1)) - F^{-1}(T(w_2,p_2))\|_X \\ &\le \lambda\|T(w_1,p_1) - T(w_2,p_2)\|_\Pi \\ &\le \lambda\|T(w_1,p_1) - T(w_1,p_2)\|_\Pi + \lambda\|T(w_1,p_2) - T(w_2,p_2)\|_\Pi \\ &\le \lambda\|T(w_1,p_1) - T(w_1,p_2)\|_\Pi + \lambda\epsilon\|w_1 - w_2\|_X.\end{aligned}$$

Rearranging this inequality, the proof is complete. ∎

Let X, Y, and $\mathcal{P}$ be metric spaces and let $w_* \in X$. Using the terminology of [3], $f : X \times \mathcal{P} \to Y$ is strictly stationary at $w = w_*$, uniformly in p near p_*, if for each $\epsilon > 0$, there exists $\delta > 0$ with the property that

$$\|f(w_1,p) - f(w_2,p)\|_Y \le \epsilon\|w_1 - w_2\|_X$$

for all $w_1, w_2 \in B_\delta(w_*)$ and $p \in B_\delta(p_*)$.

THEOREM 2.2. *Let X be a complete metric space, let Π be a linear metric space, let Y be a subset of Π, and let $\mathcal{P}$ be a metric space. Suppose that $\mathcal{F} : X \to 2^\Pi$, that $\mathcal{T} : X \times \mathcal{P} \to \Pi$, that $\mathcal{L} : X \to \Pi$ is continuous, and that for some $w_* \in X$ and $p_* \in \mathcal{P}$ we have:*

(Q1) *$\mathcal{T}(w_*, p_*) \in \mathcal{F}(w_*)$;*
(Q2) *$\mathcal{T}(w_*, \cdot\,)$ is continuous at p_*;*
(Q3) *$\mathcal{T}(w,p) - \mathcal{L}(w)$ is strictly stationary at $w = w_*$, uniformly in p near p_*;*
(Q4) *$(\mathcal{F} - \mathcal{L})^{-1}$ restricted to Y is single-valued and Lipschitz continuous, with Lipschitz constant λ;*
(Q5) *$\mathcal{T} - \mathcal{L}$ maps a neighborhood of (w_*, p_*) into Y.*

Then for each $\lambda_+ > \lambda$, there exist neighborhoods W of w_* and P of p_* such that for each $p \in P$, a unique $w \in W$ exists satisfying $\mathcal{T}(w,p) \in \mathcal{F}(w)$; moreover, for every $p_i \in P, i = 1,2$, if w_i denotes the $w \in W$ associated with p_i, then we have

$$\|w_1 - w_2\|_X \le \lambda_+\|\mathcal{T}(w_1,p_1) - \mathcal{T}(w_1,p_2)\|_\Pi.$$

Proof. By (Q5) there exist neighborhoods U' of w_* and P' of p_* such that $\mathcal{T}(w,p) - \mathcal{L}(w) \in Y$ for each $w \in U'$ and $p \in P'$. We apply Lemma 2.1 with the following identifications: X, Y, and Π are as defined in the statement of the theorem, $F(w) = \mathcal{F}(w) - \mathcal{L}(w)$, and $T(w,p) = \mathcal{T}(w,p) - \mathcal{L}(w)$. (P1) and (P4) follow immediately from (Q1) and (Q4), respectively. Choose $\epsilon > 0$ such that $\epsilon < (\lambda_+ - \lambda)/(\lambda_+\lambda)$. Since $\lambda_+ > \lambda$, it follows that for this choice of ϵ, we have $\epsilon\lambda < 1$ and $\lambda/(1-\lambda\epsilon) < \lambda_+$. By (Q3) and the identity $T(w_1,p_1) - T(w_1,p_2) = \mathcal{T}(w_1,p_1) - \mathcal{T}(w_1,p_2)$, there exist neighborhoods $P = B_r(p_*) \subset P'$ of p_* and $W = B_r(w_*) \subset U'$ of w_* such that (P3) of Lemma 2.1 holds. Let β satisfy $\lambda\beta/(1-\epsilon\lambda) \le r$, and by (Q2), choose P smaller if necessary so that (P2) holds. By Lemma 2.1, for each $p \in P$, there exists a unique $w \in W$ such that $T(w,p) \in F(w)$, and the estimate (8) holds. Since $T(w,p) \in F(w)$ if and only if $\mathcal{T}(w,p) \in \mathcal{F}(w)$, the proof is complete. ∎

A particular case of Theorem 2.2 corresponds to the well-known Robinson implicit function theorem [20] in which X is a Banach space, Π is its dual X^*, $\mathcal{F}(w) = N_\Omega(w)$, Ω is a closed, convex set in X, $N_\Omega(w)$ is the normal cone to the set Ω at the point w, $\mathcal{T}$ is differentiable with respect to w, both $\mathcal{T}$ and its derivative $\nabla_w\mathcal{T}$ are continuous in a neighborhood of (w_*, p_*), and $\mathcal{L}(w) = \mathcal{T}(w_*,p_*) + \nabla_w\mathcal{T}(w_*,p_*)(w - w_*)$ is the linearization of $\mathcal{T}$. The Robinson framework is applicable to control problems with control constraints after the range space X^* is replaced by a general Banach

space Y (see the discussion in section 5). However, for problems with state constraints, there are difficulties in applying Robinson's theory since stability results for state constrained quadratic problems, analogous to the results for control constrained problems, have not been established. In our previous paper [3], we extend Robinson's work in several different directions. For the solution map of a generalized equation in a linear metric space, we showed that Aubin's pseudo-Lipschitz property, that the existence of a Lipschitzian selection, and that local Lipschitzian invertibility are "robust" under nonlinear perturbations that are strictly stationary at the reference point. In Theorem 2.2, we focus on the latter property, giving an extension of our earlier result to nonlinear spaces. In this nonlinear setting, we are able to analyze the state constrained problem, obtaining a Lipschitzian stability result for the solution.

3. Lipschitzian stability in L^2. To prove Theorem 1.1, we apply Theorem 2.2 using the following identifications. First, we define

$$w = (x, \psi, u, \nu), \tag{9}$$

where

$$x,\ \psi \in W^{2,\infty}_{\mu} \text{ (with the } H^1 \text{ norm)},\ x(0) = x^0,\ \psi(1) = 0, \tag{10}$$

$$u,\ \nu \in \mathrm{Lip}_{\mu} \text{ (with the } L^2 \text{ norm)},\ \nu(1) \le 0 \text{ and } \dot{\nu} \ge 0 \text{ a.e.} \tag{11}$$

An appropriate value for μ is chosen later in the analysis. The space X consists of the collection of functions x, ψ, u, and ν satisfying (10) and (11) with the norm defined in (10) and (11). Observe that the norms we use are not the natural norms. For example, the u and ν components of elements in X lie in $W^{1,\infty}$, but we use the L^2 norm to measure distance. Despite the apparent mismatch of space and norm, X is complete by Lemma 3.2 below.

The functions $\mathcal{T}$ and $\mathcal{F}$ of Theorem 2.2 are selected in the following way:

$$\mathcal{T}(w,p) = \begin{pmatrix} \dot{x} - f_p(x,u) \\ \dot{\psi} + \nabla_x H_p(x,u,\psi,\nu) \\ \nabla_u H_p(x,u,\psi,\nu) \\ g_p(x) \end{pmatrix} \text{ and } \mathcal{F}(w) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ N(\nu) \end{pmatrix}. \tag{12}$$

The continuous operator $\mathcal{L}$ is obtained by linearizing the map $\mathcal{T}(\cdot, p_*)$ in L^∞ at the reference point $w_* = (x_*, \psi_*, u_*, \nu_*)$. In particular,

$$\mathcal{L}(w) = \begin{pmatrix} \dot{x} - Ax - Bu \\ \dot{\psi} + A^{\mathsf{T}}\psi + Qx + Mu - (\dot{K}^{\mathsf{T}} + A^{\mathsf{T}}K^{\mathsf{T}})\nu \\ Ru + M^{\mathsf{T}}x + B^{\mathsf{T}}\psi - B^{\mathsf{T}}K^{\mathsf{T}}\nu \\ Kx \end{pmatrix}. \tag{13}$$

Defining $\pi_* = \mathcal{T}(w_*, p_*) - \mathcal{L}(w_*)$, let a_*, s_*, r_*, and b_* denote the components of π_*:

$$\begin{aligned} a_* &= -f_*(x_*, u_*) + Ax_* + Bu_*, \\ s_* &= \nabla_x H_*(w_*) - A^{\mathsf{T}}\psi_* - Qx_* - Mu_* + (\dot{K}^{\mathsf{T}} + A^{\mathsf{T}}K^{\mathsf{T}})\nu_*, \\ r_* &= \nabla_u H_*(w_*) - Ru_* - M^{\mathsf{T}}x_* - B^{\mathsf{T}}\psi_* + B^{\mathsf{T}}K^{\mathsf{T}}\nu_*, \\ b_* &= g_*(x_*) - Kx_*. \end{aligned}$$

The space Π is the product $L^2 \times L^2 \times L^2 \times H^1$, while the elements π in Y have the form $\pi = (a, s, r, b)$, where

$$\begin{aligned} &a,\ s,\ r\ \in \mathrm{Lip} \text{ (with the } L^2 \text{ norm)},\ b \in W^{2,\infty} \text{ (with the } H^1 \text{ norm)}, \\ &\|a - a_*\|_{W^{1,\infty}} + \|r - r_*\|_{W^{1,\infty}} + \|s - s_*\|_{W^{1,\infty}} + \|b - b_*\|_{W^{2,\infty}} \le \kappa, \end{aligned} \tag{14}$$

where κ is a small positive constant chosen so that two related quadratic programs, (37) and (41), introduced later have the same solution. As we will see, the constant μ associated with the space X must be chosen sufficiently large relative to κ. Note that the inverse $(\mathcal{F}-\mathcal{L})^{-1}\pi$ is the solution (x,ψ,u,ν) of the linear variational system:

$$\dot{x} = Ax + Bu - a,\ x(0) = x^0, \tag{15}$$

$$\dot{\psi} = -A^{\mathsf{T}}\psi - Qx - Mu + (\dot{K} + A^{\mathsf{T}}K^{\mathsf{T}})\nu - s,\ \psi(1) = 0, \tag{16}$$

$$0 = Ru + M^{\mathsf{T}}x + B^{\mathsf{T}}\psi - B^{\mathsf{T}}K^{\mathsf{T}}\nu + r, \tag{17}$$

$$Kx + b \in N(\nu),\ \nu(1) \le 0,\ \dot{\nu} \ge 0 \ \text{ a.e.} \tag{18}$$

Referring to the assumptions of Theorem 2.2, (Q1) holds by the definition of X, and by the minimum principle, (Q2) follows immediately from the smoothness condition. In Lemma 3.3, we deduce (Q3) from the smoothness condition and a Taylor expansion. In Lemma 3.6, (Q5) is obtained by showing that for w near w_* and p near p_*, $\mathcal{T}(w,p) - \mathcal{L}(w)$ and its associated derivatives are near those of $\pi_* = \mathcal{T}(w_*,p_*) - \mathcal{L}(w_*)$. Finally, in a series of lemmas, (Q4) is established through manipulations of quadratic programs associated with (15)–(18).

To start the analysis, we show that X is complete using the following lemma.

LEMMA 3.1. *If $u \in \mathrm{Lip}_\mu([0,1];\mathbf{R}^1)$, then we have*

$$\|u\|_{L^\infty} \le \ \max\{\ \sqrt{3}\|u\|_{L^2},\ \sqrt[3]{3\mu}\|u\|_{L^2}^{2/3}\ \}\ .$$

Proof. Since u is continuous, its maximum absolute value is achieved at some time t_m on the interval $[0,1]$. Let $u_m = u(t_m)$ denote the associated value of u. We consider two cases.

Case 1. $u_m > \mu$. Let us examine the maximum ratio between the ∞-norm and the 2-norm:

$$\text{maximize } \{\|u\|_{L^\infty}/\|u\|_{L^2} : \|u\|_{L^\infty} = u_m, u \in \ \mathrm{Lip}_\mu\}.$$

Since $u_m > \mu$, the maximum is attained by the linear function v satisfying $v(0) = u_m$ and $\dot{v} = -\mu$. The 2-norm of this function is readily evaluated:

$$\|v\|_{L^2}^2 = u_m^2(3 - 3\alpha + \alpha^2)/3, \text{ where } \alpha = \mu/u_m.$$

Since $\alpha \in [0,1]$ and since $3 - 3\alpha + \alpha^2 \ge 1$ on this interval, we have $\|v\|_{L^2}^2 \ge u_m^2/3$. Taking square roots gives

$$\|v\|_{L^\infty}/\|v\|_{L^2} \le \sqrt{3},$$

which establishes the lemma in Case 1.

Case 2. $u_m \le \mu$. In this case, let us examine the maximum ratio between the ∞-norm and the 2-norm to the 2/3-power:

$$\text{maximize } \{\|u\|_{L^\infty}/\|u\|_{L^2}^{2/3} : \|u\|_{L^\infty} = u_m, u \in \ \mathrm{Lip}_\mu\}.$$

The maximum is attained by the piecewise linear function v satisfying $v(0) = u_m$, $\dot{v} = -\mu$ on $[0, u_m/\mu]$, and $v = 0$ elsewhere. Since

$$\|v\|_{L^2}^2 = \frac{u_m^3}{3\mu},$$

it follows that

$$\|v\|_{L^\infty}/\|v\|_{L^2}^{2/3} \le \sqrt[3]{3\mu},$$

which completes the proof of Case 2. ∎

LEMMA 3.2. *The space X of functions w satisfying* (9), (10), *and* (11) *is complete.*

Proof. Suppose that $w_k = (x_k, u_k, \psi_k, \nu_k)$ is a Cauchy sequence in X. We analyze the ν-component of w_k. The sequence ν_k is a Cauchy sequence in L^∞ by Lemma 3.1. Since L^∞ is complete, there exists a limit point $\bar\nu \in L^\infty$. Since the ν_k converge pointwise to $\bar\nu$ and since each of the ν_k is Lipschitz continuous with Lipschitz constant μ, $\bar\nu$ is Lipschitz continuous with Lipschitz constant μ. Since each of the ν_k is nondecreasing, it follows from the pointwise convergence that $\bar\nu$ is nondecreasing; hence, $\dot{\bar\nu} \ge 0$. Since $\nu_k(1) \le 0$ for each k, the pointwise convergence implies that $\bar\nu(1) \le 0$. This shows that the ν-component of X is complete. The other components can be analyzed in a similar fashion. ∎

LEMMA 3.3. *If the smoothness condition holds, then for $\mathcal{T}$ and $\mathcal{L}$ defined in* (12) *and* (13), *respectively, $\mathcal{T} - \mathcal{L}$ is strictly stationary at w_*, uniformly in p near p_*.*

Proof. Only the first component of $\mathcal{T}(w,p) - \mathcal{L}(w)$ is analyzed, since the other components are treated in a similar manner. To establish strict stationarity for the first component, we need to show that for any given $\epsilon > 0$,

$$\|(f_p(x,u) - f_p(y,v)) - A(x-y) - B(u-v)\|_{L^2} \le \epsilon\|x-y\|_{H^1} + \epsilon\|u-v\|_{L^2}, \tag{19}$$

for p near p_* and for (x,u) and $(y,v) \in W^{2,\infty}_\mu \times \mathrm{Lip}_\mu$ near (x_*,u_*) in the norm of $H^1 \times L^2$, where $A = \nabla_x f_*(x_*,u_*)$ and $B = \nabla_u f_*(x_*,u_*)$. By Lemma 3.1, (x,u) and (y,v) are also near (x_*,u_*) in L^∞. After writing the difference $f_p(x,u) - f_p(y,v)$ as an integral over the line segment connecting (x,u) and (y,v), we have

$$(f_p(x,u) - f_p(y,v)) - A(x-y) - B(u-v) = (A_p - A)(x-y) + (B_p - B)(u-v),$$

where (A_p, B_p) is the average of the gradient of f_p along the line segment connecting (x,u) and (y,v). By the smoothness condition, $\|A_p - A\|_{L^\infty} \to 0$ and $\|B_p - B\|_{L^\infty} \to 0$ as p approaches p_* and as both (x,u) and (y,v) approach (x_*,u_*) in L^∞. This completes the proof. ∎

LEMMA 3.4. *If the smoothness condition holds, then for $\mathcal{T}$ and $\mathcal{L}$ defined in* (12) *and* (13), *respectively, and for any choice of the parameter $\kappa > 0$ in* (14), *there exists $\delta > 0$ such that $\mathcal{T}(w,p) - \mathcal{L}(w) \in Y$ for all $p \in B_\delta(p_*)$ and $w \in B_\delta(w_*) \cap X$.*

Proof. Again, we focus on the first component of $\mathcal{T} - \mathcal{L}$, since the other components are treated in a similar manner. Referring to the definition of Y, we should show that

$$\|(f_p(x,u) - f_*(x_*,u_*)) - A(x-x_*) - B(u-u_*)\|_{W^{1,\infty}} \le \kappa/4 \tag{20}$$

for p near p_* and for $(x,u) \in W^{2,\infty}_\mu \times \mathrm{Lip}_\mu$ near (x_*,u_*) in the norm of $H^1 \times L^2$. The $W^{1,\infty}$ norm in (20) is composed of two norms, the L^∞ norm of the function values, and the L^∞ norm of the time derivative. By the same expansion used in Lemma 3.3, we obtain the bound

$$\|(f_p(x,u) - f_*(x_*,u_*)) - A(x-x_*) - B(u-u_*)\|_{L^\infty} \le \kappa/8$$

for p near p_* and for (x,u) near (x_*,u_*). Differentiating the expression within the norm of (20) gives

$$\begin{aligned}&\tfrac{d}{dt}\big(f_p(x,u) - f_*(x_*,u_*) - A(x-x_*) - B(u-u_*)\big)\\ &= \big(\nabla_x f_p(x,u) - A\big)\dot x + \big(\nabla_u f_p(x,u) - B\big)\dot u - \dot A(x-x_*) - \dot B(u-u_*).\end{aligned}$$

By the smoothness condition, $\dot{A}$ and $\dot{B}$ lie in L^∞, and by the definition of X, we have $\|\dot{u}\|_{L^\infty} \le \mu$. By the triangle inequality and by Lemma 3.1,

$$\|\dot{x}\|_{L^\infty} \le \|\dot{x}_*\|_{L^\infty} + \|\dot{x} - \dot{x}_*\|_{L^\infty} \le \|\dot{x}_*\|_{L^\infty} + \sqrt[3]{3\mu}\|x - x_*\|_{H^1}^{2/3}$$

for x near x_*. Moreover, by Lemma 3.1 and by the smoothness condition, $\nabla_x f_p(x,u)$ approaches A and $\nabla_u f_p(x,u)$ approaches B in L^∞ as p approaches p_* and (x,u) approaches (x_*,u_*). Hence, for p near p_* and (x,u) near (x_*,u_*), we have

$$\left\| \frac{d}{dt}\big(f_p(x,u) - f_*(x_*,u_*) - A(x-x_*) - B(u-u_*)\big)\right\|_{L^\infty} \le \kappa/8.$$

Analyzing each of the components of $\mathcal{T} - \mathcal{L}$ in this same way, the proof is complete. ∎

We now begin a series of lemmas aimed at verifying (Q4). After a technical result (Lemma 3.5) related to the constraints, a surjectivity property (Lemma 3.6) is established for the linearized constraint mapping. Then we study a quadratic program corresponding to the linear variational system (15)–(18). We show that the solution (Lemma 3.9) and the multipliers (Lemma 3.10) depend Lipschitz continuously on the parameters. And utilizing the solution regularity derived in [8], the solution and the multipliers lie in X for μ sufficiently large.

To begin, let I be any map from $[0,1]$ to the subsets of $\{1,2,\dots,k\}$ with the property that the following sets I_i are closed for every i:

$$I_i = I^{-1}(i) = \{t \in [0,1] : i \in I(t)\}\ .$$

We establish the following decomposition property for the interval $[0,1]$.

LEMMA 3.5. *If uniform independence at I holds, then for every $\alpha', 0 < \alpha' < \alpha$, there exist sets $J_1, J_2, \dots, J_l$, corresponding points $0 = \tau_1 < \tau_2 < \cdots < \tau_{l+1} = 1$, and a positive constant $\rho < \min_i(\tau_{i+1} - \tau_i)$ such that for each $t \in [\tau_i - \rho, \tau_{i+1} + \rho] \cap [0,1]$, we have $I(t) \subset J_i$, and if J_i is nonempty, then*

$$\left| \sum_{j \in J_i} v_j K_j(t) B(t) \right| \ge \alpha' |v_{J_i}| \tag{21}$$

for every choice of v. The set J_1 can always be chosen empty.

Proof. For each $t \in (0,1)$ with $I(t)^c \ne \emptyset$, there exists an open interval O centered at t with $O \subset \cap_{i \in I(t)^c} I_i^c$. If $t = 0$ or 1, then we can choose a half-open interval O, with t the closed end of the interval, such that $O \subset \cap_{i \in I(t)^c} I_i^c$. If $I(t)^c$ is empty, take $O = [0,1]$. For any fixed $t \in [0,1]$ with $I(t) \ne \emptyset$, choose O smaller if necessary so that

$$\left| \sum_{i \in I(t)} v_i K_i(s) B(s) \right| \ge \alpha' |v_{I(t)}| \tag{22}$$

for each $s \in O$ and for each choice of v. Since B and K are continuous, it is possible to choose O in this way. Observe that by the construction of O, we have $I(s) \subset I(t)$ for each $s \in O$ and (22) holds if $I(t)$ is nonempty. Given any interval O on $(0,1)$, let $O_{1/2}$ denote the open interval with the same center but with half the length; for the open intervals associated with $t = 0$ or 1, let $O_{1/2}$ denote the half-open interval with the same endpoint, 0 or 1, but with half the length. The sets $O_{1/2}$ form an open cover

of $[0,1]$. Let $O_1, O_2, \ldots, O_l$ be a finite subcover of $[0,1]$ and let $t_1, t_2, \ldots, t_l$ denote the associated centers of interior intervals, and the closed endpoint of the intervals associated with $t = 0$ or 1. It can be arranged so that no O_i is contained in the union of other elements of the subcover (by discarding these extra sets if necessary). Arrange the indices of the O_i so that the left side of O_i is to the left of the left side of O_{i+1} for each i. Let $\tau_1, \tau_2, \ldots, \tau_{l-1}$ denote the successive left sides of the O_i, and let ρ be 1/4 of the length of the smallest O_i. Defining $J_i = I(t_i)$ for $i \geq 1$, it follows from the construction of the O_i that $I(t) \subset J_i$ and (22) holds for each t in an interval associated with t_i and with length twice that of O_i. Since $(\tau_i, \tau_{i+1}) \subset O_i$, we have (21). By taking ρ smaller if necessary, we can enforce the condition $\rho < \min_i(\tau_{i+1} - \tau_i)$. □

LEMMA 3.6. *If uniform independence at I holds, then for each $a \in L^\infty$ and $b \in W^{1,\infty}$, there exist $x \in W^{1,\infty}$ and $u \in L^\infty$ such that $L(x,u) + a = 0$, $x(0) = x^0$, and*

$$K_j(t)x(t) + b_j(t) = 0 \ \text{ for each } \ j \in I(t), \ \ t \in [0,1]. \tag{23}$$

This (x,u) pair is an affine function of (a,b), and for each $\alpha \geq 1$, there exists a constant $c > 0$ such that

$$\|x_1 - x_2\|_{W^{1,\alpha}} + \|u_1 - u_2\|_{L^\alpha} \leq c(\|a_1 - a_2\|_{L^\alpha} + \|b_1 - b_2\|_{W^{1,\alpha}}) \tag{24}$$

for every $(a_i, b_i) \in L^\infty \times W^{1,\infty}$, $i = 1,2$, where (x_i, u_i) is the pair associated with (a_i, b_i).

Proof. We use the decomposition provided by Lemma 3.5 to enforce the equations

$$\dot{x}(t) - A(t)x(t) - B(t)u(t) + a(t) = 0, \ \ x(0) = x^0, \tag{25}$$

$$K_j(t)x(t) + b_j(t) = 0 \ \text{ for each } \ j \in J_i \setminus J_{i-1}, \ \ t \in [\tau_i + \rho, \tau_{i+1}], \tag{26}$$

$$K_j(t)x(t) + b_j(t) = 0 \ \text{ for each } \ j \in J_i \cap J_{i-1}, \ \ t \in [\tau_i, \tau_{i+1}], \tag{27}$$

$i = 2, 3, \ldots, l$. Since J_1 is empty, (23) holds trivially on $[\tau_1, \tau_2] = [0, \tau_2]$. Suppose that $i > 1$, and let us consider (23) on the interval $[\tau_i, \tau_{i+1}]$. Since $I(t) \subset J_i$ for $t \in [\tau_i, \tau_{i+1}]$, we conclude that any $j \in I(t)$ is contained in either $J_i \cap J_{i-1}$ or $J_i \setminus J_{i-1}$. If $j \in J_i \cap J_{i-1}$, then by (27), (23) holds. If $j \in J_i \setminus J_{i-1}$, then by the construction of the J_i, $j \notin I(t)$ for $t \in [\tau_i, \tau_i + \rho]$. Hence, (26) implies that (23) holds.

Suppose that $j \in J_i$ and let σ_j be any given Lipschitz continuous function. Observe that if

$$K_j(\tau_i)x(\tau_i) + \sigma_j(\tau_i) = 0 \ \text{ and } \ \frac{d}{dt}\big(K_j(t)x(t) + \sigma_j(t)\big) = 0 \ \text{ a.e. } \ t \in [\tau_i, \tau_{i+1}], \tag{28}$$

then $K_j(t)x(t) + \sigma_j(t) = 0$ for all $t \in [\tau_i, \tau_{i+1}]$. Carrying out the differentiation in the second relation of (28) and substituting for $\dot{x}$ using the state equation (25), we obtain a linear equation for u. By Lemma 3.5, this equation has a solution, and for fixed t and x, the minimum norm solution can be written:

$$u(t,x) = M_i(t)[-\dot{\sigma}_{J_i}(t) + K_{J_i}(t)a(t) - \dot{K}_{J_i}(t)x - K_{J_i}(t)A(t)x], \tag{29}$$

where

$$M_i(t) = (K_{J_i}(t)B(t))^{\mathsf{T}}[K_{J_i}(t)B(t)(K_{J_i}(t)B(t))^{\mathsf{T}}]^{-1}. \tag{30}$$

In the special case where J_i is empty, we simply set $u(t,x) = 0$.

These observations show how to construct x and u in order to satisfy (26) and (27). On the initial interval $[0, \tau_2]$, u is simply 0 and x is obtained from (25). Assuming x and u have been determined on the interval $[0, \tau_i]$, their values on $[\tau_i, \tau_{i+1}]$ are obtained in the following way: the control is given in feedback form by (29), where for $j \in J_i \cap J_{i-1}$,

$$\sigma_j(t) = b_j(t) \quad \text{for} \quad t \in [\tau_i, \tau_{i+1}]. \tag{31}$$

For $j \in J_i \setminus J_{i-1}$, $\sigma_j(t) = b_j(t)$ for $t \in [\tau_i + \rho, \tau_{i+1}]$, while σ_j is linear on $[\tau_i, \tau_i + \rho]$ with

$$\sigma_j(\tau_i) = -K_j(\tau_i)x(\tau_i) \quad \text{and} \quad \sigma_j(\tau_i + \rho) = b_j(\tau_i + \rho). \tag{32}$$

With this choice for σ, the first equation in (28) is satisfied, and with x and u given by (25) and (29), respectively, the second equation in (28) is satisfied. Also, by the choice of σ,

$$K_j(t)x(t) + \sigma_j(t) = K_j(t)x(t) + b_j(t) = 0$$

for each $j \in J_i \cap J_{i-1}$ and $t \in [\tau_i, \tau_{i+1}]$, and for each $j \in J_i \setminus J_{i-1}$ and $t \in [\tau_i + \rho, \tau_{i+1}]$. Hence, (26) and (27) hold, which yields (23).

For $j \in J_i$, it follows from the definition of σ that

$$|\dot\sigma_j(t)| \le c\big(|x(\tau_i)| + \|b\|_{W^{1,\infty}}\big) \quad \text{a.e.} \quad t \in [\tau_i, \tau_{i+1}].$$

When u in (29) is inserted in (25) and this bound on $|\dot\sigma_j(t)|$ is taken into account, we obtain by induction that $x \in W^{1,\infty}$ and $u \in L^\infty$. By the equations (25) for the state, (29) for the control, and (31)–(32) for σ, (x, u) is an affine function of (a, b). Moreover, the change $(\delta x, \delta u)$ in the state and control associated with the change $(\delta a, \delta b)$ in the parameters satisfies

$$\|\delta x\|_{W^{1,\alpha}([0,\tau_i])} + \|\delta u\|_{L^\alpha([0,\tau_i])} \le c\big(\|\delta a\|_{L^\alpha([0,\tau_i])} + \|\delta\dot\sigma\|_{L^\alpha([0,\tau_i])}\big), \tag{33}$$

for each i where σ is specified in (31)–(32).

To complete the proof, we need to relate the σ term of (33) to the b term of (24). For $j \in J_i$, $\delta\sigma_j(t) = \delta b_j(t)$ if $t \in [\tau_i + \rho, \tau_{i+1}]$ or if $j \in J_{i-1}$ and $t \in [\tau_i, \tau_i + \rho]$. For $j \in J_i \setminus J_{i-1}$ and $t \in [\tau_i, \tau_i + \rho]$, we have

$$\begin{aligned} |\delta\dot\sigma_j(t)| &\le \big(|\delta b_j(\tau_i + \rho)| + |K_j(\tau_i)\delta x(\tau_i)|\big)/\rho \le c\big(\|\delta b\|_{L^\infty} + |\delta x(\tau_i)|\big) \\ &\le c\big(\|\delta b\|_{W^{1,\alpha}} + |\delta x(\tau_i)|\big). \end{aligned}$$

Consequently, for almost every $t \in [\tau_i, \tau_{i+1}]$,

$$|\delta\dot\sigma(t)| \le c\big(\|\delta b\|_{W^{1,\alpha}} + |\delta\dot b(t)| + |\delta x(\tau_i)|\big). \tag{34}$$

Since $\delta x(0) = 0$, let us proceed by induction and assume that

$$|\delta x(\tau_i)| \le c(\|\delta a\|_{L^\alpha} + \|\delta b\|_{W^{1,\alpha}}) \quad \text{for} \quad i = 1,\ 2, \ldots, j.$$

Combining this with (34) and (33) for $i = j + 1$ gives

$$\|\delta x\|_{W^{1,\alpha}([0,\tau_{j+1}])} + \|\delta u\|_{L^\alpha([0,\tau_{j+1}])} \le c\ \big(\ \|\delta a\|_{L^\alpha} + \|\delta b\|_{W^{1,\alpha}}\ \big)\ .$$

Since $|\delta x(\tau_{j+1})| \le \|\delta x\|_{W^{1,\alpha}([0,\tau_{j+1}])}$, the induction step is complete. ∎

In the following lemma, we prove a pointwise coercivity result for the quadratic form $\mathcal{B}$. See [4] and [7] for more general results of this nature.

LEMMA 3.7. *If coercivity holds, then there exists a scalar $\alpha > 0$ such that*

$$\mathcal{B}(x,u) \geq \alpha[\langle x,x\rangle + \langle u,u\rangle + \langle \dot{x},\dot{x}\rangle] \quad \text{for all} \quad (x,u) \in \mathcal{M} \tag{35}$$

and

$$v^{\mathsf{T}} R(t) v \geq \alpha v^{\mathsf{T}} v \text{ for every } t \in [0,1] \text{ and } v \in \mathbf{R}^m. \tag{36}$$

Proof. If $(x,u) \in \mathcal{M}$, then $L(x,u) = 0$ and $x(0) = 0$. Hence, the L^2 norm of x and $\dot{x}$ are bounded in terms of the L^2 norm of u, and (35) follows directly from the coercivity condition. To establish (36), we consider the control u_ϵ defined by

$$u_\epsilon(s) = \begin{cases} v \text{ for } t - \epsilon/2 \leq s \leq t + \epsilon/2, \\ 0 \text{ otherwise.} \end{cases}$$

Let the state x_ϵ be the solution to $L(x_\epsilon, u_\epsilon) = 0$, $x_\epsilon(0) = 0$. For any $t \in (0,1)$, we have

$$\lim_{\epsilon \to 0} \frac{\mathcal{B}(x_\epsilon, u_\epsilon)}{\epsilon} = v^{\mathsf{T}} R(t) v \text{ and } \lim_{\epsilon \to 0} \frac{\langle u_\epsilon, u_\epsilon\rangle}{\epsilon} = v^{\mathsf{T}} v.$$

Combining this with the coercivity condition gives (36). ∎

Consider the following linear-quadratic problem involving the parameters a, s, $r \in L^\infty$ and $b \in W^{1,\infty}$:

$$\begin{aligned} &\text{minimize } \mathcal{B}(x,u) + \langle s,x\rangle + \langle r,u\rangle \\ &\text{subject to} \\ &\qquad L(x,u) + a = 0, \quad x(0) = x^0, \\ &\qquad K_{I(t)}(t)x(t) + b_{I(t)}(t) \leq 0 \text{ for all } t \in [0,1], \\ &\qquad x \in W^{1,\infty}([0,1];\mathbf{R}^n), \quad u \in L^\infty([0,1];\mathbf{R}^m). \end{aligned} \tag{37}$$

If the feasible set for (37) is nonempty, then coercivity implies the existence of a unique minimizer over $H^1 \times L^2$. Using the following lemma, we show that this minimizer lies in $W^{1,\infty} \times L^\infty$, and that it exhibits stability relative to the L^2 norm.

LEMMA 3.8. *If coercivity and uniform independence at I hold, then* (37) *has a unique solution for every $a, r, s \in L^\infty$ and $b \in W^{1,\infty}$. Moreover, the change $(\delta x, \delta u)$ in the solution to* (37) *corresponding to a change $(\delta a, \delta b, \delta s, \delta r)$ in the parameters satisfies the estimate*

$$\|\delta x\|_{H^1} + \|\delta u\|_{L^2} \leq c(\|\delta a\|_{L^2} + \|\delta b\|_{H^1} + \|\delta s\|_{L^2} + \|\delta r\|_{L^2}). \tag{38}$$

Proof. By Lemma 3.6, uniform independence at I implies that the feasible set for (37) is nonempty, while the coercivity condition implies the existence of a unique solution (x_*, u_*) in $H^1 \times L^2$. From duality theory (for example, see [10]), there exists $\lambda \in L^\infty$ with the property that $u = u_*$ is the minimum with respect to u of the expression

$$\mathcal{B}(x,u) + \langle s,x\rangle + \langle r,u\rangle + \langle \lambda, \dot{x} - Ax - Bu + a\rangle$$

over all $u \in L^\infty$. It follows that

$$R(t)u_*(t) + M(t)^{\mathsf{T}} x_*(t) + r(t) - B(t)^{\mathsf{T}} \lambda(t) = 0, \tag{39}$$

and by (36), $u_*(t)$ is uniformly bounded in t. From the equations $L(x_*, u_*) = 0$ and $x_*(0) = x^0$, $x_* \in W^{1,\infty}$.

The estimate (38) can be obtained, as in Lemma 5 in [2], by eliminating the perturbation in the constraints. Let Λ be the affine map in Lemma 3.6 relating the feasible pair (x, u) to the parameters (a, b). By making the substitution $(x, u) = (y, v) + \Lambda(a, b)$, we transform (37) to an equivalent problem of the form

$$\begin{aligned} &\text{minimize } \mathcal{B}(y, v) + \langle \sigma, y\rangle + \langle \rho, v\rangle \\ &\text{subject to} \\ &\qquad L(y, v) = 0, \quad y(0) = 0, \\ &\qquad K_{I(t)}(t)y(t) \le 0 \text{ for all } t \in [0, 1], \\ &\qquad y \in W^{1,\infty}([0, 1]; \mathbf{R}^n), \quad v \in L^\infty([0, 1]; \mathbf{R}^m). \end{aligned} \tag{40}$$

Here σ and ρ are affine functions of a, b, s, and r. Utilizing the coercivity condition and the analysis of [9, section 2], we obtain the following estimate for the change $(\delta y, \delta v)$ corresponding to the change $(\delta\sigma, \delta\rho)$:

$$\begin{aligned} \alpha(\|\delta y\|_{H^1}^2 + \|\delta v\|_{L^2}^2) &\le \|\delta\sigma\|_{L^1}\|\delta y\|_{L^\infty} + \|\delta\rho\|_{L^2}\|\delta v\|_{L^2} \\ &\le \|\delta\sigma\|_{L^1}\|\delta y\|_{H^1} + \|\delta\rho\|_{L^2}\|\delta v\|_{L^2}. \end{aligned}$$

Hence,

$$\|\delta y\|_{H^1} + \|\delta v\|_{L^2} \le c(\|\delta\sigma\|_{L^1} + \|\delta\rho\|_{L^2}).$$

Taking into account the relations between (x, u), (y, v), (σ, ρ), and (a, b, s, r), the proof is complete. □

Now let us consider the full linear-quadratic problem where the subscript I on the state constraint has been removed:

$$\begin{aligned} &\text{minimize } \mathcal{B}(x, u) + \langle s, x\rangle + \langle r, u\rangle \\ &\text{subject to} \\ &\qquad L(x, u) + a = 0, \quad x(0) = x^0, \\ &\qquad K(t)x(t) + b(t) \le 0 \text{ for all } t \in [0, 1], \\ &\qquad x \in W^{1,\infty}([0, 1]; \mathbf{R}^n), \quad u \in L^\infty([0, 1]; \mathbf{R}^m). \end{aligned} \tag{41}$$

The first-order necessary conditions for this problem are precisely (15)–(18). Observe that x_*, u_*, ψ_*, and ν_* satisfy (15)–(18) when $\pi = \pi_*$. Since the first-order necessary conditions are sufficient for optimality when coercivity holds, (x_*, u_*) is the unique solution to (41) at $\pi = \pi_*$. In addition, if uniform independence holds, we now show that the multipliers ψ and ν satisfying (16)–(18) are unique; hence, x_*, u_*, ψ_*, and ν_* are the unique solution to (15)–(18) for $\pi = \pi_*$.

To establish this uniqueness property for the multipliers, we apply Lemma 3.5 to the active constraint map $\mathcal{A}$ of section 1. Let J_i be the index sets associated with $I = \mathcal{A}$ in Lemma 3.5. Since $\mathcal{A}(t) \subset J_i$ for each $t \in [\tau_i, \tau_{i+1}]$, the complementary slackness condition $\nu_*(1)^{\mathsf{T}} g_*(1) = 0$, associated with the condition (5) of the minimum principle, implies that $(\nu_*)_{J_l^c} = 0$ on $[\tau_l, 1]$, while (21) along with (16) and (17) imply that $(\nu_*)_{J_l}$ and ψ_* are uniquely determined on $[\tau_l, 1]$. Proceeding by induction, suppose that ψ_* and ν_* are uniquely determined on the interval $[\tau_{i+1}, 1]$. Since $(\nu_*)_{J_i^c}$ is constant on $[\tau_i, \tau_{i+1}]$, it is uniquely determined by the continuity of ν_*, while $(\nu_*)_{J_i}$

and ψ_* on $[\tau_i, \tau_{i+1}]$ are uniquely determined by (21), (16), and (17). This completes the induction step.

We now use Lemma 3.8 to show that the solution to (41) depends Lipschitz continuously on the parameters when coercivity and uniform independence at $\mathcal{A}$ hold. We do this by making a special choice for the map I. Again, let J_i be the index sets associated with $I = \mathcal{A}$ by Lemma 3.5. Since $\mathcal{A}(t) \subset J_i$ for each $t \in [\tau_i, \tau_{i+1}]$, the parameter

$$\epsilon_i = -\sup\{(g_*)_j(t) : t \in [\tau_i, \tau_{i+1}], j \in J_i^c\} \tag{42}$$

is strictly positive for each i. Setting $\epsilon = .5 \min \epsilon_i$, we consider (37) in the case $I = \mathcal{A}_\epsilon$ where $\mathcal{A}_\epsilon(t)$ is the index set associated with the ϵ-active constraints for the linearized problem:

$$\mathcal{A}_\epsilon(t) = \{i : K_i(t)x_*(t) + (b_*)_i(t) \geq -\epsilon\} = \{i : (g_*)_i(t) \geq -\epsilon\}. \tag{43}$$

Since $\mathcal{A}_\epsilon(t) \subset J_i$ for each $t \in [\tau_i, \tau_{i+1}]$, Lemma 3.5 implies that uniform independence at $\mathcal{A}_\epsilon$ holds.

We now observe that the solution (x_*, u_*) of (41) at $\pi = \pi_*$ is the solution of (37) for $I = \mathcal{A}_\epsilon$ and $\pi = \pi_*$. First, (x_*, u_*) is feasible in (37) since there are fewer constraints than in (41). By the choice $I = \mathcal{A}_\epsilon$, all feasible pairs for (37) near (x_*, u_*) are also feasible in (41). Since (x_*, u_*) is optimal in (41), it is locally optimal in (37) as well, and by the coercivity condition and Lemma 3.7, (x_*, u_*) is the unique minimizer of (37) for $\pi = \pi_*$. By Lemma 3.8, we have an estimate for the change in the solution to (37) corresponding to a change in the parameters. Since $\|\delta x\|_{L^\infty} \leq \|\delta x\|_{H^1}$, it follows that for small perturbations in the data, the solution to (37) is feasible, and hence optimal, for (41). Hence, our previous stability analysis for (37) provides us with a local stability analysis for (41). We summarize this result in the following way.

LEMMA 3.9. *If coercivity and uniform independence at $\mathcal{A}$ hold, then for s, r, and a in an L^∞ neighborhood of s_*, r_*, and a_*, respectively, and for b in a $W^{1,\infty}$ neighborhood of b_*, there exists a unique minimizer of* (41)*, and the estimate* (38) *holds. Moreover, taking $I = \mathcal{A}_\epsilon$ with $\epsilon = .5 \min \epsilon_i$, where ϵ_i is defined in* (42)*, the solutions to* (37) *and* (41) *are identical in these neighborhoods.*

Now let us consider the multipliers associated with (41).

LEMMA 3.10. *If coercivity and uniform independence at $\mathcal{A}$ hold, then for s, r, and a in an L^∞ neighborhood of s_*, r_*, and a_*, respectively, and for b in a $W^{1,\infty}$ neighborhood of b_*, there exists a unique minimizer of* (41) *and associated unique multipliers satisfying the estimate:*

$$\|\delta\psi\|_{H^1} + \|\delta\nu\|_{L^2} \leq c(\|\delta a\|_{L^2} + \|\delta b\|_{H^1} + \|\delta s\|_{L^2} + \|\delta r\|_{L^2}). \tag{44}$$

Proof. Let $\mathcal{A}_\epsilon$ be the ϵ-active constraints defined by (43), where $\epsilon = .5 \min \epsilon_i$. Let J_i be the index sets and let ρ be the positive number associated with $I = \mathcal{A}$ by Lemma 3.5. Consider $\pi = \pi_* + \delta\pi$ where $\delta\pi$ is small enough that the active constraint set for (41) is a subset of $\mathcal{A}_\epsilon(t)$ for each t. By the same analysis used to establish uniqueness of (ψ_*, ν_*), there exist unique Lagrange multipliers $(\psi, \nu) = (\psi_*, \nu_*) + (\delta\psi, \delta\nu)$ corresponding to $\pi = \pi_* + \delta\pi$. We will show that

$$\|\delta\psi\|_{H^1} + \|\delta\nu\|_{L^2} \leq c(\|\delta x\|_{L^2} + \|\delta u\|_{L^2} + \|\delta s\|_{L^2} + \|\delta r\|_{L^2}). \tag{45}$$

Combining this with Lemma 3.9 yields Lemma 3.10.

We prove (45) by induction. Let us start with the interval $[\tau_l - \rho, 1]$. If $i \in J_l^c$, then $\nu_i(t) = 0$ for each $t \in [\tau_l - \rho, 1]$. Hence, $\delta\nu_{J_l^c} = 0$ on $[\tau_l - \rho, 1]$. Multiplying (17) by KB, we can solve for $\delta\nu_{J_l}$ and substitute in (16) to eliminate ν. Since $\psi(1) = 0$, it follows that

$$\|\delta\psi\|_{H^1([\sigma-\rho,1])} + \|\delta\nu\|_{L^2([\sigma-\rho,1])} \le c(\|\delta x\|_{L^2} + \|\delta u\|_{L^2} + \|\delta s\|_{L^2} + \|\delta r\|_{L^2}) \tag{46}$$

for $\sigma = \tau_l$.

Proceeding by induction, suppose that (46) holds for $\sigma = \tau_{j+1}$; we wish to show that it holds for $\sigma = \tau_j$. If $i \in J_j^c$, then $\nu_i(t)$ is constant on $[\tau_j - \rho, \tau_{j+1}]$, and we have

$$\int_{\tau_j-\rho}^{\tau_{j+1}} \delta\nu_i(t)^2 dt = \frac{\tau_{j+1} - \tau_j + \rho}{\rho} \int_{\tau_{j+1}-\rho}^{\tau_{j+1}} \delta\nu_i(t)^2 dt.$$

Combining this with (46) for $\sigma = \tau_{j+1}$, it follows that

$$\|\delta\nu_i\|_{L^2([\sigma-\rho,1])} \le c(\|\delta x\|_{L^2} + \|\delta u\|_{L^2} + \|\delta s\|_{L^2} + \|\delta r\|_{L^2})$$

for $\sigma = \tau_j$. Again, multiplying (17) by KB, we solve for $\delta\nu_{J_j}$ and substitute in (16). Since $|\delta\psi(\tau_j)| \le \|\delta\psi\|_{H^1([\tau_j,1])}$, the induction bound (46) for $\sigma = \tau_{j+1}$ coupled with the bound already established for $\delta\nu_i$, $i \in J_j^c$, gives (46) for $\sigma = \tau_j$. This completes the induction. ∎

LEMMA 3.11. *Suppose that smoothness, coercivity, and uniform independence at $\mathcal{A}$ hold and let κ be small enough that Y is contained in the neighborhoods defined in Lemmas* 3.9 *and* 3.10. *Then for some $\mu > 0$ and for each $\pi \in Y$, there exists a unique solution (x, u) to* (41) *and associated multipliers (ψ, ν) satisfying the estimates* (38) *and* (44), *$(x, \psi, u, \nu) = (\mathcal{F} - \mathcal{L})^{-1}\pi$, and we have $\dot{x}$, $\dot{\psi}$, u, $\nu \in \mathrm{Lip}_\mu$.*

Proof. If $w = (x, \psi, u, \nu)$ denotes $(\mathcal{F} - \mathcal{L})^{-1}\pi$, then w satisfies the first-order necessary conditions (15)–(18) associated with (41). Lemmas 3.9 and 3.10 tell us that the unique solution and multipliers for (41) satisfy the estimates (38) and (44) for π near π_*. Since the first-order necessary conditions are sufficient for optimality when coercivity holds, the variational system (15)–(18) has a unique solution, for π near π_*, that is identical to the solution and multipliers for (41), and the estimates (38) and (44) are satisfied.

To complete the proof, we need to show that $\dot{x}$, $\dot{\psi}$, u, $\nu \in \mathrm{Lip}_\mu$ for some constant $\mu > 0$. This follows from the regularity results of [8], where it is shown that the solution to a constant coefficient, linear-quadratic problem satisfying the uniform independence condition and with R positive definite, Q positive semidefinite, and $M = 0$ has the property that the optimal u and associated ν are Lipschitz continuous in time, while the derivatives of x and ψ are Lipschitz continuous in time. Moreover, the Lipschitz constant in time is bounded in terms of the constant α in the uniform independence condition and the smallest eigenvalue of R. Exactly the same analysis applies to a linear-quadratic problem with time-varying coefficients; however, the bound for the Lipschitz constant of the solution depends on the Lipschitz constants of the matrices of the problem and of the parameters a, r, s, and b, as well as on a uniform bound for the smallest eigenvalue of $R(t)$ on $[0, 1]$ and for the parameter α in the uniform independence condition. By Lemma 3.9 and with the choice for I given in the statement of the lemma, the quadratic programs (37) and (41) have the same solution for s, r, and a in an L^∞ neighborhood of s_*, r_*, and a_* and for b in a $W^{1,\infty}$ neighborhood of b_*. Hence, for parameters in this neighborhood of π_*, the indices of the active constraints are contained in $I(t)$ for each t, and the independence condition (21) holds. Lemma 3.7 provides a lower bound for the eigenvalues of

$R(t)$. If $(a, s, r, b) \in Y$, then the Lipschitz constants for a, s, r, and $\dot{b}$ are bounded by those for a_*, s_*, r_*, and $\dot{b}_*$ plus κ. Hence, taking μ sufficiently large, the proof is complete. □

Proof of Theorem 1.1. We apply Theorem 2.2 with the identifications given at the beginning of this section and with μ chosen sufficiently large in accordance with Lemma 3.11. The completeness of X is established in Lemma 3.2, (Q1) is immediate, (Q2) follows from smoothness, (Q3) is proved in Lemma 3.3, (Q4) follows from Lemma 3.11, and (Q5) is established in Lemma 3.4. Applying Theorem 2.2, the estimate (7) is established. Under the uniform independence condition, coercivity is a second-order sufficient condition for local optimality (see [4, Theorem 1]) which is stable under small changes in either the parameters or the solution of the first-order optimality conditions. Finally, we apply Lemma 3.1 to obtain the L^∞ estimate of Theorem 1.1. □

We note that the coercivity condition we use here is a strong form of a second-order sufficient optimality condition; it not only provides optimality, but also guarantees Lipschitz continuity of the optimal solution and multipliers when uniform independence holds. As recently proved in [6] for finite-dimensional optimization problems, Lipschitzian stability of the solution and multipliers necessarily requires a coercivity condition stronger than the usual second-order condition. For the treatment of second-order sufficient optimality under conditions equivalent to coercivity, see [18] and [21]. These sufficient conditions can be applied to state constraints of arbitrary order. For recent work concerning the treatment of second-order sufficient optimality in state constrained optimal control, see [16], [19], and [22].

4. Lipschitzian stability in L^∞. One way to sharpen the L^∞ estimate of Theorem 1.1 involves an assumption concerning the regularity of the solution to the linear-quadratic problem (41). The time t is a contact point for the ith constraint of $Kx + b \leq 0$ if $(K(t)x(t) + b(t))_i = 0$ and there exists a sequence $\{t_k\}$ converging to t with $(K(t_k)x(t_k) + b(t_k))_i < 0$ for each k.

Contact separation. There exists a finite set $I_1, I_2, \ldots, I_N$ of disjoint, closed intervals contained in $(0, 1)$ and neighborhoods of (a_*, r_*, s_*) in $W^{1,\infty}$ and of b_* in $W^{2,\infty}$ with the property that for each a, r, s, and b in these neighborhoods, and for each solution to (41), all contact points are contained in the union of the intervals I_i with exactly one contact point in each interval and with exactly one constraint changing between active and inactive at this point.

Observe that if for (1) with $p = p_*$, there are a finite number of contact points, at each contact point exactly one constraint changes between active and inactive, and each contact point in the linear-quadratic problem (41) depends continuously on the parameters, then contact separation holds. The finiteness of the contact set is a natural condition in optimal control; for example, in [5] it is proved that for a linear-quadratic problem with time invariant matrices and one state constraint, the contact set is finite when uniform independence and coercivity hold.

THEOREM 4.1. *Suppose that the problem* (1) *with $p = p_*$ has a local minimizer (x_*, u_*) and that smoothness, contact separation, and uniform independence at $\mathcal{A}$ hold. Let ψ_* and ν_* be the associated multipliers satisfying the first-order necessary conditions* (2)–(5). *If the coercivity condition holds, then there exist neighborhoods V of p_* and U of $w_* = (x_*, \psi_*, u_*, \nu_*)$ in $W^{1,\infty} \times W^{1,\infty} \times L^\infty \times L^\infty$, such that for every $p \in V$, there exists a unique solution $w = (x, \psi, u, \nu) \in U$ to the first-order necessary conditions* (2)–(5) *and (x, u) is a local minimizer of the problem* (1) *associated with p. Moreover, for every $p_i \in V, i = 1, 2$, if $w_i = (x_i, \psi_i, u_i, \nu_i)$ is the corresponding*

solution of (2)–(5)*, the following estimate holds:*

$$\|x_1 - x_2\|_{W^{1,\infty}} + \|\psi_1 - \psi_2\|_{W^{1,\infty}} + \|u_1 - u_2\|_{L^\infty} + \|\nu_1 - \nu_2\|_{L^\infty} \le cE_\infty.$$

To prove this result, we need to supplement the 2-norm perturbation estimates provided by Lemmas 3.9 and 3.10 with analogous ∞-norm estimates.

LEMMA 4.2. *If coercivity, uniform independence at $\mathcal{A}$, and contact separation hold, then there exist neighborhoods of (a_*, r_*, s_*) in $W^{1,\infty}$ and of b_* in $W^{2,\infty}$ such that for each a_i, r_i, s_i, and b_i, $i = 1, 2$, in these neighborhoods, the associated solutions (x_i, u_i) of* (41) *satisfy*

$$\begin{aligned} &\|\delta x\|_{W^{1,\infty}} + \|\delta\psi\|_{W^{1,\infty}} + \|\delta u\|_{L^\infty} + \|\delta\nu\|_{L^\infty} \\ &\le c(\|\delta a\|_{L^\infty} + \|\delta b\|_{W^{1,\infty}} + \|\delta r\|_{L^\infty} + \|\delta s\|_{L^\infty}). \end{aligned} \tag{47}$$

Proof. Letting $\mathcal{A}_\epsilon$ denote the ϵ-active set defined in (43), we again choose $\epsilon = .5 \min \epsilon_i$, where ϵ_i is defined in (42). We consider parameters a, r, s, and b chosen within the neighborhoods of the contact separation condition, and sufficiently close to a_*, r_*, s_*, and b_* that the active constraint set for the solution of the perturbed linear-quadratic problem (41) is contained in $\mathcal{A}_\epsilon(t)$ for each t. By eliminating the perturbations in the constraints, as we did in the proof of Lemma 3.8, there is no loss of generality in assuming that $a = b = 0$. We refer to the quadratic programs corresponding to the parameters (r_1, s_1) and (r_2, s_2) as Problems 1 and 2.

Let (x, u) be either (x_1, u_1) or (x_2, u_2). If $t \in (0, 1)$ is a time for which $K_i(t)x(t) = 0$ for some i, then $\frac{d}{dt}(K_i x) = \dot{K}_i x + K_i \dot{x} = 0$. Substituting for $\dot{x}$ using the state equation $\dot{x} = Ax + Bu$ and for u using the necessary condition (17) yields

$$K_i B R^{-1} (KB)^{\mathsf{T}} \nu = -\dot{K}_i x - K_i A x + K_i B R^{-1}(B^{\mathsf{T}}\psi + M^{\mathsf{T}} x + r).$$

This equation has the form

$$N_i \nu = S_i x + T_i \psi + U_i r \tag{48}$$

for suitable choices of the row vectors N_i, S_i, T_i, and U_i. Hence, at any time t where $K_i(t)x_1(t) = K_i(t)x_2(t) = 0$, the change in solution and multipliers corresponding to a change in parameters satisfies the equation

$$N_i(t)\delta\nu(t) = S_i(t)\delta x(t) + T_i(t)\delta\psi(t) + U_i(t)\delta r(t). \tag{49}$$

By the contact separation condition, Problems 1 and 2 have the same active set near $t = 1$. Since the components of ν corresponding to inactive constraints are constant and since $\nu_i(1) = 0$ if $K_i(1)x(1) < 0$, it follows that $\delta\nu_i(t) = 0$ for t near 1 when $K_i x_1(1) < 0 > K_i x_2(1)$. The relation (49) combined with uniform independence, with the L^2 estimates provided in Lemmas 3.9 and 3.10, and with a bound for the L^∞ norm in terms of the H^1 norm, gives

$$\|\delta\nu\|_{L^\infty[t,1]} \le c(\|\delta r\|_{L^\infty} + \|\delta s\|_{L^\infty}). \tag{50}$$

Using the bound (36) of Lemma 3.7 in (17) and applying Gronwall's lemma to (16), we have

$$\begin{aligned} &\|\delta x\|_{W^{1,\infty}[t,1]} + \|\delta\psi\|_{W^{1,\infty}[t,1]} + \|\delta u\|_{L^\infty[t,1]} + \|\delta\nu\|_{L^\infty[t,1]} \\ &\le c(\|\delta r\|_{L^\infty} + \|\delta s\|_{L^\infty}) \end{aligned} \tag{51}$$

for all $t < 1$ in some neighborhood of $t = 1$. As t decreases, this estimate is valid until the first contact point is reached for either Problem 1 or Problem 2. Proceeding by induction, suppose that we have established (51) up to some contact point; we now wish to show that (51) holds up to the next contact point.

Again, by the contact separation condition, there is precisely one constraint, say constraint j, that makes a transition between active and inactive at the current contact point. Suppose that on the interval (α, β), the active sets for Problems 1 and 2 differ by the element j, and let τ be the first contact point to the left of α for either Problem 1 or Problem 2. If there is no such point, we take $\tau = 0$. By the contact separation condition, the difference $\alpha - \tau$ is uniformly bounded away from zero for all choices of the parameters s and r near s_* and r_*. There are essentially two cases to consider.

Case 1. Constraint j is active in Problem 2 to the left of $t = \beta$, and constraint j is active in Problem 1 to the left of $t = \alpha$.

Case 2. Constraint j is active in Problem 2 to the right of $t = \alpha$, and constraint j is active in Problem 1 to the right of $t = \beta$.

Case 1. Since constraint j is active in both Problems 1 and 2 at $t = \alpha$, it follows from (49) and from the uniform independence condition that

$$|\delta\nu_\Gamma(\alpha)| \le c(\|\delta r\|_{L^\infty} + \|\delta s\|_{L^\infty}) + c|\delta\nu_{\Gamma^c}(\alpha)|,$$

where Γ is the set of indices of active constraints at $t = \alpha$. Since $\delta\nu_i$ is constant for $i \in \Gamma^c$ on (α, β), the induction hypothesis yields

$$|\delta\nu_{\Gamma^c}(\alpha)| = |\delta\nu_{\Gamma^c}(\beta)|_{L^\infty} \le c(\|\delta r\|_{L^\infty} + \|\delta s\|_{L^\infty}). \tag{52}$$

Hence, we have

$$|\delta\nu(\alpha)| \le c(\|\delta r\|_{L^\infty} + \|\delta s\|_{L^\infty}). \tag{53}$$

Since ν_j is constant in Problem 1 on (α, β), and since it is monotone in Problem 2, the bound (53) coupled with the bound (51) at $t = \beta$ implies that

$$\|\delta\nu_j\|_{L^\infty([\alpha,\beta])} \le c(\|\delta r\|_{L^\infty} + \|\delta s\|_{L^\infty}). \tag{54}$$

Since $\delta\nu_i$ is constant on (α, β) for $i \in \Gamma^c$, it follows from (51) that

$$\|\delta\nu_{\Gamma^c}\|_{L^\infty([\alpha,\beta])} \le c(\|\delta r\|_{L^\infty} + \|\delta s\|_{L^\infty}). \tag{55}$$

Relation (49), for $i \in \Gamma^- = \Gamma \setminus \{j\}$, along with (54) and (55) yield

$$\|\delta\nu_{\Gamma^-}\|_{L^\infty([\alpha,\beta])} \le c(\|\delta r\|_{L^\infty} + \|\delta s\|_{L^\infty}). \tag{56}$$

Combining (54)–(56) gives

$$\|\delta\nu\|_{L^\infty([\alpha,\beta])} \le c(\|\delta r\|_{L^\infty} + \|\delta s\|_{L^\infty}). \tag{57}$$

On the interval from $t = \alpha$ down to the next contact point τ, precisely the same constraints are active in both Problems 1 and 2. Again, the relation (49) combined with uniform independence, with the L^2 estimates provided in Lemmas 3.9 and 3.10, and with a bound for the L^∞ norm in terms of the H^1 norm gives

$$\|\delta\nu\|_{L^\infty([\tau,\alpha])} \le c(\|\delta r\|_{L^\infty} + \|\delta s\|_{L^\infty}). \tag{58}$$

Relation (50) for $t = \beta$, along with (57) and (58), gives

$$\|\delta\nu\|_{L^\infty([\tau,1])} \le c(\|\delta r\|_{L^\infty} + \|\delta s\|_{L^\infty}).$$

And combining this with (15)–(17) gives (51) for $t = \tau$. This completes the induction step in Case 1.

Case 2. The mean value theorem implies that for some $\gamma \in (\tau, \alpha)$, we have

$$\begin{aligned}(\alpha - \tau)\frac{d}{dt}K_j(t)\delta x(t)\big|_{t=\gamma} &= K_j(\alpha)\delta x(\alpha) - K_j(\tau)\delta x(\tau)\\ &\le 2\|K_j\|_{L^\infty}\|\delta x\|_{L^\infty} \le c(\|\delta r\|_{L^\infty} + \|\delta s\|_{L^\infty}).\end{aligned}$$

Hence, even though the derivative of $K_j x_i$ may not vanish on (τ, α), the derivative of the change $K_j\delta x$ is still bounded by the perturbation in the parameters at some $\gamma \in (\tau, \alpha)$:

$$\left|\frac{d}{dt}(K_i(t)\delta x(t))\right|_{t=\gamma} \le c(\|\delta r\|_{L^\infty} + \|\delta s\|_{L^\infty})/(\alpha - \tau). \tag{59}$$

Since α and τ lie in disjoint closed sets I_k associated with the contact separation condition, $\alpha - \tau$ is bounded away from zero by the distance between the closest pair of sets. Focusing on the left side of (59), we substitute $\delta\dot{x} = A\delta x + B\delta u$, and we substitute for δu using (17) to obtain the relation

$$N_j(\gamma)\delta\nu(\gamma) = S_j(\gamma)\delta x(\gamma) + T_j(\gamma)\delta\psi(\gamma) + U_j(\gamma)\delta r(\gamma) + \Delta_j, \tag{60}$$

where $|\Delta_j| \le c(\|\delta r\|_{L^\infty} + \|\delta s\|_{L^\infty})/(\alpha - \tau)$. Let Γ denote the set of indices of the active constraints at $t = \beta$. Combining (60) with (49) for $i \in \Gamma^- = \Gamma \setminus \{j\}$ gives

$$|\delta\nu_\Gamma(\gamma)| \le c(\|\delta r\|_{L^\infty} + \|\delta s\|_{L^\infty}) + c|\delta\nu_{\Gamma^c}(\gamma)|.$$

The analysis for Case 1 can now be applied, starting with (52) but with α replaced by γ. □

Remark 4.3. In the proof of Lemma 4.2, we needed to ensure that the difference $\alpha - \tau$, appearing in Case 2, was bounded away from zero. The contact separation condition ensures that this difference is bounded away from zero, since α and τ lie in disjoint closed intervals I_k. On the other hand, any condition that ensures a positive separation for the contact points α and τ in Case 2 can be used in place of the contact separation assumption of Theorem 4.1 and Lemma 4.2.

Proof of Theorem 4.1. The functions $\mathcal{T}$, $\mathcal{F}$, and $\mathcal{L}$ and the sets X, Π, and Y are the same as in the proof of Theorem 1.1 except that L^2 is replaced by L^∞ and H^1 is replaced by $W^{1,\infty}$ everywhere. Except for this change in norms, and the replacement of the L^2 estimates (38) and (44) referred to in Lemma 3.11 by the corresponding L^∞ estimate (47) of Lemma 4.2, the same proof used for Theorem 1.1 can be used to establish Theorem 4.1. □

5. Remarks. As mentioned in section 2, Theorem 2.2 is a generalization of Robinson's implicit function theorem [20] to nonlinear spaces. His theorem assumes that the nonlinear term is strictly differentiable and that the inverse of the linearized map is Lipschitz continuous. In optimal control, the latter condition amounts to Lipschitz continuity in L^∞ of the solution-multiplier vector associated with the linear-quadratic approximation. For problems with control constraints, this property for the solution is obtained, for example, in [1] or [4].

In this paper, we obtain Lipschitzian stability results for state constrained problems utilizing a new form of the implicit function theorem applicable to nonlinear spaces. We obtain optimal Lipschitzian stability results in L^2 and nonoptimal stability results in L^∞ under the uniform independence and the coercivity conditions. And with an additional contact separation condition, we obtain a tight L^∞ stability result. These are the first L^∞ stability results that have been established for state constrained control problems.

The uniform independence condition was introduced in [8], where it was shown that this condition together with the coercivity condition yield Lipschitz continuity in time of the solution and the Lagrange multipliers of a convex state and control constrained optimal control problem. Using Hager's regularity result, Dontchev [1] proved that the solution of this problem has a Lipschitz-type property with respect to perturbations. Various extensions of these results have been proposed by several authors. A survey of earlier results is given in [2].

In a series of papers (see [14], [15], and the references therein), Malanowski studied the stability of optimal control problems with constraints. In [15] he considers an optimal control problem with state and control constraints. His approach differs from ours in the following ways: he uses an implicit function theorem in linear spaces and a compactness argument, and the second-order sufficient condition he uses is different from our coercivity condition. Although there are some similar steps in the analysis of L^2 stability, the two approaches mainly differ in their abstract framework.

A prototype of Lemma 3.5 is given in [1, Lemma 2.5]. Lemma 3.6 is related to Lemma 3 in [2], although the analysis in Lemma 3.6 is much simpler since we ignore indices outside of $\mathcal{A}(t)$. In the analysis of the linear-quadratic problem (37), we follow the approach in [4].

Acknowledgment. The authors wish to thank both Kazimierz Malanowski for his comments on an earlier version of this paper and the reviewers for their constructive suggestions.

REFERENCES

[1] A. L. Dontchev, *Perturbations, Approximations and Sensitivity Analysis of Optimal Control Systems*, Lecture Notes in Control and Inform. Sci. 52, Springer-Verlag, New York, 1983.

[2] A. L. Dontchev and W. W. Hager, *Lipschitzian stability in nonlinear control and optimization*, SIAM J. Control Optim., 31 (1993), pp. 569–603.

[3] A. L. Dontchev and W. W. Hager, *An inverse function theorem for set-valued maps*, Proc. Amer. Math. Soc., 121 (1994), pp, 481–489.

[4] A. L. Dontchev, W. W. Hager, A. B. Poore, and B. Yang, *Optimality, stability and convergence in nonlinear control*, Appl. Math. Optim., 31 (1995), pp. 297–326.

[5] A. L. Dontchev and I. Kolmanovsky, *On regularity of optimal control*, in Proc. French-German Conf. on Optimization, Lecture Notes in Economics and Mathematical Systems 429, R. Durier and C. Michelot, eds., Springer-Verlag, New York, 1995, pp. 125–135.

[6] A. L. Dontchev and R. T. Rockafellar, *Characterizations of strong regularity for variational inequalities over polyhedral convex sets*, SIAM J. Optim., 6 (1996), pp. 1087–1105.

[7] J. C. Dunn and T. Tian, *Variants of the Kuhn–Tucker sufficient conditions in cones of nonnegative functions*, SIAM J. Control Optim., 30 (1992), pp. 1361–1348.

[8] W. W. Hager, *Lipschitz continuity for constrained processes*, SIAM J. Control Optim., 17 (1979), pp. 321–337.

[9] W. W. Hager, *Multiplier methods for nonlinear optimal control*, SIAM J. Numer. Anal., 27 (1990), pp. 1061–1080.

[10] W. W. Hager and G. D. Ianculescu, *Dual approximations in optimal control*, SIAM J. Control Optim., 22 (1984), pp. 423–465.

[11] W. W. Hager and S. K. Mitter, *Lagrange duality theory for convex control problems*, SIAM J. Control Optim., 14 (1976), pp. 843–856.

[12] R. F. Hartl, S. P. Sethi, and R. G. Vickson, *A survey of the maximum principles for optimal control problems with state constraints*, SIAM Rev., 37 (1995), pp. 181–218.
[13] A. D. Ioffe and V. M. Tihomirov, *Theory of Extremal Problems*, North-Holland, Amsterdam, 1979.
[14] K. Malanowski, *Two-norm approach in stability and sensitivity analysis of optimization and optimal control problems*, Adv. Math. Sci. Appl., 2 (1993), pp. 397–443.
[15] K. Malanowski, *Stability and sensitivity of solutions to nonlinear optimal control problems*, Appl. Math. Optim., 32 (1995), pp. 111–141.
[16] K. Malanowski, *Sufficient optimality conditions for optimal control subject to state constraints*, SIAM J. Control Optim., 35 (1997), pp. 205–227.
[17] H. Maurer, *On the Minimum Principle for Optimal Control Problems with State Constraints*, Schriftenreihe des Rechenzentrums der Univ. Münster Nr. 41, Münster, 1979.
[18] H. Maurer, *First and second order sufficient optimality conditions in mathematical programming and optimal control*, Math. Programming Stud., 14 (1981), pp. 163–177.
[19] H. Maurer and S. Pickenhain, *Second order sufficient conditions for optimal control problems with control-state constraints*, J. Optim. Theory Appl., 86 (1995), pp. 649–667.
[20] S. M. Robinson, *Strongly regular generalized equations*, Math. Oper. Res., 5 (1980), pp. 43–62.
[21] G. Sorger, *Sufficient conditions for nonconvex control problems with state constraints*, J. Optim. Theory Appl., 62 (1989), pp. 289–310.
[22] V. Zeidan, *The Riccati equation for optimal control problems with mixed state-control constraints; necessity and sufficiency*, SIAM J. Control Optim., 32 (1994), pp. 1297–1321.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 2, pp. 719–741, March 1998

012

RATES OF CONVERGENCE FOR APPROXIMATION SCHEMES IN OPTIMAL CONTROL*

PAUL DUPUIS† AND MATTHEW R. JAMES‡

Abstract. We present a simple method for obtaining rate of convergence estimates for approximations in optimal control problems. Although the method is applicable to a wide range of approximation problems, it requires in all cases some type of smoothness of the quantity being approximated. We illustrate the method by presenting a number of examples, including finite difference schemes for stochastic and deterministic optimal control problems. A general principle can be abstracted, and indeed the method may be applied to a variety of approximation problems, such as the numerical approximation of nonlinear PDEs not a priori related to control theory.

Key words. optimal control, numerical approximation, rate of convergence, finite differences, ergodic control, reflected diffusions, nonlinear PDE

AMS subject classifications. 93E20, 65N12, 65N15, 65N06, 93E25

PII. S0363012994267789

1. Introduction.

A fundamental problem in numerical analysis is the determination of the rate of convergence of approximation schemes. In general, the rate of convergence depends on the nature of the approximation and on the smoothness of the quantity being approximated. For example, the standard finite difference scheme for Laplace's equation in a smooth domain converges with a rate proportional to the discretization step size.

In optimal control theory, one is often faced with the problem of computing the minimal cost function, also referred to as the value function. In many cases, the value function can be characterized as an appropriate solution to a Hamilton–Jacobi–Bellman (HJB) equation that takes the form of a nonlinear PDE. In general, one cannot compute the value function explicitly, and instead must resort to a numerical approximation. Various approximation schemes are available (e.g., finite difference or finite element), and convergence results are either analytic (e.g., Crandall and Lions [5], Barles and Souganidis [1]) or probabilistic (e.g., Kushner [18], Kushner and Dupuis [19]). There are two types of results available on the rate of convergence. One is a global rate of convergence for deterministic problems which makes few assumptions regarding the regularity of the minimal cost function. The first paper to obtain results of this type was [5]. The rate is in the form of an upper bound on the error and is proportional to the square root of the discretization step size. Later papers considered a number of extensions, such as deterministic control problems (e.g., Capuzzo Dolcetta and Falcone [3], Capuzzo Dolcetta and Ishii [4], Gonzalez and Rofman [17]) and differential games (Souganidis [26]). In all these papers the same type of global but locally suboptimal rate estimate as in [5] is obtained. A second type of rate was obtained by Menaldi [23] in the context of control of a nondegenerate diffusion process

*Received by the editors May 16, 1994; accepted for publication (in revised form) February 4, 1997.

http://www.siam.org/journals/sicon/36-2/26778.html

†Division of Applied Mathematics, Brown University, Providence, RI 02912 (dupuis@cfm.brown.edu). The research of this author was supported in part by the Air Force Office of Scientific Research (F49620-93-1-0264) and the Army Research Office (DAAH04-93-G-0070).

‡Department of Engineering, Faculty of Engineering and Information Technology, Australian National University, Canberra, ACT 0200, Australia (Matthew.James@anu.edu.au, http://spigot.anu.edu.au/people/mat/home.html).

with discounted cost. Here the regularity of the minimal cost function was exploited to obtain sharp rates of convergence.

In this paper we present a simple method for obtaining rate of convergence estimates that is applicable to a range of approximation problems, including the important case of numerical approximation. The approach is closer in spirit to that of Menaldi rather than Crandall and Lions in that we exploit, wherever possible, smoothness of the minimal cost to obtain sharp rates and in some cases an expansion of the error in terms of the discretization parameter. A minimal requirement for the applicability of our method is local smoothness of the quantity that is being approximated. For example, in the setting of deterministic optimal control problems we obtain a rate of convergence that is proportional to the discretization step size in these regions. The practice of considering separately those regions where greater regularity applies is standard in numerical analysis, and the information so obtained is often more useful than a global but locally suboptimal rate of convergence.

The basic idea is as follows. In the problems we consider, the quantity V to be approximated is represented as a functional of some process x, and the approximating quantity V^h is analogously represented as a functional of an approximating process x^h. For a very simple (uncontrolled) example, suppose that V has a representation of the form

$$V(x) = \mathbf{E}_x \left[\int_0^\infty e^{-\lambda t} k(x_t) dt \right],$$

where x is, say, a diffusion process. Suppose also that V^h has the representation

$$V^h(x) = \mathbf{E}_x \left[\int_0^\infty e^{-\lambda t} k(x_t^h) dt \right]$$

in terms of an approximating process x^h, which for simplicity we will assume to be Markov. As we illustrate via several examples below, the assumed regularity properties of V allow one to derive a second representation for V, this time in terms of the process x^h:

$$V(x) = \mathbf{E}_x \left[\int_0^\infty e^{-\lambda t} \left(k(x_t^h) + e^h(x_t^h) \right) dt \right].$$

The function $e^h = V - V^h$ is given explicitly in terms of the function V and the generators of the processes x and x^h. Since V^h is supposed to be close to V, one would expect x^h to be close in some sense to x. In fact, one typically has the weak convergence of x^h to x. When this convergence is coupled with the explicit form for the error e^h, a comparison of the two representations of V gives a rate of convergence, and also formulas for rate coefficients when enough regularity is available.

The rate of convergence has a number of uses, the most obvious being as a guide in the selection of step sizes and the comparison of algorithms in numerical approximations. A second use in this setting is in comparing the contributions to the overall error made by various "parts" of an approximation; e.g., one can consider a problem that is posed on a bounded domain and compare the contributions made by approximations on the interior and approximations to the boundary condition. This is done for a reflecting diffusion problem with ergodic cost in section 4.

To show that the basic idea can be used in a variety of situations, we give the details for a number of problems that are quite different. In section 2, we consider

the stochastic control problem analyzed by Menaldi [23], where the value function is known to be smooth: $V \in C^{2,\alpha}$ and the (global) rate of convergence for a finite difference scheme is $O(h^\alpha)$. Our approach appears to be simpler than Menaldi's. The focus shifts in section 3 to a general class of approximations to a finite time deterministic optimal control problem, including finite difference schemes. In section 4 we treat an ergodic control problem for a reflecting diffusion process.

Other types of approximations arise in control theory. For example, a diffusion model is often a surrogate for a more realistic and more complicated controlled process. Underlying this replacement is (implicitly or explicitly) an approximation argument, in which it is supposed that the realistic process is embedded in a sequence of processes whose weak limit is the controlled diffusion process. Rate of convergence estimates are also useful in this context, and the method we discuss can in some cases be used here as well. We conclude with remarks on such possibilities and other extensions in section 5.

2. A stochastic optimal control example. In this section we introduce the basic method for calculating rates of convergence. In order to place it in perspective, it is appropriate that we begin by considering one of the few stochastic control problems for which a rate of convergence is known, namely, the stochastic control problem studied by Menaldi [23].

The dynamics are given by the controlled stochastic differential equation (SDE)

$$dx_t = b(x_t, u_t)\,dt + dw_t, \tag{2.1}$$

and the value function is defined by

$$V(x) \doteq \inf_{u\in\mathcal{U}} \mathbf{E}_x \left[\int_0^\tau e^{-\lambda t} k(x_t, u_t)\,dt\right] \tag{2.2}$$

for $x \in D$. Here, $D \subset \mathbf{R}^n$ is a smooth bounded domain, $\tau = \tau(x) = \inf\{t > 0 : x_t \notin D\}$ is the exit time from D, $\lambda > 0$, $\mathcal{U}$ is a set of admissible U-valued control processes, $U \subset \mathbf{R}^m$ is compact, $b \in C^\infty(\mathbf{R}^n \times \mathbf{R}^m, \mathbf{R}^n)$ and $k \in C^\infty(\mathbf{R}^n \times \mathbf{R}^m)$ are bounded and uniformly Lipschitz continuous, and $\mathbf{E}_x$ denotes expectation conditioned on $x_0 = x$. For the definition of admissible controls, see [23].

Although in certain cases (e.g., uncontrolled systems or one-dimensional problems) V may be more regular, it is known [16], [23, pp. 601–602] that in general $V \in C^{2,\alpha}(\bar{D})$ for some $0 < \alpha < 1$, and also that V is a classical solution of the dynamic programming or HJB equation

$$\begin{cases} \lambda V(x) = \min_{u\in U} [L_u V(x) + k(x,u)] & \text{in } D, \\ V(x) = 0 & \text{on } \partial D. \end{cases} \tag{2.3}$$

In the last display L_u is the controlled diffusion generator defined for $f \in C^2(\mathbf{R}^n)$ by

$$L_u f(x) \doteq \langle b(x,u), f_x(x)\rangle + \frac{1}{2}\mathrm{tr}[f_{xx}(x)],$$

where $\mathrm{tr}A$ denotes the trace of a square matrix A, and where $f_x(x)$ and $f_{xx}(x)$ denote the gradient and Hessian of f at x, respectively.

For a real-valued function $g(y)$ we define $g^+(y) = g(y) \vee 0$ and $g^-(y) = -(g(y) \wedge 0)$. For $h > 0$ let $h\mathbf{Z}^n = \{hz : z_i \in \mathbf{Z}, i = 1, 2, \ldots, n\}$. There are a number of ways to construct an approximation to V that has domain $h\mathbf{Z}^n$. We focus for now on the

method that is probably most familiar. Thus we replace the differential operator L_u by the finite difference operator L_u^h defined by

$$L_u^h f(x) \doteq \frac{1}{2}\sum_{i=1}^{n}\left[f(x+he_i)+f(x-he_i)-2f(x)\right]/h^2 + \sum_{i=1}^{n} b_i^{\pm}(x,u)\left[f(x\pm he_i)-f(x)\right]/h,$$

where $x \in h\mathbf{Z}^n$ and e_i, $i=1,\dots,n$, are the standard unit vectors in $\mathbf{R}^n$. Let us fix $h_0 > 0$, and suppose that D' is an open set containing the closed h_0-neighborhood of D. Given a function $f \in C^{2,\alpha}(\bar{D})$, there exists a function $f' \in C_0^{2,\alpha}(D')$ such that $f = f'$ in $\bar{D}$ (see [16, Lemma 6.37]). Henceforth, we assume $h \le h_0$ and simply write f for the extension f'. By Taylor's theorem, we see that the operator L_u^h approximates L_u in the sense that if we define the "error"

$$e_f^h(x,u) \doteq L_u f(x) - L_u^h f(x), \tag{2.4}$$

then

$$|e_f^h(x,u)| \;=\; O(h^\alpha) \text{ uniformly in } \bar{D}\times U \tag{2.5}$$

for all $f \in C^{2,\alpha}(\bar{D})$. Note that this estimate is well defined, in view of our convention of extension. The finite difference replacement of (2.3) that we consider is

$$\begin{cases} \lambda V^h(x) \;=\; \min_{u\in U}\left[L_u^h V^h(x) + k(x,u)\right] & \text{in } D^h, \\ V^h(x) \;=\; 0 & \text{on } \partial D^h, \end{cases} \tag{2.6}$$

where $D^h = D \cap h\mathbf{Z}^n$ and $\partial D^h = (\mathbf{R}^n \backslash D) \cap h\mathbf{Z}^n$.

For $v \in \mathbf{R}^n$ and $p \in [1,\infty)$ define $\|v\|_p \doteq (|v_1|^p + \cdots + |v_n|^p)^{1/p}$. The equation (2.6) can be interpreted as the HJB equation for a controlled Markov chain problem. To see this, multiply both sides of the first equation in (2.6) by $\Delta t^h(x)$, add $V^h(x)$, and then divide by $(1+\lambda\Delta t^h(x))$. We thereby obtain the equation

$$V^h(x) \;= \min_{u\in U}\left[\sum_{z\in N^h(x)} \frac{1}{1+\lambda\Delta t^h(x)}\left(p^h(x,z|u)V^h(z) + \Delta t^h(x)k(x,u)\right)\right], \tag{2.7}$$

where

$$\Delta t^h(x) \;\doteq\; \frac{h^2}{n + h\max_{u\in U}\|b(x,u)\|_1} \tag{2.8}$$

is a time interpolation scale, and the functions $p^h(x,z|u)$ are transition probabilities defined by

$$p^h(x,z|u) \;\doteq\; \begin{cases} \dfrac{h(\max_{u\in U}\|b(x,u)\|_1 - \|b(x,u)\|_1)}{n + h\max_{u\in U}\|b(x,u)\|_1} & \text{if } z = x, \\ \dfrac{1/2 + hb_i^{\pm}(x,u)}{n + h\max_{u\in U}\|b(x,u)\|_1} & \text{if } z = x \pm he_i, \text{ some } i = 1,\dots,n, \\ 0 & \text{otherwise.} \end{cases} \tag{2.9}$$

The functions p^h and $\Delta t^h(x)$ are defined as they are so that equation (2.7) can be interpreted as an HJB equation for a problem involving a controlled Markov chain. If $\{\xi_k^h, k = 0, 1, \ldots\}$ is a controlled Markov chain satisfying

$$\mathbf{P}_x^h \left(\xi_{k+1}^h = z \,|\, \xi_l^h, u_l,\, l = 0, \ldots, k\right) \;=\; p^h(\xi_k^h, z|u_k^h),$$

then V^h has the representation

$$V^h(x) \;=\; \inf_{u \in \mathcal{U}^h} \mathbf{E}_x^h \left[\sum_{k=0}^{N^h-1} \left(\prod_{l=0}^{k-1} \frac{1}{1 + \lambda \Delta t^h(\xi_l^h)} \right) k(\xi_k^h, u_k^h) \Delta t^h(\xi_k^h) \right]. \tag{2.10}$$

Here, N^h is the exit time from D^h and $\mathcal{U}^h$ is an appropriate set of control policies [19]. The minimal cost $V^h(x)$ for this problem is well defined and also the unique solution to (2.6) [18, 19]. It is easy to formally check that (2.6) is the HJB equation for (2.10). For a rigorous proof, we refer the reader to [25].

The original motivation for (2.6) is as a finite difference replacement for (2.3). The replacement of the problem (2.3) by the problem (2.6) could be viewed as an approximation at the level of the PDE. An alternative point of view is to approximate at the level of the *process*, and this is the perspective naturally associated with (2.7) and the representation (2.10). More precisely, for the transition probabilities and interpolation interval defined by (2.8) and (2.9), we have

$$\mathbf{E}_x^h \left[\xi_{k+1}^h - \xi_k^h | \, \xi_l^h, u_l,\, l = 0, \ldots, k\right] = b(\xi_k^h, u_k^h) \Delta t^h(\xi_k^h) \tag{2.11}$$

and

(2.12)

$$\mathbf{cov}_x^h \left[\xi_{k+1}^h - \xi_k^h | \, \xi_l^h, u_l,\, l = 0, \ldots, k\right] = I \Delta t^h(\xi_k^h) + O(h) \Delta t^h(\xi_k^h) + [O(\Delta t^h(\xi_k^h))]^2,$$

where **cov** stands for conditional covariance. These equations imply that an interpolated version of the chain $\{(\xi_k^h, u_k^h)\}$ that uses the interpolation intervals $\Delta t^h(\xi_k^h)$ is a good approximation to the original controlled process (2.1) in the sense of weak convergence (see [19] and section 3). Although in this section we have motivated the forms of the transition probabilities and interpolation interval by starting with finite difference approximations, in later sections we will find it more convenient to construct them directly. We refer the reader to [19, Chapter 5] for an in-depth discussion on methods for constructing these interpolation times and transition functions and for a precise statement of the conditions they should satisfy.

Using either weak convergence methods [18, 19] or viscosity solution methods [1, 14], one can show that the approximation V^h converges to V:

$$\lim_{h \to 0} \sup_{x \in D^h} |V^h(x) - V(x)| \;=\; 0.$$

A rate of convergence result was obtained by Menaldi [23], in a more general setting. Here we show that the same rate estimate can be obtained easily by employing a probabilistic representation for V in terms of the controlled Markov chain. Indeed, using (2.3) and (2.4), we see that V satisfies

$$\lambda V(x) \;=\; \min_{u \in U} \left[L_u^h V(x) + k(x, u) + e^h(x, u)\right] \quad \text{in } D^h. \tag{2.13}$$

A comparison of (2.13) with (2.6) indicates that V has a representation of exactly the same form as (2.10), save that the perturbed running cost $k + e^h$ is used; viz.,

(2.14)
$$V(x) = \inf_{u\in\mathcal{U}^h} \mathbf{E}_x^h \left[\sum_{k=0}^{N^h-1} \left(\prod_{l=0}^{k-1} \frac{1}{1+\lambda\Delta t^h(\xi_l^h)} \right) \left(k(\xi_k^h, u_k^h) + e^h(\xi_k^h, u_k^h) \right) \Delta t^h(\xi_k^h) \right].$$

Note in particular that the two representations are in terms of the same controlled Markov chain. This easily leads to the following rate of convergence result.

THEOREM 2.1. *Given that $V \in C^{2,\alpha}(\bar{D})$, we have*

$$\sup_{x\in D^h} |V^h(x) - V(x)| \;=\; O(h^\alpha) \tag{2.15}$$

as $h \downarrow 0$.

Proof. For $\varepsilon > 0$ consider an ε-optimal control policy for the right-hand side of (2.10), and let $\{(\xi_k^h, u_k^h)\}$ denote the associated controlled chain. Thus

$$V^h(x) \;\geq\; \mathbf{E}_x^h \left[\sum_{k=0}^{N^h-1} \left(\prod_{l=0}^{k-1} \frac{1}{1+\lambda\Delta t^h(\xi_l^h)} \right) k(\xi_k^h, u_k^h)\Delta t^h(\xi_k^h) \right] - \varepsilon.$$

From the representation (2.14) we obtain the estimate

$$V(x) - V^h(x) \leq \mathbf{E}_x^h \left[\sum_{k=0}^{N^h-1} \left(\prod_{l=0}^{k-1} \frac{1}{1+\lambda\Delta t^h(\xi_l^h)} \right) e^h(\xi_k^h, u_k^h)\Delta t^h(\xi_k^h) \right] + \varepsilon.$$

The definition of $\Delta t^h(x)$ implies the existence of $0 < \gamma_1 \leq \gamma_2 < \infty$ such that $\gamma_1 h^2 \leq \Delta t^h(x) \leq \gamma_2 h^2$ for all $x \in h\mathbf{Z}^h$. Using the bound (2.5), the last equation implies

$$V(x) - V^h(x) \leq \left[\sum_{k=0}^{\infty} \left(\frac{1}{1+\lambda\gamma_1 h^2} \right)^k \gamma_2 h^2 \right] O(h^\alpha) + \varepsilon = O(h^\alpha) + \varepsilon,$$

where $O(h^\alpha)$ is uniform in all admissible controls and $x \in D^h$. Since $\varepsilon > 0$ is arbitrary we have $V(x) - V^h(x) \leq O(h^\alpha)$ uniformly in $x \in D^h$. The reverse inequality $V^h(x) - V(x) \leq O(h^\alpha)$ is proved in the same way, save that we consider policies that are ε-optimal for the right-hand side of (2.14). □

Remark 2.2. The results of this section continue to hold if the SDE (2.1) has a uniformly elliptic diffusion coefficient $\sigma(x, u)$.

A formula for the rate. If V enjoys greater regularity than was used above, and if there exist a unique (in law) optimal policy u^* and process x^* for each initial condition $x \in D$, then one can derive an explicit formula for the rate of convergence. For example, if $V \in C^{3,\alpha}(\bar{D})$, then

$$\frac{e^h(x,u)}{h} \;=\; -\frac{1}{2}\sum_{i=1}^{n} |b_i(x,u)| V_{x_i x_i}(x) + O(h^\alpha).$$

To simplify, we assume that all points where the controlled chain might be stopped lie in ∂G. It turns out that the assumed uniqueness allows one to show that interpolated

versions of the optimal discrete time process and control converge weakly to the optimal control and process for the continuous time problem. This can in turn be used to show that

$$\lim_{h\to 0}\frac{V^h(x)-V(x)}{h} \;=\; \mathbf{E}_x\left[\int_0^\tau e^{-\lambda t}\frac{1}{2}\sum_{i=1}^n |b_i(x_t^*,u_t^*)|V_{x_ix_i}(x_t^*)\,dt\right]. \tag{2.16}$$

The analogous argument for a different control problem will be given in detail in section 3.

As an example of how such information might be useful, suppose that instead of the one-sided approximations to $\langle b(x,u), f_x(x)\rangle$ used above, we consider instead a central difference approximation:

$$\langle b(x,u), f_x(x)\rangle \to \sum_{i=1}^n b_i(x,u)\left(f(x+he_i)-f(x-he_i)\right)/2h.$$

In this case $V^h(x)$ has an interpretation as a functional of a controlled Markov chain if and only if $h|b_i(x,u)| \le 1$ for all i, x, and u of interest. Let us assume that this condition holds. Then, under the assumption that $V \in C^{3,\alpha}(\bar{D})$, we obtain $V^h(x)-V(x)=O(h^{1+\alpha})$. Under additional regularity one can obtain an even more refined expression in the spirit of (2.16).

Remark 2.3. Although we have used in a crucial way the fact that V solves the HJB equation (2.3), it is not actually necessary to make an analogous assumption with respect to the value functions for the approximations. This can be useful in cases where the HJB equations for the prelimit problems are not sufficiently well understood. In such cases an additional argument is needed to show that the minimal costs for the prelimit problems can be arbitrarily well approximated by problems for which the associated HJB equations are known to hold rigorously, e.g., approximation in terms of a countable state space controlled Markov chain. This can be established in wide generality by means of weak convergence techniques [19]. However, even if they hold only in a formal sense the relations between the HJB equations for the limiting and prelimit control problems are very useful in motivating the general line of reasoning that we use.

3. A deterministic optimal control example. Consider the following deterministic control system:

$$\begin{cases} \dot{x}_s \;=\; b(x_s,u_s), \quad t<s<T, \\ x_t \;=\; x, \end{cases} \tag{3.1}$$

and finite-horizon value function

$$V(x,t) \;=\; \inf_{u\in\mathcal{U}_t}\left[\int_t^T k(x_s,u_s)\,ds + g(x_T)\right], \tag{3.2}$$

where b, k, etc., are as in section 2, $\mathcal{U}_t$ consists of all measurable functions $u : [t,T]\to U$, and $g\in C^\infty(\mathbf{R}^n)$ is bounded and uniformly Lipschitz continuous. Define $L_uf(x,t)=\langle b(x,u), f_x(x,t)\rangle$. Then $V\in C(\mathbf{R}^n\times[0,T])$ is Lipschitz continuous and is the unique (viscosity) solution of the HJB equation [14]

$$\begin{cases} V_t+\min\limits_{u\in U}\left[L_uV(x,t)+k(x,u)\right] \;=\; 0 \;\text{ in } \mathbf{R}^n\times(0,T), \\ V(x,T) \;=\; g(x) \;\text{ for } x\in\mathbf{R}^n. \end{cases} \tag{3.3}$$

In this section we will consider a general class of approximations to V. In the previous section we assumed that the solution to the appropriate HJB equation was regular on the entire domain of interest. In contrast, in these deterministic examples we can consider the case where the solution may be regular only on a subset of the domain. In the first subsection we introduce a class of approximations to (3.2). In the second and third subsections we consider the cases where the value function is globally regular and regular only on a subset, respectively. Besides proving a rate result, we also show how under certain conditions the coefficients can be identified. Finally, in section 3.4 we give examples from finite difference numerical approximation.

3.1. A general approximation. The class of approximations to (3.1) and (3.3) we consider can be thought of as discrete time "small noise" approximations. Included are small noise optimal control problems associated with the large deviation theory for small noise discrete time stochastic systems, as well as explicit finite difference schemes for the numerical approximation of V. Let $\delta > 0$ denote the approximation parameter. We will restrict δ to values such that T/δ is an integer. While this is done in part just for convenience, it also turns out that this assumption plays a role in determining the specific form for the rate of convergence. See the remark after Theorem 3.3.

For each such $\delta > 0, x \in \mathbf{R}^n$, and $u \in U$, let $\mu^\delta_{x,u}$ denote a probability measure on $\mathbf{R}^n$. In order to have the processes we work with well defined, we assume that the mapping $(x,u) \to \mu^\delta_{x,u}(A)$ is Borel measurable for each Borel set $A \subset \mathbf{R}^n$. Define $t^\delta_k = k\delta$ and $N^\delta = T/\delta$. We consider controlled discrete time processes $\{\xi^\delta_i\}$ that evolve according to

$$\mathbf{P}^\delta_{x,t^\delta_k}\left(\frac{\xi^\delta_{i+1}-\xi^\delta_i}{\delta} \in A \middle| (\xi^\delta_j, u^\delta_j), j \in \{k,\dots,i\}\right) = \mu^\delta_{\xi^\delta_i,u^\delta_i}(A).$$

Here $\mathbf{P}^\delta_{x,t^\delta_k}$ denotes probability conditioned on $\xi^\delta_k = x$.

We consider the family of value functions

$$V^\delta(x,t^\delta_k) \doteq \inf_{u^\delta \in \mathcal{U}^\delta} \mathbf{E}^h_{x,t^\delta_k}\left[\sum_{i=k}^{N^\delta-1} k(\xi^\delta_i, u^\delta_i)\delta + g(\xi^\delta_{N^\delta})\right].$$

The admissible controls in this case can be taken to be the feedback control laws (i.e., each u^δ_k is simply a measurable function from $\mathbf{R}^n$ to U), which implies that the controlled process is a nonstationary Markov chain. In order for V^δ to be close to V we must impose some conditions on $\mu^\delta_{x,u}$. Define $b^\delta(x,u)$ to be the mean of $\mu^\delta_{x,u}(dy)$:

$$b^\delta(x,u) \doteq \int_{\mathbf{R}^n} y\mu^\delta_{x,u}(dy).$$

We require that

$$b^\delta(x,u) = b(x,u) + O(\delta) \quad \text{and} \quad \int_{\mathbf{R}^n} \|y\|^2 \mu^\delta_{x,u}(dy) = O(1), \tag{3.4}$$

where the $O(\delta)$ and $O(1)$ are uniform on compact subsets of $\mathbf{R}^n \times U$. For convenience we will also assume that the supports of the measures $\mu^\delta_{x,u}(dy)$ are bounded uniformly for all $\delta > 0$, $x \in \mathbf{R}^n$, and $u \in U$, which automatically implies the second part of

(3.4). A less restrictive assumption that could be used instead is a uniform bound on the moment generating functions: for each $\alpha \in \mathbf{R}^n$

$$\sup_{\delta>0, u\in U, x\in\mathbf{R}^n} \int_{\mathbf{R}^n} \exp\langle \alpha, y\rangle \mu^\delta_{x,u}(dy) < \infty. \tag{3.5}$$

Of course, this condition is implied by the assumption of a uniform bound on the supports. These conditions are usually easy to check.

The minimal costs V^δ satisfy the following HJB equation [2]:

$$\begin{aligned} \partial^\delta_t V^\delta(x, t^\delta_k) + \min_{u\in U} \left[L^\delta_u V^\delta(x, t^\delta_{k+1}) + k(x,u)\right] &= 0 \ \text{ in } \mathbf{R}^n \times \{0, \dots, N^\delta - 1\}, \\ V^\delta(x, T) &= g(x) \ \text{ for } x \in \mathbf{R}^n, \end{aligned} \tag{3.6}$$

where

$$\partial^\delta_t f(x, t^\delta_k) \doteq \left(f(x, t^\delta_{k+1}) - f(x, t^\delta_k)\right)/\delta$$

and

$$L^\delta_u f(x,t) \doteq \int_{\mathbf{R}^n} (f(x+\delta y, t) - f(x,t))\, \mu^\delta_{x,u}(dy)/\delta.$$

By weak convergence methods [18, 19] (or, alternatively, by viscosity solution methods [1, 14]), one can prove convergence for this scheme:

$$\lim_{\delta\to 0} \sup_{x\in\mathbf{R}^n,\, |x|\le C} \sup_{k=0,1,\dots,N^\delta} |V^\delta(x, t^\delta_k) - V(x, t^\delta_k)| = 0$$

for each $C < \infty$. The rate of convergence depends on the smoothness of the value function V. In general, V is merely Lipschitz continuous and may fail to be differentiable everywhere. However, when V is smooth a rate estimate can easily be established.

3.2. The globally smooth case. Assume now that $V \in C^2(\mathbf{R}^n \times [0,T])$. To obtain a rate of convergence we follow the same procedure as in section 2. Thus the first step is to obtain a representation for V in terms of the controlled chain. By Taylor's theorem, V satisfies the discrete equation

$$\partial^\delta_t V(x, t^\delta_k) + \min_{u\in U} \left[L^\delta_u V(x, t^\delta_{k+1}) + k(x,u) + e^\delta(x, u, t^\delta_k)\right] = 0 \tag{3.7}$$

in $\mathbf{R}^n \times \{0, \dots, N^\delta - 1\}$, where

$$e^\delta(x,u,t) \doteq \left(L_u V(x,t) - L^\delta_u V(x,t)\right) + \left(V_t(x,t) - \partial^\delta_t V(x,t)\right).$$

We therefore have the representation

$$V(x, t^\delta_k) = \inf_{u^\delta \in \mathcal{U}^\delta} \mathbf{E}^\delta_{x, t^\delta_k} \left[\sum_{i=k}^{N^\delta - 1} \left(k(\xi^\delta_i, u^\delta_i) + e^\delta(\xi^\delta_i, u^\delta_i, t^\delta_i)\right)\delta + g(\xi^\delta_{N^\delta}) \right]. \tag{3.8}$$

Equation (3.4) implies $e^\delta(x,u,t) = O(\delta)$ uniformly on compact subsets.

This representation leads to the following result, whose proof is exactly analogous to the proof of Theorem 2.1.

THEOREM 3.1. *Assume that* $V \in C^2(\mathbf{R}^n \times [0,T])$ *and that* (3.4) *holds. Then for each* $C < \infty$ *we have*

$$\sup_{x \in \mathbf{R}^n,\, |x| \le C} \sup_{k=0,1\ldots,N^\delta} |V^\delta(x, t_k^\delta) - V(x, t_k^\delta)| \;=\; O(\delta) \tag{3.9}$$

as $\delta \downarrow 0$.

Remark 3.2. Similar results can be obtained for differential game problems, and also for implicit schemes [19].

If V enjoys a greater degree of regularity and (3.4) is replaced by a stronger assumption, we can refine this result and obtain an explicit expression for the coefficient in the rate of convergence. Let B be any $n \times n$ symmetric matrix. In place of (3.4) we assume

$$\begin{aligned} b^\delta(x,u) &= b(x,u) + \delta s(x,u) + o(\delta), \\ \int_{\mathbf{R}^n} \langle By, y\rangle \mu^\delta_{x,u}(dy) &= q(x,u,B) + o(1), \end{aligned} \tag{3.10}$$

where $s(x,u)$ and $q(x,u,B)$ are continuous in (x,u,B), and where the $o(\delta)$ and $o(1)$ terms are uniform in compact subsets of $\mathbf{R}^n \times U$. We will also need to make the following assumption:

$$\begin{aligned} &\min_{u \in U} [b(x,u) \cdot p + k(x,u)] \quad \text{attains a unique minimum} \\ &\text{at } U^*(x,p), \text{ where } U^* \text{ is of class } C^1. \end{aligned} \tag{3.11}$$

Define

$$r(x,u,t) \;\doteq\; \langle V_x(x,t), s(x,u)\rangle + q(x,u,V_{xx}(x,t)) + \frac{1}{2} V_{tt}(x,t).$$

Note that $-e^\delta(x,u,t)/\delta \to r(x,u,t)$ uniformly on compact sets.

THEOREM 3.3. *Assume that* $V \in C^3(\mathbf{R}^n \times [0,T])$ *and* (3.10), (3.11) *hold. Then we have the explicit rate of convergence*

$$\lim_{\delta \to 0,\, x^\delta \to x,\, t_k^\delta \to t} \frac{V^\delta(x^\delta, t_k^\delta) - V(x^\delta, t_k^\delta)}{\delta} \;=\; \int_t^T r(x_s^*, u_s^*, s)\, ds, \tag{3.12}$$

uniformly on compact subsets, where x_s^* *is the optimal trajectory corresponding to the unique optimal feedback control* $u_s^* = u^*(x_s^*, s) = U^*(x_s^*, V_x(x_s^*, s))$, $t \le s \le T$, *with initial condition* $x_t^* = x$.

Proof. We first prove that u^* is the unique optimal control. Let $\tilde{u}_s$ be any control and let $\tilde{x}_s$ be the associated controlled trajectory that starts at x at time t [24]. We follow the convention of saying that $\tilde{u} = u^*$ if and only if $\tilde{u}_s = u_s^*$ for almost every (a.e.) $s \in [t,T]$. From equation (3.3) we obtain

$$V_s(x,s) + L_{\tilde{u}_s} V(x,s) + k(x, \tilde{u}_s) \ge 0,$$

with equality if and only if $\tilde{u}_s = u^*(x,s)$. Integrating along the trajectory yields

$$\int_t^T k(\tilde{x}_s, \tilde{u}_s)\, ds + g(\tilde{x}_T) \ge V(x,t), \tag{3.13}$$

with equality if and only if $\tilde{u}_s = u^*(\tilde{x}_s, s)$ for a.e. $s \in [t,T]$. Since the solution to $\dot{\phi}_s = b(\phi_s, u^*(\phi_s, s))$ is unique for any initial condition (i.e., $\tilde{x} = x^*$), we obtain equality if and only if $\tilde{u} = u^*$.

We now prove the rate result. Following the argument of Theorem 2.1, we let $\{(\xi_i^\delta, u_i^\delta)\}$ be a δ^2-optimal chain and control for the representation of $V(x^\delta, t_k^\delta)$ given in (3.8). Define interpolated state and control processes x^δ, u^δ by $x_s^\delta = \xi_i^\delta$, $u_s^\delta = u_i^\delta$ on $[t_i^\delta, t_{i+1}^\delta)$ [18, 19]. It follows from the boundedness of the supports of the measures $\mu_{x,u}^\delta$ that the random processes $\{(x^\delta, u^\delta), \delta > 0\}$ are tight (for the precise topology used on the control process, see [19]). It follows from (3.4) and an elementary martingale argument that any limit satisfies (3.1) with probability 1 (w.p.1). Since we have equality in (3.13) if and only if $\tilde{u} = u^*$, the fact that $V^\delta(x^\delta, t_k^\delta) \to V(x,t)$ (Theorem 3.1) and an argument by contradiction imply the weak convergence

$$x^\delta, u^\delta \Longrightarrow x^*, u^*$$

as $\delta \to 0$, $x^\delta \to x$, $t_k^\delta \to t$. Now since $\{(\xi_i^\delta, u_i^\delta)\}$ is δ^2-optimal in the representation (3.8),

$$\frac{V^\delta(x^\delta, t_k^\delta) - V(x^\delta, t_k^\delta)}{\delta} \le \mathbf{E}^\delta_{x^\delta, t_k^\delta}\left[\sum_{i=k}^{N^\delta - 1} \left(r(\xi_i^\delta, u_i^\delta, t_i^\delta) + o(1)\right)\delta\right] + \delta,$$

where the $o(1)$ term is uniform on compact sets. Thus, by the dominated convergence theorem,

$$\limsup_{\delta \to 0, x^\delta \to x,\, t_k^\delta \to t} \frac{V^\delta(x^\delta, t_k^\delta) - V(x^\delta, t_k^\delta)}{\delta} \le \int_t^T r(x_s^*, u_s^*, s)\, ds.$$

The opposite inequality is proven similarly, completing the proof. □

Remark 3.4. Although the assumption that T/δ is an integer is not needed for convergence or even Theorem 3.1, it is needed if we wish to identify the rate coefficient as in the last theorem.

3.3. The general case. In general, V is not smooth everywhere, and consequently one obtains a slower global rate of convergence (see the discussion in the following subsection on rates for numerical schemes). However, because V is smooth in certain regions $\mathcal{N} \subset \mathbf{R}^n \times [0,T]$ [11, 12], one might expect the rate to be faster in these regions of smoothness. We now show that the rate is $O(\delta)$ in such regions.

Let $\mathcal{N}$ be an open, bounded subset of $\mathbf{R}^n \times [0,T]$. Following [12, 13], the set $\mathcal{N}$ is called a *region of strong regularity* (RSR) provided

1. $V \in C^3(\bar{\mathcal{N}})$.
2. Assumption (3.11) holds.
3. Given $(x,t) \in \mathcal{N}$, denote by x_s^* and $u_s^* = u^*(x_s^*, s) = U^*(x_s^*, V_x(x_s^*, s))$, $t \le s \le T$, the unique optimal state trajectory and control with initial condition $x_t^* = x$. Define

$$\sigma = \sigma_{x,t} = \inf\{s > t : (x_s^*, s) \notin \mathcal{N}\},$$

$$y = y_{x,t} = x^*(\sigma), \qquad z = z_{x,t} = (y, \sigma).$$

Then $(x,t) \in \mathcal{N}$ implies $(x_s^*, s) \in \mathcal{N}$, $t \le s < \sigma$, and $\sigma_{x,t} = T$.

4. $\partial\mathcal{N} = \Gamma_1 \cup \Gamma_2$, where $\Gamma_1 = \{z_{x,t} : (x,t) \in \mathcal{N}\}$ is an open subset of $\mathbf{R}^n \times \{T\}$. For information regarding the existence of RSRs, we refer the reader to [11, 12].

THEOREM 3.5. *Assume* (3.10) *and let* $\mathcal{N}$ *be an RSR. Then*

$$\lim_{\delta \to 0,\, x^\delta \to x,\, t_k^\delta \to t} \frac{V^\delta(x^\delta, t_k^\delta) - V(x^\delta, t_k^\delta)}{\delta} \;=\; \int_t^T r(x_s^*, u_s^*, s)\, ds \tag{3.14}$$

uniformly on compact subsets of $\mathcal{N}$. *Consequently,*

$$|V^\delta - V| \;=\; O(\delta) \quad \text{in } \mathcal{N}$$

as $\delta \to 0$.

Proof (sketch). This result is proven by modifying the proof of Theorem 3.3 along the lines of [12, 13]. However, in this proof we use a slightly modified representation for $V(x,t)$. In place of (3.8) we exploit the strong Markov property to write

$$V(x, t_k^\delta) = \inf_{u^\delta \in \mathcal{U}^\delta} \mathbf{E}^\delta_{x, t_k^\delta} \left[\sum_{i=k}^{M^\delta - 1} \left(k(\xi_i^\delta, u_i^\delta) + e^\delta(\xi_i^\delta, u_i^\delta, t_i^\delta) \right) \delta + V(z^\delta) \right],$$

where $M^\delta = \inf\{i > k : (x_i^\delta, t_i^\delta) \notin \mathcal{N}\}$ is the discrete time of first exit from $\mathcal{N}$, $\sigma^\delta = t^\delta_{M^\delta}$, and $z^\delta = (\xi^\delta_{M^\delta}, \sigma^\delta)$. We can also write an analogous representation for $V^\delta(x, t_k^\delta)$ in terms of this stopping time and location. If we let $\{(\xi_i^\delta, u_i^\delta)\}$ be δ^2-optimal for V as in the proof of Theorem 3.3, then we obtain

$$\frac{V^\delta(x^\delta, t_k^\delta) - V(x^\delta, t_k^\delta)}{\delta} \le \mathbf{E}^\delta_{x^\delta, t_k^\delta} \left[\sum_{i=k}^{M^\delta - 1} \left(r(\xi_i^\delta, u_i^\delta, t_i^\delta) + o(1) \right) \delta + \frac{V^\delta(z^\delta) - V(z^\delta)}{\delta} \right] + \delta.$$

Recall that the bound (3.5) holds for the moment generating functions of the distributions $\mu^\delta_{x,u}$. Because of this bound an upper large deviation principle holds for the interpolated processes x_s^δ [7, 8]. The large deviation upper bound implies that if $x^\delta \to x$ and $t_k^\delta \to t$ as $\delta \to 0$, then given $\eta > 0$, there exists $c > 0$ such that for all sufficiently small $\delta > 0$,

$$\mathbf{P}^\delta_{x^\delta, t_k^\delta} \left(\sup_{t_k^\delta \le s \le T} |x_s^\delta - x_s^*| > \eta \right) \;\le\; e^{-c/\delta},$$

and thus by parts (iii) and (iv) of the definition of a RSR, for all sufficiently small $\eta > 0$,

$$\mathbf{P}^\delta_{x^\delta, t_k^\delta} \left(|z^\delta - z| > \eta \right) \;\le\; e^{-c/\delta}.$$

Since $V(x,t)$ and $V^\delta(x,t)$ are uniformly bounded in $(x,t) \in \mathbf{R}^n \times [0,T]$ and $\delta \in (0,1)$, $(V^\delta(z^\delta) - V(z^\delta))/\delta$ is uniformly bounded above by some constant times $1/\delta$. Therefore

$$\begin{aligned} \frac{V^\delta(x^\delta, t_k^\delta) - V(x^\delta, t_k^\delta)}{\delta} \;\le\; & \mathbf{E}^\delta_{x^\delta, t_k^\delta} \left[\left(\int_{t_k^\delta}^{T \wedge \sigma^\delta} r(x_s^\delta, u_s^\delta, s)\, ds + o(1) \right) 1_{\{|z^\delta - z| < \eta\}} \right] \\ & + O(e^{-c/\delta}) \left(1 + O(1/\delta)\right), \end{aligned}$$

and we can conclude as in the proof of Theorem 3.3. ∎

3.4. Numerical approximations. In this subsection we specialize from the previous two subsections to the case of an explicit finite difference approximation V^h to V. As in section 2, let $h > 0$ denote the space discretization step size, Δt^h denote the time discretization, etc. Select $v > 0$ such that

$$v \geq \max_{x \in \mathbf{R}^n,\, u \in U} \|b(x,u)\|_1,$$

and define the time step size

$$\Delta t^h \doteq h/v.$$

We restrict attention to values of $h > 0$ such that $N^h = T/\Delta t^h$ is an integer. We define the discrete times $t_k^h \doteq k\Delta t^h$ and consider the transition probabilities

$$p^h(x,z|u) \doteq \begin{cases} 1 - \|b(x,u)\|_1/v & \text{if } z = x, \\ b_i^{\pm}(x,u)/v & \text{if } z = x \pm he_i \text{ for some } i = 1,\dots,n, \\ 0 & \text{otherwise.} \end{cases}$$

We fit this example into the general framework by setting $\delta \doteq h/v$ and defining $\mu_{x,u}^{\delta}$ by

$$\mu_{x,u}^{\delta}(A) \doteq \sum_{w \in \mathbf{Z}^n : vw \in A} p^h(x, x + hw|u)$$

for all Borel sets $A \subset \mathbf{R}^n$. These definitions imply

$$\int_{\mathbf{R}^n} y\mu_{x,u}^{\delta}(dy) = b(x,u),$$

$$\int_{\mathbf{R}^n} \langle By, y\rangle \mu_{x,u}^{\delta}(dy) = \sum_{i=1}^{n} vB_{ii}|b_i(x,u)|,$$

where $B = (B_{ij})$. Hence equation (3.4) holds and we may apply Theorem 3.1. The function $r(x,u,t)$ in this example takes the form

$$v\sum_{i=1}^{n} V_{x_i x_i}(x,t)|b_i(x,u)| + \frac{1}{2}V_{tt}(x,t),$$

and under the appropriate conditions Theorems 3.3 and 3.5 hold as well.

As an application of the rate of convergence results, consider the following modification of the numerical approximation. It is well known that it is at least theoretically advantageous to allow the interpolation times Δt^h to depend on the state and control: $\Delta t^h = \Delta t^h(x,u)$. While it is obvious that such added flexibility in the selection of a numerical scheme can only help, it may not be the case that the additional effort required to program such schemes is worth the improvement in accuracy. The rate result allows one to estimate the improvement before implementing a more complicated scheme. Note also that the total time taken to complete the numerical computations may be reduced if the interpolation times are allowed to depend on the state and control.

We consider such a modification for the example that was just considered. The underlying reason why one expects state- and control-dependent interpolation times

to improve numerical performance is because they allow one to reduce the probability that the controlled chain remains at any given state; i.e., they allow one to design chains for which $p^h(x,x|u) = 0$ [19, Chapter 5]. With only a little effort, one can modify the proofs of the theorems stated above to allow such state and control dependency of the interpolation times. (Note that if $\Delta t^h(x)$ is state- or control-dependent then for any given x one does not know a priori which continuous times correspond to the interpolation times chosen by the discrete algorithm. Because of this, one must keep track of the interpolation times used as one iterates backward when solving the discrete HJB equation and define $V^h(x,0)$ via an interpolation.) The state-dependent interpolation times and transition probabilities that are appropriate are $\Delta t^h(x,u) = h/|b(x,u)\|_1$ and

$$p^h(x,z|u) = \begin{cases} 0 & \text{if } z = x, \\ b_i^{\pm}(x,u)/\|b(x,u)\|_1 & \text{if } z = x \pm he_i \text{ for some } i = 1,\dots,n, \\ 0 & \text{otherwise.} \end{cases}$$

The measures $\mu^{\delta}_{x,u}$ are defined as before. With these definitions, we again have $\int_{\mathbf{R}^n} y\mu^{\delta}_{x,u}(dy) = b(x,u)$, but now

$$r(x,u,t) = \|b(x,u)\|_1 \sum_{i=1}^{n} V_{x_i x_i}(x,t)|b_i(x,u)| + \frac{1}{2}V_{tt}(x,t).$$

Since $v \geq \|b(x,u)\|_1$ for all x and u, one expects the rate with the new transition probabilities and interpolation times to often be better than that of the previous setup. If v is much larger than "typical values" of $\|b(x,u)\|_1$, then the extra programming effort may indeed be worthwhile. However, if one has a bound such as $a \leq \inf_{x,u} \|b(x,u)\|_1 \leq \sup_{x,u} \|b(x,u)\|_1 \leq Ca$, where C is not very large, then it is probably not worthwhile.

4. An example with ergodic cost and a reflecting diffusion. In order to demonstrate the versatility of the approach, in this section we will consider a variation on the numerical approximation problem considered in section 2. More precisely, we treat the analogous problem where the cost to be minimized is an ergodic cost, and where a reflecting diffusion replaces the model (2.1). In order to define the reflecting diffusion model we must specify a reflection direction for each point of ∂D. Let $n(x)$ denote the inward unit normal to ∂D at $x \in \partial D$. The reflection direction will be denoted by a unit vector $r(x)$. We will assume that $\langle r(x), n(x)\rangle > 0$ for all $x \in \partial D$, and that $r \in C^{\infty}(\mathbf{R}^m)$. Since ∂D is smooth, we can assume that the function n is defined and smooth in an open neighborhood O of ∂D, and that $\langle r(x), n(x)\rangle$ is uniformly bounded below away from zero on O.

We next describe the reflected diffusion model. Since the theory of such equations is not our focus here, the description will only be heuristic. A precise definition can be found in [21] or [6]. The replacement for (2.1) takes the form

$$dx_t = b(x_t, u_t)\,dt + dw_t + dz_t, \tag{4.1}$$

where b satisfies all the assumptions used in section 2. The process z_t is a w.p.1 bounded variation function of t that constrains x_t to remain in $\bar{D}$. It acts in the following way. As long as $x_t \in D$ (recall that D is open), z_t does not affect the process x_t at all, which means that $dz_t = 0$ for all such t. If $x_t \in \partial D$, then z_t can "push" the process so as to maintain $x_t \in \bar{D}$. The requirements on the "push" are:

- that it be in the direction $r(x_t)$,
- that $x_t \in \bar{D}$ for all t w.p.1.

These requirements are formalized by the equations

$$|z|_t = \int_0^t I_{\{x_s \in \partial D\}} d|z|_s \quad \text{and} \quad z_t = \int_0^t r(x_s) d|z|_s,$$

where $|z|_t$ denotes the total variation of z on the interval $(0, t]$. Under the assumptions made above on b, r, and D, a solution to (4.1) exists and is unique. For precise statements and more discussion, we refer the reader to [21, 6, 19].

The reflecting diffusion model described above is especially useful when the controlled process is considered on an infinite time horizon, since it allows the domain on which the process is defined to be bounded without actually stopping the process when it hits ∂D. In some problems, there is a cost proportional to the constraining action of the process z_t. Because of this, we consider the minimal cost defined by

$$\gamma \doteq \inf_{u \in \mathcal{U}} \limsup_{T \to \infty} \frac{1}{T} \mathbf{E}_x \left[\int_0^T k(x_t, u_t)\, dt + \int_0^T l(x_t)\, d|z|_t \right], \tag{4.2}$$

where $l \in C^\infty(\mathbf{R}^n)$ is bounded and uniformly Lipschitz continuous. Although a priori the minimal cost might depend on the initial condition x, it turns out under our assumptions that the cost is independent of x.

The appropriate HJB equation for this problem is

$$\begin{cases} \gamma = \min_{u \in U} [L_u V(x) + k(x, u)] & \text{in } D, \\ 0 = l(x) + \langle V_x(x), r(x) \rangle & \text{on } \partial D, \end{cases} \tag{4.3}$$

where L_u is again defined by

$$L_u f(x) \doteq \langle b(x, u), f_x(x) \rangle + \frac{1}{2} \mathrm{tr}[f_{xx}(x)].$$

The solution to this equation is the pair $(\gamma, V(\cdot))$. Note that if $(\gamma, V(\cdot))$ solves (4.3), then so does $(\gamma, V(\cdot) + c)$ for any $c \in \mathbf{R}$. It turns out that this is exactly the form of nonuniqueness associated with the solutions to (4.3); i.e., if $(\gamma_1, V_1(\cdot))$ and $(\gamma_2, V_2(\cdot))$ both solve (4.3), then $\gamma_1 = \gamma_2$ and $V_1(\cdot) - V_2(\cdot)$ is a constant. We will assume, as in section 2, that $V \in C^{2,\alpha}(\bar{D})$ (see [22]).

A general reference for the Markov chain optimal control problems discussed in this section is section 5 of Chapter 7 in [19]. Recall the definition $D^h \doteq D \cap h\mathbf{Z}^n$. While the process x_t is in D it is the same as the process of section 2. This suggests that we can continue to use the transition probabilities and interpolation intervals defined by (2.9) and (2.8), respectively. However, we must still define the approximations for the boundary condition. Define the operator A by

$$Af(x) = \langle f_x(x), r(x) \rangle$$

for $f \in C^1(\mathbf{R}^n)$. Then the boundary condition can be written $0 = l(x) + AV(x)$ for $x \in \partial D$. Let ∂D^h_+ be a set that contains all points in $(\mathbf{R}^n \backslash D) \cap h\mathbf{Z}^n$ that can be reached from some point in D^h in one step for some choice of the control, i.e., all y such that

$$p^h(x, y|u) > 0 \text{ for some } x \in D^h \text{ and } u \in U.$$

We interpret ∂D_+^h as the "discrete reflecting boundary." Although one can often take ∂D_+^h to be exactly those points that can be reached in one step from D^h, the formulation as given above, which allows a bigger set, is sometimes needed. We will assume that for all h sufficiently small, $\partial D_+^h \subset O$, and remind the reader that O is an open set on which both $n(\cdot)$ and $r(\cdot)$ are defined.

We next consider the transition probabilities for $x \in \partial D_+^h$. The role of these transitions will be to "mimic" the behavior of the reflecting term z_t. The construction of the transition functions obviously depends on the shape of ∂D, $D^h \cup \partial D_+^h$, and the function $r(\cdot)$. For most problems the construction is straightforward and intuitive, since we are dealing here with only first-order boundary operators. Since it is not our goal to discuss methods for constructing these functions, we will simply assume the existence of transition probabilities that satisfy the local consistency equations (4.4) and (4.5) below, and refer the reader to Chapters 5 and 8 of [19] for further information.

Let $p^h(x,y)$ be the transition function for points $x \in \partial D_+^h$. (Note that we do not include a control for such states. This is because in our setup the reflection direction is not controlled. An interesting example where the reflection direction is controlled appears in [20].) Let $\alpha^h(x)r(x)+s^h(x)$ denote the decomposition of the mean discrete reflection $m^h(x) \doteq \sum_{y \in D^h \cup \partial D_+^h} [y-x]\, p^h(x,y)$ into the orthogonal projection onto the subspace spanned by $r(x)$ and its complement. Then the minimal type of "local consistency" we require of the functions $p^h(x,y)$ is

$$\inf_{h>0, x \in \partial D_+^h} \alpha^h(x)/h > 0,$$

$$s^h(x)/h \to 0 \text{ uniformly in } x \in \partial D_+^h, \tag{4.4}$$

and

$$c^h(x)/h \doteq \sum_{y \in D^h \cup \partial D_+^h} \left[y-x-m^h(x)\right] \left[y-x-m^h(x)\right]' p^h(x,y)/h \to 0 \tag{4.5}$$

uniformly in $x \in \partial D_+^h$. This last equation is automatic if p^h is only supported on neighboring points. The essential consequence of these conditions is that $s^h(x)/\alpha^h(x) \to 0$ and $c^h(x)/\alpha^h(x) \to 0$ uniformly in ∂D_+^h, the first of which shows that the component orthogonal to $r(x)$ vanishes faster than the component in the direction $r(x)$, and the second of which shows that the quadratic variation around the mean vanishes faster than $\alpha^h(x)$. It is often the case that one can choose the probabilities so that $s^h(x) \equiv 0$. We must also assume that the "radius" of ∂D_+^h tends to zero:

$$\sup_{x \in \partial D_+^h} \inf_{y \in D^h} \|x-y\| \to 0 \tag{4.6}$$

as $h \to 0$.

Define the operator A^h by

$$A^h f(x) = \sum_{y \in D^h \cup \partial D_+^h} [f(y)-f(x)] \frac{p^h(x,y)}{\alpha^h(x)}$$

for points $x \in \partial D_+^h$. Then the conditions given above imply for all $f \in C^1(\mathbf{R}^n)$ that

$$\left|Af(x) - A^h f(x)\right| = o(1)$$

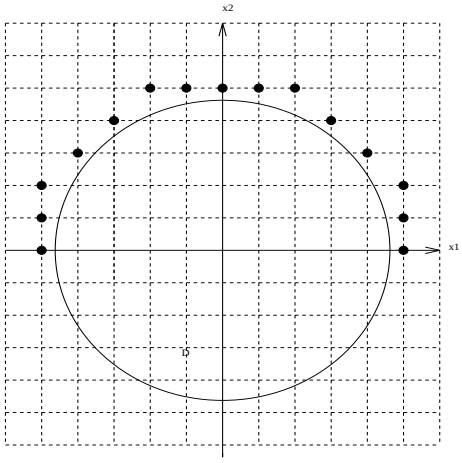


FIG. 4.1. *Boundary portion* ∂D_+^h.

uniformly in $x \in \partial D_+^h$, which is sufficient for convergence. However, in order to specify a rate of convergence, we need to be more precise in describing how fast $s^h(x)$ and $c^h(x)$ tend to zero. We will assume that

$$s^h(x) = O(h^2), \quad c^h(x) = O(h^2) \tag{4.7}$$

uniformly in $x \in \partial D_+^h$. (Note that if we want to identify coefficients, then more is needed; i.e., we need expansions of the form

$$s^h(x) = h^2 \tilde{s}(x) + o(h^2), \quad c^h(x) = h^2 \tilde{c}(x) + o(h^2)$$

for some continuous functions $\tilde{s}(x)$ and $\tilde{c}(x)$.)

Example 4.1. We consider the case $n = 2$, $D = \{x : \|x\| \le 1\}$, and $r(x) = n(x)$. Suppose $h = 1/k$, where k is an integer. In this case we can take the set ∂D_+^h to be as in Figure 4.1. The definition is

$$p^h(x,y) = \begin{cases} \dfrac{x_1^{\mp}}{|x_1| + |x_2|} & \text{if } y = x \pm h(1,0), \\ \dfrac{x_2^{\mp}}{|x_1| + |x_2|} & \text{if } y = x \pm h(0,1), \\ 0 & \text{otherwise.} \end{cases}$$

For this example, we have

$$m^h(x) = hr(x)\|x\|_2/\|x\|_1, \quad \alpha^h(x) = h\|x\|_2/\|x\|_1, \quad s^h(x) = 0,$$

and

$$c^h(x) = h^2 \frac{x_2^2|x_1| + x_1^2|x_2|}{\|x\|_1^3} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

The transition probabilities at the points of ∂D_+^h will play the role of the constraining process z in (4.1); i.e., if the process attempts to leave D^h then it is returned

instantly by a "push" in the appropriate direction [19]. Because of the instantaneous nature of the push, the interpolation interval that is correct for the points in ∂D_+^h is $\Delta t^h(x) = 0$.

The discrete replacement for (4.3) is

$$\begin{cases} \gamma^h &= \min_{u\in U}\left[L_u^h V^h(x) + k(x,u)\right] \quad \text{in } D^h, \\ 0 &= l(x) + A^h V^h(x) \quad \text{on } \partial D_+^h, \end{cases} \tag{4.8}$$

where L_u^h is as in section 2.

For an admissible control $\{u_k^h, k = 0, 1, \ldots\}$, let $\{\xi_k^h, k = 0, 1, \ldots\}$ be the corresponding controlled Markov chain; i.e.,

$$\mathbf{P}_x^h\left(\xi_{k+1}^h = z \,|\, \xi_l^h, u_l,\, l = 0, \ldots, k\right) \;=\; p^h(\xi_k^h, z|u_k^h) \quad \text{if } x \in D^h,$$

and

$$\mathbf{P}_x^h\left(\xi_{k+1}^h = z \,|\, \xi_l^h, u_l,\, l = 0, \ldots, k\right) \;=\; p^h(\xi_k^h, z) \quad \text{when } x \in \partial D_+^h.$$

Define $T_N^h = \sum_{i=0}^{N-1} \Delta t^h(\xi_i^h)$. Note that when $\xi_i^h \in \partial D_+^h$ the corresponding summand in T_N^h is zero. Equation (4.8) is the HJB equation for the Markov chain stochastic optimal control problem whose transition probabilities are those given above and for which the cost to be minimized is

$$\gamma^h = \limsup_{N\to\infty} \mathbf{E}_x^h\left[\left(\sum_{i=0}^{N-1} k(\xi_i^h, u_i^h)\Delta t^h(\xi_i^h) + \sum_{i=0}^{N-1} I_{\{\xi_i^h \in \partial D_+^h\}} l(\xi_i^h)\alpha^h(\xi_i^h)\right)/T_N^h\right]$$

(cf. [19, Chapter 7]). One can easily check that the chain is ergodic for any time independent feedback control. Because of this, the limit superior is actually a limit, and the limiting value is independent of x. The equation (4.8) exhibits the same type of nonuniqueness as the original HJB equation (4.2), namely, if $(\gamma_1^h, V_1^h(\cdot))$ and $(\gamma_2^h, V_2^h(\cdot))$ both solve (4.8), then $\gamma_1^h = \gamma_2^h$ and $V_1^h(\cdot) - V_2^h(\cdot)$ is a constant.

Define the boundary error

$$g^h(x) \doteq AV(x) - A^h V(x).$$

Thanks to our assumptions on V and (4.7), we have

$$|g^h(x)| = O(h).$$

We can then rewrite (4.3) in a form analogous to that of (4.8):

$$\begin{cases} \gamma &= \min_{u\in U}\left[L_u^h V(x) + k(x,u) + e^h(x,u)\right] \quad \text{in } D^h, \\ 0 &= l(x) + A^h V(x) + g^h(x) \quad \text{on } \partial D_+^h, \end{cases} \tag{4.9}$$

where e^h is defined as in section 2 by

$$e^h(x,u) \doteq L_u V(x) - L_u^h V(x).$$

(Recall that $|e^h(x,u)| = O(h^\alpha)$.) Thus γ has a representation as the minimal cost for the Markov chain optimal control problem whose transition probabilities are the same

as those for γ^h and for which the cost to be minimized is

$$\gamma = \limsup_{N\to\infty} \mathbf{E}_x^h \left[\left(\sum_{i=0}^{N-1} \left[k(\xi_i^h, u_i^h) + e^h(\xi_i^h, u_i^h) \right] \Delta t^h(\xi_i^h) \right.\right.$$
$$\left.\left. + \sum_{i=0}^{N-1} I_{\{\xi_i^h \in \partial D_+^h\}} \left[l(\xi_i^h) + g^h(\xi_i^h) \right] \alpha^h(\xi_i^h) \right) / T_N^h \right].$$

A comparison of these two representations allows us to prove the following rate of convergence.

THEOREM 4.2. *Assume* (4.7) *and all the smoothness conditions assumed of* b, k, ∂D, *etc., in this section and section* 2. *Given that* $V \in C^{2,\alpha}(\bar{D})$, *we have*

$$\sup_{x\in D^h} |V^h(x) - V(x)| \;=\; O(h^\alpha) \tag{4.10}$$

as $h \downarrow 0$.

Proof. We can use the same proof as that of Theorem 2.1 as soon as we show that

$$\limsup_{N\to\infty} \mathbf{E}_x^h \left[\left(\sum_{i=0}^{N-1} e^h(\xi_i^h, u_i^h) \Delta t^h(\xi_i^h) + \sum_{i=0}^{N-1} I_{\{\xi_i^h \in \partial D_+^h\}} g^h(\xi_i^h) \alpha^h(\xi_i^h) \right) / T_N^h \right] \tag{4.11}$$
$$= O(h^\alpha)$$

uniformly in all admissible controls and $x \in D^h$.

The main difficulty in proving such a bound is in dealing with the second term in the sum. Let $\theta(x)$ (a C^2 function from $\mathbf{R}^n \to \mathbf{R}$) and $\eta > 0$ be such that for all sufficiently small $h > 0$

$$\inf_{x\in\partial D_+^h} \left\langle \frac{\alpha^h(x) r(x) + s^h(x)}{\alpha^h(x)}, \theta_x(x) \right\rangle \geq \eta. \tag{4.12}$$

Then for any $k = 0, 1, 2, \ldots$,

$$\theta(\xi_k^h) - \theta(\xi_0^h) \;= \sum_{i=0}^{k-1} \left[\theta(\xi_{i+1}^h) - \theta(\xi_i^h) \right]$$
$$= \sum_{i=0}^{k-1} \langle \xi_{i+1}^h - \xi_i^h, \theta_x(\xi_i^h) \rangle + \frac{1}{2} \sum_{i=0}^{k-1} \left(\xi_{i+1}^h - \xi_i^h \right)' \theta_{xx}(\tilde{\xi}_i^h) \left(\xi_{i+1}^h - \xi_i^h \right),$$

where $\tilde{\xi}_i^h$ is an appropriately selected point between ξ_i^h and ξ_{i+1}^h. We rewrite this last equation as

$$\sum_{i=0}^{k-1} I_{\{\xi_i^h \in \partial D_+^h\}} \langle \xi_{i+1}^h - \xi_i^h, \theta_x(\xi_i^h) \rangle + \frac{1}{2} \sum_{i=0}^{k-1} I_{\{\xi_i^h \in \partial D_+^h\}} \left(\xi_{i+1}^h - \xi_i^h \right)' \theta_{xx}(\tilde{\xi}_i^h) \left(\xi_{i+1}^h - \xi_i^h \right)$$
$$= \; \theta(\xi_k^h) - \theta(\xi_0^h)$$
$$- \sum_{i=0}^{k-1} I_{\{\xi_i^h \in D^h\}} \langle \xi_{i+1}^h - \xi_i^h, \theta_x(\xi_i^h) \rangle - \frac{1}{2} \sum_{i=0}^{k-1} I_{\{\xi_i^h \in D^h\}} \left(\xi_{i+1}^h - \xi_i^h \right)' \theta_{xx}(\tilde{\xi}_i^h) \left(\xi_{i+1}^h - \xi_i^h \right).$$

By using (4.12), the fact that $c^h(x) = o(\alpha^h(x))$ uniformly in x, and equations (2.11) and (2.12), we obtain

$$\frac{\eta}{2}\mathbf{E}_x^h \sum_{i=0}^{k-1} I_{\{\xi_i^h \in \partial D_+^h\}} \alpha^h(\xi_i^h) \le 2\|\theta\|_\infty + K\mathbf{E}_x^h T_k^h,$$

for all sufficiently small $h > 0$, where $K < \infty$ is independent of both h and k.

For $T \in [0, \infty)$, define the stopping time $M_T = \min\{k : T_k^h \ge T\}$. Note that $T_{M_T}^h / T \to 1$ uniformly. It follows from the last display and the fact that M_T is a stopping time that

$$\frac{\eta}{2}\mathbf{E}_x^h \sum_{i=0}^{M_T-1} I_{\{\xi_i^h \in \partial D_+^h\}} \alpha^h(\xi_i^h) \le 2\|\theta\|_\infty + K\mathbf{E}_x^h T_{M_T}^h. \tag{4.13}$$

We can now bound (4.11). According to equation (2.5) in section 2 $|e^h(\xi_i^h, u_i^h)| = O(h^\alpha)$. Thus

$$\limsup_{N\to\infty} \mathbf{E}_x^h \left[\left(\sum_{i=0}^{N-1} e^h(\xi_i^h, u_i^h) \Delta t^h(\xi_i^h) \right) / T_N^h \right] = O(h^\alpha).$$

On the other hand, we recall that $\left|g^h(x)\right| = \left|AV(x) - A^h V(x)\right| = O(h)$. By combining this with (4.13), we obtain

$$\limsup_{N\to\infty} \mathbf{E}_x^h \left[\left(\sum_{i=0}^{N-1} I_{\{\xi_i^h \in \partial D_+^h\}} g^h(\xi_i^h) \alpha^h(\xi_i^h) \right) / T_N^h \right]$$

$$= \limsup_{T\to\infty} \mathbf{E}_x^h \left[\left(\sum_{i=0}^{M_T-1} I_{\{\xi_i^h \in \partial D_+^h\}} g^h(\xi_i^h) \alpha^h(\xi_i^h) \right) / T_{M_T}^h \right] = O(h),$$

which proves (4.11). □

An examination of the proof just given shows that the errors in the approximations to the boundary condition are of smaller order than the approximations on the interior.

As in other sections, with added regularity one can identify the coefficient of the rate of convergence. In this problem one finds two terms in the rate. One term is a functional of the x_s process and represents errors due to approximation on D, while the other is a functional of the boundary local time process z_s and represents errors due to the approximation of the boundary condition.

5. Comments and extensions. In this final section we make some general comments and discuss some extensions of our methodology.

5.1. General method. The general method we have employed can be summarized with the following heuristics.

Consider the problem of approximating the solution V to the equation

$$A(V) + k = 0 \tag{5.1}$$

by an approximation V^h given by

$$A^h(V^h) + k = 0. \tag{5.2}$$

The key to our method lies in the use of appropriate representations to solutions of equations of the type (5.2). Let us suppose that

$$V^h = \mathcal{R}^h(k), \tag{5.3}$$

for some representation operator $\mathcal{R}^h$. The operators A, A^h, and $\mathcal{R}^h$ are in general nonlinear. We have assumed that they are obtainable from linear operators via min, max, min-max, or max-min operations. Let us write

$$e^h = A(V) - A^h(V).$$

Then equation (5.1) can be rewritten as

$$A^h(V) + [e^h + k] = 0,$$

and consequently V has a representation determined by the method of approximation:

$$V = \mathcal{R}^h(e^h + k). \tag{5.4}$$

To compare V with V^h, we formally use the fact that $\mathcal{R}^h$ is obtained from a linear operator by one or more minimization or maximization operations. This allows us to write

$$V = \mathcal{R}^h(e^h + k) = \mathcal{R}^h(k) + O(|e^h|). \tag{5.5}$$

Thus if $|e^h| = O(h^\alpha)$, depending on the smoothness of V, this yields the rate of convergence estimate

$$V^h - V = O(h^\alpha). \tag{5.6}$$

More detailed information is available with stronger assumptions. Suppose that $e^h = h^\alpha \phi + O(h^{\alpha+\delta})$, for some $\delta > 0$, and (5.5) is improved:

$$V = \mathcal{R}^h(e^h + k) = \mathcal{R}^h(k) + \mathcal{R}_1^h(e^h). \tag{5.7}$$

Then we have the explicit limit

$$\lim_{h \to 0} \frac{V^h - V}{h^\alpha} = \mathcal{R}_1(\phi). \tag{5.8}$$

5.2. Partial differential equations. Our approach is applicable to PDEs which need not have any a priori connection to control theory. The simplest instance is that of linear equations. For example, consider a linear uniformly elliptic PDE with smooth coefficients, boundary, and boundary data. Such a boundary value problem has a smooth solution, of sufficient regularity to apply our theory and obtain rate of convergence estimates for a variety of approximation methods. The representation for linear equations and their approximations is quite simple, in that no minimizations or maximizations are required (cf. Feynman–Kac formulas).

A second instance of interest is the case of quasi-linear or even fully nonlinear uniformly elliptic/parabolic PDE. Smooth classical solutions are often available; see [16]. To apply our approach, a representation is needed, and indeed this can be obtained in a great many cases using control or game theory.

To illustrate, let us consider an example similar to the problem of section 2. We wish only to communicate the general idea, and omit technical details. Suppose that the fully nonlinear equation

$$\begin{cases} \lambda V(x) \;=\; F(V_{xx}(x)) + k(x) \;\text{ in } D, \\ \;\; V(x) \;=\; 0 \;\text{ on } \partial D, \end{cases} \tag{5.9}$$

has a unique classical solution $V \in C^{2,\alpha}(\bar{D})$, where F is a smooth nonlinear function with bounded gradient satisfying

(i) $\xi' F_{XX}(X)\xi \geq c|\xi|^2$, $c > 0$, and

(ii) $\lim_{|X|\to\infty} |F(X)|/|X| = 0$.

We have not assumed that F is convex, nor any other specific form. Following [9] (see also [10]), F admits a max-min representation:

$$F(X) = \max_{v\in\mathbf{R}^{n^2}} \min_{u\in\mathbf{R}^{n^2}} \left[\sum_{i,j=1}^{n} \left(\int_0^1 \frac{\partial F}{\partial X_{ij}}((1-r)v + ru)\, dr \right) (X_{ij} - v_{ij}) + F(v) \right]. \tag{5.10}$$

In view of this, let us write

$$F(X) + k = \max_{v\in\mathbf{R}^{n^2}} \min_{u\in\mathbf{R}^{n^2}} \left[\sum_{i,j=1}^{n} a_{ij}(u,v) X_{ij} + \hat{k}(x,u,v) \right],$$

where the matrix $a_{ij}(u,v)$ is defined from (5.10) and $\hat{k}(x,u,v) = k(x) - \sum_{i,j=1}^{n} a_{ij}(u,v)v_{ij} + F(v)$. Suppose that we can write $a(u,v) = \frac{1}{2}\sigma(u,v)\sigma(u,v)'$ for some Lipschitz matrix function σ.

The desired game theoretic representation for V is

$$V(x) \;=\; \inf_{u_\cdot} \sup_{v_\cdot} \mathbf{E}_x \left[\int_0^\tau e^{-\lambda t} \hat{k}(x_t, u_t, v_t)\, dt \right], \tag{5.11}$$

where

$$dx_t = \; \sigma(u_t, v_t)\, dw_t.$$

For precise information concerning games and their strategies, see [15].

A finite difference approximation V^h can be constructed, along the same lines as in section 2, which will have a game representation. Note that the various quantities will depend on the additional control variable v. Then a straightforward modification of the proof of Theorem 2.1 yields the rate of convergence estimate

$$\sup_{x\in D^h} |V^h(x) - V(x)| \;=\; O(h^\alpha)$$

as $h \downarrow 0$.

REFERENCES

[1] G. Barles and P. E. Souganidis, *Convergence of approximation schemes for fully nonlinear second order equations*, J. Asymptotic Anal., 4 (1991), pp. 271–283.

[2] D. Bertsekas and S. Shreve, *Stochastic Optimal Control: The Discrete Time Case*, Academic Press, New York, 1978.
[3] I. Capuzzo Dolcetta and M. Falcone, *Discrete dynamic programming and viscosity solutions of the Bellman equation*, Ann. Inst. H. Poincaré, Anal. Non Linéaire, 6 (1989), pp. 161–184.
[4] I. Capuzzo Dolcetta and H. Ishii, *Approximate solution of the Hamilton-Jacobi equation of deterministic control theory*, Appl. Math. Optim., 11 (1984), pp. 161–181.
[5] M. G. Crandall and P. L. Lions, *Two approximations of solutions of Hamilton-Jacobi equations*, Math. Comp., 43 (1984), pp. 1–19.
[6] P. Dupuis and H. Ishii, *SDEs with oblique reflection on nonsmooth domains*, Ann. Probab., 21 (1993), pp. 554–580.
[7] P. Dupuis and H. J. Kushner, *Stochastic approximation and large deviations: Upper bounds and w.p.*1 *convergence*, SIAM J. Control Optim., 27 (1989), pp. 1108–1135.
[8] P. Dupuis, R. S. Ellis, and A. Weiss, *Large deviations for Markov processes with discontinuous statistics* I: *General upper bounds*, Ann. Probab., 19 (1991), pp. 1280–1297.
[9] L. C. Evans, *On solving certain nonlinear partial differential equations by accretive operator methods*, Israel J. Math., 36 (1980), pp. 225–247.
[10] W. H. Fleming, *The Cauchy problem for degenerate parabolic equations*, J. Math. Mech., 13 (1964), pp. 987–1008.
[11] W. H. Fleming, *The Cauchy problem for a nonlinear first order partial differential equation*, J. Differential Equations, 5 (1969), pp. 515–530.
[12] W. H. Fleming, *Stochastic control for small noise intensities*, SIAM J. Control Optim., 9 (1971), pp. 473–517.
[13] W. H. Fleming and M. R. James, *Asymptotic series and exit time probabilities*, Ann. Probab., 20 (1992), pp. 1369–1384.
[14] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.
[15] W. H. Fleming and P. E. Souganidis, *On the existence of value functions of two-player zero sum stochastic differential games*, Indiana Univ. Math. J., 38 (1989), pp. 293–314.
[16] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer-Verlag, New York, 1983.
[17] R. Gonzalez and E. Rofman, *On deterministic control problems: An approximation procedure for the optimal cost, parts* I *and* II, SIAM J. Control Optim., 23 (1985), pp. 242–285.
[18] H. J. Kushner, *Probability Methods for Approximations in Stochastic Control and for Elliptic Equations*, Academic Press, New York, 1977.
[19] H. J. Kushner and P. G. Dupuis, *Numerical Methods for Stochastic Control Problems in Continuous Time*, Springer-Verlag, New York, 1992.
[20] H. J. Kushner and J. Yang, *A numerical method for controlled routing in large trunk line networks via stochastic control theory*, ORSA J. Comput., to appear.
[21] P. L. Lions and A. S. Sznitman, *Stochastic differential equations with reflecting boundary conditions*, Comm. Pure Appl. Math., 37 (1984), pp. 511–553.
[22] P. L. Lions and N. S. Trudinger, *Linear oblique derivative boundary problems for the uniformly elliptic Hamilton-Jacobi-Bellman equation*, Math. Z., 191 (1986), pp. 1–15.
[23] J. Menaldi, *Some estimates for finite difference approximations*, SIAM J. Control Optim., 27 (1989), pp. 579–607.
[24] R. K. Miller and A. N. Michel, *Ordinary Differential Equations*, Academic Press, New York, 1982.
[25] M. L. Puterman, *Markov decision processes*, in Stochastic Models, Vol. 2, D. P. Heyman and M. J. Sobel, eds., North–Holland, Amsterdam, 1991.
[26] P. E. Souganidis, *Approximation schemes for viscosity solutions of Hamilton-Jacobi equations*, J. Differential Equations, 59 (1985), pp. 1–43.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 2, pp. 742–759, March 1998

013

A NEW VALUE ITERATION METHOD FOR THE AVERAGE COST DYNAMIC PROGRAMMING PROBLEM*

DIMITRI P. BERTSEKAS†

Abstract. We propose a new value iteration method for the classical average cost Markovian decision problem, under the assumption that all stationary policies are unichain and that, furthermore, there exists a state that is recurrent under all stationary policies. This method is motivated by a relation between the average cost problem and an associated stochastic shortest path problem. Contrary to the standard relative value iteration, our method involves a weighted sup-norm contraction, and for this reason it admits a Gauss–Seidel implementation. Computational tests indicate that the Gauss–Seidel version of the new method substantially outperforms the standard method for difficult problems.

Key words. dynamic programming, average cost, value iteration

AMS subject classifications. 90C35, 49L20

PII. S0363012995291609

1. Introduction.

We consider a controlled discrete-time dynamic system with n states, denoted $1,\dots,n$. At each time, if the state is i, a control u is chosen from a given finite constraint set $U(i)$, and the next state is j with given probability $p_{ij}(u)$. An admissible policy is a sequence of functions from states to controls, $\pi = \{\mu_0, \mu_1, \dots\}$, where $\mu_k(i) \in U(i)$ for all i and k. The average cost corresponding to π and initial state i is

$$J_\pi(i) = \limsup_{N\to\infty} \frac{1}{N} E\left\{ \sum_{k=0}^{N-1} g\big(x_k, \mu_k(x_k)\big) \,\middle|\, x_0 = i \right\},$$

where x_k is the state at time k and g is a given cost function. A *stationary policy* is an admissible policy of the form $\pi = \{\mu, \mu, \dots\}$, and its corresponding cost function is denoted by $J_\mu(i)$. For brevity, we refer to $\{\mu, \mu, \dots\}$ as the stationary policy μ. We want to solve the classical problem of finding an optimal policy, that is, an admissible policy π such that $J_{\pi^*}(i) = \min_\pi J_\pi(i)$ for all i.

A stationary policy is called *unichain* if it gives rise to a Markov chain with a single recurrent class. Throughout the paper, we assume the following.

Assumption 1: All stationary policies are unichain. Furthermore, state n is recurrent in the Markov chain corresponding to each stationary policy.

It is well known that under Assumption 1, the optimal cost $J^*(i)$ has a common value for all initial states, which is denoted by λ^*,

$$J^*(i) = \lambda^*, \qquad i = 1, \dots, n.$$

Furthermore, λ^* together with a differential cost vector $h = \big(h(1), \dots, h(n)\big)$ satisfies Bellman's equation

$$\lambda^* + h(i) = \min_{u\in U(i)} \left[g(i,u) + \sum_{j=1}^{n} p_{ij}(u) h(j) \right], \qquad i = 1, \dots, n. \tag{1}$$

*Received by the editors September 11, 1995; accepted for publication (in revised form) February 4, 1997. This research was supported by NSF grant 9300494-DMI.
http://www.siam.org/journals/sicon/36-2/29160.html

†Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 (bertseka@lids.mit.edu).

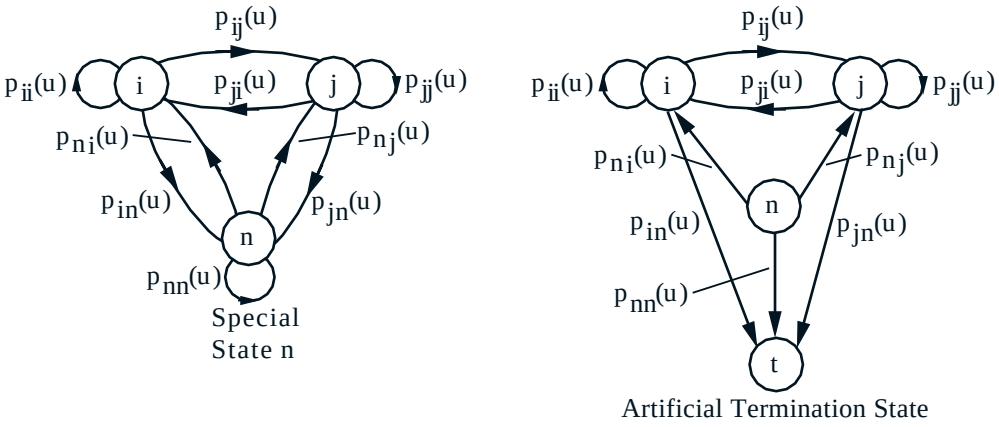


FIG. 1. *Transition probabilities for an average cost problem and its associated stochastic shortest path problem. The latter problem is obtained by introducing, in addition to* $1, \ldots, n$, *an artificial termination state* t *to which we move from each state* i *with probability* $p_{in}(u)$, *by setting all transition probabilities* $p_{in}(u)$ *to* 0, *and by leaving unchanged all other transition probabilities.*

In addition, a stationary policy μ is optimal if and only if $\mu(i)$ attains the minimum in the above equation for all i. These results can be shown under the assumption that all stationary policies are unichain, without requiring the additional condition that there is a common recurrent state to all stationary policies. However, for the methods of this paper, the existence of a common recurrent state is essential, at least for the purposes of analysis. From the computational point of view, the existence of a common recurrent state is less significant, as long as all stationary policies are unichain. One may modify the problem so that Assumption 1 holds by adding a very small positive ϵ to all transition probabilities of the form $p_{in}(u)$. The effect on the average cost per stage of each stationary policy will be $O(\epsilon)$.

Under Assumption 1 we can make an important connection of the average cost problem with an associated stochastic shortest path problem, which has been the basis for a recent textbook treatment of the average cost problem [Ber95, Vol. I, section 7.4]. This problem is obtained by leaving unchanged all transition probabilities $p_{ij}(u)$ for $j \neq n$, by setting all transition probabilities $p_{in}(u)$ to 0, and by introducing an artificial cost-free and absorbing termination state t to which we move from each state i with probability $p_{in}(u)$; see Fig. 1. The expected stage cost at state i of the stochastic shortest path problem is $g(i,u) - \lambda$, where λ is a scalar parameter. Let $h_{\mu,\lambda}(i)$ be the cost of stationary policy μ for this stochastic shortest path problem, starting from state i; that is, $h_{\mu,\lambda}(i)$ is the total expected cost incurred starting from state i up to reaching the termination state t. We refer to this problem as λ-SSP. Let $h_\lambda(i) = \min_\mu h_{\mu,\lambda}(i)$ be the corresponding optimal cost of the λ-SSP. Then the following can be shown (see Fig. 2).

(a) For all μ and λ, we have

$$h_{\mu,\lambda}(i) = h_{\mu,\lambda_\mu}(i) + (\lambda_\mu - \lambda)N_\mu(i), \qquad i = 1, \ldots, n, \tag{2}$$

where $N_\mu(i)$ is the average number of steps required to reach n under μ starting from state i, and λ_μ is the average cost corresponding to μ.

(b) The functions

$$h_\lambda(i) = \min_\mu h_{\mu,\lambda}(i), \qquad i = 1, \ldots, n, \tag{3}$$

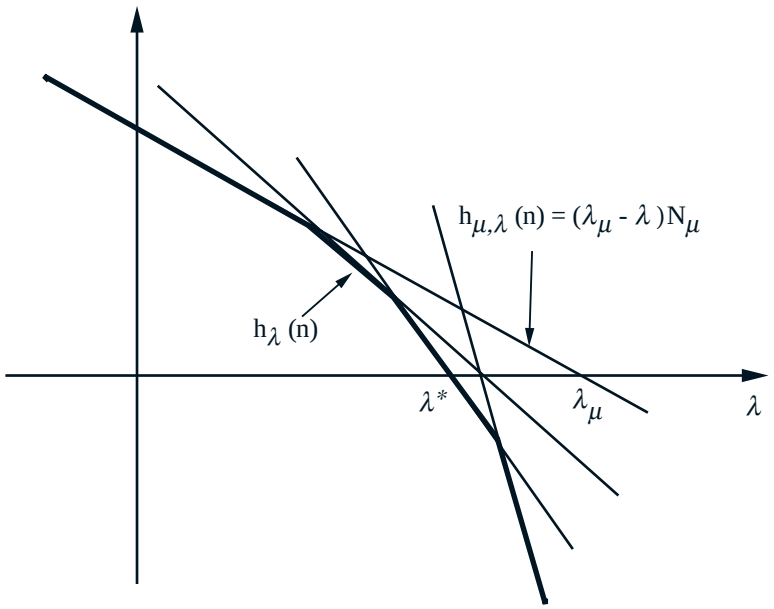


FIG. 2. *Relation of the costs of stationary policies in the average cost problem and the associated stochastic shortest path problem.*

are concave, monotonically decreasing, and piecewise linear as functions of λ, and

$$h_\lambda(n) = 0 \qquad \text{if and only if} \qquad \lambda = \lambda^*. \tag{4}$$

Furthermore, the vector h_{λ^*} satisfies Bellman's equation (1) together with λ^*.

From Fig. 2, it can be seen that λ^* can be obtained by a one-dimensional search procedure that brackets λ^* within a sequence of nested and diminishing intervals; see [Ber95, Vol. II, Fig. 4.5.2]. This method is probably inefficient because it requires the (exact) solution of several λ-SSPs, corresponding to several different values of λ. An alternative method, which is also inefficient because it requires the exact solution of several λ-SSPs, is to update λ by an iteration of the form

$$\lambda^{k+1} = \lambda^k + \gamma^k h_{\lambda^k}(n), \tag{5}$$

where γ^k is a positive stepsize parameter. This iteration is motivated by Fig. 2, where it is seen that $\lambda < \lambda^*$ (or $\lambda > \lambda^*$) if and only if $h_\lambda(n) > 0$ (or $h_\lambda(n) < 0$, respectively). Indeed, it can be seen from Fig. 2 that the sequence $\{\lambda^k\}$ generated by (5) converges to λ^* provided the stepsize γ^k is the same for all iterations and does not exceed the threshold value $1/\max_\mu N_\mu(n)$. Such a stepsize is sufficiently small to guarantee that the difference $\lambda - \lambda^*$ does not change sign during the algorithm (5). Note that each λ-SSP can be solved by value iteration, which has the form

$$h^{k+1}(i) = \min_{u \in U(i)} \left[g(i,u) + \sum_{j=1}^{n-1} p_{ij}(u) h^k(j) \right] - \lambda, \qquad i = 1, \ldots, n, \tag{6}$$

with λ kept fixed throughout the value iteration method.

In this paper we propose algorithms based on the λ-SSP, which are more efficient than the algorithms mentioned above. In particular, we change λ during the value iteration process (6) by using an iteration of the form (5), but with $h_{\lambda^k}(n)$ replaced by an approximation, the current value iterate $h^{k+1}(n)$. Such an algorithm may be viewed as a *value iteration algorithm for a slowly varying stochastic shortest path problem*. It has the form

$$h^{k+1}(i) = \min_{u \in U(i)} \left[g(i,u) + \sum_{j=1}^{n-1} p_{ij}(u) h^k(j) \right] - \lambda^k, \qquad i = 1, \ldots, n, \tag{7}$$

$$\lambda^{k+1} = \lambda^k + \gamma^k h^{k+1}(n), \tag{8}$$

where γ^k is a positive stepsize. We prove convergence of this method for the case where γ^k is a sufficiently small constant. Convergence can also be similarly proved for a variety of other stepsize rules.

Our method should be contrasted with the standard relative value iteration method for average cost problems due to [Whi63], which takes the form (see, e.g., [Ber95], [Put94])

$$\lambda^{k+1} = \min_{u \in U(n)} \left[g(n,u) + \sum_{j=1}^{n} p_{nj}(u) h^k(j) \right], \tag{9}$$

$$h^{k+1}(i) = \min_{u \in U(i)} \left[g(i,u) + \sum_{j=1}^{n} p_{ij}(u) h^k(j) \right] - \lambda^{k+1}, \qquad i = 1, \ldots, n. \tag{10}$$

If we use (7) to write iteration (8) in the equivalent form

$$\lambda^{k+1} = (1 - \gamma^k)\lambda^k + \gamma^k \min_{u \in U(n)} \left[g(n,u) + \sum_{j=1}^{n-1} p_{nj}(u) h^k(j) \right],$$

we see that if $\gamma^k = 1$ for all k, the new value iteration (7)–(8) becomes similar to the standard value iteration (9)–(10): the updating formulas are the same in both methods (because we have $h^k(n) = 0$ for all $k \geq 1$ in the iteration (9)–(10)), but the order of updating λ is just reversed relative to the order of updating h. Despite the similarity of the two methods, the proof of convergence of the standard method (9)–(10) (as given, for example, in [Ber95, Vol. II] or [Put94]) does not seem to be applicable to the new method. The line of proof given in the next section is substantially different, and makes essential use of Assumption 1 and the connection with the stochastic shortest path problem. Furthermore, one can construct examples where Assumption 1 is violated because state n is transient under some stationary policy, and where the new method (7)–(8) does not converge while the known method (9)–(10) converges. Conversely, it can be seen that the standard aperiodicity assumption required for convergence of the standard method (9)–(10) (see, e.g., [Ber95], [Put94]) is not needed for the new method. We note also that there is a variant of the standard method (9)–(10) that does not require an aperiodicity assumption and involves interpolations

between h^k and h^{k+1} according to a stepsize parameter (see [Sch71, [Pla77], [Var78], [PBW79], [Put94], [Ber95]). However, the new method does not seem as closely related to this variant.

A significant improvement in the algorithm, which guarantees that bounded iterates will be generated for any choice of stepsize, is to calculate upper and lower bounds on λ^* from iteration (7) and then modify iteration (8) to project the iterate $\lambda^k + \gamma^k h^k(n)$ on the interval of the bounds. In particular, based on the Odoni bounds [Odo69] for the relative value iteration method (see, e.g., [Ber95, Vol. II, p. 209], it can be seen that

$$\underline{\beta}^k \leq \lambda^* \leq \overline{\beta}^k,$$

where

$$\underline{\beta}^k = \lambda^k + \min\left[\min_{i \neq n}\big[h^{k+1}(i) - h^k(i)\big],\, h^{k+1}(n)\right], \tag{11}$$

$$\overline{\beta}^k = \lambda^k + \max\left[\max_{i \neq n}\big[h^{k+1}(i) - h^k(i)\big],\, h^{k+1}(n)\right]. \tag{12}$$

Thus we may replace the iteration $\lambda^{k+1} = \lambda^k + \gamma^k h^{k+1}(n)$ (cf. (8)) by

$$\lambda^{k+1} = \Pi_k\big[\lambda^k + \gamma^k h^{k+1}(n)\big], \tag{13}$$

where $\Pi_k[c]$ denotes the projection of a scalar c on the interval

$$\left[\max_{m=0,\ldots,k} \underline{\beta}^m,\ \min_{m=0,\ldots,k} \overline{\beta}^m\right]. \tag{14}$$

We note that the issue of stepsize selection is crucial for the success of our algorithm. In particular, if γ^k is a chosen constant but very small, or diminishing at the rate of $1/k$ (as is common in stochastic approximation algorithms), then λ changes slowly relative to h, and the iteration (8) essentially becomes identical to iteration (5) but with a very small stepsize, which leads to slow convergence. On the other hand, if γ^k is too large, λ^k will oscillate and diverge. One may keep the stepsize γ^k constant at a value found by trial and error, but there are some better alternatives. One possibility that has worked quite reliably and efficiently in our tests is to start with a fairly large γ^k and gradually diminish it if $h^k(n)$ changes sign frequently; for example, we may use

$$\gamma^k = m(\hat{k})\gamma, \tag{15}$$

where

(a) γ is the initial stepsize (a positive constant),

(b) $m(\hat{k})$ is a decreasing function of $\hat{k}$, which is defined as the number of indexes $t \leq k$ such that $h^{t-1}(n)h^t(n) < 0$ and $|h^t(n)|$ is greater than some fixed threshold θ.

Examples of functions $m(\cdot)$ that we tried are

$$m(\hat{k}) = \frac{1}{\hat{k}+1} \tag{16a}$$

and

$$m(\hat{k}) = \xi^{\hat{k}}, \tag{16b}$$

where ξ is a fixed scalar from the range $(0,1)$, so that γ^k is decreased by a factor ξ each time $\hat{k}$ is incremented. Our experience indicates that it is best to choose the initial stepsize γ in the range $[1,5]$. Typically, the stepsize is reduced quickly according to (15) to an appropriate level (which depends on the problem) and then stays constant for the remaining iterations. In our experiments, we have used the preceding choices of γ^k with $\gamma = 1$, $\xi = 0.95$, and $\theta = 1$.

The motivation for our method is that value iteration for stochastic shortest path problems involves a contraction. In particular, consider the mapping $F : \Re^n \to \Re^n$ with components given by

$$F_i(h) = \min_{u \in U(i)} \left[g(i,u) + \sum_{j=1}^{n-1} p_{ij}(u) h(j) \right], \qquad i = 1, \ldots, n.$$

It is known (see, e.g., [BeT89, p. 325] or [Tse90]) that, under Assumption 1, F is a contraction mapping with respect to some weighted sup-norm; that is, for some positive scalars $v_1, \ldots, v_n$, and some scalar $\alpha \in (0,1)$, we have

$$\max_{i=1,\ldots,n} \frac{|F_i(h) - F_i(\overline{h})|}{v_i} \leq \alpha \max_{i=1,\ldots,n} \frac{|h(i) - \overline{h}(i)|}{v_i} \qquad \forall\ h,\ \overline{h} \in \Re^n. \tag{17}$$

Note here that while there is coupling between the iteration of h as per (7) and the iteration for λ as per (8), the latter iteration can be made much slower than the former through the use of the stepsize γ, so that the weighted sup-norm contraction character of the iteration (7) is preserved. Furthermore, even when the stepsize γ is not small, the contraction property of F is analytically convenient, as will be seen, for example, in the analysis of section 3. By contrast, the standard relative value iteration method (9)–(10) does not involve a weighted sup-norm contraction, and in fact it may not involve a contraction of any kind, unless an additional aperiodicity assumption on the Markov chains corresponding to the stationary policies is imposed. We speculate that the sup-norm contraction structure may be helpful in other contexts, beyond those discussed in this paper; for example, in Q-learning (stochastic approximation) variants of the method and when parallel asynchronous variations are considered. In fact, an analysis of Q-learning variants of our method that admit a parallel asynchronous implementation is the subject of a forthcoming report [ABB97].

The new method (7)–(8) can be viewed as a Jacobi type of method, since all the components of h are simultaneously updated. A particularly interesting fact is that the weighted sup-norm contraction property of the mapping F can also be exploited to construct valid Gauss–Seidel variants, where the components of h are updated sequentially in some order. In particular, the method of proof of the next section can be used to show convergence for the Gauss–Seidel version of the method, given by

$$h^{k+1}(i) = G_i(h^k, \lambda^k), \qquad i = 1, \ldots, n,$$

$$\lambda^{k+1} = \lambda^k + \gamma^k h^{k+1}(n),$$

where $G : \Re^{n+1} \to \Re^n$ is the Gauss–Seidel mapping based on F, having components given by

$$G_1(h, \lambda) = \min_{u \in U(1)} \left[g(1,u) + \sum_{j=1}^{n-1} p_{1j}(u) h(j) \right] - \lambda,$$

$$G_i(h,\lambda) = \min_{u\in U(i)} \left[g(i,u) + \sum_{j=1}^{i-1} p_{ij}(u) G_j(h,\lambda) + \sum_{j=i}^{n-1} p_{ij}(u) h(j) \right] - \lambda, \qquad i = 2,\ldots,n.$$

By contrast, we do not know of any convergent Gauss–Seidel version of the standard value iteration (9)–(10). In fact, simple counterexamples show that the straightforward Gauss–Seidel variant of the standard method may diverge.

Note that the Odoni bounds (11)–(12) are not available when the Gauss–Seidel variant is used. However, it is still possible to use the projection (13)–(14) by performing once in a while (say, every 10 iterations) the regular (Jacobi) version (7)–(8) of the method, and obtain corresponding Odoni bounds that can be used for projection at all subsequent iterations. This device proved to be very effective in our experiments.

Regarding a theoretical comparison of the performance of the new methods and the standard method, it can be seen with simple examples that neither type of method dominates the other. Suppose, for instance, that there is only one policy and that the corresponding transition probability matrix is

$$\begin{pmatrix} \epsilon & 1-\epsilon \\ 1-\epsilon & \epsilon \end{pmatrix},$$

where ϵ is a scalar from $[0,1]$. Then both methods (7)–(8) and (9)–(10) become linear iterations, and their rate of convergence is governed by the eigenvalues of the corresponding iteration matrix. The eigenvalues corresponding to the standard relative value iteration (9)–(10) can be shown to be 0 and $1-2\epsilon$, so that the method converges very fast for $\epsilon \sim 1/2$ and slowly for $\epsilon \sim 0$ or $\epsilon \sim 1$. It can also be verified that, for a constant but well-chosen value of γ, the eigenvalue structure of the new value iteration method (7)–(8) is worse than the one for the standard method for $\epsilon \sim 1/2$, more favorable for $\epsilon \sim 0$, and comparably unfavorable for $\epsilon \sim 1$.

Our limited computational experiments also indicate that the Jacobi version (7)–(8) of the new method, when properly implemented with the adaptive stepsize rule (15) and the projection scheme of (11)–(14), is competitive with the relative value iteration method of (9)–(10). There are problems where one method outperforms the other and vice versa. When the initial stepsize γ in (15) is equal to 1, the performance of the two methods appears to be quite similar for many problems (see, e.g., Tables 2 and 3 in section 4). On the other hand, our computational results indicate that the Gauss–Seidel variant of the new method substantially outperforms the standard method for relatively difficult problems. This is not surprising, since Gauss–Seidel methods are known to have better performance than their Jacobi counterparts when a weighted sup-norm contraction is involved. Both the standard method and the new methods can be very slow on unfavorably structured problems. This is to be expected, since these methods exhibit convergence rate behavior similar to linear iterations and are subject to ill-conditioning.

The paper is organized as follows. In the next section we prove a convergence result for the Jacobi version of the new method. In section 3 we extend this result to apply to the Gauss–Seidel variant. The method of proof can also be used to prove convergence of a variety of other variants involving different orders of updating the components of the vector h, as well as asynchronous versions. All this flexibility is possible thanks to the weighted sup-norm contraction property of the mapping F. Finally, in section 4 we describe some of our computational experience. In particular, we compare the standard method (9)–(10) with implementations of the Jacobi and Gauss–Seidel versions of our method, which involve an adaptive stepsize rule like the

one of (15) and the projection scheme of (11)–(14). We find that the Gauss–Seidel method outperforms the other methods on the more difficult problems.

2. Convergence analysis. We now investigate the convergence of the new value iteration algorithm. For convenience, let us denote by $\|\cdot\|$ the weighted sup-norm with respect to which the contraction property of (17) holds; that is,

$$\|h\| = \max_{i=1,\ldots,n} \frac{|h(i)|}{v_i} \qquad \forall\ h \in \Re^n.$$

Let us also normalize the vector v so that its last coordinate is equal to 1; that is,

$$v_n = 1.$$

Note that since h_λ is the optimal cost vector of the λ-SSP, we have that h_λ is the unique fixed point of the contraction mapping $F(h) - \lambda e$; that is,

$$h_\lambda = F(h_\lambda) - \lambda e \qquad \forall\ \lambda \in \Re. \tag{18}$$

By writing for all stationary policies μ, states i, and scalars λ and λ',

$$h_{\mu,\,\lambda}(i) = h_{\mu,\,\lambda'}(i) + N_\mu(i)(\lambda' - \lambda),$$

and by using the definition $h_\lambda(i) = \min_\mu h_{\mu,\lambda}(i)$, we obtain the following relation:

$$h_{\lambda'}(i)+\underline{N}(\lambda'-\lambda) \le h_\lambda(i) \le h_{\lambda'}(i)+\overline{N}(\lambda'-\lambda) \qquad \forall\, i = 1,\ldots,n, \text{ and } \lambda,\, \lambda' \in \Re, \tag{19}$$

where $\underline{N}$ and $\overline{N}$ are the positive scalars

$$\underline{N} = \min_\mu \min_{i=1,\ldots,n} N_\mu(i), \qquad \overline{N} = \max_\mu \max_{i=1,\ldots,n} N_\mu(i).$$

We can write (19) in the equivalent form

$$N|\lambda' - \lambda| \le |h_\lambda(i) - h_{\lambda'}(i)| \le \overline{N}|\lambda' - \lambda|, \qquad \forall\ i \text{ and } \lambda, \lambda' \in \mathcal{R}. \tag{20}$$

We can interpret $\underline{N}$ and $\overline{N}$ as uniform lower and upper bounds on the slope of the piecewise linear function $h_\lambda(i)$, viewed as a function of λ (see Fig. 2).

The following is our main result.

PROPOSITION 1. *There exists a positive scalar $\overline{\gamma}$ such that if*

$$\underline{\gamma} \le \gamma^k \le \overline{\gamma} \tag{21}$$

for some positive scalar $\underline{\gamma}$ and all k, the sequence (h^k, λ^k) generated by iteration (7), (8) converges to $(h_{\lambda^*}, \lambda^*)$ at the rate of a geometric progression.

Proof. We will show that there exists a threshold value $\overline{\gamma} > 0$ and a continuous function $c(\gamma)$ with $0 \le c(\gamma) < 1$ for all $\gamma \in (0, \overline{\gamma}]$ such that for any $B > 0$, the relations

$$\|h^k - h_{\lambda^k}\| \le B \quad \text{and} \quad |\lambda^k - \lambda^*| \le \frac{B}{\underline{N}} \tag{22}$$

imply that

$$\|h^{k+1} - h_{\lambda^{k+1}}\| \le c(\gamma^k)B \quad \text{and} \quad |\lambda^{k+1} - \lambda^*| \le \frac{c(\gamma^k)B}{\underline{N}}. \tag{23}$$

This implies that for a stepsize sequence satisfying the assumptions of the proposition, the sequence $|\lambda^k - \lambda^*|$ converges to zero at the rate of a geometric progression, and the same is true of the sequence $\|h^k - h_{\lambda^k}\|$. Since, using (20), we have

$$\|h^k - h_{\lambda^*}\| \le \|h^k - h_{\lambda^k}\| + \|h_{\lambda^k} - h_{\lambda^*}\| \le \|h^k - h_{\lambda^k}\| + O\big(|\lambda^k - \lambda^*|\big),$$

we see that $\|h^k - h_{\lambda^*}\|$ also converges to zero at the rate of a geometric progression.

We first show two preliminary relations. We have, using (18),

$$\begin{aligned}\|h_{\lambda^{k+1}} - h_{\lambda^k}\| &= \|F\big(h_{\lambda^{k+1}}\big) - \lambda^{k+1}e - F\big(h_{\lambda^k}\big) + \lambda^k e\| \\ &\le \|F\big(h_{\lambda^{k+1}}\big) - F\big(h_{\lambda^k}\big)\| + \|(\lambda^{k+1} - \lambda^k)e\| \\ &\le \alpha\|h_{\lambda^{k+1}} - h_{\lambda^k}\| + |\lambda^{k+1} - \lambda^k|\|e\|.\end{aligned}$$

Thus

$$\|h_{\lambda^{k+1}} - h_{\lambda^k}\| \le \frac{\|e\|}{1-\alpha}|\lambda^{k+1} - \lambda^k|. \tag{24}$$

Also, by subtracting the relations

$$h^{k+1}(n) = F_n(h^k) - \lambda^k,$$

$$h_{\lambda^k}(n) = F_n(h_{\lambda^k}) - \lambda^k,$$

we have

$$|h^{k+1}(n) - h_{\lambda^k}(n)| = |F_n(h^k) - F_n(h_{\lambda^k})| \le \alpha\|h^k - h_{\lambda^k}\|. \tag{25}$$

Using this relation and (19), we obtain

$$|h^{k+1}(n)| \le |h^{k+1}(n) - h_{\lambda^k}(n)| + |h_{\lambda^k}(n)| \le \alpha\|h^k - h_{\lambda^k}\| + \overline{N}|\lambda^k - \lambda^*|. \tag{26}$$

We will now derive functions $c_1(\cdot)$ and $c_2(\cdot)$ for which the first and the second relations in (22), respectively, hold. We will then use $c(\gamma) = \max\big[c_1(\gamma), c_2(\gamma)\big]$ in (22). Regarding the first relation in (23), we note that

$$\begin{aligned}\|h^{k+1} - h_{\lambda^{k+1}}\| &= \|F(h^k) - \lambda^k e - F\big(h_{\lambda^{k+1}}\big) + \lambda^{k+1}e\| \\ &\le \|F(h^k) - F\big(h_{\lambda^{k+1}}\big)\| + |\lambda^{k+1} - \lambda^k|\,\|e\| \\ &\le \alpha\|h^k - h_{\lambda^{k+1}}\| + |\lambda^{k+1} - \lambda^k|\,\|e\| \\ &\le \alpha\|h^k - h_{\lambda^k}\| + \alpha\|h_{\lambda^k} - h_{\lambda^{k+1}}\| + |\lambda^{k+1} - \lambda^k|\,\|e\|.\end{aligned} \tag{27}$$

Using the above inequality and (22), (24), and (26), we obtain

$$\begin{aligned}\|h^{k+1} - h_{\lambda^{k+1}}\| &\le \alpha B + \left(\frac{\alpha}{1-\alpha} + 1\right)|\lambda^{k+1} - \lambda^k|\,\|e\| \\ &= \alpha B + \frac{\|e\|\gamma^k}{1-\alpha}|h^{k+1}(n)| \\ &\le \alpha B + \frac{\|e\|\gamma^k}{1-\alpha}\big(\alpha\|h^k - h_{\lambda^k}\| + \overline{N}|\lambda^k - \lambda^*|\big) \\ &\le \alpha B + \frac{\|e\|\gamma^k}{1-\alpha}\left(\alpha B + \frac{\overline{N}B}{\underline{N}}\right) \\ &= c_1(\gamma^k)B,\end{aligned} \tag{28}$$

where $c_1(\cdot)$ is the function

$$c_1(\gamma) = \alpha + \frac{\gamma\|e\|(\alpha + \overline{N}/\underline{N})}{1-\alpha}. \tag{29}$$

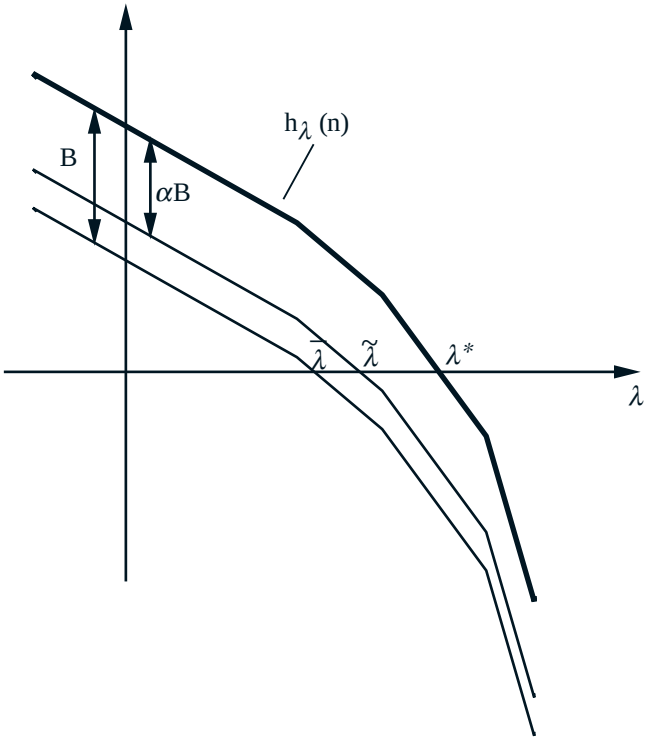


FIG. 3. *Definition of* $\overline{\lambda}$ *and* $\tilde{\lambda}$ *in the proof of Proposition* 1.

Note that if

$$\gamma < \frac{(1-\alpha)^2}{\|e\|(\alpha + \overline{N}/\underline{N})},$$

we have $c_1(\gamma) < 1$.

We now turn to the second relation in (23); that is, we show that

$$|\lambda^{k+1} - \lambda^*| \leq \frac{c_2(\gamma^k)B}{\underline{N}}$$

for an appropriate continuous function $c_2(\gamma)$. Let $\overline{\lambda}$ and $\tilde{\lambda}$ be the unique scalars such that

$$h_{\overline{\lambda}}(n) = B, \qquad h_{\tilde{\lambda}}(n) = \alpha B \tag{30}$$

(see Fig. 3). Also let $\hat{\lambda}$ be the midpoint between $\overline{\lambda}$ and $\tilde{\lambda}$:

$$\hat{\lambda} = \frac{\overline{\lambda} + \tilde{\lambda}}{2}. \tag{31}$$

Note that from (19), we have

$$\frac{(1-\alpha)B}{\overline{N}} \leq \tilde{\lambda} - \overline{\lambda} \leq \frac{(1-\alpha)B}{\underline{N}} \tag{32}$$

and that

$$\frac{\alpha B}{\overline{N}} \leq \lambda^* - \tilde{\lambda} \leq \frac{\alpha B}{\underline{N}},$$

$$\frac{B}{\overline{N}} \leq \lambda^* - \overline{\lambda} \leq \frac{B}{\underline{N}}.$$

From the last three relations, we also obtain

$$\frac{(1+\alpha)B}{2\overline{N}} \leq \lambda^* - \hat{\lambda} \leq \frac{(1+\alpha)B}{2\underline{N}}, \tag{33}$$

$$\frac{(1-\alpha)B}{2\overline{N}} \leq \tilde{\lambda} - \hat{\lambda} \leq \frac{(1-\alpha)B}{2\underline{N}}. \tag{34}$$

We assume that $\lambda^k \leq \lambda^*$; the complementary case where $\lambda^k \geq \lambda^*$ is handled similarly. We distinguish between two cases:

(a) $\lambda^k \leq \hat{\lambda}$,

(b) $\hat{\lambda} < \lambda^k \leq \lambda^*$.

In the case where $\lambda^k \leq \hat{\lambda}$, we have, using (19) and (30)–(32),

$$h_{\lambda^k}(n) \geq h_{\hat{\lambda}}(n) \geq h_{\tilde{\lambda}}(n) + \underline{N}(\tilde{\lambda} - \hat{\lambda}) = \alpha B + \underline{N}(\tilde{\lambda} - \hat{\lambda}) \geq \alpha B + \frac{(1-\alpha)B\underline{N}}{2\overline{N}}. \tag{35}$$

On the other hand, from (22) and (25), we have $|h^{k+1}(n) - h_{\lambda^k}(n)| \leq \alpha B$ so that

$$h^{k+1}(n) \geq h_{\lambda^k}(n) - \alpha B. \tag{36}$$

By combining (35) and (36), we obtain

$$h^{k+1}(n) \geq \frac{(1-\alpha)B}{2\overline{N}^2}.$$

We now have, using the above equation,

$$\lambda^* - \lambda^{k+1} = \lambda^* - \lambda^k - \gamma^k h^{k+1}(n) \leq \frac{B}{\underline{N}} - \frac{\gamma^k(1-\alpha)B\underline{N}}{2\overline{N}} = \frac{B}{\underline{N}^2}\left(1 - \frac{\gamma^k(1-\alpha)\underline{N}}{2\overline{N}}\right), \tag{37}$$

and we also have, using (25), (22), and (19)

$$\lambda^* - \lambda^{k+1} = \lambda^* - \lambda^k - \gamma^k h^{k+1}(n) \geq \lambda^* - \lambda^k - \gamma^k\big(h_{\lambda_k}(n) + \alpha B\big) \geq (1 - \gamma^k\overline{N})(\lambda^* - \lambda^k) - \gamma^k\alpha B. \tag{38}$$

It can be seen now from (38) that for $\gamma^k \in (0, 1/\overline{N}]$, we have $\lambda^* - \lambda^{k+1} \geq -\gamma^k\alpha B$, and it follows using also (37) that

$$|\lambda^* - \lambda^{k+1}| \leq \frac{c_2(\gamma^k)B}{\underline{N}},$$

where $c_2(\cdot)$ is the continuous function

$$c_2(\gamma) = \max\left[1 - \frac{\gamma(1-\alpha)\underline{N}^2}{2\overline{N}},\ \gamma\alpha\underline{N}\right].$$

Since there exists a threshold value $\overline{\gamma} > 0$ such that the continuous function $c_2(\gamma)$ satisfies $0 < c(\gamma) < 1$ for all $\gamma \in (0, \overline{\gamma}]$, the desired relation (23) is proved in the case $\lambda^k \leq \hat{\lambda}$.

In the case where $\hat{\lambda} < \lambda^k \leq \lambda^*$, there are two possibilities.

(1) $h^{k+1}(n) \geq 0$. Then $\lambda^k \leq \lambda^{k+1}$, and by also using (33), we have

$$\lambda^* \leq \hat{\lambda} + \frac{(1+\alpha)B}{2\underline{N}} \leq \lambda^k + \frac{(1+\alpha)B}{2\underline{N}} \leq \lambda^{k+1} + \frac{(1+\alpha)B}{2\underline{N}}. \tag{39}$$

Furthermore, from (22) and (26), we have

$$\lambda^{k+1} = \lambda^k + \gamma^k h^{k+1}(n) \leq \lambda^* + \gamma^k \left(\alpha B + \frac{\overline{N}B}{\underline{N}} \right).$$

Thus, by choosing γ^k sufficiently small, we can guarantee that

$$\lambda^{k+1} \leq \lambda^* + \frac{(1+\alpha)B}{2\underline{N}}. \tag{40}$$

From (39) and (40), it follows that for γ^k less than some positive constant, we have

$$|\lambda^{k+1} - \lambda^*| \leq \frac{(1+\alpha)B}{2\underline{N}},$$

proving the second relation in (23), with $c_2(\gamma) = (1+\alpha)/2$.

(2) $h^{k+1}(n) < 0$. In this case, since from (22) and (25) we have

$$h_{\lambda^k}(n) \leq h^{k+1}(n) + \alpha B \leq \alpha B, \tag{41}$$

and since $h_{\tilde{\lambda}}(n) = \alpha B$ and $h_\lambda(n)$ is monotonically decreasing in λ, it follows that $\tilde{\lambda} \leq \lambda^k$. Since $\lambda^k \leq \lambda^*$, we also have $0 \leq h_{\lambda^k}(n) \leq \alpha B$, so that by using (41) and the fact $h_{\lambda^k}(n) \geq 0$, we obtain $|h^{k+1}(n)| \leq \alpha B$ and

$$|\gamma^k h^{k+1}(n)| \leq \gamma^k \alpha B.$$

By choosing

$$\gamma^k \in \left(0, \frac{1-\alpha}{2\alpha\overline{N}} \right], \tag{42}$$

the above inequality, together with (34), yields

$$|\gamma^k h^{k+1}(n)| \leq \frac{(1-\alpha)B}{2\overline{N}} \leq \tilde{\lambda} - \hat{\lambda} \leq \lambda^k - \hat{\lambda}.$$

Thus, we have

$$\lambda^{k+1} = \lambda^k + \gamma^k h^{k+1}(n) \geq \hat{\lambda},$$

and from (33), using also the fact $\lambda^{k+1} \leq \lambda^k \leq \lambda^*$, we obtain for γ^k satisfying (42),

$$|\lambda^{k+1} - \lambda^*| \leq \frac{(1+\alpha)B}{2\underline{N}},$$

proving the second relation in (23) for the case $h^{k+1}(n) < 0$ as well.

Thus, (23) holds with $c(\cdot)$ given by

$$c(\gamma) = \max \left[\alpha + \frac{\gamma \|e\| (\alpha + \overline{N}/\underline{N})}{1-\alpha},\ 1 - \frac{\gamma(1-\alpha)\underline{N}^2}{2\overline{N}},\ \gamma\alpha\underline{N},\ \frac{1+\alpha}{2} \right]. \qquad \square$$

3. Convergence analysis of the Gauss–Seidel version. In this section, we prove the result of Proposition 1 for the Gauss–Seidel version of the method, given by

$$h^{k+1}(i) = G_i(h^k, \lambda^k), \qquad i = 1, \ldots, n, \tag{43}$$

$$\lambda^{k+1} = \lambda^k + \gamma^k h^{k+1}(n), \tag{44}$$

where the components of the mapping $G = (G_1, \ldots, G_n)$ are given by

$$G_1(h, \lambda) = \min_{u \in U(1)} \left[g(1, u) + \sum_{j=1}^{n-1} p_{1j}(u) h(j) \right] - \lambda, \tag{45}$$

(46)

$$G_i(h, \lambda) = \min_{u \in U(i)} \left[g(i, u) + \sum_{j=1}^{i-1} p_{ij}(u) G_j(h, \lambda) + \sum_{j=i}^{n-1} p_{ij}(u) h(j) \right] - \lambda, \qquad i = 2, \ldots, n.$$

The proof of Proposition 1 essentially carries through with the aid of the following result.

PROPOSITION 2. *The mapping G of* (45) *and* (46) *satisfies for all $h \in \Re^n$, $\overline{h} \in \Re^n$, $\lambda \in \Re$, and $\overline{\lambda} \in \Re$:*

$$\frac{|G_i(h, \lambda) - G_i(\overline{h}, \overline{\lambda})|}{v_i} \leq \alpha \|h - \overline{h}\| + \delta_i |\lambda - \overline{\lambda}| \qquad \forall\ i = 1, \ldots, n, \tag{47}$$

where α is the contraction modulus of F, $v_1, \ldots, v_n$ are the weights of the sup-norm $\|\cdot\|$ with respect to which F is a contraction (cf. (17)*), and $\delta_1, \ldots, \delta_n$ are defined recursively by*

$$\delta_1 = \frac{1}{v_1}, \qquad \delta_i = \frac{1 + \max_{j=1,\ldots,i-1} \delta_j}{v_i}, \quad i = 2, \ldots, n. \tag{48}$$

In particular, by taking the maximum over i in (47)*, we obtain*

$$\|G(h, \lambda) - G(\overline{h}, \overline{\lambda})\| \leq \alpha \|h - \overline{h}\| + \delta |\lambda - \overline{\lambda}| \qquad \forall\ h \in \Re^n,\ \overline{h} \in \Re^n,\ \lambda \in \Re,\ \overline{\lambda} \in \Re, \tag{49}$$

where

$$\delta = \max_{i=1,\ldots,n} \delta_i.$$

Proof. We prove (47) by induction. For the case where $i = 1$, we have from the contraction property of the mapping F (cf. (17)):

$$\frac{|G_1(h, \lambda) - G_1(\overline{h}, \lambda)|}{v_1} \leq \alpha \max_{i=1,\ldots,n} \frac{|h(i) - \overline{h}(i)|}{v_i} = \alpha \|h - \overline{h}\|.$$

Therefore,

$$\begin{aligned} \frac{G_1(h, \lambda)}{v_1} &\leq \frac{G_1(\overline{h}, \lambda)}{v_1} + \alpha \|h - \overline{h}\| \\ &\leq \frac{G_1(\overline{h}, \overline{\lambda})}{v_1} + \alpha \|h - \overline{h}\| + \frac{|\lambda - \overline{\lambda}|}{v_1}. \end{aligned}$$

Similarly, we obtain

$$\frac{G_1(\overline{h},\overline{\lambda})}{v_1} \le \frac{G_1(h,\lambda)}{v_1} + \alpha\|h-\overline{h}\| + \frac{|\lambda-\overline{\lambda}|}{v_1}.$$

By combining the last two relations, we see that

$$\frac{|G_1(h,\lambda)-G_1(\overline{h},\overline{\lambda})|}{v_1} \le \alpha\|h-\overline{h}\| + \delta_1|\lambda-\overline{\lambda}|,$$

so that (47) is proved for $i=1$.

Assume that (47) holds for $i=1,\dots,m-1$. We will show that it holds for $i=m$. We have from the contraction property of the mapping F and the induction hypothesis

$$\begin{aligned}
&\frac{|G_m(h,\lambda)-G_m(\overline{h},\lambda)|}{v_m} \\
&\quad\le \alpha\max\left\{\max_{i=1,\dots,m-1}\frac{|G_i(h,\lambda)-G_i(\overline{h},\lambda)|}{v_i}, \max_{i=m,\dots,n}\frac{|h(i)-\overline{h}(i)|}{v_i}\right\} \\
&\quad\le \alpha\|h-\overline{h}\|.
\end{aligned}$$

Using this relation and the induction hypothesis, we obtain

$$\begin{aligned}
\frac{G_m(h,\lambda)}{v_m} &\le \frac{G_m(\overline{h},\lambda)}{v_m} + \alpha\|h-\overline{h}\| \\
&= \frac{1}{v_m}\min_{u\in U(m)}\left[g(m,u)+\sum_{j=1}^{m-1}p_{mj}(u)G_j(\overline{h},\lambda)+\sum_{j=m}^{n-1}p_{mj}(u)\overline{h}(j)\right] \\
&\quad -\frac{\lambda}{v_m}+\alpha\|h-\overline{h}\| \\
&\le \frac{1}{v_m}\min_{u\in U(m)}\left[g(m,u)+\sum_{j=1}^{m-1}p_{mj}(u)G_j(\overline{h},\overline{\lambda})+\sum_{j=m}^{n-1}p_{mj}(u)\overline{h}(j)\right]-\frac{\overline{\lambda}}{v_m} \\
&\quad +\frac{|\lambda-\overline{\lambda}|}{v_m}+\max_{j=1,\dots,m-1}\delta_j\frac{|\lambda-\overline{\lambda}|}{v_m}+\alpha\|h-\overline{h}\| \\
&= \frac{G_m(\overline{h},\overline{\lambda})}{v_m}+\delta_m|\lambda-\overline{\lambda}|+\alpha\|h-\overline{h}\|.
\end{aligned}$$

Similarly, we obtain

$$\frac{G_m(\overline{h},\overline{\lambda})}{v_m} \le \frac{G_m(h,\lambda)}{v_m} + \delta_m|\lambda-\overline{\lambda}| + \alpha\|h-\overline{h}\|,$$

thus proving (47) for $i=m$. This completes the induction. □

Note that Proposition 1 implies that for any λ, $G(\cdot,\lambda)$ is a weighted sup-norm contraction when viewed as a function of h. It can be easily verified that

$$h_\lambda = G(h_\lambda,\lambda) \qquad \forall\ \lambda\in\Re,$$

so it follows that for all λ, the mapping $G(\cdot,\lambda)$ has h_λ as its unique fixed point. The following result proves convergence of the Gauss–Seidel method and parallels Proposition 1.

TABLE 1

n	Sparsity	STANDARD	SSP-JACOBI	SSP-Gauss–Seidel
10	0.5	16	39	40
20	0.5	9	39	75
30	0.5	9	48	105
40	0.5	8	46	55
50	0.5	8	56	90
10	0.1	674	727	185
20	0.1	202	203	160
30	0.1	38	66	130
40	0.1	36	77	75
50	0.1	21	63	110
10	0.05	114	294	70
20	0.05	131	145	100
30	0.05	49	53	235
40	0.05	259	226	205
50	0.05	313	313	325

TABLE 2

n	STANDARD	SSP-JACOBI	SSP-Gauss–Seidel
10	211	211	180
20	2658	2658	2070
30	29638	29647	20615
40	286550	286765	222855
50	13219	13217	9035

PROPOSITION 3. *There exists a positive scalar $\overline{\gamma}$ such that if*

$$\underline{\gamma} \leq \gamma^k \leq \overline{\gamma}$$

for some positive scalar $\underline{\gamma}$ and all k, the sequence (h^k, λ^k) generated by the Gauss–Seidel iteration (43), (44) converges to $(h_{\lambda^*}, \lambda^*)$ at the rate of a geometric progression.

Proof. The proof is essentially identical to the one of Proposition 1. The only difference is that the three relations (27), (28), and (29) must be modified to involve the mapping G and to make use of Proposition 2. In particular, (27) becomes

$$\|h^{k+1} - h_{\lambda^{k+1}}\| \leq \alpha\|h^k - h_{\lambda^k}\| + \alpha\|h_{\lambda^k} - h_{\lambda^{k+1}}\| + \delta|\lambda^{k+1} - \lambda^k|,$$

and (28) becomes

$$\|h^{k+1} - h_{\lambda^{k+1}}\| \leq c_1(\gamma^k)B,$$

where the function $c_1(\cdot)$ of (29) is now given by

$$c_1(\gamma) = \alpha + \gamma\left(\frac{\alpha\|e\|}{1-\alpha} + \delta\right)\left(\alpha + \frac{\overline{N}}{\underline{N}}\right).$$

The remainder of the proof goes through with no modification. □

4. Implementation and experimentation. In this section we describe some of our computational experience with the standard method (9)–(10) and with the new Jacobi and Gauss–Seidel methods. The latter methods were implemented with an adaptive stepsize rule of the form $\gamma^k = m(\hat{k})\gamma$ (cf. (15)), using an initial stepsize γ equal to 1. We used the function $m(\hat{k})$ of (16a) for the test results of Tables 1–3 and the function $m(\hat{k})$ of (16b) for the test results of Table 4. The projection

TABLE 3

n	STANDARD	SSP-JACOBI	SSP-Gauss–Seidel
10	121	119	80
20	826	825	545
30	18020	18026	13465
40	2186	2186	1360
50	5942	5941	4770
75	7978	7984	5000
100	9035	9028	6880
125	10306	10323	7440
150	9011	9015	6870

TABLE 4

n	STANDARD	SSP-JACOBI	SSP-Gauss–Seidel
250	939	940	420
500	4724	4725	470
750	1257	1257	740
1000	710	711	1040
1250	1693	1693	1425
1500	2870	2870	1890
1750	5605	5609	4230
2000	4691	4693	3180

scheme of (11)–(14) was also used. To obtain error bounds on which to project in the Gauss–Seidel method, we performed one Jacobi iteration following nine consecutive Gauss–Seidel iterations. Each Jacobi iteration yielded an upper and a lower bound for λ^*, and the λ-iterate obtained by each iteration was projected on the interval of the best upper and lower bounds obtained so far. For each problem, the three methods were initialized with $h = 0$ and (for the case of the new methods) $\lambda = n/2$. Note that because of the device of projection on the error bound range, the initial choice of λ is not critical.

Our computational results with randomly generated problems are summarized in Tables 1–4 for the three methods labeled STANDARD (which is the known iteration (9)–(10)), SSP-JACOBI (which is the Jacobi version of the new method (7)–(8)), and SSP-Gauss–Seidel (which is the Gauss–Seidel version of the new method (43)–(44)). Let us describe how the test problems were generated. Regarding cost structure, in all problems and for each pair (i, u), the one-stage cost at state i was randomly selected from the range $(0, n)$ according to a uniform distribution. Regarding the transition probabilities, in all the problems, we specified the structure of the transition probability graph by specifying for each state-control pair (i, u), according to some (possibly random) rule, the states j for which the transition probability $p_{ij}(u)$ is nonzero. We then generated each of the nonzero transition probabilities by randomly selecting a corresponding number from the interval $(0, 1)$ according to uniform probability distribution, and by normalizing so that $\sum_{j=1}^{n} p_{ij}(u) = 1$ for all pairs (i, u). The test problems were generated as follows.

(1) Problems of Table 1. Here there is only one control available at each state. The sparsity of the transition probability graph is controlled by a parameter $q \in (0, 1)$. In particular, each possible transition probability is selected to be nonzero with a given probability q. We used sparsity parameters $q = 0.5$, $q = 0.1$, and $q = 0.05$ in our tests.

(2) Problems of Table 2. Here also there is only one control available at each state. At states i with $1 < i < n$, the nonzero transition probabilities are the ones to

the states $i-1$, i, and $i+1$. At state 1 the nonzero transition probabilities are to states 1 and 2, and at state n the nonzero transition probabilities are to states $n-1$ and n. This type of transition probability graph arises in queueing systems.

(3) Problems of Table 3. Here there are two controls available at each state, call them u_1 and u_2. Under u_1, the transition probabilities are specified in the same way as for the problems of Table 2. Under u_2, at each state i with $1 < i < n$, the nonzero transition probabilities are the ones to the states $i-1$ and $i+1$. At state 1 the only nonzero transition probabilities are the ones to the states 1 and 2, and at state n the only nonzero transition probabilities are the ones to the states $n-1$ and n.

(4) Problems of Table 4. Here there are three controls available at each state, call them u_1, u_2, and u_3. Under u_1, the transition probabilities are specified in the same way as for the problems of Table 2. Under u_2, at each state i with $1 < i < n-10$, the nonzero transition probabilities are the ones to the states $i-1$ and $i+10$. At state 1 the only nonzero transition probabilities are the ones to the states 1 and 11, and at states $i = n-10, n-9, \ldots, n$ the only nonzero transition probabilities are the ones to the states $i-1$ and n. Under u_3, at each state i with $10 < i < n$, the nonzero transition probabilities are the ones to the states $i-10$ and $i+1$. At states $i = 1, \ldots, 10$, the only nonzero transition probabilities are the ones to the states 1 and $i+1$, and at state n the only nonzero transition probabilities are the ones to the states $n-10$ and n.

Tables 1–4 give the number of iterations required by each method for the difference between the upper and lower bounds to become smaller than 10^{-3}. Each entry of the tables represents the average of two problems. We should note here that the number of iterations varies a great deal from one problem to another, so the variance of the number of iterations for a given type of problem is very large. For example, one of the two problems in the fourth entry of Table 2 is extremely difficult and requires a much larger number of iterations than the other. However, it is generally true that if a problem is difficult for one method (requires a lot of iterations), it is also difficult for all the other methods.

It can be seen that the problems of Table 1 are generally much easier than the problems of Tables 2–4. Generally, it appears that these problems become more difficult as the sparsity of the transition probability graph increases. On some of these problems (generally the easier ones), the standard method performs extremely well and much better than the new methods. This is probably due to the need for stepsize selection in the new methods. The adaptive stepsize rule that we used generally works well, but on occasion may end up with a stepsize that is either too large or too small for optimal performance. We believe that the subject of appropriate stepsize selection method is a potential topic for theoretical or empirical research.

On the more difficult problems of Tables 2 and 3, the Gauss–Seidel version of the new method is uniformly faster than the other methods. In fact, the Gauss–Seidel method has substantially outperformed the other methods on every single problem with the queueing structure that we tried. The Jacobi version of the new method performs comparably to the standard method on the problems of Tables 2 and 3. What happens here is that for the problems of Tables 2 and 3, the difference between the iterations (7)–(8) and (9)–(10) are minor, particularly when the number of states is large (see the discussion following (9)–(10)).

For the larger problems of Table 4, again the Jacobi version of the new method performs comparably to the standard method. The Gauss–Seidel version of the new method is generally faster than the other methods, but the factor of superiority is problem dependent and its variance is substantial.

5. Conclusions. The methods of this paper were derived by exploiting the connection between average cost and stochastic shortest path problems. We developed a new value iteration method that involves the same type of weighted sup-norm contraction that arises in stochastic shortest path problems. This method is the first, to our knowledge, that admits a convergent Gauss–Seidel implementation. We also believe that the weighted sup-norm contraction property inherent in our method is likely to prove useful in other related contexts.

REFERENCES

[ABB97] J. Abounadi, D. P. Bertsekas, and V. Borkar, *Q-Learning Algorithms for the Average Cost Markovian Decision Problem*, in preparation, 1997.

[BeT89] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1989.

[Ber95] D. P. Bertsekas, *Dynamic Programming and Optimal Control, Vols.* I *and* II, Athena Scientific, Belmont, MA, 1995.

[Odo69] A. R. Odoni *On finding the maximal gain for Markov decision processes*, Oper. Res., 17 (1969), pp. 857–860.

[PBW79] J. L. Popyack, R. L. Brown, and C. C. White, III, *Discrete versions of an algorithm due to Varaiya*, IEEE Trans. Automat. Control, 24 (1969), pp. 503–504.

[Pla77] L. Platzman, *Improved conditions for convergence in undiscounted Markov renewal programming*, Oper. Res., 25 (1977), pp. 529–533.

[Put94] M. L. Puterman, *Markovian Decision Problems*, Wiley, New York, 1994.

[Sch71] P. J. Schweitzer, *Iterative solution of the functional equations of undiscounted Markov renewal programming*, J. Math. Anal. Appl., 34 (1971), pp. 495–501.

[Tse90] P. Tseng, *Solving H-horizon, stationary Markov decision problems in time proportional to* $\log(H)$, Oper. Res. Lett., 9, (1990), pp. 287–297.

[Var78] P. P. Varaiya, *Optimal and suboptimal stationary controls of Markov chains*, IEEE Trans. Automat. Control, AC-23 (1978), pp. 388–394.

[Whi63] D. J. White, *Dynamic programming, Markov chains, and the method of successive approximations*, J. Math. Anal. Appl., 6 (1963), pp. 373–376.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 2, pp. 760–779, March 1998

014

MULTIDIMENSIONAL SYSTEMS WITH FINITE SUPPORT BEHAVIORS: SIGNAL STRUCTURE, GENERATION, AND DETECTION*

ETTORE FORNASINI† AND MARIA ELENA VALCHER†

Abstract. The main features of finite multidimensional behaviors are introduced as properties of the trajectories supports and connected with the polynomial matrices adopted for their description.

Observability and local detectability are shown to be equivalent to the kernel representation of a behavior via some parity check matrix H^T. The main properties of locally undetectable behaviors as well as their connections with the notion of constrained variables are investigated, and a general representation result for finite support behaviors is derived.

The input-output representation via generator matrices is finally discussed, and some connections between matrix primeness and the constraints every trajectory imposes on the support of the corresponding input are analyzed.

Key words. multidimensional systems, behavior theory, polynomial matrices, parity checks, observability, local detectability/undetectability

AMS subject classifications. 93C35, 93B25, 93A25, 94B10

PII. S0363012995292044

1. Introduction.

Behavior theory is the study of the trajectories a dynamical system produces according to its evolution laws. It originated in the analysis of one-dimensional (1D) systems and was developed in a complete and useful form by J. C. Willems in the past two decades. In a series of papers [18, 19, 20], Willems provided a thorough description of the ways a system interacts with its environment, as well as a clear conceptual apparatus for analyzing and identifying the attributes a family of trajectories possibly exhibits. Perhaps the most important of the notions he introduced is external controllability, which displays the way memory function operates and hence constitutes a powerful tool for obtaining state space models of infinite behaviors.

Recently, purely ring-theoretic extensions of Willems theory have been obtained by F. Fagnani, S. K. Mitter, and S. Zampieri in [3, 22]. The new field of research is relevant for the investigation of many classes of systems and makes it quite clear how several concepts of behavior theory depend on the nature of the underlying algebraic structures. Nevertheless, a certain continuity with Willems's former results is apparent, if for no other reason than that the analysis is normally developed and thought of in a standard 1D time domain.

A second stage in the development of behavior theory, initiated by P. Rocha and J. C Willems at the end of the 1980s [13, 14], resulted in the absorption of two-dimensional (2D) signals into the theory. The analysis of 2D behaviors has led to new insights into the classical theory of 2D systems and to new investigations of Laurent polynomial operators, centering around the algebra of matrix pairs and various primeness conditions for polynomial matrices.

*Received by the editors September 20, 1995; accepted for publication (in revised form) February 4, 1997.

http://www.siam.org/journals/sicon/36-2/29204.html

†Dipartimento di Elettronica ed Informatica, Università di Padova, via Gradenigo 6a, 35131 Padova, Italy (fornasini@dei.unipd.it, meme@dei.unipd.it).

Another development in behavior theory is the work of G. D. Forney and M. D. Trott on the behavioral approach to group systems [8]. Like the original work on minimal bases of rational spaces [7], Forney's papers find several applications in the theory of convolutional codes. At the same time, however, they draw on duality theory and suggest new problems on observability and memory span. Also, they emphasize the importance of topological groups in behavior and coding theory.

During the last few years, there has been an increasing interest in convolutional coding of multidimensional (nD) data [4, 16], motivated to a large extent by the possibility of investigating code performances and properties in a behavior context. Also, multidimensional convolutional codes have been a fruitful source of problems and conjectures, both in polynomial modules algebra and in signal processing of discrete data arrays [15].

The aim of this paper is to present, in as self-contained a manner as possible, the behavior theory of finite support multidimensional signals. The finite support assumption is motivated by the fact that in several applications the independent variables represent spatial coordinates, and the phenomenon one aims to model regards only a finite region of the space. So, infinite behaviors, which constitute the core of Willems's theory, are only marginally touched on here. A detailed analysis of the main connections between finite and infinite nD signals falls within the scope of duality theory, and as far as the 2D case is concerned, has been carried on in [16].

Not intending to be inclusive of all aspects of the subject, we have concentrated on what seem to be the most interesting topics to be investigated and have included some preliminary material, as necessary for the discussion. Particular attention has been devoted to the supports of the signals and to certain elementary operations (restriction, extension, and concatenation) which have a concrete meaning from the signal processing standpoint. Actually, several "internal" properties of a behavior have been introduced in terms of these operations and expressed as possibilities of "cutting and pasting together" pieces of different trajectories into a new one.

As each of these features mirrors a particular polynomial matrix representation, an explicit link between the parity checks description of an nD behavior and the concept of observability is derived; indeed, the support of the parity check matrix measures the range of action of the system laws and provides useful bounds on the region where parity checks apply when detecting if some signal is a legal sequence.

The trajectories of an observable behavior can be expressed as the solutions of a system of multidimensional difference equations, and hence can be recognized by means of local testing procedures. Locally undetectable behaviors, instead, exhibit opposite properties, because every finite signal can be completed into a legal trajectory and no local recognition procedure can be successfully implemented. Interestingly enough, these two classes of behaviors allow every finite behavior to be described via intersection operations.

A point of view somewhat complementary to detection calls for an input-output analysis of the way behavior trajectories are generated, and the supports of the trajectories are related to the corresponding inputs. This problem appears particularly relevant when the behavior sequences are injectively generated, and hence a given trajectory is produced by a unique input. Although no general statement can be made about the way these supports are related, specific assumptions on the structure of the generating matrices allow us to uniformly confine the support of each input signal into a suitable extension of the support of the associated output trajectory.

The use of nD Laurent polynomial (L-polynomial) matrices is pervasive throughout the paper; no attempt has been made, however, to give a complete account of

their algebraic properties. For these we may refer the reader to recent books [2] and articles [5, 17, 21] dealing with that part of abstract algebra. A certain attention, however, has been paid to the analysis of the supports of nD L-polynomial vectors, and some results obtained in this context seem to be original.

The paper is organized as follows. The first part introduces the basic definitions and properties of nD finite behaviors; in particular, operations which involve only the supports are sufficient to define the notions of (external) controllability and observability. While controllability is well established [8, 14], and in the context of finite support signals, it follows from linearity and shift invariance, the observability definition we will adopt comes from duality issues and is fully justified when a parity-checks description of the behavior trajectories is adopted. Actually, as shown in section 3, an observable behavior is characterized by a finite set of parity checks one has to apply in order to recognize its trajectories. This result allows observable behaviors to be identified with kernels of polynomial matrix operators or, in more abstract terms, as maximal submodules of given rank in the module of all finite support signals.

In section 4 the notions of unconstrained variables and locally undetectable behaviors are introduced. A general representation result is then provided, showing that every finite behavior can be expressed as the intersection of an observable and a (generally not unique) locally undetectable behavior.

In the last part of the paper we develop the theory of input-output generation of nD behaviors and present some relevant connections between support conditions on the input-output pairs and primeness requirements on the generator matrices.

2. Finite support behaviors: Preliminary definitions and basic properties. Let $\mathbb{F}$ be an arbitrary field and denote by $\mathbf{z}$ the n-tuple $(z_1, z_2, \ldots, z_n)$, so that $\mathbb{F}[\mathbf{z}]$ and $\mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]$ are shorthand notations [1] for the polynomial and the Laurent polynomial (L-polynomial) rings in the indeterminates $z_1, \ldots, z_n$, respectively.

For any sequence $\mathbf{w} = \{\mathbf{w}(\mathbf{h})\}_{\mathbf{h}\in\mathbb{Z}^n}$, taking values in $\mathbb{F}^p$, the *support* of $\mathbf{w}$ is the set of points where $\mathbf{w}$ is nonzero, i.e., $\mathrm{supp}(\mathbf{w}) := \{\mathbf{h} = (h_1, h_2, \ldots, h_n) \in \mathbb{Z}^n : \mathbf{w}(\mathbf{h}) \neq 0\}$. Also, $\mathbf{w}$ can be represented via a formal power series

$$\sum_{h_i\in\mathbb{Z}} \mathbf{w}(h_1, h_2, \ldots, h_n)\; z_1^{h_1} z_2^{h_2} \cdots z_n^{h_n} = \sum_{\mathbf{h}\in\mathbb{Z}^n} \mathbf{w}(\mathbf{h})\; \mathbf{z}^{\mathbf{h}},$$

where $\mathbf{h}$ stands for the n-tuple $(h_1, h_2, \ldots, h_n)$ and $\mathbf{z}^{\mathbf{h}}$ for the term $z_1^{h_1} z_2^{h_2} \ldots z_n^{h_n}$. On the other hand, power series can be viewed as representing vectors with entries in $\mathcal{F}_\infty := \mathbb{F}^{\mathbb{Z}^n}$, thus setting a bijective map between nD sequences taking values in $\mathbb{F}^p$ and formal power series with coefficients in $\mathbb{F}^p$. This allows us to identify nD sequences with the associated power series, in particular, finite support nD signals with L-polynomial vectors, and to denote both of them with the same symbol $\mathbf{w}$. Sometimes, mostly when a power series $\mathbf{w}$ is obtained as a Cauchy product, it will be useful to denote the coefficient of $\mathbf{z}^{\mathbf{h}}$ in $\mathbf{w}$ as $(\mathbf{w}, \mathbf{z}^{\mathbf{h}})$.

Linear operators on the sequence space are represented by appropriate matrices with elements in $\mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]$, whose primeness features find a counterpart in terms of properties of the associated operators. The main primeness notions which arise in the nD context are the following.

DEFINITION 2.1. *An L-polynomial matrix $G \in \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^{p\times m}$, $p \geq m$, is*

- unimodular *if $p = m$ and $\det G$ is a unit in $\mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]$, i.e., $\det G = c\mathbf{z}^{\mathbf{h}}$ for some nonzero $c \in \mathbb{F}$ and some $\mathbf{h} \in \mathbb{Z}^n$;*
- right factor prime *(rFP) if in every factorization $G = \bar{G}T$, with $\bar{G} \in \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^{p\times m}$ and $T \in \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^{m\times m}$, T is a unimodular matrix;*

- right minor prime *(rMP) if its maximal order minors have no common factors;*
- right variety prime *(rVP) if the ideal $\mathcal{I}_G$, generated by its maximal order minors, includes (nonzero) L-polynomials in $\mathbb{F}[z_i, z_i^{-1}]$ for every $i = 1, 2, \ldots, n$;*
- right zero prime *(rZP) if the ideal $\mathcal{I}_G$ is the ring $\mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]$ itself.*

The support of a matrix $G \in \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^{p\times m}$ is the union of the supports of its elements.

An nD (finite) *behavior $\mathfrak{B}$ with p components* is a set of finite support signals (trajectories) taking values in $\mathbb{F}^p$ and endowed with the following properties.

(L) [Linearity]. *If $\mathbf{w}_1$ and $\mathbf{w}_2$ belong to $\mathfrak{B}$, then $\alpha\mathbf{w}_1 + \beta\mathbf{w}_2 \in \mathfrak{B}$ for all α, β in $\mathbb{F}$.*

(SI) [Shift-invariance]. *$\mathbf{w} \in \mathfrak{B}$ implies $\mathbf{v} = \mathbf{z}^{\mathbf{h}}\mathbf{w} \in \mathfrak{B}$ for every $\mathbf{h} \in \mathbb{Z}^n$; i.e., $\mathfrak{B}$ is invariant with respect to the shifts along the coordinate axes in $\mathbb{Z}^n$.*

As every nD behavior $\mathfrak{B}$ can be viewed as an $\mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]$-submodule of $\mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^p$, which is a Noetherian module [11], $\mathfrak{B}$ is finitely generated; i.e., there exists a finite set of column vectors $\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_m$ in $\mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^p$ such that

$$\mathfrak{B} \equiv \left\{ \sum_{i=1}^{m} \mathbf{g}_i u_i : u_i \in \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}] \right\} = \{\mathbf{w} = G\mathbf{u} : \mathbf{u} \in \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^m\} =: \mathrm{Im}G. \tag{2.1}$$

The L-polynomial matrix $G := \mathrm{row}\{\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_m\}$ is called the *generator matrix* of $\mathfrak{B}$.

$G_1 \in \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^{p\times m_1}$ and $G_2 \in \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^{p\times m_2}$ are generator matrices of the same behavior if and only if there exist $P_1 \in \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^{m_1\times m_2}$ and $P_2 \in \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^{m_2\times m_1}$ such that $G_1P_1 = G_2$ and $G_2P_2 = G_1$. Consequently, G_1 and G_2 have the same rank r over the field of rational functions $\mathbb{F}(\mathbf{z})$. Being an invariant with respect to all generator matrices of $\mathfrak{B}$, r is called the *rank of* $\mathfrak{B}$. It somehow represents a complexity index of the behavior, as r independent trajectories can be found in $\mathfrak{B}$, while $r + 1$ trajectories $(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{r+1})$ always satisfy an autoregressive equation $\mathbf{w}_1p_1 + \mathbf{w}_2p_2 + \cdots + \mathbf{w}_{r+1}p_{r+1} = \mathbf{0}$, with at least one nonzero $p_i \in \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]$.

A behavior $\mathfrak{B}$ of rank r is *free* if it admits a full column rank generator matrix, that is, a generator matrix G with r columns. This amounts to saying that each trajectory $\mathbf{w}$ in $\mathfrak{B}$ is uniquely expressed as a linear combination $\mathbf{w} = \mathbf{g}_1u_1 + \mathbf{g}_2u_2 + \cdots + \mathbf{g}_ru_r$, $u_i \in \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]$, of the columns of G.

The main properties of a finite behavior $\mathfrak{B}$ are connected with certain elementary operations we can perform on the system trajectories. These operations essentially reduce to "pasting" pieces of different trajectories into legal elements of $\mathfrak{B}$ or to "cutting" a set of samples out of a given trajectory, so as to obtain a new behavior sequence.

One of the pillars of Willems's behavior theory is the notion of (external) controllability. For 1D controllable behaviors the past has no lasting implications about the future [19], which means that the restriction of a 1D trajectory to $(-\infty, t]$ does not provide any information about the values the trajectory takes on $[t + \delta, +\infty)$, when $\delta > 0$ is properly chosen. In a multidimensional context the notions of "past" and "future" are quite elusive and, in many cases, unsuitable for classifying and processing the available data. What seems more reasonable, instead, is to investigate to what extent the values a trajectory $\mathbf{w}$ assumes on a subset $\mathcal{S}_1 \subset \mathbb{Z}^n$ influence the values on a subset $\mathcal{S}_2$, disjoint from $\mathcal{S}_1$, and to check if there exists a lower bound on the distance

$$d(\mathcal{S}_1, \mathcal{S}_2) := \min \left\{ \sum_{i=1}^{n} |h_i - k_i| : (h_1, h_2, \ldots, h_n) \in \mathcal{S}_1, (k_1, k_2, \ldots, k_n) \in \mathcal{S}_2 \right\}, \tag{2.2}$$

which guarantees that $\mathbf{w}|\mathcal{S}_2$, the *restriction to* $\mathcal{S}_2$ *of the sequence* $\mathbf{w}$, is independent of $\mathbf{w}|\mathcal{S}_1$. This point of view led to the following definition [13].

(C_1) [Controllability]. *A finite behavior* $\mathfrak{B}$ *is* controllable *if there exists an integer* $\delta > 0$ *such that, for any pair of nonempty subsets* $\mathcal{S}_1, \mathcal{S}_2$ *of* $\mathbb{Z}^n$*, with* $d(\mathcal{S}_1, \mathcal{S}_2) \geq \delta$*, and any pair of trajectories* $\mathbf{w}_1$ *and* $\mathbf{w}_2 \in \mathfrak{B}$ *, there exists* $\mathbf{v} \in \mathfrak{B}$ *such that*

$$\mathbf{v}|\mathcal{S}_1 = \mathbf{w}_1|\mathcal{S}_1 \qquad \textit{and} \qquad \mathbf{v}|\mathcal{S}_2 = \mathbf{w}_2|\mathcal{S}_2. \tag{2.3}$$

While definition (C_1) requires pasting together different signals into a new one, the following statement refers to the possibility of finding a legal extension for every portion $\mathbf{w}|\mathcal{S}$ of a behavior trajectory $\mathbf{w}$ by adjusting the sample values in a small area surrounding $\mathcal{S}$. More precisely, by introducing for $\varepsilon \geq 0$ the ε-*extension* of the set $\mathcal{S}$

$$\mathcal{S}^{\varepsilon} := \{\mathbf{h} \in \mathbb{Z}^n : d(\mathbf{h}, \mathcal{S}) \leq \varepsilon\},$$

one can give the following definition.

(C_2) [Zero-controllability]. *A finite behavior* $\mathfrak{B}$ *is* zero-controllable *if there exists an integer* $\varepsilon > 0$ *such that, for any nonempty set* $\mathcal{S}$ *of* $\mathbb{Z}^n$ *and any* $\mathbf{w} \in \mathfrak{B}$*, there exists* $\mathbf{v} \in \mathfrak{B}$ *satisfying*

$$\mathbf{v}|\mathcal{S} = \mathbf{w}|\mathcal{S} \qquad \textit{and} \qquad \mathrm{supp}(\mathbf{v}) \subseteq \mathcal{S}^{\varepsilon}. \tag{2.4}$$

Properties (C_1) and (C_2) make sense for both finite and infinite support behaviors, and the proof of (C_1) $\Leftrightarrow$ (C_2) given below holds for both of them. However, while conditions (C_1) and (C_2) are always met by a finite behavior $\mathfrak{B}$, and essentially follow from the module structure of $\mathfrak{B}$, for an infinite behavior, controllability constitutes an additional constraint with respect to linearity and shift-invariance [13, 14].

PROPOSITION 2.2. *Controllability and zero-controllability are equivalent.*

Proof. (C_1) $\Rightarrow$ (C_2) Assume that $\mathfrak{B}$ meets condition (C_1). Given $\mathbf{w} \in \mathfrak{B}$ and $\mathcal{S} \subset \mathbb{Z}^n$, take in ($C_1$) $\mathbf{w}_1 = \mathbf{w}$, $\mathbf{w}_2 = \mathbf{0}$, $\mathcal{S}_1 = \mathcal{S}$, and $\mathcal{S}_2 = \mathcal{C}\mathcal{S}^{\delta}$, where $\mathcal{C}\mathcal{S}$ denotes the complementary set of $\mathcal{S}$. Then the trajectory $\mathbf{v}$ which fulfills (2.3) satisfies (2.4) with $\varepsilon = \delta$.

(C_2) $\Rightarrow$ (C_1) Assume that $\mathfrak{B}$ satisfies condition (C_2). Given $\mathbf{w}_1$ and $\mathbf{w}_2 \in \mathfrak{B}$ and $\mathcal{S}_1, \mathcal{S}_2 \subset \mathbb{Z}^n$, with $d(\mathcal{S}_1, \mathcal{S}_2) > \varepsilon$, by ($C_2$) there exist $\mathbf{v}_1$ and $\mathbf{v}_2 \in \mathfrak{B}$ such that

$$\mathbf{v}_i|\mathcal{S}_i = \mathbf{w}_i|\mathcal{S}_i, \quad \mathrm{supp}(\mathbf{v}_i) \subset \mathcal{S}_i^{\varepsilon}, \qquad i = 1, 2.$$

Thus $\mathbf{v} := \mathbf{v}_1 + \mathbf{v}_2 \in \mathfrak{B}$ satisfies $\mathbf{v}|\mathcal{S}_i = \mathbf{w}_i|\mathcal{S}_i$, $i = 1, 2$, and (C_1) holds for $\delta = \varepsilon + 1$. □

PROPOSITION 2.3. *A finite behavior* $\mathfrak{B}$ *is controllable.*

Proof. Suppose that $G \in \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^{p \times m}$ is a generator matrix of $\mathfrak{B}$, and let η be a positive integer such that $B(\mathbf{0}, \eta)$, the ball of radius η and center in the origin, includes $\mathrm{supp}(G)$. Consider any set $\mathcal{S} \subset \mathbb{Z}^n$ and $\mathbf{w} = G\mathbf{u} \in \mathfrak{B}$. If $\bar{\mathbf{u}}$ is the sequence which coincides with $\mathbf{u}$ on $\mathcal{S}^{\eta}$ and is zero elsewhere, the trajectory $\mathbf{v} := G\bar{\mathbf{u}}$ satisfies $\mathbf{v}|\mathcal{S} = \mathbf{w}|\mathcal{S}$ and has support which does not exceed $\mathcal{S}^{2\eta}$. So (C_2) is met with $\varepsilon = 2\eta$. □

Given two disjoint sets $\mathcal{S}_1$ and $\mathcal{S}_2$ which are far enough apart, controllability expresses the possibility of steering any behavior sequence known in $\mathcal{S}_1$ into another element of $\mathfrak{B}$ assigned on $\mathcal{S}_2$, meanwhile producing a legal trajectory. Like controllability, observability will also be introduced without reference to the concept of state, according to some recent works of Forney et al. [8, 12]. Observability formalizes the

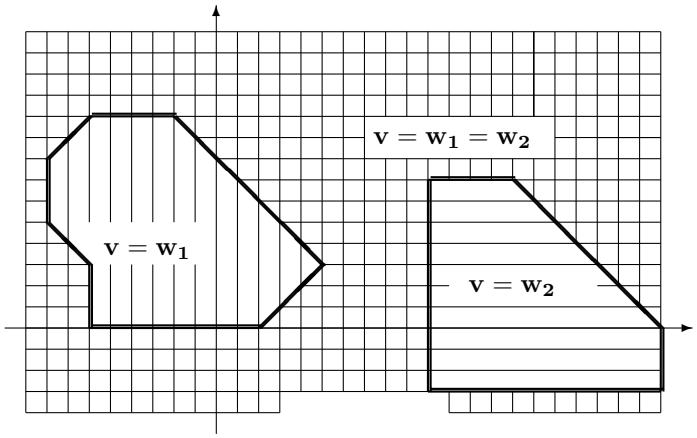


FIG. 2.1.

possibility of pasting into a legal sequence any pair of trajectories that take the same values on a sufficiently large subset of $\mathbb{Z}^n$. This is equivalent to saying that, however a sequence $\mathbf{w} \in \mathfrak{B}$ and a subset $\mathcal{S} \subset \mathbb{Z}^n$ are chosen, the possible extensions of $\mathbf{w}|\mathcal{S}$ only depend on the values of $\mathbf{w}$ on a boundary region of $\mathcal{S}$.

Under this viewpoint, observability endows a behavior with a "separation property" that allows us to take into account only a small amount of data in order to extend a portion of the behavior sequence. Furthermore, once we think of the samples in $\mathcal{S}$ as the information about the past dynamics of the system, observability enables us to design the "future" evolution by considering only the most "recent" data (those on the boundary), thus reminding us of the notion of state.

(O_1) [Observability]. *A finite behavior* $\mathfrak{B}$ *is* observable *if there exists an integer* $\delta > 0$ *such that, for any pair of nonempty subsets* $\mathcal{S}_1, \mathcal{S}_2$ *of* $\mathbb{Z}^n$*, with* $d(\mathcal{S}_1, \mathcal{S}_2) \geq \delta$*, and any pair of trajectories* $\mathbf{w}_1$*,* $\mathbf{w}_2 \in \mathfrak{B}$*, satisfying* $\mathbf{w}_1|\mathcal{C}(\mathcal{S}_1 \cup \mathcal{S}_2) = \mathbf{w}_2|\mathcal{C}(\mathcal{S}_1 \cup \mathcal{S}_2)$*, the trajectory*

$$\mathbf{v}(\mathbf{h}) = \begin{cases} \mathbf{w}_1(\mathbf{h}), & \mathbf{h} \in \mathcal{S}_1, \\ \mathbf{w}_1(\mathbf{h}) = \mathbf{w}_2(\mathbf{h}), & \mathbf{h} \in \mathcal{C}(\mathcal{S}_1 \cup \mathcal{S}_2), \\ \mathbf{w}_2(\mathbf{h}), & \mathbf{h} \in \mathcal{S}_2, \end{cases} \tag{2.5}$$

is an element of $\mathfrak{B}$ (*see Fig.* 2.1).

Observability can be equivalently restated as follows: if the support of a behavior sequence $\mathbf{w}$ can be partitioned into a pair of disjoint subsets, which are far enough apart, the restrictions of $\mathbf{w}$ to each subset represent legal trajectories.

(O_2) [Zero-observability]. *A finite behavior* $\mathfrak{B}$ *is* zero-observable *if there exists an integer* $\varepsilon > 0$ *such that for any* $\mathbf{w} \in \mathfrak{B}$ *satisfying* $\mathbf{w}|(\mathcal{S}^\varepsilon \setminus \mathcal{S}) = \mathbf{0}$*,* $\mathcal{S}$ *a nonempty set in* $\mathbb{Z}^n$*, the sequence*

$$\mathbf{v}(\mathbf{h}) = \begin{cases} \mathbf{w}(\mathbf{h}), & \mathbf{h} \in \mathcal{S}, \\ 0 & \textit{elsewhere} \end{cases} \tag{2.6}$$

belongs to $\mathfrak{B}$ (*see Fig.* 2.2).

PROPOSITION 2.4. *Observability and zero-observability are equivalent.*

Proof. (O_1) $\Rightarrow$ (O_2) Assume that $\mathfrak{B}$ fulfills condition (O_1). Given $\mathcal{S} \subset \mathbb{Z}^n$ and $\mathbf{w} \in \mathfrak{B}$ such that $\mathbf{w}|(\mathcal{S}^\delta \setminus \mathcal{S}) = \mathbf{0}$, take in ($O_1$) $\mathbf{w}_1 = \mathbf{w}$, $\mathbf{w}_2 = \mathbf{0}$, $\mathcal{S}_1 = \mathcal{S}$, and $\mathcal{S}_2 = \mathcal{C}\mathcal{S}^\delta$. The trajectory $\mathbf{v} \in \mathfrak{B}$ satisfying (2.5), also satisfies (2.6), with $\varepsilon = \delta$.

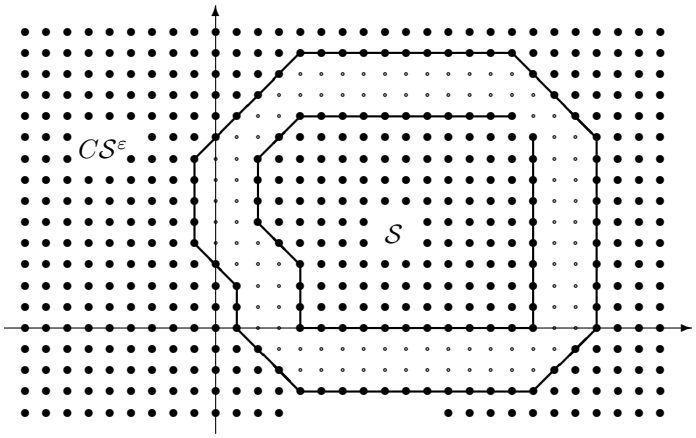


FIG. 2.2.

$(O_2) \Rightarrow (O_1)$ Assume that $\mathfrak{B}$ fulfills condition (O_2). Given $\mathcal{S}_1, \mathcal{S}_2 \subset \mathbb{Z}^n$, with $d(\mathcal{S}_1, \mathcal{S}_2) > \varepsilon$, and $\mathbf{w}_1, \mathbf{w}_2 \in \mathfrak{B}$ satisfying $\mathbf{w}_1|\mathcal{C}(\mathcal{S}_1 \cup \mathcal{S}_2) = \mathbf{w}_2|\mathcal{C}(\mathcal{S}_1 \cup \mathcal{S}_2)$, the sequence $\mathbf{w}_1 - \mathbf{w}_2 \in \mathfrak{B}$ satisfies $(\mathbf{w}_1 - \mathbf{w}_2)|\mathcal{C}(\mathcal{S}_1 \cup \mathcal{S}_2) = \mathbf{0}$. As a consequence, the sequence $\mathbf{w}$ given by

$$\mathbf{w}(\mathbf{h}) = \begin{cases} \mathbf{w}_1(\mathbf{h}) - \mathbf{w}_2(\mathbf{h}), & \mathbf{h} \in \mathcal{S}_1, \\ 0 & \text{elsewhere}, \end{cases}$$

is in $\mathfrak{B}$, and $\mathbf{v} := \mathbf{w} + \mathbf{w}_2 \in \mathfrak{B}$ fulfills (2.5). So, (O_1) holds for $\delta = \varepsilon + 1$. □

3. Parity checks and trajectory recognition. Underlying the definition of controllability is the idea of driving a portion of a trajectory into another one, provided that there is room enough for adjustments. In rough terms, one's objective is to manipulate the control variables to cause the system to behave in $\mathcal{S}_2$ in a more desirable manner than is expected by watching the system trajectory on $\mathcal{S}_1$. So, controllability is naturally connected with the generation of $\mathfrak{B}$ as the image of some matrix G, acting on the input space.

Observability is somehow related to the "dual" issue of recognizing whether a given sequence $\mathbf{v} \in \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^p$ is an element of $\mathfrak{B}$. This problem, which typically arises in fault detection and convolutional encoding contexts, can be managed by resorting to a linear filter (residual generator or syndrome former) that produces an identically zero output signal when the input is an admissible trajectory of $\mathfrak{B}$. From a mathematical point of view, this requires us to find a set of sequences (parity checks) endowed with the property that their convolution with every element of $\mathfrak{B}$ is zero.

So, for a given behavior $\mathfrak{B} \subseteq \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^p$, a (finite) *parity check* is a column vector $\mathbf{s} \in \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^p$ that satisfies $\mathbf{s}^T\mathbf{w} = \mathbf{0}$ for all $\mathbf{w} \in \mathfrak{B}$. The set $\mathfrak{B}^\perp$ of all finite parity checks of $\mathfrak{B}$ is the *orthogonal behavior*, and as a submodule of $\mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^p$, it is generated by the columns of some matrix $H \in \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^{p \times q}$, that is,

$$\mathfrak{B}^\perp = \{\mathbf{s} \in \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^p : \mathbf{s} = H\mathbf{x}, \mathbf{x} \in \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^q\} = \mathrm{Im}H. \tag{3.1}$$

Condition $\mathbf{s}^T\mathbf{w} = \mathbf{0}$, $\forall\ \mathbf{s} \in \mathfrak{B}^\perp$, however, need not imply $\mathbf{w} \in \mathfrak{B}$. In general,

$$\mathfrak{B}^{\perp\perp} := \{\mathbf{w} \in \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^p : \mathbf{s}^T\mathbf{w} = \mathbf{0} \forall\ \mathbf{s} \in \mathfrak{B}^\perp\} \tag{3.2}$$

properly includes $\mathfrak{B}$ and is the set of all L-polynomial vectors obtained by combining the columns of G over the field of rational functions $\mathbb{F}(\mathbf{z})$. It is clear that $\mathfrak{B}$ can be identified via a finite set of parity checks if and only if $\mathfrak{B} = \mathfrak{B}^{\perp\perp}$ or, equivalently,

$$\mathfrak{B} = \ker H^T := \{\mathbf{w} \in \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^p : H^T\mathbf{w} = \mathbf{0}\}. \tag{3.3}$$

In this setting, observability finds a somewhat more substantial interpretation. Actually, if $\mathfrak{B} = \ker H^T$, the restriction of a trajectory to a set $\mathcal{S}$ still provides a legal signal every time the distance between $\mathcal{S}$ and the remaining support of the trajectory exceeds the range of action of the parity check matrix H.

Proposition 3.1 below shows that kernel representations are possible, as can be expected, only for observable behaviors, and makes it clear that observability induces further constraints on the structure of $\mathfrak{B}$, in addition to linearity and shift-invariance.

PROPOSITION 3.1. *A behavior $\mathfrak{B} \subseteq \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^p$ is observable if and only if there exist an integer $h > 0$ and an L-polynomial matrix $H^T \in \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^{h\times p}$ such that $\mathfrak{B} = \ker H^T$.*

The proof of the proposition depends on a couple of technical lemmas.

LEMMA 3.2. *Let R be an integral domain, and consider the polynomial in $R[z]$,*

$$m(z) = \alpha_0\ z^r - \alpha_1\ z^{r-1} + \alpha_2\ z^{r-2} + \cdots + (-1)^r\alpha_r.$$

For any $\rho \geq 0$ there is a $p(z) \in R[z]$ such that $p(z)m(z) \in R[z^{\rho+1}]$.

Proof. Let Q be the field of fractions of R, and L the algebraic closure of Q. Then $m(z)$ can be written as $m(z) = \alpha_0 \prod_{i=1}^r (z - \xi_i)$, where $\xi_i \in L,\ i = 1, 2, \ldots, n$, and

$$\sum_i \xi_i = \alpha_1/\alpha_0, \qquad \sum_{i<j} \xi_i\xi_j = \alpha_2/\alpha_0, \quad \ldots \quad \xi_1\xi_2\cdots\xi_r = \alpha_r/\alpha_0. \tag{3.4}$$

Consider the following polynomial in $L[z]$:

$$\begin{aligned}\tilde{p}(z) &= \prod_{i=1}^r (z^\rho + \xi_i z^{\rho-1} + \xi_i^2 z^{\rho-2} + \cdots + \xi_i^{\rho-1} z + \xi_i^\rho) \\ &= \sum_{0\leq i_1,i_2,\ldots,i_r\leq\rho} z^{r\rho - i_1 - i_2 - \cdots - i_r} \xi_1^{i_1}\xi_2^{i_2}\cdots\xi_r^{i_r} = \sum_{t=0}^{r\rho} z^{r\rho-t} \sum_{\substack{i_1+i_2+\cdots+i_r=t \\ 0\leq i_1,i_2,\ldots,i_r\leq\rho}} \xi_1^{i_1}\xi_2^{i_2}\cdots\xi_r^{i_r}.\end{aligned}$$

Each coefficient of $\tilde{p}(z)$ is a symmetric polynomial in the indeterminates $\xi_1, \xi_2, \ldots, \xi_r$, with integer coefficients, and hence it is expressible [11] as a polynomial in the elementary symmetric polynomials defined in (3.4), again with integer coefficients. Thus $\tilde{p}(z)$ is in $Q[z]$, the denominators of its coefficients are powers of α_0, and there exists a positive integer ν such that $p(z) := \alpha_o^\nu\, \tilde{p}(z)$ belongs to $R[z]$. To conclude the proof, we note that $p(z)m(z)$ is an element of $R[z^{\rho+1}]$ since

$$\begin{aligned}p(z)m(z) &= \alpha_o^\nu \prod_{i=1}^r [(z^\rho + \xi_i z^{\rho-1} + \xi_i^2 z^{\rho-2} + \cdots + \xi_i^\rho)(z - \xi_i)] \\ &= \alpha_o^\nu \prod_{i=1}^r (z^{\rho+1} - \xi_i^{\rho+1}). \qquad \square\end{aligned}$$

LEMMA 3.3. *Let $m(\mathbf{z})$ be in $\mathbb{F}[\mathbf{z}]$. For any integer $\rho > 0$ there is $p(\mathbf{z}) \in \mathbb{F}[\mathbf{z}]$ s.t.*

$$m(\mathbf{z})p(\mathbf{z}) \ \in\ \mathbb{F}[\mathbf{z}^\rho] := \mathbb{F}[z_1^\rho, \ldots, z_n^\rho]. \tag{3.5}$$

Proof. As $m(\mathbf{z}) = m(z_1, \ldots, z_n)$ can be viewed as an element of $\mathbb{F}[z_1, \ldots, z_{n-1}][z_n]$, by Lemma 3.2 there exists $p_1(z_1, \ldots, z_n) \in \mathbb{F}[z_1, \ldots, z_{n-1}][z_n]$ such that

$$m_1(z_1, \ldots, z_n^\rho) := m(z_1, \ldots, z_n) p_1(z_1, \ldots, z_n) \in \mathbb{F}[z_1, \ldots, z_{n-1}][z_n^\rho].$$

Looking at $m_1(z_1, \ldots, z_n^\rho)$ as a polynomial in $\mathbb{F}[z_1, \ldots, z_{n-2}, z_n^\rho][z_{n-1}]$, we know that there exists $p_2(z_1, \ldots, z_{n-1}, z_n^\rho)$ such that

$$m_2(z_1, \ldots, z_{n-1}^\rho, z_n^\rho) := m_1(z_1, \ldots, z_{n-1}, z_n^\rho) p_2(z_1, \ldots, z_{n-1}, z_n^\rho) \in \mathbb{F}[z_1, \ldots, z_{n-2}, z_n^\rho][z_{n-1}^\rho].$$

In n steps we end up with a polynomial

$$\begin{aligned} m_n(z_1^\rho, z_2^\rho, \ldots, z_n^\rho) : = & \, m(z_1, \ldots, z_n) p_1(z_1, \ldots, z_n) \\ & \cdot \; p_2(z_1, z_2, \ldots, z_{n-1}, z_n^\rho) \ldots p_n(z_1^\rho, z_2^\rho, \ldots, z_n^\rho) \in \mathbb{F}[z_1^\rho, z_2^\rho, \ldots, z_n^\rho], \end{aligned}$$

and (3.5) holds with $p = p_1 \; p_2 \; \ldots \; p_n$. ∎

Proof of Proposition 3.1. Assume that $\mathfrak{B} = \mathrm{Im}G$, $G \in \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^{p \times m}$, is an observable behavior, and let $\mathfrak{B}^\perp = \mathrm{Im}\; H$, $H \in \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^{p \times q}$, denote the orthogonal behavior introduced in (3.1). We will show that $\mathfrak{B} \equiv \ker H^T$. Since $H^T G = \mathbf{0}$, it is clear that $\ker H^T \supseteq \mathfrak{B}$. To prove the converse, express $\mathbf{w} \in \ker H^T$ as $\mathbf{w} = G\mathbf{n}/d(\mathbf{z})$, $d \in \mathbb{F}[\mathbf{z}], \mathbf{n} \in \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^{m \times 1}$. By Lemma 3.3, for every integer $\rho > 0$, there is a suitable polynomial $p(\mathbf{z})$ such that $p(\mathbf{z})d(\mathbf{z}) \in \mathbb{F}[z_1^\rho, \ldots, z_n^\rho]$. If property (O_2) holds for $\varepsilon > 0$, and $r > 0$ is an integer such that $\mathrm{supp}(\mathbf{w}) \subseteq B(\mathbf{0}, r)$, we choose $\rho > 2r + \varepsilon$. So, the behavior sequence $p(\mathbf{z})d(\mathbf{z})\mathbf{w} = G\mathbf{n}p(\mathbf{z})$ can be written as

$$\sum_{i_1, i_2, \ldots, i_n} c_{i_1, i_2, \ldots, i_n} z_1^{\rho i_1} z_2^{\rho i_2} \cdots z_n^{\rho i_n} \; \mathbf{w}$$

and thus is the sum of disjoint shifted copies of $\mathbf{w}$, and the distance between two arbitrary copies exceeds ε. So, by (O_2), each copy of $\mathbf{w}$, and hence $\mathbf{w}$ itself, is in $\mathfrak{B}$.

Conversely, let $\mathfrak{B} = \ker H^T$ and set $\varepsilon = 2s$, with $s > 0$ an integer s.t. $B(\mathbf{0}, s) \supseteq \mathrm{supp}(H^T)$. If $\mathcal{S}$ is a subset of $\mathbb{Z}^n$ and $\mathbf{w} \in \mathfrak{B}$ satisfies $\mathbf{w}|(\mathcal{S}^\varepsilon \setminus \mathcal{S}) = \mathbf{0}$, the sequence

$$\mathbf{v}(\mathbf{h}) = \begin{cases} \mathbf{w}(\mathbf{h}), & \mathbf{h} \in \mathcal{S}, \\ 0 & \text{elsewhere} \end{cases}$$

is in $\ker H^T$ and hence in $\mathfrak{B}$. Consequently, $\mathfrak{B}$ is zero-observable. ∎

The kernel description given in Proposition 3.1 leads to new insights into the internal structure of an observable behavior. Observability, indeed, expresses a sort of "localization" of the system laws or, equivalently, the existence of a bound on the size of all windows (in $\mathbb{Z}^n$) we have to look at when deciding whether a signal belongs to $\mathfrak{B}$. Denoting by $\mathfrak{B}|\mathcal{S} := \{\mathbf{w}|\mathcal{S} : \mathbf{w} \in \mathfrak{B}\}$ the set of all restrictions to $\mathcal{S}$ of behavior trajectories, the above localization property finds a formal statement as follows.

(O_3) [Local detectability]. *A finite behavior* $\mathfrak{B}$ *is* locally detectable *if there is an integer* $\nu > 0$ *such that every signal* $\mathbf{w}$ *satisfying* $\mathbf{w}|\mathcal{S} \in \mathfrak{B}|\mathcal{S}$ *for every* $\mathcal{S} \subset \mathbb{Z}^n$ *with* $\mathrm{diam}(\mathcal{S}) \leq \nu$ *is in* $\mathfrak{B}$.

PROPOSITION 3.4. *Local detectability and observability are equivalent.*

Proof. Assume that $\mathfrak{B}$ satisfies (O_3) for a certain $\nu > 0$. Given $\mathcal{S} \subset \mathbb{Z}^n$ and $\mathbf{w} \in \mathfrak{B}$ such that $\mathbf{w}|(\mathcal{S}^\nu \setminus \mathcal{S}) = \mathbf{0}$, define $\mathbf{v}$ as follows:

$$\mathbf{v}(\mathbf{h}) = \begin{cases} \mathbf{w}(\mathbf{h}), & \mathbf{h} \in \mathcal{S}^\nu, \\ 0 & \text{elsewhere.} \end{cases} \tag{3.6}$$

Consider any window $\mathcal{W}$, with $\operatorname{diam}(\mathcal{W}) \leq \nu$. If $\mathcal{W}$ is included in $\mathcal{S}^\nu$, then $\mathbf{v}|\mathcal{W} = \mathbf{w}|\mathcal{W} \in \mathfrak{B}|\mathcal{W}$; otherwise we have $\mathcal{W} \cap \mathcal{S} = \emptyset$, and therefore

$$\mathbf{v}|\mathcal{W} = \mathbf{0}|\mathcal{W} \in \mathfrak{B}|\mathcal{W}. \tag{3.7}$$

So, by (O_3), $\mathbf{v}$ is a legal trajectory, and (O_2) holds for $\varepsilon = \nu$.

Conversely, assume that $\mathfrak{B}$ is observable. By Proposition 3.1, there exists an L-polynomial matrix $H \in \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^{p\times q}$ such that $\mathfrak{B} = \ker H^T$. Let $\nu > 0$ be an integer such that $\operatorname{supp}(H^T) \subseteq B(\mathbf{0}, \nu)$, and suppose that $\mathbf{v}$ is any signal satisfying $\mathbf{v}|\mathcal{S} \in \mathfrak{B}|\mathcal{S}$ for every $\mathcal{S} \subset \mathbb{Z}^n$ with $\operatorname{diam}(\mathcal{S}) \leq 2\nu$. If $\bar{\mathcal{S}} := -\operatorname{supp}(H^T)$, the computation of the coefficient of $\mathbf{z}^{\mathbf{k}}$ in $H^T\mathbf{v}$ involves only samples of $\mathbf{v}$ indexed in

$$\mathbf{k} + \bar{\mathcal{S}} := \{\mathbf{h} \in \mathbb{Z}^n : \mathbf{h} - \mathbf{k} \in \bar{\mathcal{S}}\} = -\operatorname{supp}(\mathbf{z}^{\mathbf{k}} H^T). \tag{3.8}$$

On the other hand, since $\operatorname{diam}(\mathbf{k} + \bar{\mathcal{S}}) \leq 2\nu$, there exists $\mathbf{w}_{\mathbf{k}} \in \mathfrak{B}$ which satisfies $\mathbf{v}|(\mathbf{k} + \bar{\mathcal{S}}) = \mathbf{w}_{\mathbf{k}}|(\mathbf{k} + \bar{\mathcal{S}})$, and this result holds for every $\mathbf{k} \in \mathbb{Z}^n$. So, the coefficient of $\mathbf{z}^{\mathbf{k}}$ in $H^T\mathbf{v}$ is the same as in $H^T\mathbf{w}_{\mathbf{k}} \equiv \mathbf{0}$, and hence $\mathbf{v} \in \ker H^T = \mathfrak{B}$. □

The equivalent descriptions of observability given in $(\mathrm{O}_1) \div (\mathrm{O}_3)$ rely on the trajectories' supports, whereas Proposition 3.1 involves parity checks and kernel representations. A different approach to this notion consists of regarding behaviors with p components as elements in the lattice of submodules of $\mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^p$, and investigating whether observable elements enjoy some special ordering properties.

In keeping with the same spirit, we may investigate how an observable behavior is affected by certain "extension operations" that merge lattice elements into larger ones. There are essentially two natural ways to perform these extensions: one consists of embedding $\mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^p$, and therefore each of its submodules, in the rational vector space $\mathbb{F}(\mathbf{z})^p$, the other of considering $\mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^p$ as a submodule of $\mathcal{F}_\infty^p$, the set of nD trajectories with p components, whose supports possibly extend to the whole space $\mathbb{Z}^n$.

Once a behavior $\mathfrak{B}$ with p components is given, in the first case we have to consider the smallest vector subspace of $\mathbb{F}(\mathbf{z})^p$ including $\mathfrak{B}$,

$$\mathfrak{B}_{\mathrm{rat}} := \left\{ \sum_{i=1}^{r} \mathbf{w}_i a_i : \mathbf{w}_i \in \mathfrak{B},\ a_i \in \mathbb{F}(\mathbf{z}),\ r \in \mathbb{N} \right\}, \tag{3.9}$$

and restrict our attention to the submodule $\mathfrak{B}_{\mathrm{rat}} \cap \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^p$ of finite support sequences. In general, this properly includes $\mathfrak{B}$, and hence is a larger element of the lattice. In the other case, we merge $\mathfrak{B}$ in

$$\mathfrak{B}_\infty := \left\{ \sum_{i=1}^{r} \mathbf{w}_i a_i : \mathbf{w}_i \in \mathfrak{B},\ a_i \in \mathcal{F}_\infty,\ r \in \mathbb{N} \right\}, \tag{3.10}$$

the smallest $\mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]$-submodule of $\mathcal{F}_\infty^p$ which includes $\mathfrak{B}$. Again, we have to confine ourselves to the set of its finite elements $\mathfrak{B}_\infty \cap \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^p$, which clearly includes all trajectories of $\mathfrak{B}$.

PROPOSITION 3.5. *Let $\mathfrak{B} \subseteq \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^p$ be a behavior of rank r. The following statements are equivalent:*

i) $\mathfrak{B}$ *is observable;*
ii) $\mathfrak{B} \equiv \mathfrak{B}_\infty \cap \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^p$*;*
iii) $\mathfrak{B} \equiv \mathfrak{B}_{\mathrm{rat}} \cap \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^p$*;*
iv) $\mathfrak{B}$ *is maximal in the set of all submodules of $\mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^p$ of rank r;*
v) $s\mathbf{w} \in \mathfrak{B} \Rightarrow \mathbf{w} \in \mathfrak{B}$ *for every $\mathbf{w} \in \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^p$ and every nonzero $s \in \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]$;*
vi) $\mathfrak{B} = \mathfrak{B}^{\perp\perp}$.

Proof. i) ⇒ ii) As $\mathfrak{B}$ is observable, there exists $H \in \mathbb{F}[\mathbf{z},\mathbf{z}^{-1}]^{p\times q}$ such that $\mathfrak{B} = \ker H^T = \{\mathbf{w} \in \mathbb{F}[\mathbf{z},\mathbf{z}^{-1}]^p : H^T\mathbf{w} = \mathbf{0}\}$. If $\mathbf{w} \in \mathfrak{B}_\infty \cap \mathbb{F}[\mathbf{z},\mathbf{z}^{-1}]^p$, then $\mathbf{w} = \sum_i \mathbf{w}_i a_i$, $a_i \in \mathcal{F}_\infty, \mathbf{w}_i \in \mathcal{B}$, and therefore $H^T\mathbf{w} = H^T(\sum_i \mathbf{w}_i a_i) = \sum_i (H^T\mathbf{w}_i)a_i = \mathbf{0}$. Thus $\mathbf{w} \in \ker H^T = \mathfrak{B}$, which implies $\mathfrak{B} \supseteq \mathfrak{B}_\infty \cap \mathbb{F}[\mathbf{z},\mathbf{z}^{-1}]^p$. The reverse inclusion is obvious.

ii) ⇒ iii) This follows immediately from $\mathfrak{B} \subseteq \mathfrak{B}_{\rm rat} \cap \mathbb{F}[\mathbf{z},\mathbf{z}^{-1}]^p \subseteq \mathfrak{B}_\infty \cap \mathbb{F}[\mathbf{z},\mathbf{z}^{-1}]^p$.

iii) ⇒ iv) If $\mathfrak{B}' \supseteq \mathfrak{B}$ and $\text{rank}\mathcal{B}' = \text{rank}\mathfrak{B}$, it is clear that $\mathfrak{B}$ and $\mathfrak{B}'$ generate the same $\mathbb{F}(\mathbf{z})$-subspace of $\mathbb{F}(\mathbf{z})^p$ and, consequently, $\mathfrak{B}_{\rm rat} \cap \mathbb{F}[\mathbf{z},\mathbf{z}^{-1}]^p = \mathfrak{B}'_{\rm rat} \cap \mathbb{F}[\mathbf{z},\mathbf{z}^{-1}]^p$. So, the inclusions chain $\mathfrak{B}_{\rm rat} \cap \mathbb{F}[\mathbf{z},\mathbf{z}^{-1}]^p \supseteq \mathfrak{B}' \supseteq \mathfrak{B}$ and assumption iii) together imply $\mathfrak{B}' = \mathfrak{B}$, which means that $\mathfrak{B}$ is maximal.

iv) ⇒ v) Suppose $s\mathbf{w} \in \mathfrak{B}$, $s \in \mathbb{F}[\mathbf{z},\mathbf{z}^{-1}]$. The behavior $\mathfrak{B}'$ generated by $\mathfrak{B}$ and $\mathbf{w}$ has the same rank of $\mathfrak{B}$, and hence, by the maximality assumption, coincides with $\mathfrak{B}$.

v) ⇒ vi) As $\mathfrak{B}$ and $\mathfrak{B}^{\perp\perp}$ have the same rank r and $\mathfrak{B}^{\perp\perp} \supseteq \mathfrak{B}$, both behaviors generate the same $\mathbb{F}(\mathbf{z})$-subspace of $\mathbb{F}(\mathbf{z})^p$. In particular, $\mathbf{w} \in \mathfrak{B}^{\perp\perp}$ implies $\mathbf{w} \in (\mathfrak{B}^{\perp\perp})_{\rm rat} = \mathfrak{B}_{\rm rat}$. So, there exist $p_i, s_i \in \mathbb{F}[\mathbf{z},\mathbf{z}^{-1}]$ and $\mathbf{w}_i \in \mathfrak{B}$, such that $\mathbf{w} = \sum_{i=1}^r \mathbf{w}_i\, p_i/s_i$, which implies $s\mathbf{w} \in \mathfrak{B}$, where s is the least common multiple (ℓ.c.m.) of the s_i's. By assumption v), also, $\mathbf{w}$ is in $\mathfrak{B}$.

vi) ⇒ i) Since $\mathfrak{B}^\perp$ is a submodule of $\mathbb{F}[\mathbf{z},\mathbf{z}^{-1}]^p$, there exists a suitable L-polynomial matrix H such that $\mathfrak{B}^\perp = \text{Im}H$. So

$$\mathfrak{B}^{\perp\perp} = (\mathfrak{B}^\perp)^\perp = \{\mathbf{w} \in \mathbb{F}[\mathbf{z},\mathbf{z}^{-1}]^p : \mathbf{v}^T\mathbf{w} = \mathbf{0} \forall \mathbf{v} \in \text{Im}H\} = \ker H^T.$$

By assumption vi), $\mathfrak{B}$ coincides with $\ker H^T$ and hence is observable. □

4. Behavior decomposition. In this section we take a first step toward a structural analysis of finite support behaviors. The scope of structure theory is to describe general behaviors in terms of some simpler ones—simpler in some perceptible way, perhaps concreteness, or tractability. Of essential importance, after one has decided upon these simpler objects, is to find a method of passing down to them and to discover how they weave together to yield the general behavior with which we began.

Observable behaviors constitute good candidates for these simpler objects, as each behavior can be embedded into an observable one. In order to represent a general behavior $\mathfrak{B}$, then, we have to slice out of its enveloping observable behavior $\mathcal{B}^{\perp\perp}$ a certain part. This can be done by intersecting $\mathcal{B}^{\perp\perp}$ with a suitable, not necessarily unique, element of a behavior class that exhibits properties which are as far as possible from observability and hence from local detectability. The definition of this class depends on the notion of constrained variables which we now introduce.

DEFINITION 4.1. *Let $\mathfrak{B} \subseteq \mathbb{F}[\mathbf{z},\mathbf{z}^{-1}]^p$ be a finite support behavior and $\{i_1, i_2, \ldots, i_r\}$, $r < p$, a subset of $\{1, 2, \ldots, p\}$. We call $w_{i_1}, w_{i_2}, \ldots, w_{i_r}$* constrained variables *of $\mathfrak{B}$ if for every pair of trajectories $\mathbf{v}, \mathbf{v}' \in \mathfrak{B}$, $v_j = v'_j$ for every $j \notin \{i_1, i_2, \ldots, i_r\}$ implies $\mathbf{v} = \mathbf{v}'$.*

As shown in the following lemma, the maximum number of constrained variables of a behavior $\mathfrak{B}$ in $\mathbb{F}[\mathbf{z},\mathbf{z}^{-1}]^p$ can be expressed in terms of the rank and the number of components of $\mathfrak{B}$.

LEMMA 4.2. *Let $\mathfrak{B} \subseteq \mathbb{F}[\mathbf{z},\mathbf{z}^{-1}]^p$ be a behavior of rank r. The maximum number of constrained variables of $\mathfrak{B}$ is $p - r$.*

Proof. Let $G \in \mathbb{F}[\mathbf{z},\mathbf{z}^{-1}]^{p\times m}$ be a generator matrix of $\mathfrak{B}$ and suppose, for sake of simplicity, that the first r rows of G are linearly independent, so that in

$$G = \begin{bmatrix} G_1 \\ G_2 \end{bmatrix} \begin{matrix} \}r \\ \}p-r \end{matrix},$$

G_1 has full row rank. The components w_i, $i = r+1, r+2, \ldots, n$, are constrained variables. If not, there would be a trajectory $\mathbf{w} = \left[\begin{smallmatrix} \mathbf{0} \\ \mathbf{w}_2 \end{smallmatrix}\right]$ in $\mathfrak{B}$, with $\mathbf{w}_2 \neq 0$, and hence an L-polynomial vector $\mathbf{u} \in \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^m$ s.t. $\left[\begin{smallmatrix} G_1 \\ G_2 \end{smallmatrix}\right]\mathbf{u} = \left[\begin{smallmatrix} \mathbf{0} \\ \mathbf{w}_2 \end{smallmatrix}\right]$. This is a contradiction, however, because

$$\text{rank } G_1 = \text{rank} \begin{bmatrix} G_1 \\ G_2 \end{bmatrix} \qquad \Rightarrow \qquad \ker G_1 = \ker \begin{bmatrix} G_1 \\ G_2 \end{bmatrix}.$$

It remains to prove that the number of constrained variables cannot exceed $p - r$. Suppose, instead, that $k > p - r$ variables of $\mathfrak{B}$, say the last k, are constrained, and partition the generator matrix G into

$$G = \begin{bmatrix} \hat{G}_1 \\ \hat{G}_2 \end{bmatrix} \begin{matrix} \}p-k \\ \}k \end{matrix}.$$

As $r = \text{rank } G > \text{rank}\hat{G}_1$, $\ker\hat{G}_1$ properly includes $\ker G$. Consequently, there exists $\mathbf{u}$ s.t. $\hat{G}_2\mathbf{u} \neq \mathbf{0}$ and both $\left[\begin{smallmatrix} \mathbf{0} \\ \mathbf{0} \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} \mathbf{0} \\ \hat{G}_2\mathbf{u} \end{smallmatrix}\right], \hat{G}_2\mathbf{u} \neq \mathbf{0}$, are in $\mathfrak{B}$, which contradicts the assumption that the last k components are constrained. □

A behavior $\mathfrak{B}$ devoid of constrained variables exhibits the very peculiar feature that for every finite set $\mathcal{S} \subset \mathbb{Z}^n$, $\mathfrak{B}|\mathcal{S}$ coincides with $\mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^p \mid \mathcal{S}$. This property, which appears somehow opposite to local detectability, makes it impossible to recognize the trajectories of $\mathfrak{B}$ by resorting to a local checking procedure.

(LU) [Local undetectability]. *A behavior* $\mathfrak{B} \subseteq \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^p$ *is* locally undetectable *if there exists* $\delta > 0$ *s.t. for every sequence* $\mathbf{v} \in \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^p$ *and every set* $\mathcal{S} \subset \mathbb{Z}^n$, *a trajectory* $\mathbf{w} \in \mathfrak{B}$ *can be found satisfying*

$$\mathbf{w}|\mathcal{S} = \mathbf{v}|\mathcal{S} \qquad \textit{and} \qquad \text{supp}(\mathbf{w}) \subseteq \mathcal{S}^\delta. \tag{4.1}$$

PROPOSITION 4.3. *Let* $\mathfrak{B} \subseteq \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^p$ *be a finite support behavior. The following facts are equivalent:*

i) $\mathfrak{B}$ *is devoid of constrained variables;*

ii) $\mathfrak{B}$ *is the image of some L-polynomial matrix* $G \in \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^{p\times m}$ *with rank* p*;*

iii) $\mathfrak{B}$ *is locally undetectable.*

Proof. i) $\Leftrightarrow$ ii) The proof is immediate from Lemma 4.1.

ii) $\Rightarrow$ iii) Let $\mathbf{v}$ be an arbitrary sequence in $\mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^p$ and $\mathcal{S}$ a finite set. If $\mathfrak{B} = \text{Im}G$, for some $G \in \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^{p\times m}$ of rank p, $\mathbf{v}$ can be obtained as the image of some rational vector $\mathbf{u} \in \mathbb{F}(\mathbf{z})^p$, i.e., $\mathbf{v} = G\mathbf{u}$. Consider a power series expansion of $\mathbf{u}$ with support in a suitable cone of $\mathbb{Z}^n$, and introduce the finite sequence

$$\bar{\mathbf{u}}(\mathbf{h}) := \begin{cases} \mathbf{u}(\mathbf{h}), & \mathbf{h} \in \mathcal{S}^\varepsilon, \\ \mathbf{0} & \text{elsewhere,} \end{cases}$$

where ε is the radius of a ball centered in the origin and including the support of G. The behavior sequence $\bar{\mathbf{v}} := G\bar{\mathbf{u}}$ coincides with $\mathbf{v}$ on $\mathcal{S}$ and has support included in $\mathcal{S}^{2\varepsilon}$. So, (4.1) holds with $\delta = 2\varepsilon$.

iii) $\Rightarrow$ ii) Suppose that $\mathfrak{B}$ is locally undetectable and assume, by contradiction, that $\mathfrak{B} = \text{Im}G$, for some $G \in \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^{p\times m}$ with rank less than p. Then there exists a nonzero L-polynomial vector $\mathbf{h} \in \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^p$ satisfying $\mathbf{h}^T G = 0$. Consider a sequence $\mathbf{v} \in \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^p$ s.t. $\mathbf{h}^T\mathbf{v} \neq 0$ and a set $\mathcal{T} \subset \mathbb{Z}^n$ which includes both $\text{supp}(\mathbf{v})$ and $\text{supp}(\mathbf{h}^T\mathbf{v})$, and define $\mathcal{S} := \mathcal{T}^\rho$, where ρ is the radius of a ball, centered in the origin, which includes $\text{supp}(\mathbf{h})$. If property (LU) holds for some $\delta > 0$, there is a trajectory $\mathbf{w} \in \mathfrak{B}$ that can be expressed as $\mathbf{w} = \mathbf{v} + \mathbf{r}$ for some $\mathbf{r}$ with support in $\mathcal{S}^\delta \setminus \mathcal{S}$ (see Fig. 4.1). As $\mathbf{w} = G\mathbf{a}$, for some $\mathbf{a} \in \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^m$, it follows that

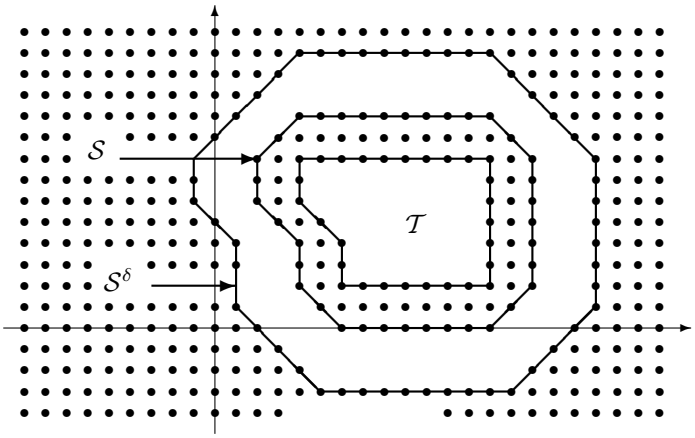


FIG. 4.1.

$$0 = \mathbf{h}^T G\mathbf{a} = \mathbf{h}^T\mathbf{w} = \mathbf{h}^T\mathbf{v} + \mathbf{h}^T\mathbf{r}.$$

This is not possible, however, since $\mathcal{T}$ includes the support of $\mathbf{h}^T\mathbf{v}$ without intersecting $\text{supp}(\mathbf{h}^T\mathbf{r})$. □

PROPOSITION 4.4. *For every behavior $\mathfrak{B} \subseteq \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^p$ there exist an observable behavior $\mathfrak{B}_0$ and a locally undetectable behavior $\mathfrak{B}_{\text{lu}}$ in $\mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^p$ s.t.*

$$\mathfrak{B} = \mathfrak{B}_0 \cap \mathfrak{B}_{\text{lu}}. \tag{4.2}$$

Moreover, $\mathfrak{B}_0$ is uniquely determined as $\mathfrak{B}^{\perp\perp}$, the smallest observable behavior including $\mathfrak{B}$.

Proof. Let $\mathfrak{B} = \text{Im}G$ and $\mathfrak{B}_0 := \mathcal{B}^{\perp\perp} = \ker H^T$. Clearly, $\mathfrak{B}_0$ is an observable behavior including $\mathfrak{B}$. If G has rank r, we can assume, for the sake of simplicity, that its first r rows are linearly independent. So, G can be partitioned as

$$G = \begin{bmatrix} G_1 \\ G_2 \end{bmatrix} \begin{matrix} \}r \\ \}p-r \end{matrix},$$

where G_1 is a full rank matrix. Let

$$G_{\text{lu}} := \begin{bmatrix} G_1 & 0 \\ 0 & I_{p-r} \end{bmatrix}$$

and $\mathfrak{B}_{\text{lu}} := \text{Im}G_{\text{lu}}$. Clearly, $\mathfrak{B}_{\text{lu}}$ is a locally undetectable behavior, and it includes $\mathfrak{B}$ as

$$G = \begin{bmatrix} G_1 \\ G_2 \end{bmatrix} = \begin{bmatrix} G_1 & 0 \\ 0 & I_{p-r} \end{bmatrix} \begin{bmatrix} I \\ G_2 \end{bmatrix}.$$

So, one obviously gets $\mathfrak{B} \subset \mathfrak{B}_0 \cap \mathfrak{B}_{\text{lu}}$.

To prove the reverse inclusion, consider $\mathbf{w} \in \mathfrak{B}_0 \cap \mathcal{B}_{\text{lu}}$. Clearly, $\mathbf{w}$ satisfies $H^T\mathbf{w} = 0$ and can be expressed as

$$\mathbf{w} = \begin{bmatrix} G_1\mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix}.$$

Factoring G into the product of a (full column rank) right factor prime matrix $\bar{G}$ and a full row rank rational matrix Q [17], one gets

$$\begin{bmatrix} G_1 \\ G_2 \end{bmatrix} = G = \bar{G}Q = \begin{bmatrix} \bar{G}_1 \\ \bar{G}_2 \end{bmatrix} Q.$$

As the columns of $\bar{G}$ generate the $\mathbb{F}(\mathbf{z})$-vector space orthogonal to the rows of H^T, there exists $\mathbf{v} \in \mathbb{F}(\mathbf{z})^r$ s.t. $\mathbf{w} = \bar{G}\mathbf{v}$. But then $\bar{G}_1\mathbf{v} = G_1\mathbf{u}_1 = \bar{G}_1 Q\mathbf{u}_1$ implies $\mathbf{v} = Q\mathbf{u}_1$, and thus $\mathbf{u}_2 = \bar{G}_2\mathbf{v} = \bar{G}_2 Q\mathbf{u}_1$ and

$$\mathbf{w} = \begin{bmatrix} G_1\mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix} = \begin{bmatrix} \bar{G}_1 \\ \bar{G}_2 \end{bmatrix} Q\mathbf{u}_1 = G\mathbf{u}_1.$$

This implies that $\mathbf{w}$ is in $\mathfrak{B}$.

It remains to prove the uniqueness of $\mathfrak{B}_0$ in the above representation. To this end we need the following technical lemma.

LEMMA 4.5. *Let $\mathfrak{B}_i \subset \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^p$, $i = 1, 2$, be finite support behaviors with p components. If $\mathfrak{B} = \mathfrak{B}_1 \cap \mathfrak{B}_2$, then*

$$\mathfrak{B}_{\text{rat}} = (\mathfrak{B}_1)_{\text{rat}} \cap (\mathfrak{B}_2)_{\text{rat}}. \tag{4.3}$$

Proof. Let G, G_1, and G_2 be generator matrices of $\mathfrak{B}$, $\mathcal{B}_1$ and $\mathfrak{B}_2$, respectively. Clearly, $\mathfrak{B}_{\text{rat}} = \text{Im}_{\mathbb{F}(\mathbf{z})} G := \{\mathbf{v} \in \mathbb{F}(\mathbf{z})^p : \mathbf{v} = G\mathbf{u}, \mathbf{u} \text{ rational}\}$, and similarly, $(\mathfrak{B}_i)_{\text{rat}} = \text{Im}_{\mathbb{F}(\mathbf{z})} G_i$, $i = 1, 2$. Therefore, $\mathbf{v} \in \mathfrak{B}_{\text{rat}}$ implies $\mathbf{v} = G\mathbf{n}/d$, for some L-polynomial vector $\mathbf{n}$ and some L-polynomial d, and hence $d\mathbf{v} \in \mathfrak{B} = \mathfrak{B}_1 \cap \mathfrak{B}_2$. But then $d\mathbf{v} = G_1\mathbf{u}_1 = G_2\mathbf{u}_2$ for suitable L-polynomial vectors $\mathbf{u}_1$ and $\mathbf{u}_2$, which implies $\mathbf{v} \in (\mathfrak{B}_1)_{\text{rat}} \cap (\mathfrak{B}_2)_{\text{rat}}$.

Conversely, if $\mathbf{v} \in (\mathfrak{B}_1)_{\text{rat}} \cap (\mathfrak{B}_2)_{\text{rat}}$, it can be expressed as $\mathbf{v} = G_1\mathbf{n_1}/d_1 = G_2\mathbf{n_2}/d_2$, for suitable L-polynomial vectors $\mathbf{n}_i$ and L-polynomials d_i, $i = 1, 2$. If $d := \ell\text{.c.m.}\ (d_1, d_2)$, it is clear that $d\mathbf{v}$ is an element of $\mathfrak{B}_1 \cap \mathfrak{B}_2$, and hence of $\mathfrak{B}$. Consequently, $\mathbf{v}$ is in $\mathfrak{B}_{\text{rat}}$. ∎

We now return to the proof of the uniqueness of $\mathfrak{B}_0$. Suppose, by contradiction, that $\mathfrak{B} = \hat{\mathfrak{B}}_0 \cap \hat{\mathfrak{B}_{\text{lu}}}$ for some observable behavior $\hat{\mathfrak{B}}_0 \neq \mathfrak{B}^{\perp\perp}$ and some locally undetectable behavior $\hat{\mathfrak{B}_{\text{lu}}}$. As $\mathfrak{B}^{\perp\perp}$ is the smallest observable behavior including $\mathfrak{B}$ and is maximal in the class of modules of rank r, $\hat{\mathfrak{B}}_0$ must have rank greater than r. Consequently, $(\hat{\mathfrak{B}}_0)_{\text{rat}} \supsetneq (\mathfrak{B}^{\perp\perp})_{\text{rat}}$. On the other hand

$$(\mathfrak{B}_{\text{lu}})_{\text{rat}} = (\hat{\mathfrak{B}_{\text{lu}}})_{\text{rat}} = \mathbb{F}(\mathbf{z})^p,$$

and therefore $(\mathfrak{B}^{\perp\perp})_{\text{rat}} \cap (\mathfrak{B}_{\text{lu}})_{\text{rat}} = (\mathfrak{B}^{\perp\perp})_{\text{rat}} \subsetneq (\hat{\mathfrak{B}}_0)_{\text{rat}} = (\hat{\mathfrak{B}}_0)_{\text{rat}} \cap (\hat{\mathfrak{B}_{\text{lu}}})_{\text{rat}}$. But this is not possible, as $\mathfrak{B}^{\perp\perp} \cap \mathfrak{B}_{\text{lu}} = \mathfrak{B} = \hat{\mathfrak{B}}_0 \cap \hat{\mathfrak{B}_{\text{lu}}}$ should imply, by the above lemma, $(\mathfrak{B}^{\perp\perp})_{\text{rat}} \cap (\mathfrak{B}_{\text{lu}})_{\text{rat}} = (\hat{\mathfrak{B}}_0)_{\text{rat}} \cap (\hat{\mathfrak{B}_{\text{lu}}})_{\text{rat}}$. ∎

5. Input-output description and trajectory generation. The analysis we carried out in the previous sections focused on the properties of behavior trajectories without concern for the way they are generated. Once a behavior $\mathfrak{B}$ is represented via a finite set of generators $\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_m$, however, it is natural to look at $G := [\mathbf{g}_1\ \mathbf{g}_2\ \cdots\ \mathbf{g}_m]$ as a transfer matrix, and hence to consider $\mathfrak{B}$ as the image of an input-output map acting on $\mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^m$. This point of view seems particularly appropriate when $\mathfrak{B}$ is a convolutional code [4, 16], as it is customary to regard it as the result of an encoding process, and, consequently, its trajectories (codewords)

as the outputs of a dynamical encoder. In a wider context, the trajectories of $\mathfrak{B}$ are obtained from certain processing operations applied to multidimensional data, or from different transformations (desired or not) performed on the original signal. In both cases the analysis of the algebraic properties of the generator matrices makes possible a detailed knowledge of the behavior structure.

When an input-output description is adopted, it is often imperative to associate trajectories of $\mathfrak{B}$ and input sequences bijectively. In data transmission the meaning of this requirement is clear, as input signals represent information messages to be retrieved from the received codewords, and an unambiguous decision at the decoding stage is possible when each codeword encodes a unique information sequence. This amounts to saying that the *encoder* G induces a 1-1 map.

Throughout this section we steadily assume that $\mathfrak{B}$ has a full column rank generator matrix G, and hence is free. Under this assumption, G admits (possibly infinitely many) rational left inverses G^{-1}. Each of them, when applied to a behavior trajectory $\mathbf{w} = G\mathbf{u}$, uniquely retrieves the (finite) input sequence $\mathbf{u}$. When $\mathfrak{B}$ represents a finite convolutional code, this implies that every estimate $\hat{\mathbf{w}} \in \mathfrak{B}$ of the codeword $\mathbf{w}$ produces a finite error $\mathbf{e_u} := \mathbf{u} - G^{-1}\hat{\mathbf{w}} = G^{-1}(\mathbf{w} - \hat{\mathbf{w}})$ in reconstructing the information sequence $\mathbf{u}$. Consequently, when a codeword estimator is available, no catastrophic error can arise [4, 6]. However, if we apply the "decoder" G^{-1} directly to the noisy sequence $\mathbf{v} = \mathbf{w} + \mathbf{r}$, as $\mathbf{r}$ generally is not an element of $\mathfrak{B}$, the decoding algorithm possibly gives an infinite support sequence, which differs from the correct input in infinitely many points and clearly is not even an admissible information sequence. This drawback can be avoided if and only if G^{-1} is an L-polynomial matrix.

Proposition 5.1 below provides equivalent conditions for the existence of an L-polynomial inverse, and in particular shows that such an inverse exists if and only if G is left zero prime.

DEFINITION 5.1. *Let G be in $\mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^{p\times m}$ and $\hat{G} = \mathbf{z^h}G = z_1^{h_1}\cdots z_n^{h_n}G$ in $\mathbb{F}[\mathbf{z}]^{p\times m}$ for some $\mathbf{h} \in \mathbb{N}^n$. If $\mathbb{K}$ denotes the algebraic closure of $\mathbb{F}$, the* L-variety $\mathcal{V}^L(G)$ of the maximal order minors of G *is the algebraic set*

$$\mathcal{V}^L(G) := \mathcal{V}(\hat{G}) \setminus \left\{(k_1, k_2, \ldots, k_n) : k_i \in \mathbb{K},\ \prod_i k_i = 0\right\}, \tag{5.1}$$

where $\mathcal{V}(\hat{G})$ denotes the variety (in $\mathbb{K}$) of the maximal order minors of $\hat{G}$.

The above definition is well posed, as (5.1) does not depend on the choice of $\hat{G}$.

PROPOSITION 5.2. *Let G be a $p \times m$ L-polynomial matrix. The following statements are equivalent:*

i) *G is rzp;*

ii) *there exists an L-polynomial matrix $P \in \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^{m\times p}$ such that $PG = I_m$;*

iii) *$\mathcal{V}^L(G)$ is empty;*

iv) *Im $G^T = \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^m$.*

Proof. i) $\Rightarrow$ ii) Let $m_i(G)$ denote the ith maximal order minor of G, $i = 1, 2, \ldots, \binom{p}{m}$. By the zero primeness assumption, there exist L-polynomials $h_i \in \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]$ such that $\sum_i h_i m_i(G) = 1$. If S_i is the $m \times p$ matrix which selects in G the m rows corresponding to $m_i(G)$, from $I_m = \sum_i h_i m_i(G) I_m = \sum_i h_i\big(\mathrm{adj}(S_iG)\big)(S_iG)$, we find that $P := \sum_i h_i\big(\mathrm{adj}(S_iG)\big)S_i$ is a left inverse of G with elements in $\mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]$.

ii) $\Rightarrow$ iii) Let $\mathbf{z^r} = z_1^{r_1}\cdots z_n^{r_n}$ be a suitable term such that $\hat{P} = \mathbf{z^r}P$ is in $\mathbb{F}[\mathbf{z}]^{m\times p}$. By applying the Binet–Cauchy formula [9] to equation $\hat{P}\hat{G} = z_1^{h_1+r_1}\cdots z_n^{h_n+r_n} I_m$, we

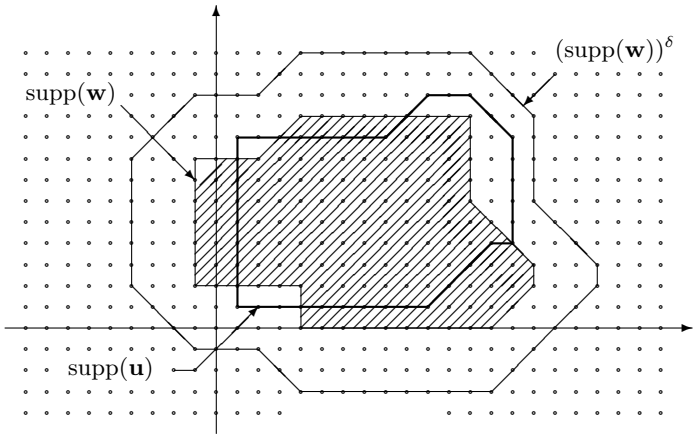


FIG. 5.1.

get

$$\sum_i m_i(\hat{P}) m_i(\hat{G}) = z_1^{m(h_1+r_1)} z_2^{m(h_2+r_2)} \cdots z_n^{m(h_n+r_n)}, \tag{5.2}$$

where $m_i(\hat{P})$ and $m_i(\hat{G})$ are corresponding maximal order minors of $\hat{P}$ and $\hat{G}$, respectively. Then $\mathcal{V}(\hat{G})$ is included in the variety of $z_1^{m(h_1+r_1)} z_2^{m(h_2+r_2)} \cdots z_n^{m(h_n+r_n)}$, which is a subset of $\mathcal{K} := \{(k_1, k_2, \ldots, k_n) : k_i \in \mathbb{K},\ \prod_i k_i = 0\}$.

iii) $\Rightarrow$ i) As $\mathcal{K}$ is the variety of $\mathbf{z} = z_1 \cdots z_n$, by assumption iii), $\mathcal{V}(\hat{G})$ is included in the variety of $\mathbf{z}$. So, by Hilbert's Nullstellensatz [11], an integer $r > 0$ exists such that $z_1^r \cdots z_n^r$ belongs to the ideal generated in $\mathbb{F}[\mathbf{z}]$ by the maximal order minors of $\hat{G}$:

$$z_1^r \cdots z_n^r = \sum_i \bar{h}_i\ m_i(\hat{G}), \qquad \bar{h}_i \in \mathbb{F}[\mathbf{z}]. \tag{5.3}$$

As each maximal order minor $m_i(\hat{G})$ differs from $m_i(G)$ in a unit of $\mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]$, the zero primeness of G easily follows after dividing both members of (5.3) by $z_1^r \cdots z_n^r$.

ii) $\Leftrightarrow$ iv) This is obvious. □

When the generator matrix G has an L-polynomial inverse, a uniform bound can be found on the support of the input sequences which correspond to the behavior trajectories. Actually, if P is such an inverse, $\mathbf{w} \in \mathfrak{B}$ is generated by the input signal $\mathbf{u} = P\mathbf{w}$ whose support cannot exceed "too much" that of $\mathbf{w}$ (see Fig. 5.1). This feature, which we will refer to as the *wrapping input property*, is quite appealing, as the mere recognition of the support of a trajectory allows the derivation of a uniformly tight bound on the support of the corresponding input sequence. In particular, in the context of finite convolutional codes, the above property guarantees that small errors in the codeword estimate reflect into small errors in the information sequence reconstruction.

(WI) [Wrapping input property]. *A finite behavior* $\mathfrak{B}$ *has the* wrapping input property *if there exist a full column rank generator matrix* G *and a positive integer* δ *such that* $\mathbf{w} = G\mathbf{u}$ *implies*

$$\operatorname{supp}(\mathbf{u}) \subseteq \big(\operatorname{supp}(\mathbf{w})\big)^{\delta}. \tag{5.4}$$

It is worthwhile to notice that property (WI) does not depend on the particular full column rank generator matrix of $\mathfrak{B}$ we are considering. In fact, it is easily seen that if (5.4) holds for any one of these generator matrices, then it holds for all of them (in general, for a different δ). On the other hand, when noninjective generator matrices of $\mathfrak{B}$ are considered and the uniqueness of the input sequence producing a given trajectory is lost, a particular input can be found whose support satisfies (5.4), as shown by the following proposition.

PROPOSITION 5.3. *Assume that $\mathfrak{B}$ has the (WI) property for some full column rank matrix G and some integer $\delta > 0$. Then, for every generator matrix $\bar{G} \in \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^{p\times q}$, an integer $\bar{\delta} > 0$ can be found s.t. each trajectory $\mathbf{w} \in \mathfrak{B}$ can be expressed as $\mathbf{w} = \bar{G}\bar{\mathbf{u}}$ for some input $\bar{\mathbf{u}}$ with* $\operatorname{supp}(\bar{\mathbf{u}}) \subseteq \big(\operatorname{supp}(\mathbf{w})\big)^{\bar{\delta}}$.

Proof. Since G and $\bar{G}$ are generator matrices of the same behavior, there exists a full column rank L-polynomial matrix Q such that $G = \bar{G}Q$. Let τ be the radius of a ball, with center in the origin, including $\operatorname{supp}(Q)$, and consider $\mathbf{w} \in \mathfrak{B}$. By property (WI), there is $\mathbf{u}$ such that $\mathbf{w} = G\mathbf{u}$ and $\operatorname{supp}(\mathbf{u}) \subseteq \big(\operatorname{supp}(\mathbf{w})\big)^{\delta}$. So, $\bar{\mathbf{u}} := Q\mathbf{u}$ satisfies $\mathbf{w} = G\mathbf{u} = \bar{G}Q\mathbf{u} = G\bar{\mathbf{u}}$, and $\operatorname{supp}(\bar{\mathbf{u}}) = \operatorname{supp}(Q\mathbf{u}) \subseteq \big(\operatorname{supp}(\mathbf{u})\big)^{\tau} \subseteq \big(\operatorname{supp}(\mathbf{w})\big)^{\tau+\delta}$. Consequently, the proposition holds for $\bar{\delta} = \tau + \delta$. □

Interestingly enough, the zero primeness of G is not only sufficient but also necessary for property (WI). So, free behaviors satisfying property (WI) can be identified with behaviors that are generated by ℓ zero prime matrices.

PROPOSITION 5.4. *A finite behavior $\mathfrak{B}$ has the (WI) property if and only if it admits an rzp generator matrix.*

Proof. The "if" part has already been proved. To show the converse, we need the following characterization of rZP matrices.

LEMMA 5.5. *Let $G \in \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^{p\times q}$ be a Laurent polynomial matrix and denote by $\mathbb{F}[[\mathbf{z}, \mathbf{z}^{-1}]]$ the space of bilateral scalar formal power series in the indeterminates $z_1, \ldots, z_n$. Then G is rzp if and only if*

$$G\mathbf{s} = \mathbf{0} \tag{5.5}$$

for some sequence $\mathbf{s} \in \mathbb{F}[[\mathbf{z}, \mathbf{z}^{-1}]]^q$ implies $\mathbf{s} = \mathbf{0}$.

Proof. Introduce in $\mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^q \times \mathbb{F}[[\mathbf{z}, \mathbf{z}^{-1}]]^q$ the following nondegenerate bilinear form

$$\langle \cdot, \cdot \rangle_q : \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^q \times \mathbb{F}[[\mathbf{z}, \mathbf{z}^{-1}]]^q \to \mathbb{F}$$

defined by : $\langle \mathbf{u}, \mathbf{v} \rangle_q = (\mathbf{u}\mathbf{v}^T, 1) = \sum_{\mathbf{h}\in\mathbb{Z}^n} u(\mathbf{h})v^T(-\mathbf{h})$.

With this position, the space $\mathbb{F}[[\mathbf{z}, \mathbf{z}^{-1}]]^q$ is naturally viewed as $L(\mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^q)$, the algebraic dual of $\mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^q$ [10, 16]. In fact, we can associate with every $\mathbf{v} \in \mathbb{F}[[\mathbf{z}, \mathbf{z}^{-1}]]^q$ the linear functional on $\mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^q$ defined by

$$f_{\mathbf{v}}(\cdot) = \langle \cdot, \mathbf{v} \rangle_q, \tag{5.6}$$

and, conversely, every linear functional on $\mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^q$ can be represented as in (5.6) for an appropriate choice of $\mathbf{v} \in \mathbb{F}[[\mathbf{z}, \mathbf{z}^{-1}]]^m$. Upon identifying $\mathbb{F}[[\mathbf{z}, \mathbf{z}^{-1}]]^q$ with $L(\mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^q)$, we can resort to the well-known relation

$$\begin{aligned} \ker_\infty G &:= \{\mathbf{s} \in \mathbb{F}[[\mathbf{z}, \mathbf{z}^{-1}]]^q : G\mathbf{s} = \mathbf{0}\} \\ &\equiv \{\mathbf{s} \in \mathbb{F}[[\mathbf{z}, \mathbf{z}^{-1}]]^q : \mathbf{s}^T\mathbf{v} = \mathbf{0}\ \forall \mathbf{v} \in \operatorname{Im}G^T\} =: (\operatorname{Im}G^T)^{\perp\!\!\!\perp}. \end{aligned}$$

If $\operatorname{Im}G^T = \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^q$, all canonical vectors $\mathbf{e}_i$ and all monomial vectors $z_1^h z_2^k \mathbf{e}_i$ belong to $\operatorname{Im}G^T$, and therefore $\mathbf{s} \in (\operatorname{Im}G^T)^{\perp\!\!\!\perp}$ implies $\langle z_1^h z_2^k \mathbf{e}_i, \mathbf{s} \rangle_q = 0$, $h, k \in \mathbb{Z}$, and $i =$

$1, 2, \ldots, q$, and hence $\mathbf{s} = 0$. So, it is clear that $\ker_\infty G = \{\mathbf{0}\}$ if and only if $\mathrm{Im}G^T = \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^q$, and this happens if and only if G is rZP. □

Suppose, now, that $\mathfrak{B}$ has the (WI) property with respect to some positive integer δ and some full column rank generator matrix G. We aim to prove that G is rZP. If not, there would be a sequence $\mathbf{s} \in \mathbb{F}[[\mathbf{z}, \mathbf{z}^{-1}]]^q$ satisfying (5.5). Let η be the radius of a ball, $B(\mathbf{0}, \eta)$, centered in the origin and including $\mathrm{supp}(G)$. If $\mathbf{k}$ is an element of $\mathrm{supp}(\mathbf{s})$, the finite support sequence

$$\mathbf{u}(\mathbf{h}) := \begin{cases} \mathbf{s}(\mathbf{h}), & \mathbf{h} \in B(\mathbf{k}, 2\delta + \eta), \\ \mathbf{0} & \text{elsewhere} \end{cases}$$

generates a behavior sequence $\mathbf{w} := G\mathbf{u}$ that does not fulfill (5.4). □

The (WI) property introduces very severe constraints on the supports of the input sequences which produce the behavior trajectories. So, it is not unexpected that it reflects into the strongest primeness property a generator matrix can be endowed with, i.e., zero primeness. Obviously, weaker requirements on the supports of the generating sequences correspond to weaker primeness properties of G. In particular, minor primeness guarantees that the signal producing a behavior sequence $\mathbf{w}$ exhibits a support which slightly exceeds a parallelepipedal box including $\mathrm{supp}(\mathbf{w})$, whereas variety primeness ensures that each projection of $\mathbf{u}$ and $\mathbf{w}$ onto a coordinate hyperplane gives a pair of signals with the (WI) property.

A standpoint which proves to be quite fruitful in analyzing the above-mentioned connections is to regard an arbitrary finite support sequence $\mathbf{w} \in \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^p$ as a vector with entries in certain L-polynomial rings that properly include $\mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]$. Actually, $\mathbf{w}$ can be thought of as an element of $\mathbb{F}(z_i^c)[z_i, z_i^{-1}] := \mathbb{F}(z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n)[z_i, z_i^{-1}]$,

$$\mathbf{w} = \sum_{h_i \in \mathbb{Z}} \mathbf{w}_{h_i}(z_i^c) z_i^{h_i},$$

or as an element of $\mathbb{F}(z_i)[z_i^c, (z_i^c)^{-1}] := \mathbb{F}(z_i)[z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n, z_1^{-1}, \ldots, z_{i-1}^{-1}, z_{i+1}^{-1}, \ldots, z_n^{-1}]$

$$\mathbf{w} = \sum_{h_1, \ldots, h_{i-1}, h_{i+1}, \ldots, h_n \in \mathbb{Z}} \mathbf{w}_{h_1, \ldots, h_{i-1}, h_{i+1}, \ldots, h_n}(z_i) z_1^{h_1} \cdots z_{i-1}^{h_{i-1}} z_{i+1}^{h_{i+1}} \cdots z_n^{h_n}.$$

Correspondingly, we are led to introduce the following support sets:

$$\mathrm{supp}_i(\mathbf{w}) := \{(h_1, \ldots, h_n) \in \mathbb{Z}^n : \mathbf{w}_{h_i}(z_i^c) \neq 0\},$$
$$\mathrm{supp}_{i^c}(\mathbf{w}) := \{(h_1, \ldots, h_n) \in \mathbb{Z}^n : \mathbf{w}_{h_1, \ldots, h_{i-1}, h_{i+1}, \ldots, h_n}(z_i) \neq 0\}.$$

LEMMA 5.6 (see [17]). *Let $G \in \mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]^{p \times m}$ be a full column rank matrix. Then*

i) *G is rMP if and only if G is right (zero) prime in $\mathbb{F}(z_i^c)[z_i, z_i^{-1}]$ for every $i = 1, 2, \ldots, n$;*

ii) *G is rVP if and only if G is rZP in $\mathbb{F}(z_i)[z_i^c, (z_i^c)^{-1}]$ for every $i = 1, 2, \ldots, n$.*

PROPOSITION 5.7. *Let $\mathfrak{B}$ be a finite behavior. Then*

i) *$\mathfrak{B}$ has a rMP generator matrix if and only if there exist an integer $\delta > 0$ and a full column rank generator matrix G, such that $\mathbf{w} \in \mathfrak{B}$ implies $\mathbf{w} = G\mathbf{u}$ with*

$$\mathrm{supp}(\mathbf{u}) \subseteq \bigcap_{i=1}^{n} \Big(\mathrm{supp}_i(\mathbf{w})\Big)^{\delta}; \tag{5.7}$$

ii) $\mathfrak{B}$ *has a rVP generator matrix if and only if there exist an integer* $\delta > 0$ *and a full column rank generator matrix* G*, such that* $\mathbf{w} \in \mathfrak{B}$ *implies* $\mathbf{w} = G\mathbf{u}$ *with*

$$\operatorname{supp}(\mathbf{u}) \subseteq \bigcap_{i=1}^{n} \Big(\operatorname{supp}_{i^c}(\mathbf{w})\Big)^{\delta}. \tag{5.8}$$

Proof. i) It is easy to realize that condition (5.7) is equivalent to the set of conditions $\operatorname{supp}_i(\mathbf{u}) \subseteq (\operatorname{supp}_i(\mathbf{w}))^{\delta}$, $i = 1, 2, \ldots, n$. These hold true if and only if G is an rZP matrix in $\mathbb{F}(z_i^c)[z_i, z_i^{-1}]$ for every $i = 1, 2, \ldots, n$; namely, G is rMP in $\mathbb{F}[\mathbf{z}, \mathbf{z}^{-1}]$.

ii) The result is shown along the same lines as i), after replacing $\mathbb{F}(z_i^c)[z_i, z_i^{-1}]$ with $\mathbb{F}(z_i)[z_i^c, (z_i^c)^{-1}]$. □

6. Conclusions. In this paper we have focused on some features of finite support nD behaviors which are relevant for multidimensional signal generation and recognition. Two opposite situations have been considered, namely the case when a local testing procedure suffices to decide whether a given signal belongs to the behavior, and the case when every finite signal can be completed into a legal trajectory, and hence behavior sequences cannot be recognized by means of local checks.

Observable and locally undetectable behaviors, which correspond to these two situations, have been characterized in terms of both their internal properties and their polynomial matrix descriptions. Any finite support behavior, being the intersection of an observable and an unconstrained behavior, exhibits intermediate properties.

Finally, adopting an input-output point of view, the connections between the support of a behavior trajectory and that of its generating input have been enlightened.

Acknowledgments. The authors are indebted to S. Zampieri for shortening the proof of Proposition 5.3.

REFERENCES

[1] T. Becker and V. Weispfenning, *Gröbner Bases*, Springer-Verlag, New York, 1993.
[2] N. K. Bose, *Multidimensional Systems Theory*, D. Reidel, Boston, MA, 1985.
[3] F. Fagnani and S. Zampieri, *Matrix shifts operators and controllable systems over principal ideal domains*, Sc. Norm. Sup. Pisa, preprints di Matematica, 36 (1994), pp. 1–25.
[4] E. Fornasini and M. E. Valcher, *Algebraic aspects of* 2*D convolutional codes*, IEEE Trans. Inform. Theory, 40 (1994), pp. 1068–1082.
[5] E. Fornasini and M. E. Valcher, *On some primeness properties in the analysis of* n*D finite support behaviors*, in Proc. IMACS Intern. Symp. on Sign. Proc., Robotics, and Neural Networks, Lille, France, 1994, pp. 89–92.
[6] G. D. Forney, *Convolutional codes* I*: Algebraic structure*, IEEE Trans. Inform. Theory, 16 (1970), pp. 720–738.
[7] G. D. Forney, *Minimal bases of rational vector spaces, with applications to multivariable linear systems*, SIAM J. Control, 13 (1975), pp. 493–521.
[8] G. D. Forney and M. D. Trott, *The dynamics of group codes: State spaces, trellis diagrams and canonical encoders*, IEEE Trans. Inform. Theory, 39 (1993), pp. 1491–1513.
[9] F. R. Gantmacher, *Theory of Matrices*, Chelsea, New York, 1959.
[10] W. Greub, *Linear Algebra*, Springer-Verlag, New York, 1975.
[11] S. Lang, *Algebra*, Addison-Wesley, Menlo Park, CA, 1993.
[12] H. A. Loeliger, G. D. Forney, T. Mittelholzer, and M. D. Trott, *Minimality and observability of group systems*, Linear Algebra Appl., 205–206 (1994), pp. 937–963.
[13] P. Rocha, *Structure and Representation of* 2*D Systems*, Ph.D. thesis, Rijksuniversiteit Groningen, the Netherlands, 1990.
[14] P. Rocha and J. C. Willems, *Controllability of* 2*D systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 413–423.

[15] S. Sakata, *On determining the independent point set for doubly periodic arrays and encoding two-dimensional cyclic codes and their duals*, IEEE Trans. Inform. Theory, 27 (1981), pp. 556–565.

[16] M. E. Valcher and E. Fornasini, *On* 2*D finite support convolutional codes: An algebraic approach*, Multidimens. Systems Signal Process., 5 (1994), pp. 231–243.

[17] M. E. Valcher, *Modellistica ed Analisi dei Sistemi* 2*D con Applicazioni alla Codifica Convoluzionale*, Ph.D. thesis, University of Padova, I, Italy, 1995.

[18] J. C. Willems, *From time series to linear systems, part* I*: Finite dimensional linear time invariant systems*, Automatica, 22 (1986), pp. 561–580.

[19] J. C. Willems, *Models for Dynamics*, in Dynamics Reported, Vol. 2, U. Kirchgraber and H. O. Walther, eds., J. Wiley and B. G. Teubner, New York, Leipzig, 1989, pp. 171–269.

[20] J. C. Willems, *Paradigms and puzzles in the theory of dynamical systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 259–294.

[21] D. C. Youla and G. Gnavi, *Notes on n-dimensional system theory*, IEEE Trans. Circuits Systems, 26 (1979), pp. 105–111.

[22] S. Zampieri and S. K. Mitter, *Linear systems over Noetherian rings in the behavioral approach*, J. Math. Systems, Estim. Control, 6 (1996), pp. 235–238.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 2, pp. 780–796, March 1998

015

STABILITY AND EULER APPROXIMATION OF ONE-SIDED LIPSCHITZ DIFFERENTIAL INCLUSIONS*

TZANKO DONCHEV† AND ELZA FARKHI‡

Abstract. Ordinary differential and functional-differential inclusions with compact right-hand sides are considered. Stability theorems of Filippov's type in the convex and nonconvex case are proved under a one-sided Lipschitz condition, which extends the notions of Lipschitz continuity, dissipativity, and the uniform one-sided Lipschitz condition for set-valued mappings. The accuracy of approximation of the solution sets by means of the Euler discretization scheme for both types of inclusions is estimated.

Key words. one-sided Lipschitz continuity, differential inclusions, Filippov theorem, Euler method

AMS subject classifications. 34A60, 49J24, 93B40

PII. S0363012995293694

1. Introduction.

Let I be the interval [0,1], $X = \mathbf{R}^n$, and K_0 be a compact subset of X. We consider a multifunction F from $I \times X$ to the set of all convex compact subsets of X. We study the following initial value problem:

$$\dot{x}(t) \in F(t, x(t)) \quad \text{for a.e. } t \in I, \quad x(0) \in K_0 \ , \tag{1.1}$$

where $x(\cdot)$ is an absolutely continuous (AC) vector function from I to X.

Stability properties of the solutions of differential equations and inclusions with respect to various perturbations are of great importance for their qualitative and quantitative analysis. They are closely related to existence and relaxations theory, on the one hand, and to sensitivity and approximations analysis, on the other.

In this sense the stability theorem, published by Filippov [11], and sometimes referred to as the Gronwall–Filippov–Ważewski theorem (see [3] for a more general setting), takes a central place in the theory of differential inclusions. It gives an estimate of sensitivity of the reachable set of (1.1) with respect to perturbations in the initial condition and the right-hand side. The assumptions, for which the theorem was originally proven in [11], are

1°. $F(\cdot, \cdot)$ is Hausdorff continuous with nonempty closed values.

2°. $F(t, \cdot)$ is Lipschitz continuous; i.e., there is a nonnegative integrable function $L : I \to \mathbf{R}$ such that

$$haus(F(t, x'), F(t, x'')) \le L(t)|x' - x''| \quad \text{for any} \quad x', x'' \quad \text{and a.e.} \quad t \in I.$$

FILLIPPOV'S THEOREM [11]. *Let $y : I \to X$ be absolutely continuous and $dist(\dot{y}(t),\ F(t, y(t))) \le g(t)$ for a.e. $t \in I$, where $g(\cdot)$ is summable. Let $1°, 2°$ be satisfied in the region $\{(t, x) : t \in I, |x - y(t)| \le b\}$, $K_0 = \{x_0\}$ be such that $|x_0 - y(0)| < b$,*

*Received by the editors October 25, 1995; accepted for publication (in revised form) February 8, 1997. The work of this author was partially supported by the National Fund for Scientific Research at the Bulgarian Ministry of Science, Education, and Technology under contract MM-408/94.
http://www.siam.org/journals/sicon/36-2/29369.html

†Department of Mathematics, University of Mining and Geology, 1100 Sofia, Bulgaria (donchev@staff.mgu.bg).

‡School of Mathematical Sciences, Sackler Faculty of Exact Sciences, Tel Aviv University, 69978 Tel Aviv, Israel (elza@math.tau.ac.il). On leave from Institute of Mathematics, Bulgarian Academy of Sciences.

and

$$m(t) = \int_0^t L(r)dr, \quad v(t) = e^{m(t)} \left(|x_0 - y(0)| + \int_0^t e^{-m(s)} g(s)ds \right).$$

Then there is a solution $x(\cdot)$ *of* (1.1) *on the interval* $\Delta = [t : t \in I, v(t) \leq b]$ *that satisfies*

$$|x(t) - y(t)| \leq v(t), \quad \textit{for all } t \in \Delta, \quad x(0) = x_0,$$

$$|\dot{x}(t) - \dot{y}(t)| \leq L(t)v(t) + g(t) \quad \textit{for a.e. } t \in \Delta.$$

It is worth stressing the wide range of applications of this theorem. It is a basic tool in studying the relations between the original and the convexified problem (or so-called "relaxed problem") (cf. [2], [5]), implying density of the solution set of (1.1) in the solution set of the relaxed problem. Naturally, it is very helpful in perturbations analysis and studying various discrete approximations (see, e.g., [7], [22]).

We should mention that this work was inspired by Veliov's paper [22], where a refinement of the above theorem had been proved and applied to singularly perturbed differential inclusions.

The aim of this paper is to weaken the continuity assumptions of the theorem in a manner that admits inference of the same estimate for the trajectories. The Lipschitz continuity (LC) of $F(t, \cdot)$ is replaced by a one-sided Lipschitz (OSL) condition and upper semicontinuity. Convexity of the right-hand side compensates for the lack of continuity of $F(t, \cdot)$ and ensures existence of a solution to (1.1). A growth condition appears to bound the derivatives, which are not necessarily bounded anymore. One should note that for the OSL right-hand side the estimate holds true for the state variables, but not for the velocities.

The OSL condition we use is introduced in [8], [10] and naturally generalizes the existing notions of LC, of dissipativity (or monotonicity; cf., e.g., [2], [5]), and the (uniform) OSL condition [15], [7] of set-valued maps. It does not, however, imply continuity of $F(t, \cdot)$, as the Lipschitz condition, nor uniqueness of the trajectory starting from a given point, as the latter (uniform) OSL condition.

Nevertheless, it assures stability of the attainable set, which is important for the sensitivity and approximations analysis of (1.1). This gives a positive answer to Artstein's question [1] of whether there is some general condition other than the Lipschitz one leading to stability of the attainable set.

We give an implementation of the above-mentioned theorem to obtain error estimates for Euler discrete approximation of differential inclusions.

Let us now formulate the Euler approximation for (1.1).

For any natural N divide I into N equal subintervals by $t_i = ih$, $i = 0, 1, \ldots, N$, where $h = \frac{1}{N}$. Approximate the solutions of (1.1) by piecewise linear functions $x : I \to X$ satisfying

$$\text{(1.2)} \qquad \begin{cases} x(t) = x(t_i) + (t - t_i)f_i \quad \text{for all} \quad t \in [t_i, t_{i+1}], \quad x(0) \in K_0, \\ \quad \text{where} \quad f_i \in F(t_i, x(t_i)), \quad i = 0, 1, \ldots, N-1. \end{cases}$$

This discretization has been studied by different authors (see, e.g., the survey [7]). We note here papers [6], [19], where an error of order $\mathcal{O}(h)$ is obtained for Lipschitz continuous $F(t, \cdot)$; Artstein's paper [1], containing error estimates in terms of the modulus of continuity of $F(\cdot, \cdot)$ provided it is uniformly continuous and the

attainable set satisfies a stability condition; and the result of Lempio [18], achieving rate of convergence $\mathcal{O}(h)$ for a "strengthened one-sided Lipschitz continuous" $F(t,\cdot)$, which may be discontinuous with respect to the state variable.

Our estimate involves moduli of continuity of the right-hand side that are uniform with respect to the state variable and averaged with respect to the time. It implies an error of order $\mathcal{O}(h^\alpha)$ for $F(t,\cdot)$ Hölder continuous of degree α and with uniformly bounded $1/\alpha$-variation with respect to the time variable.

A notion of OSL continuity for the following system with time lag is defined:

$$\dot{x}(t) \in F(t, x_t), \quad x_0 = \varphi, \tag{1.3}$$

where x_t, φ belong to the space $E = C([-\theta, 0], X)$ of the continuous functions from the interval $[-\theta, 0]$ to X, $x_t : [-\theta, 0] \to X$ is defined by $x_t(s) = x(t+s)$, and F is a set-valued function defined for every $t \in I$ and $x_t \in E$. Assuming that $F(t,x)$ is nonconvex-valued, Hausdorff continuous in x, and OSL, an "asymptotic"-type stability theorem is proved for (1.3): for every $\varepsilon > 0$ there exists a solution of (1.3), which satisfies Filippov's inequality for the trajectories with a defect ε. This result holds true for nonconvex-valued ordinary differential inclusion (1.1) as a special case of (1.3). Error estimates for the Euler approximation of the system (1.3) are also obtained.

The paper is organized as follows: section 2 is devoted to the OSL condition. Comparisons with the existing notions and some examples are given there. The main theorem for OSL inclusion (1.1) is proven in section 3. The Euler discrete approximations of (1.1) are studied in section 4. All the respective results for the functional-differential inclusion (1.3) are grouped in section 5. Section 6 contains some additional examples with applications of the OSL constants.

2. OSL continuous multifunctions. Let $X = \mathbf{R}^n$ and denote by $\langle \cdot, \cdot \rangle$ the inner product, and by $|\cdot|$ the norm in X, and let U be the unit ball and S the unit sphere in X.

Let $\mathcal{K}$ be the set of all nonempty compacts and $\mathcal{KK}$ the family of all convex compacts in X. Let $A, B \in \mathcal{K}$ and $x \in X$. The distance from x to A is given by $dist(x, A) = \inf\{|x - a| : a \in A\}$, the one-sided excess of A from B is

$$ex(A, B) = \inf\{\alpha > 0 : A \subset B + \alpha U\},$$

and Hausdorff metrics in $\mathcal{K}$ are defined as

$$haus(A, B) = \max\{ex(A, B), ex(B, A)\}.$$

Endowed with the Hausdorff distance, $\mathcal{K}$ becomes a complete separable metric space. We denote $|A| = \sup\{|a| : a \in A\}$. Denote the support function of the set $A \in \mathcal{K}$ by

$$\sigma(x, A) = \max\{\langle x, a\rangle : a \in A\}.$$

Note that σ can be defined for nonconvex sets. For convex A, $\sigma(\cdot, A)$ uniquely determines A. Remember that for $A, B \in \mathcal{KK}$

$$haus(A, B) = \max\{|\sigma(e, A) - \sigma(e, B)| : e \in S\}.$$

Let Y be a normed space. A multifunction $F : Y \to \mathcal{K}$ is called upper-semicontinuous (USC) [resp., lower-semicontinuous (LSC)] in a point $x_o \in Y$, if for every $\varepsilon > 0$

there exists $\delta > 0$ such that $F(x) \subset F(x_0) + \varepsilon U$ (resp., $F(x_0) \subset F(x) + \varepsilon U$) $\forall x \in x_0 + \delta U$. It is called continuous in x_o if it is USC and LSC in x_o, i.e., continuous with respect to the Hausdorff distance. $F : I \times Y \to \mathcal{K}$ is called almost USC (resp., almost LSC, or almost continuous) if for every $\varepsilon > 0$ there is a compact $I_\varepsilon \subset I$ with $\mathrm{meas}(I \backslash I_\varepsilon) < \varepsilon$ such that F is USC (LSC, continuous) on $I_\varepsilon \times Y$.

DEFINITION 2.1. *The set-valued mapping $F : I \times X \to \mathcal{K}$ is called one-sided Lipschitz* (OSL) *continuous (with respect to x) if there is an integrable function $L : I \to \mathbf{R}$ such that for every $x, y \in X$, $t \in I$, and $v \in F(t, x)$, there exists $w \in F(t, y)$ such that*

$$\langle x - y, v - w \rangle \le L(t)|x - y|^2.$$

Let us give an equivalent definition in terms of the support function: there exists integrable $L(\cdot)$ such that for any $x, y \in X$ and $t \in I$

$$\sigma(x - y, F(t, x)) - \sigma(x - y, F(t, y)) \le L(t)|x - y|^2.$$

In other words, there is an integrable function $L(\cdot)$ such that for each direction $e \in S$ and every $x \in X$, the scalar function $\phi : I \times \mathbf{R} \to \mathbf{R}$, defined by $\phi(t, s) = \sigma(e, F(t, x + se))$, is OSL continuous, i.e., satisfies the inequality

$$(\phi(t, s) - \phi(t, r))(s - r) \le L(t)(s - r)^2$$

for every $s, r \in \mathbf{R}$ and $t \in I$.

Note that an OSL mapping is characterized by its support function in each direction, i.e., only by the behavior of its support faces. Besides, for Lipschitz convex-valued maps the absolute value of the support functions difference is bounded, while for OSL mappings the bound is one-sided and the support function may decrease in an arbitrary way in some directions.

Remark 2.1. It is easy to prove that F is OSL iff $\overline{co}F$ is OSL.

Remark 2.2. As follows from Definition 2.1, the OSL condition extends the notion of OSL continuity for vector-valued functions (cf. [4]) and also the notions of Lipschitz continuity, dissipativity (monotonicity; cf. [2], [5]), and (uniform) OSL continuity for set-valued mappings [15], [7]. Remember that $G : X \to \mathcal{K}$ is dissipative if $\langle x - y, v - w \rangle \le 0$ for all $x, y \in X$ and $v \in F(x)$, $w \in F(y)$. Then the map $-G$ is called monotone. The dissipativity notion was weakened by Kastner-Maresch [15], replacing the zero in the definition by $L|x - y|^2$. Such mappings were called (uniformly) one-sided Lipschitz (UOSL) continuous in [15], [7]. By simple use of Gronwall's inequality one can see that UOSL inclusions have at most one trajectory starting from a given point [15]. The uniqueness of the trajectory is the main difference between UOSL and OSL inclusions. An OSL system may have many solutions. Error estimates for implicit Runge–Kutta approximations of UOSL inclusions are presented in [15], provided the unique trajectory possesses suitable smoothness properties. Lempio [18] introduced a strengthened UOSL condition that ensures an $O(h)$ rate of convergence for Euler's method, even for right-hand side, discontinuous in the state variable.

The following simple examples show that the OSL continuity is an essential extension of LC, the uniform OSL continuity, and the dissipativity.

Example 2.1 (discontinuous OSL interval function which is neither LC nor UOSL).

$$F(x) = \begin{cases} [0, 1] & \text{for } x < 0, \\ [-1, 1] & \text{for } x \ge 0. \end{cases}$$

Example 2.2. Consider the following control system:

$$\dot{x}(t) = f(t, x, u),\ x(0) = x_0;\ u \in U,$$

where $f : \mathbf{R}^{1+n+m} \to \mathbf{R}^n$ is continuous in u, U is compact in $\mathbf{R}^m$, and $f(t, \cdot, u)$ is an OSL continuous vector function, uniformly in $u \in U$. Then $F(t, \cdot) = f(t, \cdot, U)$ is an OSL continuous mapping in the sense of Definition 2.1 but may not be UOSL in the sense of [15], [7], as the next example shows.

Example 2.3 (continuous OSL interval function, neither Lipschitz nor UOSL). Let $m = n = 1$, $U = [-1, 1]$, and $f(x, u) = -x^{\frac{1}{3}} + u$. Then the map

$$F(x) = f(x, U) = -x^{\frac{1}{3}} + [-1, 1]$$

is a continuous OSL mapping that does not satisfy the UOSL condition of [15], [7].

3. Filippov Theorem for convex one-sided Lipschitz differential inclusions. The following basic assumptions are made.

A1. F is defined on $I \times X$, with nonempty compact and convex values, $F(\cdot, x)$ measurable for each x, and $F(t, \cdot)$ USC for all t.

A2. (integrably linear growth of F). There is an integrable function $\lambda : I \to \mathbf{R}_+$ such that $|F(t, x)| \le \lambda(t)(1 + |x|)$ for all $x \in X$ and almost all $t \in I$.

A3. F is OSL continuous with an integrable function $L(\cdot)$.

We first prove boundedness of the approximate trajectories set. Naturally, if the growth condition A2 is satisfied, an easy application of the Gronwall inequality will give the needed boundedness. We show here that the OSL condition implies boundedness of the state variable, provided the set $F(\cdot, K_0) = \bigcup_{x \in K_0} F(\cdot, x)$ is integrably bounded.

LEMMA 3.1. *Let F satisfy* A1, A3 *and the function $|F(\cdot, K_0)|$ be bounded by an integrable function $\mu(\cdot) : I \to \mathbf{R}_+$, and let $g : I \to \mathbf{R}_+$ be summable. Then all the solutions of the inclusion*

$$\dot{x}(t) \in F(t, x(t)) + g(t)U, \quad x(0) \in K_0 + d_o U,$$

are contained in a ball of radius $M = |K_0| + \max_{t \in I} |v(t)|$, centered in the origin.

Here $v(t) = \{d_o e^{m(t)} + \int_0^t e^{m(t)-m(s)}(\mu(s) + g(s))ds\}$ and $m(t) = \int_0^t L(r)dr$.

Proof. Given a solution $x(\cdot)$ of the given inclusion and $t \in I$, choose $x_0 \in K_0$ such that $|x(0) - x_0| \le d_o$ and $z \in g(t)U$ for which $\dot{x}(t) + z \in F(t, x(t))$.

By A3 there is a $w \in F(t, x_0)$ satisfying

$$\langle x(t) - x_0, \dot{x}(t) + z - w \rangle \le L(t)|x(t) - x_0|^2.$$

Therefore

$$\begin{aligned}\langle x(t) - x_0, \dot{x}(t) \rangle &\le \langle x(t) - x_0, w - z \rangle + L(t)|x(t) - x_0|^2 \\ &\le (|F(t, x_0)| + g(t))|x(t) - x_0| + L(t)|x(t) - x_0|^2.\end{aligned}$$

Denoting $s(t) = |x(t) - x_0|$, it is easy to verify that $s(\cdot)$ is an absolutely continuous function. At every $t \in I$, for which $s(\cdot)$ is differentiable, we have the inequality

$$s(t) \cdot \dot{s}(t) = \frac{1}{2}\frac{d}{dt}s^2(t) = \langle x(t) - x_0, \dot{x}(t) \rangle \le L(t)s^2(t) + (\mu(t) + g(t))s(t).$$

Define the set $T = \{t \in I : s(t) = 0\}$ and let T' be the set of all density points of T. It is well known that $meas(T') = meas(T)$. If $t \notin T$, then $\dot{s}(t) \le L(t)s(t) + \mu(t) + g(t)$, since $s(t) > 0$. If $t \in T'$ and if $\dot{s}(t)$ exists, then $\dot{s}(t) = 0$. Hence $\dot{s}(t) \le L(t)s(t) + \mu(t) + g(t)$ for a.e. $t \in I$. We calculate $\frac{d}{dt}[e^{-m(t)}(s(t) - v(t))] \le 0$. Therefore $e^{-m(t)}(s(t) - v(t)) \le s(0) - v(0) \le 0$; i.e., $s(t) \le v(t)$ for all $t \in I$. Hence $|x(t)| \le |x_0| + s(t) \le M$. □

Remark 3.1. Note that by the OSL condition we estimate only the norm of the solutions and not the norm of the derivatives. Generally, the velocities may be unbounded, and we need some growth condition so as to ensure existence of a trajectory. Clearly, if in addition A2 holds, then for every solution $x(\cdot)$ of (1.1), $|\dot{x}(t)| \le \lambda(t)(1 + M)$. These bounds provide extendability of the solutions on the whole interval I (the proof of this fact is standard and reiterated in Proposition 5.3).

The main result of this section follows.

THEOREM 3.2. *Suppose* A1, A2, A3 *hold and* $y : I \to X$ *is an AC function satisfying* $dist(\dot{y}(t), F(t, y(t))) \le g(t)$ *for a.e.* $t \in I$, *where* g *is integrable. Then for every* $K_0 \in \mathcal{K}$ *there exists a solution* $x(\cdot)$ *of* (1.1) *on the interval* I, *such that*

$$|x(t) - y(t)| \le de^{m(t)} + \int_0^t e^{m(t)-m(s)} g(s)ds, \tag{3.1}$$

where $d = dist(y(0), K_0)$, $m(t) = \int_0^t L(s)ds$.

Proof. Consider the ordinary differential inclusion

$$\dot{x}(t) \in G(t, x(t)), \quad x(0) = x_0, \tag{3.2}$$

where $x_0 \in K_0$ satisfies $|y(0) - x_0| = dist(y(0), K_0)$, $G(t, x) = F(t, x) \cap H(t, x)$, and $H(t, x)$ is determined by

$$H(t, x) = \{v \in X | \langle y(t) - x, \dot{y}(t) - v\rangle \le L(t)|y(t) - x|^2 + g(t)|y(t) - x|\}.$$

We will prove that (3.2) has a solution on I that satisfies the desired inequality by use of a known existence theorem for USC differential inclusions (cf. [5, Theorem 5.2]). It suffices to verify that

(i) $G(t, x)$ is nonempty, convex, and closed-valued, measurable in t;
(ii) for each t, $G(t, \cdot)$ has a closed graph, which implies that $G(t, \cdot)$ is USC;
(iii) $G(\cdot, \cdot)$ satisfies the growth condition A2.

We first prove that $G(t, x) \ne \emptyset$ for each t, x. Given t, x, let $w \in F(t, y(t))$ be such that $|\dot{y}(t) - w| = dist(y(t), F(t, y(t)))$. By the OSL condition we can choose $v \in F(t, x)$ so as to satisfy $\langle y(t) - x, w - v\rangle \le L(t)|y(t) - x|^2$. Hence

$$\begin{aligned}\langle y(t) - x, \dot{y}(t) - v\rangle &= \langle y(t) - x, \dot{y}(t) - w\rangle + \langle y(t) - x, w - v\rangle \\ &\le |y(t) - x|g(t) + L(t)|y(t) - x|^2, \quad \text{i.e., } v \in G(t, x)\ .\end{aligned}$$

Obviously $G(t, x)$ is convex, closed-valued, measurable in t, and satisfies A2 since F does. Moreover,

$$Graph\ G(t, \cdot) = Graph\ F(t, \cdot) \bigcap Graph\ H(t, \cdot),$$

and it is closed because the graphs of $F(t, \cdot)$ and $H(t, \cdot)$ are closed. From Theorem 5.2 of [5] we conclude that there exists a solution $x(\cdot)$ of (3.2) on I. Denoting $s(t) =$

$|x(t)-y(t)|$, one gets that $s(\cdot)$ is an AC function. Furthermore, if $\dot{s}(t)$ exists, then

$$s(t)\cdot\dot{s}(t)=\frac{1}{2}\frac{d}{dt}s^2(t)=\langle x(t)-y(t),\dot{x}(t)-\dot{y}(t)\rangle\le L(t)s^2(t)+g(t)s(t).$$

As in the previous lemma, we infer that $\dot{s}(t)\le L(t)s(t)+g(t)$ for a.e. $t\in I$, and the proof is completed by a simple comparison argument that repeats the one of Lemma 3.1. □

Let $0\le t_0\le t\le 1$. Denote by $A(t,t_0,K)$ the attainable set of (1.1) with initial condition $x(t_0)\in K$.

A direct consequence of Theorem 3.2 is the following stability condition, used by Artstein [1] to get a global error estimate for the attainable set in the Euler method.

COROLLARY 3.3. *Under the conditions of Theorem* 3.2, *if F is OSL with a constant L, then for given $a\le t_0\le t\le b$ and $K_1,K_2\in\mathcal{K}$*

$$haus(A(t,t_0,K_1),A(t,t_0,K_2))\le(1+Le^{L(b-a)}(t-t_0))haus(K_1,K_2).$$

Proof. The proof comes directly from Theorem 3.2 and the inequality $e^{L(t-t_0)}\le 1+Le^{L(b-a)}(t-t_0)$ for $a\le t_0\le t\le b$. □

Remark 3.2. The function L in OSL and (3.1) can take negative values. This advantage is used in some examples in section 6 for getting better estimates via the OSL constant.

Remark 3.3. Let us note that this theorem does not generalize Filippov's Theorem. As was mentioned in the Introduction, Filippov proved it without convexity of the right-hand side, assuming LC of $F(t,\cdot)$ and getting convergence of the successive approximations method for the velocities and trajectories and thus the existence of a solution. We replace here the continuity of $F(t,\cdot)$ by convexity and upper-semicontinuity, to ensure existence of a solution of (3.2). Note that for an OSL system a solution may exist without convergence of the successive approximations method (see the example in [16, p. 37]).

As is well known (cf. [5]), there are differential inclusions, discontinuous with respect to the state variable, that have no solution. The case when $F(t,\cdot)$ is continuous and nonconvex-valued is considered in section 5 (Theorem 5.4). The result proved there holds for ordinary differential inclusions as well.

Remark 3.4. Note that measurability of the function $dist(\dot{y}(t),F(t,y(t)))$ (cf. [3]) is not necessary, if we need only an estimate from above. Its integrable boundedness is sufficient.

Remark 3.5. A general form of Filippov's theorem, that is less known, was proven by Pliś [20] for measurable $F(\cdot,x)$, that satisfies a Kamke-type continuity condition with respect to x, instead of the Lipschitz one. A one-sided extension of Pliś' theorem can be obtained for $F(\cdot,\cdot)$ satisfying a general one-sided Kamke-type condition [10] by obvious modifications in the proof.

4. Euler method. Suppose that $A\subset X$ is compact and $F: I\times A\to\mathcal{KK}$ is uniformly bounded, measurable in the first argument, and continuous in the second one, i.e., almost continuous (cf. [5]). For given $t\in I$, $x\in A$ we define two local moduli of continuity of $F(\cdot,\cdot)$ with respect to each argument:

$$\chi(F,A,t,h)=\sup\{haus(F(t,x),F(t,y)):\ |y-x|\le h,x,y\in A\},$$

$$\tau(F,A,t,h)=\sup\{\sup(haus(F(s,x),F(r,x)):s,r\in[t-h/2,t+h/2]\cap I):x\in A\}.$$

These two moduli are measurable functions of t since for a.e. $t \in I$ they can be represented as supremum of countably many measurable functions of t, obtained for values of x in a countable dense subset of X.

Note that these moduli are local only in t and global in $x \in A$. The corresponding averaged L_p-moduli are

$$\chi(F,A,h)_p = \left(\int_0^1 \chi(F,A,t,h)^p dt\right)^{\frac{1}{p}}, \quad \tau(F,A,h)_p = \left(\int_0^1 (\tau(F,A,t,h))^p dt\right)^{\frac{1}{p}},$$

where $1 \le p < \infty$. For $p = 1$ we denote $\tau(F,A,h) = \tau(F,A,h)_1$, $\chi(F,A,h) = \chi(F,A,h)_1$. For additional information about averaged moduli of continuity we refer the reader to [21], [6], and [9].

Let us note that if F is integrably Hölder continuous in x of degree α (i.e., there exists an integrable function $L : I \to X$ such that for any $x, y \in A$, $haus(F(t,x), F(t,y)) \le L(t)|x-y|^\alpha$), then $\chi(F,A,h)_p = O(h^\alpha)$. If F has bounded p-variation in t, uniformly in $x \in A$, i.e.,

$$W_p(F) = \sup_k \left\{ \sum_{i=1}^{k-1} \sup_{x\in A} (haus(F(x,t_{i+1}), F(x,t_i)))^p, 0 \le t_1 \le \cdots \le t_k \le 1 \right\} < \infty,$$

then $\tau(F,A,h)_p = O(h^{\frac{1}{p}})$ [14], [9].

Consider the discrete approximation (1.2) of the initial problem (1.1). Obviously, if F has compact convex images, then for all $t \in I$ the attainable set of (1.2) coincides with the attainable set of the differential inclusion

$$\begin{cases} \dot{x}(t) \in F(t_i, x(t_i)) \quad \text{for a.e. } t \in (t_i, t_{i+1}),\ i = 0,1,\ldots,N-1\ , \\ x(0) \in K_0. \end{cases} \tag{4.1}$$

This technical note means that one can work with trajectories of (4.1) instead of (1.2) and vice versa. It is used in some further proofs.

Denote by R_1, R_2 the solution sets of (1.1) and (1.2), respectively, metrized by the C-norm of the solutions and the Hausdorff metrics for the solution sets.

We prove first boundedness of the discrete trajectories.

LEMMA 4.1. *Let* A2 *hold with a Riemann integrable* $\lambda(\cdot)$. *Then every trajectory* $x(\cdot)$ *of* (1.2) *is bounded by*

$$\max_{t\in I} |x(t)| \le e^{\Lambda}(|K_0| + \Lambda), \quad \textit{where} \quad \Lambda = \sup_{N\in\mathbf{N}} \frac{1}{N} \sum_{i=0}^{N} \lambda(t_i)\ .$$

Proof. Note that Λ is finite because of the Riemann integrability of $\lambda(\cdot)$. Let $x(\cdot)$ be a solution of (1.2). Denote for $i = 0,1,\ldots,N$, $x_i = x(t_i)$, $\lambda_i = \lambda(t_i)$. Clearly, it is sufficient to estimate x_i, $i = 1,\ldots,N$. From the equality $x_{i+1} = x_i + hf_i$, where $|f_i| \le \lambda_i(1 + |x_i|)$, and A2 we obtain

$$\begin{aligned} |x_{i+1}| &\le (1 + h\lambda_i)|x_i| + h\lambda_i \le \cdots \\ &\le \prod_{k=0}^{i} (1 + h\lambda_k)|x_0| + h \sum_{k=0}^{i-1} \lambda_k \prod_{j=k+1}^{i} (1 + h\lambda_j)\ . \end{aligned}$$

The simple observation that $1+h\lambda_j \le e^{h\lambda_j}$ implies $\prod_{j=1}^{N}(1+h\lambda_j) \le \exp(h\sum_{j=1}^{N}\lambda_j) \le e^{\Lambda}$. Then

$$\max\{|x_i| \mid 0 \le i \le N\} \le e^{\Lambda}\left(|K_0| + h\sum_{k=1}^{N-1}\lambda_k\right) \le e^{\Lambda}(|K_0|+\Lambda). \qquad \square$$

Remark 4.1. Lemmas 3.1 and 4.1 give boundedness of all "almost" trajectories, being continuous or discrete, i.e., satisfying the inclusions (1.1) or (1.2) with some integrably bounded defect $g(\cdot)$ (see Lemma 3.1). In the discrete case we require Riemann integrability of $g(\cdot)$. If $\lambda(\cdot)$ is integrable (or Riemann integrable in the discrete case), then the velocities are integrably bounded. If λ is a constant, then all the trajectories of (1.1), (1.2), and (4.1) are contained in a bounded set A, and all the velocities, in a bounded set B.

LEMMA 4.2. *Suppose* A1, A2, A3 *are satisfied with a constant* λ, *and* $F(t,\cdot)$ *is continuous. Then, for every solution* $x(\cdot)$ *of* (1.1), *there exists a trajectory* y *of* (4.1) *such that*

$$\max_{t\in[0,1]}|x(t)-y(t)| \le c(\tau(F,A,h)+\chi(F,A,h)),$$

where $c = e^{m(1)}\max(2,|B|)$, $m(t)$ *is defined in Lemma* 3.1, *and* A, B *are defined in the previous remark.*

Proof. Let $y(0)=x(0)$ and suppose $y(\cdot)$ exists on $[0,t_i]$. We prove inductively that it exists on $I_i=[t_i,t_{i+1}]$ $(i=0,1,\dots,N-1)$.

For given $t\in I_i$ and $u\in X$ consider the map $G(t,u)=F(t_i,y(t_i))\cap H(t,u)$, where

$$H(t,u)=\{v\in X|\langle v-\dot{x}(t),u-x(t)\rangle \le |x(t)-u|[L(t)|x(t)-u| + ex(F(t,u),F(t_i,y(t_i))]\},$$

where $ex(A,B)$ is the one-sided excess of A from B. As in Theorem 3.2 we get existence of a solution of the initial problem

$$\dot{y}(t)\in G(t,y), \quad y(t_i) \text{ known.}$$

In order to apply the existence theorem ([5, Theorem 5.2]), we verify that $G(\cdot,\cdot)$ satisfies (i), (ii), (iii).

(i) G is nonempty for every t,u. Indeed, there is a $z\in F(t,u)$ such that $\langle x(t)-u,\dot{x}(t)-z\rangle \le L(t)|x(t)-u|^2$. Further, for z we find $v\in F(t_i,y(t_i))$ satisfying

$$|z-v| = dist(z,F(t_i,y(t_i))) \le ex(F(t,u),F(t_i,y(t_i))).$$

Then $\langle x(t)-u,\dot{x}(t)-v\rangle \le L(t)|x(t)-u|^2+|x(t)-u|ex(F(t,u),F(t_i,y(t_i)))$; i.e., $v\in G(t,u)$. Obviously $G(\cdot,\cdot)$ is convex and closed-valued, measurable in t on I_i.

(ii) For fixed t, $G(t,\cdot)$ has a closed graph, as $H(t,\cdot)$ has. This fact follows from the upper-semicontinuity of the function $ex(F(t,\cdot),A)$ when $F(t,\cdot)$ is USC.

(iii) $G(\cdot,\cdot)$ trivially satisfies a growth condition being a subset of $F(t_i,x(t_i))$.

This way we infer existence of $y(\cdot)$ on I_i, $i=0,1,\dots,N-1$. Denote $\Delta(t)=x(t)-y(t)$. Then for $t\in I_i$

$$\frac{1}{2}\frac{d}{dt}\Delta^2(t)=\langle\Delta(t),\dot{\Delta}(t)\rangle \le |\Delta(t)|\,[L(t)|\Delta(t)| + haus(F(t,y(t)),F(t,y(t_i))) + haus(F(t,y(t_i)),F(t_i,y(t_i)))].$$

We repeat the arguments from Theorem 3.2 to obtain that for a.e. $t \in I$

$$\frac{d}{dt}|\Delta(t)| \le L(t)|\Delta(t)| + \chi(F, A, t, |B|h) + \tau(F, A, t, 2h),\ \Delta(0) = 0\ .$$

The proof is completed by a comparison argument and the observation that $\chi(F, A, |B|h) \le |B|\chi(F, A, h)$ and $\tau(F, A, 2h) \le 2\tau(F, A, h)$ (cf. [21]). ∎

THEOREM 4.3. *Under the conditions of the previous lemma*

$$haus(R_1, R_2) \le c_o(\tau(F, A, h) + \chi(F, A, h) + h),$$

where $c_o = \max\{e^{m_+(1)} \max(2, |B|), |B|\}$ *and* $m_+(t) = \int_0^t \max\{L(s), 0\}\, ds$.

Proof. Let $x(\cdot)$ be a solution of the discrete system (1.2). Then

$$\begin{aligned} &dist(\dot{x}(t), F(t, x(t))) \le ex(F(t_i, x(t_i)), F(t, x(t))) \\ &\le haus(F(t_i, x(t_i)), F(t, x(t_i))) + haus(F(t, x(t_i)), F(t, x(t))) \\ &\le \tau(F, A, t, 2h) + \chi(F, A, t, |B|h)\ . \end{aligned}$$

Applying Theorem 3.2, we obtain

$$\begin{aligned} &dist(x(\cdot), R_1) \\ &\le \int_0^1 e^{m(1)-m(t)}(\tau(F, A, t, 2h) + \chi(F, A, t, |B|h))dt \\ &\le e^{m_+(1)}(2\tau(F, A, h) + |B|\chi(F, A, h))\ . \end{aligned}$$

Let $y(\cdot)$ be a solution of (1.1). The previous lemma gives the existence of a solution $z(\cdot)$ of (4.1) with the desired property.

Now, take the piecewise-linear approximation of $z(\cdot)$,

$$x(t) = z(t_i) + (t - t_i)\frac{(z(t_{i+1}) - z(t_i))}{h} \quad \text{for} \quad t \in [t_i, t_{i+1}], \quad i = 0, 1, \ldots, N-1.$$

Obviously, $\max_{t\in I} |x(t) - z(t)| \le |B|h$, and by the triangle inequality we complete the proof. ∎

COROLLARY 4.4. *If additionally* $F(t, x)$ *is integrably Hölder continuous in* x *of degree* $\frac{1}{p}$, $1 \le p < \infty$, *and has uniformly bounded* p*-variation in* t, *then* $haus(R_1, R_2) = \mathcal{O}(h^{\frac{1}{p}})$.

Remark 4.2. If $L < 0$ and $\tau(F, A, t, h) \le Ph^\alpha$, $\chi(F, A, t, h) \le Qh^\beta$ $(0 < \alpha, \beta \le 1)$, the following inequality holds:

$$haus(R_1, R_2) \le \frac{1}{|L|}(1 - e^L)[P(2h)^\alpha + Q(|B|h)^\beta] + |B|h.$$

If, additionally, F does not depend on t, then

$$haus(R_1, R_2) \le \frac{1}{|L|}(1 - e^L)Q(|B|h)^\beta + |B|h.$$

5. The functional-differential case. In this section we will extend our results to the case of functional-differential inclusion (1.3). We will follow the style of [13].

The extension of OSL condition in this case possesses some specific problems. The state variable x_t and the velocities are not elements of the same space. These problems are overcome in the case of functional-differential equations in [17], where a Razumikhin-type Lipschitz condition is exploited.

Denote $E_0 = \{\alpha \in E : |\alpha(0)| = \max\{|\alpha(s)| : s \in [-\theta, 0]\}\}$. The mapping $F : E \to K$ is Lipschitz continuous (in Razumikhin sense) iff $haus(F(\alpha), F(\beta)) \le L|\alpha - \beta|_E$ for all $\alpha, \beta \in E$ such that $\alpha - \beta \in E_0$. We introduce a Razumikhin-type OSL condition, which will be called OSL here for the reader's convenience.

DEFINITION 5.1. *The multifunction $F : I \times E \to K$ is said to be OSL when there exists an integrable function $L(\cdot)$ such that*

$$\sigma(\alpha(0) - \beta(0), F(t, \alpha)) - \sigma(\alpha(0) - \beta(0), F(t, \beta)) \le L(t)|\alpha(0) - \beta(0)|^2$$

for every $\alpha, \beta \in E$ such that $\alpha - \beta \in E_0$.

Remark 5.1. Note that the conditions hold true not for all α, β, but only when $\alpha - \beta \in E_0$. In the proof of the stability theorem we cannot use the same approach to obtain existence as in the previous sections, since the mapping $G(\cdot, \cdot)$, defined in Theorem 3.2 would not be USC anymore. In order to apply the existence theory for LSC mappings, we suppose that $F(t, .)$ is continuous. In this case we cannot use negative OSL constants.

The main assumptions in this section follow.

H1. F is nonempty, compact-valued, and integrably bounded on bounded sets.

H2. F is OSL with integrable function $L(t)$, and $F(\cdot, \alpha)$ is measurable, while $F(t, \cdot)$ is continuous.

As was noted, in this case we suppose that $L(t) \ge 0$ or replace it by $L_+(t) = \max\{0, L(t)\}$.

Denote for $A \in \mathcal{K}$ the set $F(I, A) = \bigcup_{t \in I, x \in A} F(t, x)$.

PROPOSITION 5.2. *Let F be bounded on the bounded sets and OSL with a constant L. Then the multifunction $\alpha \to F(I, \alpha + U)$ is also OSL with constant L.*

Proof. Let $\alpha - \beta \in E_0$. Fix $\varepsilon > 0$. Therefore there exist $\tilde{t} \in I$ and $\tilde{l} \in U$ such that $\sigma(\alpha(0) - \beta(0), F(I, \alpha + U)) \le \sigma(\alpha(0) - \beta(0), F(\tilde{t}, \alpha + \tilde{l})) + \varepsilon$. The following inequalities complete the proof, since $\varepsilon > 0$ is arbitrary:

$$\begin{aligned} &\sigma(\alpha(0) - \beta(0), F(I, \alpha + U)) - \sigma(\alpha(0) - \beta(0), F(I, \beta + U)) \\ &\le \sigma(\alpha(0) - \beta(0), F(\tilde{t}, \alpha + \tilde{l})) - \sigma(\alpha(0) - \beta(0), F(\tilde{t}, \beta + \tilde{l})) + \varepsilon \\ &\le L|\alpha(0) - \beta(0)|^2 + \varepsilon. \qquad \square \end{aligned}$$

We approximate the solutions of (1.3) by polygonal functions

$$\begin{cases} x(t) = x(t_i) + (t - t_i) f_i \quad \text{for } t \in [t_i, t_{i=1}], \\ \text{where } f_i \in F(t_i, x_{t_i}), \quad i = 0, 1, \ldots, N-1, \quad x(0) \in K_0. \end{cases} \tag{5.1}$$

As in section 4, if F has compact convex images, the attainable set of (5.1) coincides with the attainable set of the inclusion

$$\begin{cases} \dot{x}(t) \in F(t_i, x_{t_i}) \quad \text{for } t \in (t_i, t_{i+1}),\ i = 0, 1, \ldots, N-1, \\ x(0) \in K_0. \end{cases} \tag{5.2}$$

Denote by R_3, R_4 the solution sets of (1.3) and (5.1), respectively, metrized by the C-norm of the solutions and the Hausdorff metrics for the solution sets.

If $L(.)$ is not a constant, the conclusion of Proposition 5.2 need not be valid. Next we prove the following proposition.

PROPOSITION 5.3. *Under* H1, H2 *the solutions' set of* (1.3) *is bounded and every solution is extendable on the whole* I.

Proof. Let $x(\cdot)$ be a solution of (1.3). Denote $s(t) = |x_t|_E$. Let $x_t \notin E_0$. Then there exists $h > 0$ such that $|x_{t+h}(0)| < |x_{t+h}|_E$, since $x(.)$ is continuous. So in this case $\dot{s}(t) \le 0$. If $x_t \in E_0$, then $\langle x(t), \dot{x}(t)\rangle \le \sigma(x(t), F(t, x_t)) \le L(t)|x(t)|^2 + |F(t,0)||x(t)|$. Therefore $\dot{s}(t) \le L_+(t)s(t) + |F(t,0)|$, $s(0) = |\phi|_E$. Hence $s(t) \le e^{m(t)}(|\phi| + \int_0^1 e^{-m(\tau)}|F(\tau,0)|\,d\tau)$. Thus $x(\cdot)$ is bounded. Let $x(\cdot)$ be defined on $[0,T)$ with $T < 1$. Then $x(\cdot)$ is AC, since F is integrably bounded on the bounded sets. Therefore $x(T) = \lim_{t\to T^-} x(t)$ exists and hence $x(\cdot)$ may be extended on some larger interval. Applying the Zorn lemma, one proves that $x(.)$ exists on the whole I. □

Remark 5.2. Let $\dot{x}(t) \in F(I, x_t + U)$. If the conditions of Proposition 5.2 hold, then it is easy to show as above that

$$\frac{d}{dt}|x(t)| \le L_+|x(t)| + |F(I,U)|, \text{ therefore } |x(t)| \le e^{L_+ t}\{|\phi| + |F(I,U)|\} = M.$$

Denote $V = MU, W = |F(I, MU)|$. Then $x(.)$ is Lipschitz with a constant W. If $Wh < 1$, then every solution of (5.2) is also a solution of

$$\dot{x}(t) \in F(I, x_t + U), \quad x_0 \in K_0.$$

Hence, if the requirements of Proposition 5.2 hold, then for every solution of (1.3) or of (5.2), $x(t) \in V$, $|F(t, x_t)| \le W$. Therefore, without loss of generality, we can suppose that the right-hand side F is bounded by the constant W.

Now we will prove a Filippov-type theorem for the nonconvex case.

THEOREM 5.4. *Suppose* H1 *and* H2 *hold and that* $y(\cdot)$ *is an AC function with* $dist(\dot{y}(t), F(t, y_t)) \le f(t)$ *for some integrable* $f(\cdot)$. *Then for every* $\varepsilon > 0$ *and every* $x_0 \in X$ *there exists a solution* $x(\cdot)$ *of* (1.3) *such that*

$$|x(t) - y(t)| \le de^{m(t)} + \int_0^t e^{m(t)-m(s)} f(s)\,ds + \varepsilon, \quad x(0) = x_0, \tag{5.3}$$

where $d = |x_0 - y(0)|$, $m(t) = \int_0^t L_+(s)\,ds$, $L_+(t) = \max\{L(t), 0\}$.

Proof. For $\delta > 0$ define the multifunction

$$G_\delta(t,u) = \begin{cases} F(t,u), & u - y_t \notin E_o, \\ \{v \in F(t,y_t) : |v - \dot{y}(t)| = dist(\dot{y}(t), F(t,y_t))\}, & u = y_t, \\ \mathrm{cl}\{v \in F(t,u) : \langle y(t) - u(0), \dot{y}(t) - v\rangle \\ < L(t)|y(t) - u(0)|^2 + |y(t) - u(0)|(f(t) + \delta)\} & \text{elsewhere.} \end{cases}$$

We are going to prove that the differential inclusion

$$\dot{x}(t) \in G_\delta(t, x_t)$$

has a solution. In order to use the existence theorem of [12], we prove that $G_\delta(\cdot,\cdot)$ has closed values and is almost LSC. Observe that, as in the proof of Theorem 3.2, $G_\delta(\cdot,\cdot)$ is nonempty, compact-valued.

Further, we prove that $G_\delta(\cdot,\cdot)$ is almost LSC. Since E is separable, one can apply the Scorza–Dragoni theorem. Fix $\nu > 0$. From the Scorza–Dragoni theorem and

Lusin's theorem there exists $I_\nu \subset I$ with $meas(I \setminus I_\nu) < \nu$ such that $\dot{y}(\cdot), f(\cdot), L(\cdot)$ are continuous on I_ν and $F(\cdot,\cdot)$ is continuous on $I_\nu \times E$. Let $t \in I_\nu, u - y_t \in E_o, u \neq y_t, v \in G_\delta(t,u)$. Suppose $t = \lim_{i\to\infty} t_i, u = \lim_{i\to\infty} u_i$. Hence $f(t) = \lim_{i\to\infty} f(t_i)$, $F(t,u) = \lim_{i\to\infty} F(t_i,u_i)$, $\dot{y}(t) = \lim_{i\to\infty} \dot{y}(t_i)$. Let

$$\langle y(t) - u(0), \dot{y}(t) - v\rangle = L(t)|y(t) - u(0)|^2 + |y(t) - u(0)|(f(t) + \delta - \mu)$$

where $\mu > 0$. Since $y_t - u \in E_o$ and $y_t \neq u$, it follows that $|y(t) - u(0)| > 0$. Therefore

$$\langle y(t) - u(0), \dot{y}(t) - v\rangle = L(t)|y(t) - u(0)|^2 + |y(t) - u(0)|(f(t) + \delta) - c,$$

where $c > 0$. Define $v_i \in F(t_i,u_i)$ such that $|v - v_i| = dist(v, F(t_i,u_i))$. The continuity of $F(\cdot,\cdot)$ implies $\lim_{i\to\infty} v_i = v$. Moreover, $v_i \in G_\delta(t_i,u_i)$ for sufficiently large i. If $u_i - y_t \notin E_o$, then obviously $\lim_{i\to\infty} G_\delta(t_i,u_i) = \lim_{i\to\infty} F(t_i,u_i) \supset G_\delta(t,u)$. If $u = y_t$, let $v_i \in F(t_i,u_i)$ be such that $|v - v_i| = dist(v, F(t_i,u_i))$. Because of the continuity of $F(\cdot,\cdot), \dot{y}(\cdot), f(\cdot), L(\cdot)$, we get $\lim_{i\to\infty} v_i = v$. Thus $G_\delta(.,.)$ is almost LSC. From Theorem 1 of [12] we infer that the upper inclusion admits a solution $x(\cdot)$. As in the previous proof, by use of Lemma 2.1 of [17], one can show that

$$|x(t) - y(t)| \leq de^{m(t)} + \int_o^t e^{m(t)-m(s)}(\delta + f(s))ds.$$

Evidently choosing appropriate small δ, one completes the proof. □

COROLLARY 5.5. *Under the conditions of Theorem* 5.4, *if F is also convex-valued, one can replace ε by zero.*

Proof. The solution set of (1.3) is compact, since F is convex compact-valued. Let $x_n(\cdot)$ be a sequence of solutions of (1.3), satisfying the conditions of Theorem 5.4 with $\varepsilon = \frac{1}{n}$. Passing to subsequences, if necessary, one concludes that $\lim_{n\to\infty} x_n(t) = x(t)$, where $x(.)$ obviously satisfies the conclusion. □

Remark 5.3. Obviously, the proof of Theorem 5.4 is valid for ordinary differential inclusions. Generally, without the convexity assumption, the defect ε cannot be removed. Nevertheless, this result is sufficient for many applications.

In the rest of this section we will suppose as follows.

H3. $L(\cdot)$ is constant and F is convex-valued and bounded on bounded sets.

The following lemma corresponds to Lemma 4.2 for the functional-differential case.

LEMMA 5.6. *Suppose $F : I \times E \to \mathcal{KK}$ satisfy* H1, H2, H3. *If $x(\cdot)$ is a solution of* (1.3), *then there exists an AC solution $y(\cdot)$ of* (5.2), *such that $|x(t) - y(t)| \leq r(t)$, where $r(\cdot)$ is a nonnegative function satisfying*

$$\dot{r}(t) = L_+ r(t) + 2\chi(F,V,t,h) + W\tau(F,V,t,h),$$

and V, W are as defined in Remark 5.2.

Proof. Let $0 = t_0 < t_1 < \cdots < t_N = 1$ be a uniform grid of I with $h = \frac{1}{N}$. Suppose $y(\cdot)$ is constructed on $[0,t_i]$ and $t \in [t_i, t_{i+1}]$. Fix $\varepsilon > 0$ and for given y_i, t, u define the map

$$G_\varepsilon(t,u) = \begin{cases} F(t_i,y_i), & u - x_t \notin E_o, \\ \{v \in F(t_i,y_i) : |v - \dot{x}(t)| = dist(\dot{x}(t), F(t_i,y_i))\}, & u = x_t, \\ \mathrm{cl}\{v \in F(t_i,y_i) : \langle x(t) - u(0), \dot{x}(t) - v\rangle < \psi(t,u,y_i)\} & \text{elsewhere,} \end{cases}$$

where $y_i = y_{t_i}, \psi(t,u,y_i) = L_+|x(t)-u(0)|^2+|x(t)-u(0)|[\varepsilon+haus(F(t,u),F(t,y_i))+haus(F(t_i,y_i),F(t,y_i))]$. As above, one can show that $G_\varepsilon(\cdot,\cdot)$ is compact-valued and almost LSC. Since $F(t_i,y_i)$ is bounded, it follows that the differential inclusion

$$\dot{y}(t) \in G_\varepsilon(t,y_t), \quad y(t_i) \text{ fixed}, \quad t \in [t_i, t_{i+1}]$$

has a solution $y^\varepsilon(\cdot)$ in $[t_i,t_{i+1}]$. If we denote $r^\varepsilon(t) = \max_{-\theta\le s\le\theta}|x(t+s)-y^\varepsilon(t+s)|$ and suppose $y_o^\varepsilon = \phi$, then

$$\dot{r}^\varepsilon(t) \le Lr^\varepsilon(t) + \varepsilon + haus(F(t_i,y_i^\varepsilon),F(t,y_t^\varepsilon)), \quad r^\varepsilon(0) = 0 \text{ or } \dot{r}^\varepsilon(0) = 0.$$

Or $\dot{r}^\varepsilon(t) \le 0$. Now it is easy to show by induction that

$$r^\varepsilon(t) \le e^{L_+t}\left\{\varepsilon t + \sum_{k=0}^{i-1}\int_{t_k}^{t_{k+1}} haus(F(t_k,y_k^\varepsilon),F(s,y_s^\varepsilon))ds + \int_{t_i}^{t} haus(F(t_i,y_i^\varepsilon),F(s,y_s^\varepsilon))ds\right\}.$$

Since the solution set of the upper inclusions is compact and decreasing for $\varepsilon > 0$ decreasing, one gets existence of a trajectory $y(\cdot)$ such that $|x(t)-y(t)| \le r(t)$, where

$$r(t) \le e^{L_+t}\left\{\sum_{k=0}^{i-1}\int_{t_k}^{t_{k+1}} haus(F(t_k,y_k),F(s,y_s))\,ds + \int_{t_i}^{t} haus(F(t_i,y_i),F(s,y_s))\,ds\right\}.$$

By induction one can extend this $y(.)$ on the whole I. Furthermore, $haus(F(t_k,y_k),F(s,y_s)) \le haus(F(t_k,y_k),F(s,y_k)) + haus(F(s,y_k),F(s,y_s)) \le \tau(F,V,t,2h)+\chi(F,V,t,Wh)$. Hence

$$\dot{r}(t) = L_+r(t) + \chi(F,V,t,Wh) + \tau(F,V,t,2h). \qquad \square$$

Now it is easy to prove the main result of this section.

THEOREM 5.7. *Under the conditions of the previous lemma*

$$haus(R_3,R_4) \le c_1(2\tau(F,V,h) + W\chi(F,V,h) + h),$$

where $c_1 = \max\{e^{L_+}\max(2,W), W\}$.

The proof is similar to the proof of Theorem 4.3 and is omitted.

Remark 5.4. As the reader has noted, the exposition of this section is similar to the previous one. Some specific technical difficulties have been overcome by modifying the definitions and proofs. Moreover, Theorem 5.4 also holds for infinite-dimensional spaces, assuming some compactness conditions to get the existence of solutions. In this case the OSL condition should be defined as in [8], where it is called "a dissipative type condition." Besides, Lemma 5.6 and Theorem 5.7 are valid in Hilbert spaces (under some additional compactness assumptions again). We will not go into details of the infinite-dimensional case, since it is more complicated and goes beyond the scope of this paper. Finally, we present two examples.

Example 5.1. Consider the following system with maximum:

$$\dot{x}(t) \in -x + -x^{3/5} + \max_{t\in[-\theta,0]}|x^{3/5}(t+s)| + y + [-1,1], \quad x(s) = 0, s \in [-\theta,0],$$

$$\dot{y}(t) \in -y + [0,1], \quad y(0) = 0.$$

Since the OSL condition holds for $x_t \in E_0$, this system is OSL with constant $L = -\sqrt{2}$. So we have to replace L by 0. Moreover, this system is Hölder continuous of degree $\alpha = 3/5$. Therefore there exists a constant C such that $haus(R_3,R_4) \le Ch^{3/5}$.

Example 5.2. This example presents an integrodifferential inclusion

$$\dot{x}(t) \in -x^{5/7} + \int_{-\tau}^{0} x(t+s)\,ds + [-1,1], \quad x(s) = 0, s \in [-\tau, 0].$$

Obviously, the right-hand side is OSL with constant less than 1 and Hölderian of degree 5/7. Hence $haus(R_3, R_4) \le Ch^{5/7}$, where C can be estimated by Theorem 5.7.

6. Examples. Let $l, m \in \mathbf{N}$, $\alpha = \frac{2l+1}{2m+1}$. First, we prove the useful fact that if $|x-y| = h$, then $|x^\alpha - y^\alpha| \le 2^{1-\alpha}h^\alpha$. Let $f(x) = (x+h)^\alpha - x^\alpha$. Then $f'(x)$ exists for every $x \ne 0$. Calculating $f(0) = f(-h) = h^\alpha$, we see that $f(.)$ has a maximal value when $f'(x) = 0$, i.e., for $x = -\frac{h}{2}$.

Let us give some examples.

Example 6.1 (continuous OSL system that is not Lipschitz).

$$\dot{x}(t) \in -x - x^{3/5} + y + [-1,1], \quad x(0) = 0,$$
$$\dot{y}(t) \in -y + [0,1], \quad y(0) = 0.$$

It is easy to calculate $0 \le y(t) \le 1 - \exp(-t)$. Besides, for $|x| \le 1$, one has $|x| \le |x|^{3/5}$. Thus $\dot{x}(t) \le -2x + y + 1$. Hence $x(t) \le 1 - \exp(-t)$. Thus $0 \le y(t) \le 1 - \exp(-t)$ and $-1 + \exp(-t) \le x(t) \le 1 - \exp(-t)$. Obviously, every discrete trajectory also approximately satisfies the same inequalities. Hence, without loss of generality, one can suppose

$$|F(x,y)| \le \sqrt{[(1-e^{-1})^{3/5} + (3-2e^{-1})]^2 + (2-e^{-1})^2}$$
$$< \sqrt{(1-e^{-1})^{6/5} + 2(1-e^{-1}) + (3-2e^{-1})^2 + (2-e^{-1})^2}$$
$$\le \sqrt{20 - 25e^{-1} + 9e^{-2}} < 3.7.$$

Denote $z = (x, y)$. Then

$$haus(F(z_1), F(z_2)) \le \sqrt{3(x_1-x_2)^2 + 3(x_1^{3/5} - x_2^{2/5})^2 + 4(y_1-y_2)^2}$$
$$\le 2|z_1 - z_2| + \sqrt{3} \cdot 2^{3/5}|z_1 - z_2|^{3/5}.$$

Indeed, we showed in the beginning of this section that $|a^{3/5} - b^{3/5}| \le 2|\frac{a-b}{2}|^{3/5}$. Clearly, also,

$$\sigma(z_1 - z_2, F(z_1)) - \sigma(z_1 - z_2, F(z_2))$$
$$= -(x_1-x_2)^2 - (x_1-x_2)(x_1^{3/5} - x_2^{3/5}) + (x_1-x_2)(y_1-y_2) - (y_1-y_2)^2$$
$$\le -\frac{1}{2}[|x_1-x_2|^2 + |y_1-y_2|^2] = -\frac{1}{2}|z_1 - z_2|^2.$$

By these facts and Theorem 4.3 we obtain

$$haus(R_1, R_2) \le 3.7(3h + \sqrt{3} \cdot 2^{2/5}h^{3/5}).$$

Example 6.2 (OSL system that is not Lipschitz on a dense set). Let $\{x_k\}_{k=1}^{\infty}$ be the set of all rational numbers in $[0,1]$ ordered in sequence. Consider the following

system:

$$\dot{x}(t) \in -2x + \sum_{k=1}^{\infty} \frac{1}{2^k}(x_k - x)^{5/7} + [0,1], \quad x(0) = 0.$$

Then

$$\sigma(x-y, F(x)) - \sigma(x-y, F(y)) \le -2|x-y|^2.$$

Moreover, for every measurable $g(t) \in [0,1]$,

$$2 - 2x \ge -2x + \sum_{k=1}^{\infty} \frac{1}{2^k}(x_k - x)^{5/7} + g(t) > -3x,$$

i.e., for every solution $0 \le x(t) \le 1 - e^{-2t}$. The last inequality is also valid for every discrete trajectory. Thus $|F(x)| \le 2$. As in the beginning of this section one can prove that $|x^{5/7} - y^{5/7}| \le 2|\frac{x-y}{2}|^{5/7}$. Therefore $haus(F(x), F(y)) < 2|x-y| + 2^{2/7}|x-y|^{5/7}$. Thus, by Remark 4.2,

$$haus(R_1, R_2) \le \frac{1}{2}(1 - e^{-2})(4h + 2h^{5/7}) + 2h \le 4h + h^{5/7}.$$

Even when the right-hand side F is Lipschitz, sometimes the OSL constant gives a better estimate.

Example 6.3 (Lipschitz continuous multifunction with better OSL constant).

$$\dot{x}(t) \in -6x + y + G(t), \quad x(0) = 0,$$
$$\dot{y}(t) \in x - 6y + H(t), \quad y(0) = 0.$$

Denote again $z = (x, y)$. Therefore

$$haus(F(z_1), F(z_2)) = \sqrt{37(x_1 - x_2)^2 + 37(y_1 - y_2)^2 - 24(x_1 - x_2)(y_1 - y_2)}.$$

Hence the Lipschitz constant is $L = 7$. Furthermore

$$\sigma(z_1 - z_2, F(z_1)) - \sigma(z_1 - z_2, F(z_2))$$
$$-6(x_1 - x_2)^2 + 2(x_1 - x_2)(y_1 - y_2) - 6(y_1 - y_2)^2 \le -5|z_1 - z_2|^2.$$

Let $G(t) \subset [0,1]$ and let $H(t) \subset [0,1]$. It is easy to calculate $0 \le \dot{x}(t) \le 1 - 6x + y$ and $0 \le \dot{y}(t) \le 1 + x - 6y$. Now one can conclude that $0 \le x(t) \le 1/5$, $0 \le y(t) \le 1/5$. Hence $|F(x,y)| \le \sqrt{(1 + 1/5)^2 + (1 + 1/5)^2} = \frac{6\sqrt{2}}{5} \le 2$. If, moreover, $H(t) \equiv G(t) \equiv [0,1]$, then by Remark 4.2, $haus(R_1, R_2) \le \frac{14}{5}(1 - e^{-5})h + 2h \le 5h$. This estimate is much better than the one obtained with the ordinary Lipschitz constant $L = 7$: $haus(R_1, R_2) \le (e^7 - 1)h$.

Acknowledgments. We wish to thank the anonymous referees of the first version of this paper for their valuable comments and suggestions, which contributed much to improve the presentation. We are also grateful to Mrs. Diana Yellin of the School of Mathematical Sciences at Tel-Aviv University for her help in preparing the LaTeX version of this manuscript.

REFERENCES

[1] Z. Artstein, *First order approximations for differential inclusions*, Set-Valued Anal., 2 (1994), pp. 7–18.

[2] J.-P. Aubin and A. Cellina, *Differential Inclusions*, Springer-Verlag, Berlin, Heidelberg, New York, Tokyo, 1984.

[3] J.-P. Aubin and H. Frankowska, *Set-Valued Analysis*, Birkhäuser, Boston, Basel, Berlin, 1990.

[4] G. Birkhoff and G.-C. Rota, *Ordinary Differential Equations*, John Wiley, New York, Chicago, 1978.

[5] K. Deimling, *Multivalued Differential Equations*, De Gruyter, Berlin, 1992.

[6] A. Dontchev and E. Farkhi, *Error estimates for discretized differential inclusions*, Computing, 41 (1989), pp. 349–358.

[7] A. Dontchev and F. Lempio, *Difference methods for differential inclusions: A survey*, SIAM Rev., 34 (1992), pp. 263–294.

[8] T. Donchev, *Functional differential inclusions with monotone right hand side*, Nonlinear Anal., 16 (1991), pp. 543–552.

[9] T. Donchev and E. Farkhi, *Moduli of smoothness of vector-valued functions of a real variable and applications*, Numer. Funct. Anal. Optim., 11 (1990), pp. 497–509.

[10] T. Donchev and R. Ivanov, *On the existence of solutions of differential inclusions in uniformly convex Banach spaces*, Math. Balkanica, 6 (1992), pp. 13–24.

[11] A. F. Filippov, *Classical solutions of differential equations with multivalued right-hand side*, SIAM J. Control Optim., 5 (1967), pp. 609–621.

[12] A. Fryszkowski, *Existence of solutions of functional differential inclusions in nonconvex case*, Ann. Polon. Math., 45 (1989), pp. 121–124.

[13] J. Hale, *Theory of Functional Differential Equations,* Springer-Verlag, New York, 1977.

[14] V. Hristov, *On the coefficients of Fourier-Lagrange*, Pliska - Bulgarian Mathematical Studies, 5 (1983), pp. 23–31 (in Russian).

[15] A. Kastner-Maresch, *Implicit Runge-Kutta methods for differential inclusions*, Numer. Funct. Anal. Optim., 11 (1990), pp. 937–958.

[16] V. Lakshmikantham and S. Leela, *Nonlinear Differential Equations in Abstract Spaces*, Pergamon, Oxford, UK, 1981.

[17] V. Lakshmikantham, A. Mitchel, and R. Mitchel, *On the existence of solutions of differential equations of retarded type in a Banach space*, Ann. Polon. Math., 35 (1977), pp. 253–260.

[18] F. Lempio, *Euler's method revisited*, Proc. Steklov Inst. Math., Moscow, 211 (1995), pp. 473–494.

[19] M. Nikolskii, *On a method for approximation of the reachable set for a differential inclusion*, J. Vychisl. Mat. i Mat. Phys., 28 (1988), pp. 1252–1254 (in Russian).

[20] A. Pliś, *On trajectories of orientor fields*, Bull. Acad. Pol. Sci. Ser. Sci. Math. Astron. Phys., 13 (1965), pp. 571–573.

[21] B. Sendov and V. Popov, *The Averaged Moduli of Smoothness*, John Wiley, New York, 1988.

[22] V. Veliov, *Differential inclusions with stable subinclusions*, Nonlinear Anal., 23 (1994), pp. 1027–1038.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 3, pp. 797–813, May 1998

EQUILIBRIUM CONDITIONS FOR YOUNG MEASURES*

PABLO PEDREGAL†

Abstract. We explore variations of Young measures in order to establish equilibrium conditions for minimizers of generalized variational principles where Young measures enter the minimization problem. The slicing measure decomposition is extremely useful in rendering equilibrium conditions tractable; they usually yield information on the support of generalized minimizers. It also enables us to describe new, equivalent variational principles where these equilibrium constraints are taken into account. The only case where computations can be made explicit is the scalar, one-dimensional case.

Key words. generalized variational principles, necessary conditions for equilibrium, slicing measures

AMS subject classifications. 40K27, 35D05

PII. S036301299630080X

1. Introduction.

Nonconvex variational problems are important from the point of view of applications. For emerging, interesting problems in optimization and the calculus of variations, lack of convexity leads to the analysis of generalized or relaxed principles that provide information on the behavior of minimizing sequences for the original problem. Young measures are fundamental to this approach since they furnish the competing objects for these new variational problems. Some basic references for nonconvexity, Young measures, and applications to several situations in continuum mechanics and in the theory of optimal control (where Young measures were originally introduced) are [2], [5], [6], [7], [8], [9], [14], [15], [16], [17], [18], [19], [27], [28], [29].

Let us consider the problem of finding a minimizer for

$$J(u) = \int_\Omega W(x, u(x), \nabla u(x))\, dx, \quad u \in W^{1,p}(\Omega), u - U \in W_0^{1,p}(\Omega),$$

where $\Omega \subset \mathbf{R}^N$ is a regular domain and U is a given function in $W^{1,p}(\Omega)$. The integrand

$$W : \Omega \times \mathbf{R}^m \times \mathbf{M}^{m\times N} \to \mathbf{R}$$

is assumed to be as smooth as we may need to avoid technicalities in the derivation of equilibrium conditions, and verifies the bounds

$$\begin{aligned} c\left(|A|^p + |\lambda|^p - 1\right) \le W(x,\lambda,A) &\le C\left(|A|^p + |\lambda|^p + 1\right), \\ \left|\frac{\partial W}{\partial A}(x,\lambda,A)\right| &\le C(|A|^{p-1} + |\lambda|^{p-1} + 1), \\ \left|\frac{\partial^2 W}{\partial A^2}(x,\lambda,A)\right| &\le C(|A|^{p-2} + |\lambda|^{p-2} + 1) \end{aligned} \tag{1.1}$$

for $p > 1$ for the bounds on the first derivative, $p > 2$ for the bounds on the second derivative, and $0 < c < C$. The condition that guarantees existence of minimizers for

*Received by the editors March 20, 1996; accepted for publication (in revised form) February 8, 1997. This work was supported by DGICYT (Spain) grant PB93-0070.
http://www.siam.org/journals/sicon/36-3/30080.html
†ETSI Industriales, Universidad de Castilla–La Mancha, 13071 Ciudad Real, Spain (ppedregal@ind-cr.uclm.es).

J is the property of quasiconvexity (for the vector case $m > 1$) or convexity (for the scalar case $m = 1$) [1], [3], [4], [10], [12], [13], [24]. W is said to be quasiconvex if

$$W(x, \lambda, F) \leq \frac{1}{|\Omega|} \int_\Omega W(x, \lambda, F + \nabla w(y))\, dy \tag{1.2}$$

for every pair $(x, \lambda) \in \Omega \times \mathbf{R}^m$, every matrix $F \in \mathbf{M}^{m\times N}$, and every test function w. When the integrand W enjoys this property, the direct method of the calculus of variations provides minimizers for our problem. The quasiconvexity condition (1.2) together with the bounds (1.1) ensures the weak lower semicontinuity property of the functional J in $W^{1,p}(\Omega)$ [1]. For the scalar case ($m = 1$) (1.2) reduces to plain convexity.

If the quasiconvexity condition (1.2) fails, the problem might not have minimizers; minimizing sequences converge weakly, but weak limits may not be minimizers. In fact, in many situations of interest minimizers do not exist. In some others, existence of minimizers may be established with different techniques (there is a rather large amount of papers on this issue, but we do not include any specific reference because it is not relevant to our discussion here). We would like to place ourselves in the situation where we do not have minimizers so that the quasiconvexity condition (1.2) fails to hold. At this stage, the analysis may proceed in two ways. We may consider the so-called relaxed problem, defined by

$$QJ(u) = \int_\Omega QW(x, u(x), \nabla u(x))\, dx, \quad u \in W^{1,p}(\Omega), u - U \in W_0^{1,p}(\Omega),$$

where QW is the quasiconvexification of W with respect to the gradient variable

$$QW(x, \lambda, F) = \inf_w \frac{1}{|\Omega|} \int_\Omega W(x, \lambda, F + \nabla w(y))\, dy.$$

The infimum is taken over the set of test functions w [11], [12], [13], [22], [23]. Alternately, we may retain the functional J and allow Young measures to enter the minimization problem; this viewpoint is extensively studied in [26]. The crucial point here is to make sure that the admissible Young measures for our problem are generated by sequences of gradients. Under the bounds (1.1), we call those $W^{1,p}$-Young measures: they are generated as the parametrized measure corresponding to the gradients of a bounded sequence in $W^{1,p}(\Omega)$ [5]. We define J on families of probability measures $\nu = \{\nu_x\}_{x\in\Omega}$ by putting

$$J(\nu) = \int_\Omega \int_{\mathbf{M}^{m\times N}} W(x, u(x), A)\, d\nu_x(A)\, dx,$$

where ν is a $W^{1,p}$-Young measure such that

$$\nabla u(x) = \int_{\mathbf{M}^{m\times N}} A\, d\nu_x(A), \quad \text{a.e. } x \in \Omega, \tag{1.3}$$

$u \in W^{1,p}(\Omega)$, and $u - U \in W_0^{1,p}(\Omega)$. Equation (1.3) gives the link between ν and u.

$W^{1,p}$-Young measures are intimately connected to the quasiconvexity condition. Indeed, this class of Young measures are characterized by Jensen's inequality for quasiconvex functions [20], [21]. Precisely, we have the following theorem. For notational convenience, set

$$\mathcal{A} = \left\{ \nabla u : u \in W^{1,p}(\Omega), u - U \in W_0^{1,p}(\Omega) \right\},$$
$$\overline{\mathcal{A}} = \left\{ \nu = \{\nu_x\}_{x\in\Omega} : \nu \text{ is associated with a sequence in } \mathcal{A} \right\}.$$

THEOREM 1.1. *$\nu = \{\nu_x\}_{x\in\Omega}$ belongs to $\overline{\mathcal{A}}$ if and only if*
1. *$\nabla u(x) = \int_{\mathbf{M}^{m\times N}} A\, d\nu_x(A)$ for some $u \in \mathcal{A}$;*
2.

$$\varphi(\nabla u(x)) \le \int_{\mathbf{M}^{m\times N}} \varphi(A)\, d\nu_x(A)$$

for a.e. $x \in \Omega$ and every quasiconvex function φ with growth of order less than p at infinity;

3.

$$\int_\Omega \int_{\mathbf{M}^{m\times N}} |A|^p \; d\nu_x(A)\, dx < \infty.$$

The significance of the different variational principles discussed above and the relationship between them is established in the following important relaxation theorem [10], [12], [13], [21], [26].

THEOREM 1.2.

1.

$$\inf_{\mathcal{A}} J = \inf_{\mathcal{A}} QJ = \inf_{\overline{\mathcal{A}}} J;$$

2. *the last two infima are attained;*
3. *if $\nu = \{\nu_x\}_{x\in\Omega}$ is a minimizer for J in $\overline{\mathcal{A}}$, $u \in \mathcal{A}$ determined by*

$$\nabla u(x) = \int_{\mathbf{M}^{m\times N}} A\, d\nu_x(A), \quad \text{a.e. } x \in \Omega, \tag{1.4}$$

is a minimizer for QJ and

$$QW(x, u(x), \nabla u(x)) = \int_{\mathbf{M}^{m\times N}} W(x, u(x), A)\, d\nu_x(A), \quad \text{a.e. } x \in \Omega; \tag{1.5}$$

conversely, if $u \in \mathcal{A}$ is a minimizer for QJ and $\nu = \{\nu_x\}_{x\in\Omega}$ is such that (1.4) *and* (1.5) *hold, ν is a minimizer for J in $\overline{\mathcal{A}}$;*

4. *if ν is a minimizer for J in $\overline{\mathcal{A}}$,*

$$\operatorname{supp}(\nu_x) \subset \{W(x, u(x), \cdot) = QW(x, u(x), \cdot)\}, \quad \text{a.e. } x \in \Omega.$$

This result is the main motivation for our present analysis. It says that under the bounds (1.1), the functional J always admits minimizers in $\overline{\mathcal{A}}$ regardless of the convexity properties of the integrand W. We would like to derive necessary (equilibrium) conditions that these minimizers should verify, and that may help in understanding their properties. In particular, new, equivalent variational principles may be considered incorporating, as part of admissibility, these equilibrium conditions. Notice that by (1.4) and (1.5), those equilibrium conditions should essentially yield information on the quasiconvexification (or convexification) of functions. Our point of view, however, does not preclude any a priori information on the convex hulls of the integrands involved. We believe that our main contribution here is the way to understand and manipulate "variations" of Young measures (section 2) so that analytical equilibrium conditions can be derived more or less explicitly. On the other hand we provide a clear variational interpretation of some well-known facts about convex hulls of scalar functions.

We concentrate in this paper especially on the scalar ($m = 1$), one-dimensional ($N = 1$) case where the equilibrium conditions lead to specific and concrete conclusions. Notice that in this case quasiconvexity reduces to convexity in the usual sense.

The vector case or even the scalar higher-dimensional case are much more delicate. It is not clear whether equilibrium conditions may provide explicit, nontrivial, helpful information in either of these two situations.

Our analysis is close to that considered in [9]. In this work, the authors talk about Young measures minimizers and Young measures equilibria. The restrictions under which they pursue their analysis (nonlinear elasticity, vector case $m > 1$, rotation, and symmetry invariance) make hard to describe equilibrium as we intend to do here. They pay attention to variations of the domain, and get some helpful equilibrium criteria.

We proceed in several steps of increasing complexity. In section 3, we restrict attention to the scalar, one-dimensional case, where W depends only upon the derivative variable. Spatial dependence does not play a role in this case, and the analysis of the first and second variations lead to restrictions on the support of the minimizer ν. The next step (section 4) allows for the full generality for W depending on x and u as well, but still in the scalar, one-dimensional case. The final section consists of a description of the difficulties with the scalar, higher-dimensional case and the vector case. Some examples to illustrate our results are considered.

2. Variations of Young measures. The basic issue in deriving necessary conditions for equilibrium is to understand "variations" of Young measures: a continuous family of admissible Young measures depending on one parameter. Assume that

$$\nu^{(0)} = \left\{\nu_x^{(0)}\right\}_{x\in\Omega} \in \overline{\mathcal{A}}$$

is a minimizer for J in $\overline{\mathcal{A}}$

$$J(\nu^{(0)}) = \inf_{\overline{\mathcal{A}}} J.$$

Let $\{\nabla u_j\}$ be a generating sequence for $\nu^{(0)}$ so that $\{u_j\} \subset \mathcal{A}$ is minimizing for J in $\mathcal{A}$. For any sequence $\{w_j\}$ such that $w_j \in W_0^{1,p}(\Omega)$ and any t we may consider $\{u_j + tw_j\} \subset \mathcal{A}$. Therefore, the Young measure associated with $\{\nabla(u_j + tw_j)\}$, $\nu^{(t)} = \{\nu_x^{(t)}\}_{x\in\Omega}$, is admissible, and hence

$$J(\nu^{(t)}) \geq J(\nu^{(0)}) \quad \text{for all } t.$$

The function $g(t) = J(\nu^{(t)})$ has a minimum for $t = 0$. This is the classical idea of equilibrium. The issue here is how to manipulate this function g.

Let $\mu = \{\mu_x\}_{x\in\Omega}$ be the Young measure corresponding to the sequence of pairs $\{(\nabla u_j, \nabla w_j)\}$. Clearly, $\nu_x^{(t)}$ is defined via the formula

$$\langle \nu_x^{(t)}, \varphi\rangle = \int_{\mathbf{M}^{m\times N}\times\mathbf{M}^{m\times N}} \varphi(A_0 + tA_1)\, d\mu_x(A_0, A_1).$$

The clue to finding interesting equilibrium conditions is to consider the slicing measure decomposition for μ_x. In general terms, this is a way of decomposing any measure (not just product measures) supported on a product space. In this regard it may be considered as a generalization of Fubini's theorem. The following theorem can be found in [15].

THEOREM 2.1. *Let μ be a nonnegative, finite, Radon measure on $\mathbf{R}^{n+m}$, and let σ be its canonical projection onto $\mathbf{R}^n$ ($\sigma(E) = \mu(E \times \mathbf{R}^m)$). For σ-a.e $x \in \mathbf{R}^n$ there exists a probability measure ν_x on $\mathbf{R}^m$ such that*

i) *the map*

$$x \mapsto \int_{\mathbf{R}^m} f(x,y)\, d\nu_x(y)$$

is σ-measurable;

ii) *for every bounded, continuous function f,*

$$\int_{\mathbf{R}^{n+m}} f(x,y)\, d\mu(x,y) = \int_{\mathbf{R}^n} \left(\int_{\mathbf{R}^m} f(x,y)\, d\nu_x(y) \right) d\sigma(x).$$

A useful way to shorten the statement in ii) is to write

$$\mu(x,y) = \nu_x(y) \otimes \sigma(x).$$

An interesting remark is that this theorem can be the starting point of a treatment of Young measures based on slicing measure decompositions, as was the motivation in [15].

If we apply Theorem 2.1 to each μ_x, we can write

$$\mu_x(A_0, A_1) = \mu_x^{(A_0)}(A_1) \otimes \nu_x^{(0)}(A_0),$$

so that if $\Psi : \mathbf{M}^{m\times N} \times \mathbf{M}^{m\times N} \to \mathbf{R}$ is bounded and continuous, then

$$\int_{\mathbf{M}^{m\times N}\times\mathbf{M}^{m\times N}} \Psi(A_0,A_1)\, d\mu_x(A_0,A_1) = \int_{\mathbf{M}^{m\times N}} \left(\int_{\mathbf{M}^{m\times N}} \Psi(A_0,A_1)\, d\mu_x^{(A_0)}(A_1) \right) d\nu_x^{(0)}(A_0).$$

This same identity holds if Ψ is integrable with respect to μ_x; it is enough to consider bounded truncations of Ψ and conclude by monotone convergence.

Going back to our derivation of equilibrium conditions, we have that $g(t) = J(\nu^{(t)})$ becomes

$$\int_\Omega \int_{\mathbf{M}^{m\times N}} \left(\int_{\mathbf{M}^{m\times N}} W(x, u_0(x) + tw(x), A_0 + tA_1)\, d\mu_x^{(A_0)}(A_1) \right) d\nu_x^{(0)}(A_0)\, dx, \tag{2.1}$$

where, again,

$$\nabla u_0(x) = \int_{\mathbf{M}^{m\times N}} A\, d\nu_x^{(0)}(A) = \int_{\mathbf{M}^{m\times N}\times\mathbf{M}^{m\times N}} A_0\, d\mu_x(A_0,A_1)$$

and

$$\nabla w(x) = \int_{\mathbf{M}^{m\times N}\times\mathbf{M}^{m\times N}} A_1\, d\mu_x(A_0,A_1), \quad w \in W_0^{1,p}(\Omega). \tag{2.2}$$

Notice that the first projection of μ_x is precisely $\nu_x^{(0)}$ and the second one is the parametrized measure generated by $\{\nabla w_j\}$ so that ∇w in (2.2) is its weak limit.

Under suitable smoothness assumptions on W the function g is smooth and the equilibrium conditions (first and second variations) become $g'(0) = 0$ and $g''(0) \geq 0$. The condition on the first derivative (first variation) is written explicitly in the form (the hypotheses assumed on W allow derivation under the integral sign)

$$\int_\Omega \int_{\mathbf{M}^{m\times N}} \left[\frac{\partial W}{\partial \lambda}(x, u_0(x), A_0)\, w(x) + \frac{\partial W}{\partial A}(x, u_0(x), A_0) \int_{\mathbf{M}^{m\times N}} A_1\, d\mu_x^{(A_0)}(A_1) \right] d\nu_x^{(0)}(A_0)\, dx = 0$$

for any such measure μ. If we let

$$\Upsilon_1(x,A_0)=\int_{\mathbf{M}^{m\times N}} A_1\,d\mu_x^{(A_0)}(A_1) \tag{2.3}$$

so that

$$\nabla w(x)=\int_{\mathbf{M}^{m\times N}\times\mathbf{M}^{m\times N}} A_1\,d\mu_x(A_0,A_1)=\int_{\mathbf{M}^{m\times N}}\Upsilon_1(x,A_0)\,d\nu_x^{(0)}(A_0) \tag{2.4}$$

and $w\in W_0^{1,p}(\Omega)$, the first necessary condition may be put in the following way:

$$\begin{aligned}\int_\Omega\int_{\mathbf{M}^{m\times N}}\Big[&\frac{\partial W}{\partial\lambda}\,(x,u_0(x),A_0)\,w(x)\\ &+\frac{\partial W}{\partial A}\,(x,u_0(x),A_0)\,\Upsilon_1(x,A_0)\Big]\,d\nu_x^{(0)}(A_0)\,dx=0.\end{aligned} \tag{2.5}$$

The connection between w and Υ_1 is given in (2.4). Equation (2.5) should be valid for all possible Υ_1 coming from all admissible μ corresponding to all pairs $\{(\nabla u_j,\nabla w_j)\}$ as indicated. The analysis and characterization of the possible test fields Υ_1 are vital.

All this formalism based on the slicing measure decomposition is nothing more than a convenient way of manipulating the function

$$g(t)=\lim_{j\to\infty}\int_\Omega W(x,u_j(x)+tw_j(x),\nabla u_j(x)+t\nabla w_j(x))\,dx.$$

As we have seen, the derivative of g at the origin can be computed by differentiating formally the above limit, interchanging the limit and the derivative

$$\begin{aligned}g'(0)&=\int_\Omega\int_{\mathbf{M}^{m\times N}\times\mathbf{M}^{m\times N}}\Big[\frac{\partial W}{\partial\lambda}\,(x,u_0(x),A_0)\,w(x)\\ &\qquad+\frac{\partial W}{\partial A}\,(x,u_0(x),A_0)\,A_1\Big]\,d\mu_x(A_0,A_1)\,dx\\ &=\lim_{j\to\infty}\int_\Omega\Big[\frac{\partial W}{\partial\lambda}\,(x,u_j(x),\nabla u_j(x))\,w_j(x)\\ &\qquad+\frac{\partial W}{\partial A}\,(x,u_j(x),\nabla u_j(x))\,\nabla w_j(x)\Big]\,dx.\end{aligned}$$

For the second variation we get something similar but more complicated to write. The condition is

$$\begin{aligned}\int_\Omega\int_{\mathbf{M}^{m\times N}}\Big[&\frac{\partial^2 W}{\partial\lambda^2}\,(x,u_0(x),A_0)\,w(x)w(x)\\ &+2\frac{\partial^2 W}{\partial\lambda\partial A}\,(x,u_0(x),A_0)\,w(x)\Upsilon_1(x,A_0)\\ &+\frac{\partial^2 W}{\partial A^2}\,(x,u_0(x),A_0)\,\Upsilon_2(x,A_0)\Big]\,d\nu_x^{(0)}(A_0)\,dx\ge 0,\end{aligned} \tag{2.6}$$

where Υ_1 and w are as in (2.3) and (2.4) and $\Upsilon_2(x,A_0)$ is the tensor of second moments of $\mu_x^{(A_0)}$, in short form,

$$\Upsilon_2(x,A_0)=\int_{\mathbf{M}^{m\times N}} A_1A_1\,d\mu_x^{(A_0)}(A_1). \tag{2.7}$$

Understanding (2.6) would require comprehending the conditions on the tensor of second moments Υ_2 for all admissible μ as before, and its relationship to Υ_1 and w.

We pursue in subsequent sections the analysis of (2.5) and (2.6) in cases of increasing complexity. This amounts to characterizing the test fields Υ_1 and Υ_2 in each case. Remember that these are associated with the slicing decomposition of μ and this in turn is the Young measure associated to pairs of sequences $\{(\nabla u_j, \nabla w_j)\}$ where $u_j \in \mathcal{A}$ and $w_j \in W_0^{1,p}(\Omega)$. Our main task consists of understanding restrictions on Young measures associated with pairs of gradients.

3. The simplest problem. We start by looking at the equilibrium necessary conditions for the simplest case: the scalar, one-dimensional problem when the integrand W depends only on the derivative. Our functional is

$$J(\nu) = \int_\Omega \int_{\mathbf{R}} W(A)\, d\nu_x(A)\, dx,$$

where $\Omega = (0,1)$, W is assumed to be sufficiently smooth and verifying

$$\begin{aligned} c\left(|A|^p - 1\right) \leq W(A) &\leq C\left(|A|^p + 1\right), \\ |W'(A)| &\leq C\left(|A|^{p-1} + 1\right), \\ |W''(A)| &\leq C\left(|A|^{p-2} + 1\right) \end{aligned} \tag{3.1}$$

for $p > 1$ or $p > 2$, $0 < c < C$. The admissibility conditions on ν reduce in this case to requiring

$$\int_\Omega \int_{\mathbf{R}} A\, d\nu_x(A)\, dx = \lambda, \quad \int_\Omega \int_{\mathbf{R}} |A|^p\, d\nu_x(A)\, dx < \infty.$$

The number λ is given and determined by the boundary conditions. Notice that quasiconvexity reduces to convexity and therefore Jensen's inequality does not place any restriction on ν. Let $\overline{\mathcal{A}}$ denote again the class of admissible measures ν.

By Theorem 1.2, J admits minimizers in $\overline{\mathcal{A}}$. An elementary observation is that, since there is no dependence on x, there always exist minimizers $\nu^{(0)}$ not depending on x. Said differently, if $\nu^{(0)} = \{\nu_x^{(0)}\}_{x\in\Omega}$ is a minimizer, the probability measure obtained by "averaging" (see [21]) will also be a minimizer. The converse is also true: if $\nu^{(0)}$ is a minimizer with no spatial dependence, any admissible Young measure whose "average" is $\nu^{(0)}$ will also be a minimizer. The real issue is to find the homogeneous minimizers. We simplify accordingly the setup of the problem:

$$\begin{aligned} J(\nu) &= \int_{\mathbf{R}} W(A)\, d\nu(A), \quad \nu \in \overline{\mathcal{A}}, \\ \overline{\mathcal{A}} &= \left\{ \nu : \lambda = \int_{\mathbf{R}} A\, d\nu(A), \int_{\mathbf{R}} |A|^p\, d\nu(A) < \infty \right\}. \end{aligned}$$

Clearly, we are looking for $CW(\lambda)$, where CW is the convexification of W and $\nu^{(0)}$ is such that

$$CW(\lambda) = J(\nu^{(0)}).$$

In this case the restriction on the pth powers in $\overline{\mathcal{A}}$ does not play a real role. Indeed, by Carathéodory's theorem (see, for instance, [13]), we know that

$$CW(\lambda) = \inf\left\{ tW(\alpha) + (1-t)W(\beta) : t\alpha + (1-t)\beta = \lambda, t \in [0,1] \right\}.$$

Under the coerciveness assumption (3.1) there always exists a minimizer of the form

$$\nu^{(0)} = t\delta_\alpha + (1-t)\delta_\beta.$$

We would like, however, to find necessary equilibrium conditions that must verify all minimizers, not only those supported on two points, and possibly some other elements in $\overline{\mathcal{A}}$.

In this case the first necessary condition (2.5) becomes (recall that we do not have dependence on x)

$$\int_{\mathbf{R}} W'(A_0)\Upsilon_1(A_0)\, d\nu^{(0)}(A_0) = 0 \tag{3.2}$$

for all first moments

$$\Upsilon_1(A_0) = \int_{\mathbf{R}} A_1\, d\mu^{(A_0)}(A_1),$$

where

$$\mu(A_0, A_1) = \mu^{(A_0)}(A_1) \otimes \nu^{(0)}(A_0)$$

is the slicing decomposition of μ generated by a sequence $\left\{(u_j', w_j')\right\}$, $u_j, w_j \in W^{1,p}(\Omega)$, $u_j(1) - u_j(0) = \lambda$, $w_j(1) = w_j(0)$. Let z_j denote the pairs $\left\{(u_j', w_j')\right\}$ so that $z_j : \Omega \subset \mathbf{R} \to \mathbf{R}^2$. $\{z_j\}$ is not a sequence of scalar-valued functions. Yet, since the dimension of Ω is one ($N = 1$), Jensen's inequality does not place any restriction on Young measures generated by derivatives because quasiconvexity collapses to convexity in this case as well (for a full discussion see [13] and [26]). Accordingly, the only constraints on μ to be generated by such sequence of pairs are

$$\begin{aligned} &\int_{\mathbf{R}\times\mathbf{R}} A_0\, d\mu(A_0, A_1) = \lambda, \\ &\int_{\mathbf{R}\times\mathbf{R}} A_1\, d\mu(A_0, A_1) = 0, \\ &\int_{\mathbf{R}\times\mathbf{R}} |A_i|^p\, d\mu(A_0, A_1) < \infty, \quad i = 0, 1. \end{aligned} \tag{3.3}$$

The restrictions that we obtain on the test fields $\Upsilon_1(A_0)$ are

$$\begin{aligned} &\int_{\mathbf{R}} \Upsilon_1(A_0)\, d\nu^{(0)}(A_0) = 0, \\ &\int_{\mathbf{R}} |\Upsilon_1(A_0)|^p\, d\nu^{(0)}(A_0) < \infty. \end{aligned} \tag{3.4}$$

If (3.4) holds, the measure

$$\mu(A_0, A_1) = \delta_{\Upsilon_1(A_0)}(A_1) \otimes \nu^{(0)}(A_0)$$

verifies (3.3).

PROPOSITION 3.1 (first necessary condition). *Equation* (3.2) *is equivalent to the existence of a constant* $k \in \mathbf{R}$ *such that*

$$\operatorname{supp}(\nu^{(0)}) \subset \{W' = k\}.$$

Proof. The proof reduces to examining (3.2) under the constraints (3.4). Let $\Upsilon(A_0)$ be any function such that

$$\int_{\mathbf{R}} |\Upsilon(A_0)|^p\, d\nu^{(0)}(A_0) < \infty.$$

In this case, $\Upsilon_1(A_0) = \Upsilon(A_0) - \overline{\Upsilon}$ satisfies (3.4), where

$$\overline{\Upsilon} = \int_{\mathbf{R}} \Upsilon(A)\, d\nu^{(0)}(A).$$

Put

$$k = \int_{\mathbf{R}} W'(A)\, d\nu^{(0)}(A) \in \mathbf{R}.$$

After some manipulation (3.2) becomes, for this choice of Υ_1,

$$\int_{\mathbf{R}} (W'(A_0) - k)\, \Upsilon(A_0)\, d\nu^{(0)}(A_0) = 0$$

for all such Υ. To conclude, take any Υ such that

$$\Upsilon(A) = \frac{W'(A) - k}{|W'(A) - k|^{(p-2)/(p-1)}}$$

if $W'(A) - k \neq 0$, which is admissible.

The converse is immediate. □

The treatment of the second variation follows along the same lines. Equation (2.6) is simplified to

$$\int_{\mathbf{R}} W''(A_0)\Upsilon_2(A_0)\, d\nu^{(0)}(A_0) \geq 0 \tag{3.5}$$

for all second moments

$$\Upsilon_2(A_0) = \int_{\mathbf{R}} (A_1)^2\ d\mu^{(A_0)}(A_1),$$

where, again,

$$\mu(A_0, A_1) = \mu^{(A_0)}(A_1) \otimes \nu^{(0)}(A_0).$$

The constraints on $\Upsilon_2(A_0)$ simply are

$$\Upsilon_2 \geq 0, \quad \int_{\mathbf{R}} |\Upsilon_2(A_0)|^{p/2}\ d\nu^{(0)}(A_0) < \infty. \tag{3.6}$$

This is elementary to derive. Notice that in order to consider the second variation we take $p \geq 2$ so that the power function with exponent $p/2$ is convex. If (3.6) holds for Υ_2, the measure

$$\mu(A_0, A_1) = \frac{1}{2}\left(\delta_{\sqrt{\Upsilon_2(A_0)}}(A_1) + \delta_{-\sqrt{\Upsilon_2(A_0)}}(A_1)\right) \otimes \nu^{(0)}(A_0)$$

satisfies (3.3).

PROPOSITION 3.2 (second necessary condition). *Equation* (3.5) *is equivalent to*

$$\operatorname{supp}(\nu^{(0)}) \subset \{W'' \geq 0\}.$$

Proof. It is enough to take

$$\Upsilon_2(A_0) = \chi_{\{W''<0\}}(A_0)$$

(χ stands for the characteristic function of the corresponding set) in (3.5). This choice is permitted according to (3.6). □

We gathered the two preceding propositions in the following statement.

THEOREM 3.3. *Let $\nu^{(0)}$ be a minimizer for J in $\overline{\mathcal{A}}$. There exists a constant k such that*

$$\operatorname{supp}(\nu^{(0)}) \subset \{W' = k\} \cap \{W'' \geq 0\}\,. \tag{3.7}$$

It is interesting to notice that these conditions on the support of minimizers do not determine the minimizers themselves so that there might be many equilibrium Young measures. Consider the typical example

$$W(A) = \left(A^2 - 1\right)^2, \quad p = 4,$$

and take $\lambda = 0$. A few easy computations show that (3.7) is verified by a one-parameter family of probability measures

$$\nu_\alpha = t(\alpha)\delta_\alpha + (1 - t(\alpha))\,\delta_{\beta(\alpha)}, \tag{3.8}$$

where $\alpha \in [1/\sqrt{3}, t_0]$, $t_0^3 - t_0 = 2/(3\sqrt{3})$, $\beta^3 - \beta = \alpha^3 - \alpha$, $\beta \in [-t_0, -1/\sqrt{3}]$, and t is determined by $0 = t\alpha + (1-t)\beta$. Of all those, the unique minimizer corresponds to $\alpha = 1$

$$\nu_1 = \frac{1}{2}\delta_1 + \frac{1}{2}\delta_{-1}.$$

We also notice, in view of this example, that there are variations not captured by our scheme in section 2. The family given in (3.8) cannot be reproduced by a variation of the form discussed in section 2. The reason is that the one-parameter family of probability measures obtained as variations of $\nu^{(0)}$, $\nu^{(t)}$, comes from a measure on the product $\mathbf{R} \times \mathbf{R}$, μ, and is such that the two projections are $\nu^{(0)}$ and any other admissible Young measure $\nu \in \overline{\mathcal{A}}$. In this simplified situation, unless $\nu = \nu^{(0)}$, μ has to be a product measure and therefore $\nu^{(t)}$ is supported in four different points. It can never be supported in two points, as is the case with the probabilities in (3.8).

Our conclusions for the case we are analyzing in this section are elementary and well known. They can be easily derived by basic geometric and/or analytical arguments. Indeed, under our coerciveness assumptions on W,

$$CW(\lambda) = \min\left\{tW(\alpha) + (1-t)W(\beta) : \lambda = t\alpha + (1-t)\beta\right\}. \tag{3.9}$$

By Theorem 3.3,

$$\begin{aligned} CW(\lambda) = \min\{&tW(\alpha) + (1-t)W(\beta) : \lambda = t\alpha + (1-t)\beta,\\ &W'(\alpha) = W'(\beta), W''(\alpha) \geq 0, W''(\beta) \geq 0\}\,. \end{aligned}$$

The equilibrium condition $W'(\alpha) = W'(\beta)$ is actually recovered by solving the constrained problem (3.9). This is a good calculus exercise. After some algebra we find

$$CW(\lambda) = tW(\alpha) + (1-t)W(\beta),$$

where

$$\begin{gathered} \frac{W(\alpha) - W(\beta)}{\alpha - \beta} = W'(\alpha) = W'(\beta),\\ W''(\alpha) \geq 0, \quad W''(\beta) \geq 0,\\ \lambda = t\alpha + (1-t)\beta. \end{gathered}$$

It is remarkable that these facts can be interpreted as some sort of equilibrium criterion of a clear variational nature.

4. The scalar, one-dimensional case. We consider in this section the general scalar, one-dimensional case. The functional J is

$$J(\nu) = \int_\Omega \int_{\mathbf{R}} W(x, u(x), A)\, d\nu_x(A)\, dx,$$

where $\Omega = (0,1)$ and the set of admissible measures, ν, is

$$\overline{\mathcal{A}} = \left\{ \nu = \{\nu_x\}_{x\in\Omega} : u'(x) = \int_{\mathbf{R}} A\, d\nu_x(A), u \in W^{1,p}(\Omega), u(0) = 0, \right.$$
$$\left. u(1) = \lambda, \int_\Omega \int_{\mathbf{R}} |A|^p\, d\nu_x(A)\, dx < \infty \right\}.$$

W is assumed to satisfy the coerciveness hypothesis (1.1), and λ is given a priori. The first necessary condition (2.5) becomes in this situation

$$\int_\Omega \int_{\mathbf{R}} \left[\frac{\partial W}{\partial \lambda}(x, u_0(x), A_0)\, w(x) + \frac{\partial W}{\partial A}(x, u_0(x), A_0)\, \Upsilon_1(x, A_0) \right] d\nu_x^{(0)}(A_0)\, dx = 0. \tag{4.1}$$

Remember that

$$u_0'(x) = \int_{\mathbf{R}} A_0\, d\nu_x^{(0)}(A_0).$$

Υ_1 and w are given by

$$\Upsilon_1(x, A_0) = \int_{\mathbf{R}} A_1\, d\mu_x^{(A_0)}(A_1),$$
$$w'(x) = \int_{\mathbf{R}\times\mathbf{R}} A_1\, d\mu_x(A_0, A_1) = \int_{\mathbf{R}} \Upsilon_1(x, A_0)\, d\nu_x^{(0)}(A_0),$$

where $\mu = \{\mu_x\}_{x\in\Omega}$ can be any measure supported on $\mathbf{R}\times\mathbf{R}$ with first projection $\nu^{(0)} = \{\nu_x^{(0)}\}_{x\in\Omega}$ and verifying

$$\int_\Omega \int_{\mathbf{R}\times\mathbf{R}} A_0\, d\mu_x(A_0, A_1)\, dx = \lambda,$$
$$\int_\Omega \int_{\mathbf{R}\times\mathbf{R}} A_1\, d\mu_x(A_0, A_1)\, dx = 0, \tag{4.2}$$
$$\int_\Omega \int_{\mathbf{R}\times\mathbf{R}} |A_i|^p\, d\mu_x(A_0, A_1)\, dx < \infty, \quad i = 0, 1.$$

The restrictions that we obtain on the test fields $\Upsilon_1(x, A_0)$ are

$$\int_\Omega \int_{\mathbf{R}} \Upsilon_1(x, A_0)\, d\nu_x^{(0)}(A_0)\, dx = 0,$$
$$\int_\Omega \int_{\mathbf{R}} |\Upsilon_1(x, A_0)|^p\, d\nu_x^{(0)}(A_0)\, dx < \infty. \tag{4.3}$$

If (4.3) holds, the family of measures

$$\mu_x(A_0, A_1) = \delta_{\Upsilon_1(x,A_0)}(A_1) \otimes \nu_x^{(0)}(A_0)$$

verifies (4.2) trivially.

Relying on this description of the test fields Υ_1 we can identify the first necessary condition.

PROPOSITION 4.1. *Let*

$$F(x) = \int_{\mathbf{R}} \frac{\partial W}{\partial A}\,(x, u_0(x), A_0)\; d\nu_x^{(0)}(A_0),$$
$$G(x) = \int_{\mathbf{R}} \frac{\partial W}{\partial \lambda}\,(x, u_0(x), A_0)\; d\nu_x^{(0)}(A_0)$$

for $x \in \Omega$. Equation (4.1) *is equivalent to*

1. $G - F' = 0$ *(in a weak sense);*
2. $$\operatorname{supp}(\nu_x^{(0)}) \subset \left\{ \frac{\partial W}{\partial A}(x, u_0(x), \cdot) = F(x) \right\}, \quad \text{a.e. } x \in \Omega.$$

Proof. Let $\Upsilon(x, A_0)$ be any function such that

$$\int_\Omega \int_{\mathbf{R}} |\Upsilon(x, A_0)|^p \; d\nu_x^{(0)}(A_0)\, dx < \infty.$$

Consider

$$\Upsilon_1(x, A_0) = \eta(x)\left[\Upsilon(x, A_0) - \overline{\Upsilon}(x)\right] + \xi'(x),$$

which is admissible in (4.1) for any arbitrary smooth function η and any test function ξ, where

$$\overline{\Upsilon}(x) = \int_{\mathbf{R}} \Upsilon(x, A_0)\, d\nu_x^{(0)}(A_0).$$

Equation (4.1) reduces in this case to

$$\int_\Omega \eta(x) \int_{\mathbf{R}} \frac{\partial W}{\partial A}(x, u_0(x), A_0)\left[\Upsilon(x, A_0) - \overline{\Upsilon}(x)\right]\; d\nu_x^{(0)}(A_0)\; dx$$
$$+ \int_\Omega \xi(x)\,(G(x) - F'(x))\; dx = 0.$$

Taking $\eta \equiv 0$, the arbitrariness of ξ leads to 1 in the statement of the proposition, and taking $\xi \equiv 0$, the arbitrariness of η implies

$$\int_{\mathbf{R}} \frac{\partial W}{\partial A}(x, u_0(x), A_0)\left[\Upsilon(x, A_0) - \overline{\Upsilon}(x)\right]\; d\nu_x^{(0)}(A_0) = 0, \quad \text{a.e. } x \in \Omega.$$

This identity can be rewritten as

$$\int_{\mathbf{R}} \left[\frac{\partial W}{\partial A}(x, u_0(x), A_0) - \int_{\mathbf{R}} \frac{\partial W}{\partial A}(x, u_0(x), A)\, d\nu_x^{(0)}(A)\right] \Upsilon(x, A_0)\, d\nu_x^{(0)}(A_0) = 0.$$

If we choose $\Upsilon(x, A_0)$ such that

$$\Upsilon(x, A_0) = \frac{\dfrac{\partial W}{\partial A}(x, u_0(x), A_0) - \displaystyle\int_{\mathbf{R}} \frac{\partial W}{\partial A}(x, u_0(x), A)\, d\nu_x^{(0)}(A)}{\left|\dfrac{\partial W}{\partial A}(x, u_0(x), A_0) - \displaystyle\int_{\mathbf{R}} \frac{\partial W}{\partial A}(x, u_0(x), A)\, d\nu_x^{(0)}(A)\right|^{(p-1)/(p-2)}}$$

whenever the denominator does not vanish, we are led to 2.

The converse is immediate. ∎

For the second variation, when $p \geq 2$ we get

$$\begin{aligned}\int_\Omega \int_{\mathbf{R}} \Big[& \frac{\partial^2 W}{\partial \lambda^2}(x, u_0(x), A_0)\, w(x) w(x) \\ & + 2 \frac{\partial^2 W}{\partial \lambda \partial A}(x, u_0(x), A_0)\, w(x) \Upsilon_1(x, A_0) \\ & + \frac{\partial^2 W}{\partial A^2}(x, u_0(x), A_0)\, \Upsilon_2(x, A_0) \Big]\, d\nu_x^{(0)}(A_0)\, dx \geq 0,\end{aligned} \tag{4.4}$$

where Υ_1 and w are as before, and $\Upsilon_2(x, A_0)$ is the second moment of $\mu_x^{(A_0)}$

$$\Upsilon_2(x, A_0) = \int_{\mathbf{R}} A_1 A_1\, d\mu_x^{(A_0)}(A_1).$$

The relationship between Υ_1 and Υ_2 is clear. Indeed, the constraints on the pair (Υ_1, Υ_2) to be admissible are

$$\begin{gathered}\int_\Omega \int_{\mathbf{R}} \Upsilon_1(x, A)\, d\nu_x^{(0)}(A)\, dx = 0, \\ \int_\Omega \int_{\mathbf{R}} |\Upsilon_2(x, A)|^{p/2}\, d\nu_x^{(0)}(A)\, dx < \infty, \\ \Upsilon_2 \geq \Upsilon_1^2 \geq 0.\end{gathered}$$

To show this, it is enough to consider the family of probability measures

$$\mu_x(A_0, A_1) = \left[\delta_{\Upsilon_1(x, A_0)}(A_1) + \frac{1}{2} \left(\delta_{\Upsilon(x, A_0)}(A_1) + \delta_{-\Upsilon(x, A_0)}(A_1) \right) \right] \otimes \nu_x^{(0)}(A_0),$$

where $\Upsilon(x, A_0) = \sqrt{\Upsilon_2(x, A_0) - \Upsilon_1(x, A_0)^2}$, which is clearly admissible.

PROPOSITION 4.2. *Equation* (4.4) *is equivalent to*

1.

$$\operatorname{supp}(\nu_x^{(0)}) \subset \left\{ \frac{\partial^2 W}{\partial A^2}(x, u_0(x), \cdot) \geq 0 \right\}, \quad \textit{a.e. } x \in \Omega;$$

2. *for all admissible* Υ_1

$$\begin{aligned}\int_\Omega \int_{\mathbf{R}} \Big[& \frac{\partial^2 W}{\partial \lambda^2}(x, u_0(x), A_0)\, w(x) w(x) \\ & + 2 \frac{\partial^2 W}{\partial \lambda \partial A}(x, u_0(x), A_0)\, w(x) \Upsilon_1(x, A_0) \\ & + \frac{\partial^2 W}{\partial A^2}(x, u_0(x), A_0)\, \Upsilon_1^2(x, A_0) \Big]\, d\nu_x^{(0)}(A_0)\, dx \geq 0,\end{aligned}$$

where, again,

$$w'(x) = \int_{\mathbf{R}} \Upsilon_1(x, A_0)\, d\nu_x^{(0)}(A_0)$$

is a test function.

The proof consists in examining (4.4), keeping in mind the restrictions on the pairs (Υ_1, Υ_2). Notice that part 1 is the usual condition yielding further information on the support of $\nu^{(0)}$. Condition 2 expresses the interplay of the different partial derivatives of W. It cannot be easily simplified.

THEOREM 4.3. *Let $\nu^{(0)}$ be a minimizer for J in $\overline{\mathcal{A}}$ and put*

$$F(x) = \int_{\mathbf{R}} \frac{\partial W}{\partial A}\,(x, u_0(x), A_0)\; d\nu_x^{(0)}(A_0),$$
$$G(x) = \int_{\mathbf{R}} \frac{\partial W}{\partial \lambda}\,(x, u_0(x), A_0)\; d\nu_x^{(0)}(A_0)$$

for $x \in \Omega$, where

$$u_0'(x) = \int_{\mathbf{R}} A_0\, d\nu_x^{(0)}(A_0).$$

Then $G - F' = 0$ in a weak sense and

$$\operatorname{supp}(\nu_x^{(0)}) \subset \left\{ \frac{\partial W}{\partial A}(x, u_0(x), \cdot) = F(x) \right\} \cap \left\{ \frac{\partial^2 W}{\partial A^2}(x, u_0(x), \cdot) \geq 0 \right\}, \quad \text{a.e. } x \in \Omega.$$

Let us examine several examples. Take first

$$W(x, A) = \left(A^2 + 2x - 1\right)^2, \quad x \in \Omega = (0, 1),$$

independent of u. In this case $G \equiv 0$ and F is constant. The Young measures equilibria, $\nu = \{\nu_x\}_{x \in \Omega}$, satisfy

$$\operatorname{supp}(\nu_x) \subset \left\{ \frac{\partial W}{\partial A}(x, \cdot) = k \right\} \cap \left\{ \frac{\partial^2 W}{\partial A^2}(x, \cdot) \geq 0 \right\}$$

for some constant k that varies with ν. For this example

$$\operatorname{supp}(\nu_x) \subset \left\{ A \in \mathbf{R} : 4\left(A^2 + 2x - 1\right) A = k \right\} \cap \left\{ A \in \mathbf{R} : 3A^2 + 2x - 1 \geq 0 \right\}.$$

Since W is a polynomial of degree four, the conditions on the support of ν imply that all Young measures equilibria are a convex combination of two deltas

$$\nu_x = t(x)\delta_{\alpha_1(x)} + (1 - t(x))\delta_{\alpha_2(x)},$$

where for some k

$$\begin{aligned} 4\left(\alpha_i(x)^2 + 2x - 1\right)\alpha_i(x) = k, \quad i = 1, 2, \\ 3\alpha_i(x)^2 + 2x - 1 \geq 0, \quad i = 1, 2, \end{aligned} \tag{4.5}$$

and we take $\alpha_1 \leq \alpha_2$. Finally, $t(x) \in [0, 1]$ is determined by the condition

$$\int_\Omega \left[t(x)\alpha_1(x) + (1 - t(x))\alpha_2(x)\right] dx = \lambda. \tag{4.6}$$

The Young measure minimizer would correspond to particular values of k and $t(x)$ in such a way that we can set up a new, equivalent (in the sense that it shares the same minimizers) variational principle where pairs $(k, t(x))$ are the competing objects. In fact set

$$I(k, t) = \int_\Omega \left[t(x)\left(\alpha_1(x)^2 + 2x - 1\right)^2 + (1 - t(x))\left(\alpha_2(x)^2 + 2x - 1\right)^2 \right] dx,$$

where α_i are determined uniquely by (4.5) and the function t verifies the constraint (4.6).

We claim that if

$$|\lambda| \leq \int_0^{1/2} \sqrt{1 - 2x}\, dx, \tag{4.7}$$

the minimizers are of the form

$$\nu_x = \delta_0, \quad x \geq \frac{1}{2},$$
$$\nu_x = t(x)\delta_{\sqrt{1-2x}} + (1-t(x))\delta_{-\sqrt{1-2x}}, \quad x \leq \frac{1}{2},$$

where $t(x)$ is such that

$$\int_0^{1/2} t(x)\sqrt{1-2x}\,dx = \frac{1}{2}\left(\lambda + \int_0^{1/2}\sqrt{1-2x}\,dx\right).$$

Observe that (4.7) permits us to have $t(x) \in [0,1]$. The value of the minimum, m, is

$$m = \int_{1/2}^1 (2x-1)^2\,dx.$$

These facts are easy to derive because α_i have been chosen to be pointwise a minimum of W.

Another example is

$$W(x,u,A) = \left(A^2-1\right)^2 + (u-f(x))^2,$$

where f is a given, smooth function. All of our conditions hold for $p=4$. In this case the conditions for equilibrium are

$$F(x) = \int_{\mathbf{R}} 4A\left(A^2-1\right)\,d\nu_x(A),$$
$$G(x) = 2\left(u(x)-f(x)\right),$$
$$u'(x) = \int_{\mathbf{R}} A\,d\nu_x(A),$$
$$\operatorname{supp}(\nu_x) \subset \left\{4A(A^2-1) = F(x)\right\} \cap \left\{\left(-\infty, -\frac{1}{\sqrt{3}}\right] \cup \left[\frac{1}{\sqrt{3}}, \infty\right)\right\}.$$

The equation $G - F' = 0$ is

$$2\left(u(x)-f(x)\right) - F' = 0,$$

so that

$$u(x) = f(x) + \frac{1}{2}F'(x)$$

and

$$u'(x) = f'(x) - \frac{1}{2}F''(x)$$

must be the first moment of ν_x. Let I_k stand for the convex hull of the set of roots of the equation

$$4A\left(A^2-1\right) = k.$$

For the field F we get the restrictions

$$\begin{aligned} &f'(x) + \frac{1}{2}F''(x) \in I_{F(x)}, \\ &F'(0) = -2f(0), \\ &F'(1) = 2\left(\lambda - f(1)\right). \end{aligned} \tag{4.8}$$

For any such F there exists a unique $\nu = \{\nu_x\}_{x\in\Omega}$, equilibrium measure, which is of the form

$$\nu_x = t(x)\delta_{\alpha_1(x)} + (1-t(x))\delta_{\alpha_2(x)},$$

where α_i are uniquely determined by the conditions

$$\begin{gathered} 4\alpha_i\left(\alpha_i^2-1\right) = F, \quad i=1,2, \\ \alpha_1 \in \left(-\infty, -\frac{1}{\sqrt{3}}\right], \ \alpha_2 \in \left[\frac{1}{\sqrt{3}}, \infty\right), \end{gathered} \tag{4.9}$$

and t is chosen so that

$$t(x)\alpha_1(x) + (1-t(x))\alpha_2(x) = f'(x) + \frac{1}{2}F''(x). \tag{4.10}$$

Consider the functional

$$I(F) = \int_\Omega \left[t(x)\left(\alpha_1(x)^2-1\right)^2 + (1-t(x))\left(\alpha_2(x)^2-1\right)^2 + \frac{1}{4}\left(F'(x)\right)^2\right] dx,$$

under the admissibility conditions (4.8). This problem has a minimizer, F_0, that corresponds exactly to the minimizer $\nu^{(0)} = \{\nu_x^{(0)}\}_{x\in\Omega}$ of J. Notice that by the basic properties of $I_{F(x)}$ we should have

$$4\left(f'(x) - \frac{1}{2}F''(x)\right)\left(\left(f'(x) - \frac{1}{2}F''(x)\right)^2 - 1\right) = F(x)$$

whenever $|F(x)| > 8/(3\sqrt{3})$. In a subsequent work, [25], we will pursue the analysis of this type of variational principle and how it can help in finding $\nu^{(0)}$.

5. Final remarks. We would like to point out the main difficulties with the scalar, higher-dimensional and the vector cases. Concerning the latter, the difficulties are clear. Jensen's inequality with respect to quasiconvex functions places a real and fundamental restriction on Young measures generated by gradients that we do not really understand. What is quite surprising is that almost the same troubles arise in the scalar, higher-dimensional case even though quasiconvexity reduces to convexity in this case.

The whole point is that the class of variations we are considering are associated to Young measures, μ, generated by "pairs" of sequences of gradients $\{(\nabla u_j, \nabla w_j)\}$. Even though each u_j and w_j are scalar-valued, the pair is vector-valued, so that there are restrictions on μ in the form of Jensen's inequality for quasiconvex functions unless the domain Ω is one dimensional. The description of the test fields Υ_1 is the same:

$$\begin{aligned} \Upsilon_1(x, A_0) &= \int_{\mathbf{R}^2} A_1 \, d\mu_x^{(A_0)}(A_1), \\ \mu_x(A_0, A_1) &= \mu_x^{(A_0)}(A_1) \otimes \nu_x^{(0)}(A_0). \end{aligned}$$

But now the constraints we have on Υ_1 are not simply

$$\begin{gathered} \int_{\mathbf{R}^2} \Upsilon_1(x, A_0)\, d\nu_x^{(0)}(A_0) = \nabla w(x), \quad w \in W_0^{1,p}(\Omega), \\ \int_\Omega \int_{\mathbf{R}^2} |\Upsilon_1(x, A_0)|^p \, d\nu_x^{(0)}(A_0)\, dx < \infty, \end{gathered}$$

but also the fact that μ has to be generated by gradients. As we know, this is a profound restriction on μ and the analysis should go deeper than the one carried out in this work.

REFERENCES

[1] E. Acerbi and N. Fusco, *Semicontinuity problems in the calculus of variations*, Arch. Rational Mech. Anal., 86 (1984), pp. 125–145.

[2] E. J. Balder, *A general approach to lower semicontinuity and lower closure in optimal control theory*, SIAM J. Control Optim., 22 (1984), pp. 570–598.

[3] J. M. Ball, *Convexity conditions and existence theorems in nonlinear elasticity*, Arch. Rational Mech. Anal., 63 (1977), pp. 337–403.

[4] J. M. Ball, *Minimizers and the Euler–Lagrange equations*, in Trends and Applications of Pure Mathematics to Mechanics, P. G. Ciarlet and M. Roseau, eds., Springer-Verlag, Berlin, pp. 1–4.

[5] J. M. Ball, *A version of the fundamental theorem for Young measures*, in PDE's and Continuum Models of Phase Transitions, Lecture Notes in Phys., 344, M. Rascle, D. Serre, and M. Slemrod, eds., Springer-Verlag, Berlin, 1989, pp. 207–215.

[6] J. M. Ball and R. D. James, *Fine phase mixtures as minimizers of energy*, Arch. Rational Mech. Anal., 100 (1987), pp. 13–52.

[7] J. M. Ball and R. D. James, *Proposed experimental tests of a theory of fine microstructure and the two well problem*, Philos. Trans. Roy. Soc. London Ser. A, 338 (1992), pp. 389–450.

[8] E. Bonnetier and C. Conca, *Approximation of Young measures by functions and application to a problem of optimal design for plates with variable thickness*, Proc. Roy. Soc. Edinburgh Sect. A, 124 (1994), pp. 843–878.

[9] M. Chipot and D. Kinderlehrer, *Equilibrium configurations of crystals*, Arch. Rational Mech. Anal., 103 (1988), pp. 237–277.

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

[11] B. Dacorogna, *A relaxation theorem and its applications to the equilibrium of gases*, Arch. Rational Mech. Anal., 77 (1981), pp. 359–386.

[12] B. Dacorogna, *Quasiconvexity and relaxation of non convex variational problems*, J. Funct. Anal., 46 (1982), pp. 102–118.

[13] B. Dacorogna, *Direct methods in the Calculus of Variations*, Springer-Verlag, Berlin, 1989.

[14] I. Ekeland, *Nonconvex minimization problems*, Bull. Amer. Math. Soc., 1 (1979), pp. 443–475.

[15] L. C. Evans, *Weak Convergence Methods for Nonlinear Partial Differential Equations*, CBMS 74, American Mathematical Society, Providence, RI, 1990.

[16] R. D. James and D. Kinderlehrer, *Theory of diffusionless phase transitions*, in PDE's and Continuum Models of Phase Transitions, Lecture Notes in Phys. 344, M. Rascle, D. Serre, and M. Slemrod, eds., Springer-Verlag, Berlin, 1989, pp. 51–84.

[17] R. D. James and D. Kinderlehrer, *Frustration in ferromagnetic materials*, Contin. Mech. Thermodyn., 2 (1990), pp. 215–239.

[18] R. D. James and D. Kinderlehrer, *A theory of magnetostriction with application to TbDyFe2*, Phil. Mag., B 68 (1993), pp. 237–274.

[19] R. D. James and D. Kinderlehrer, *Frustration and microstructure: An example in magnetostriction*, in Proc. First Europ. Conf. Elliptic Parab. Prob., Pont a Mousson, 1990.

[20] D. Kinderlehrer and P. Pedregal, *Characterizations of Young measures generated by gradients*, Arch. Rational Mech. Anal., 115 (1991), pp. 329–365.

[21] D. Kinderlehrer and P. Pedregal, *Gradient Young measures generated by sequences in Sobolev spaces*, J. Geom. Anal., 4 (1994), pp. 59–90.

[22] P. Marcellini, *Nonconvex integrals of the calculus of variations*, in Methods of Nonconvex Analysis, Lecture Notes in Math. 1446, Springer-Verlag, Berlin, 1990, pp. 16–57.

[23] P. Marcellini and C. Sbordone, *Relaxation of non-convex variational problems*, Alti Acad. Naz. Lincei, 63 (1977), pp. 341–344.

[24] Ch. B. Morrey, *Quasiconvexity and the lower semicontinuity of multiple integrals*, Pacific J. Math., 2 (1952), pp. 25–53.

[25] J. Muñoz and P. Pedregal, *Explicit Solutions of Nonconvex Variational Problems in Dimension One*, Appl. Math. Optim., submitted.

[26] P. Pedregal, *Parametrized Measures and Variational Principles*, Birkhäuser-Verlag, Basel, Switzerland, 1997.

[27] J. Warga, *Relaxed variational problems*, J. Math. Anal. Appl., 4 (1962), pp. 111–128.

[28] L. C. Young, *Generalized curves and the existence of an attained absolute minimum in the calculus of variations*, Comptes Rendus de la Société des Sciences et des Lettres de Varsovie, classe III, 30 (1937), pp. 212–234.

[29] L. C. Young, *Generalized surfaces in the calculus of variations,* I *and* II, Ann. Math., 43 (1942), pp. 84–103, 530–544.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 3, pp. 814–839, May 1998

002

DIFFERENTIAL GAMES WITH UNBOUNDED VERSUS BOUNDED CONTROLS*

FRANCO RAMPAZZO†

Abstract. A zero-sum differential game with an unbounded control and no coercivity assumptions is investigated. By means of suitable reparameterization techniques one is able to avoid the serious drawbacks which derive from the lack of a sufficiently fast growth hypothesis. In particular, one achieves a remarkable regularization of the two related Hamilton–Jacobi boundary value problems. One also proves that the latter admit unique continuous solutions, which coincide necessarily with the upper and lower values of the game.

Key words. slow growth, differential games, upper and lower values

AMS subject classifications. 90D25, 49L20, 49N25, 49L25

PII. S0363012995294602

1. Introduction.

In this paper dynamic programming is investigated for a zero-sum, finite horizon, differential game defined as follows: the dynamics is given by the differential equation

$$\begin{cases} \dot{x} = f(t, x, a, c), \\ x(\bar{t}) = \bar{x}, \end{cases}$$

while the payoff has the form

$$P[\bar{t}, \bar{x}; a, c] \doteq \int_{\bar{t}}^{T} l(t, x, a, c)\ dt + g(x(T)),$$

where the antagonist controls a and c take values in a compact subset $A \subset \mathbb{R}^q$ and a cone $C \subset \mathbb{R}^m$, respectively. Though no bounds are imposed on the values of the control c, the latter is subject to an integral bound of the form

$$\int_{\bar{t}}^{T} |c(t)|\ dt \leq K - \bar{k},$$

where, just like $\bar{t}$ and $\bar{x}$, $\bar{k}$ ($\in [0, K]$) has to be considered as an initial data of the problem. Besides some mild regularity conditions, we assume that the dynamics f and the Lagrangian l are *sublinear* in the unbounded control c, in a sense made precise in section 2. The player maneuvering the *conventional* control a, say, the a-Player, wishes to minimize the functional P, while the c-Player pursues the goal of maximizing P. Of course, the policies of the two players are described as *strategies* (see section 2).

Let us remark that in most cases involving unbounded controls (but an exception is represented by [5]), a superlinear growth assumption involving f and h allows one to exclude the occurrence of controls c which take values larger than a fixed constant which depends on the sole initial data $(\bar{t}, \bar{x})$ (see, e.g., [27]). This makes the problem locally equivalent to the one where only bounded controls are allowed.

*Received by the editors December 6, 1995; accepted for publication (in revised form) February 11, 1997.

http://www.siam.org/journals/sicon/36-3/29460.html

†Dipartimento di Matematica Pura e Applicata, Università di Padova, via Belzoni 7, 35131 Padova, Italy (rampazzo@math.unipd.it).

On the contrary, the crucial slow growth assumption on f and l can make the exploitation of larger and larger values of c necessary. This is clearly a phenomenon of the same kind as those occurring in connection with problems of calculus of variations with slow growth.

A motivation for studying slow growth differential games (with integral constraints) is trivially represented by the fact that in some practical situations it happens that (the dynamics and the Lagrangian are sublinear and) one of the two controls coincides with the (unbounded) derivative (i.e., the *rate of expenditure*) of some stocked quantity (see, e.g., [10], [12], [15], [17], [23], [26]). A further motivation comes from $\mathcal{H}_\infty$ control problems, with a sublinear performance index replacing the usual quadratic one (see, e.g., [28], [6] for a general account on $\mathcal{H}_\infty$ control problems).

The first question to be answered concerns the regularity of the upper value $\mathcal{U}$ and the lower value $\mathcal{V}$, here defined according to Roxin, Varaya, Elliott, and Kalton (see, e.g., [13]). This question has been treated in a preliminary paper [24], where some continuity properties of both $\mathcal{V}$ and $\mathcal{U}$ have been proved.

In the present paper we complete the program started in [24] by establishing upper value boundary value problems UVBVP and lower value boundary value problems LVBVP (see section 6), whose unique continuous solutions turn out to coincide with $\mathcal{U}$ and $\mathcal{V}$, respectively. It is already known from analogous questions arising in control theory (see [21], [22]) that in the general case a formal extension of the dynamic programming approach finds serious drawbacks. For instance, because of the lack of sufficient growth assumptions, the formal Hamiltonians turn out to be highly discontinuous. Moreover, unless strong commutativity hypotheses are assumed on f and l, no measure-theoretical interpretation of the dynamics accounts for the limits of trajectories corresponding to controls c which take larger and larger values (see, e.g., [8]).

Let us mention incidentally that differential games with measures, appearing linearly with constant coefficients, has been investigated by Yong [29] within the framework of impulse control problems introduced by Bensoussans and Lions [7]. It is easy to see that the involved trajectories are limits of trajectories with unbounded ordinary controls. However, though the dynamics considered in Yong's work represents the simplest case in the class of dynamics we study here, the associated payoff is not a Boltz functional. In particular, Yong's problem cannot be regarded as a *representation* of a standard problem with unbounded controls and slow growth.

The present paper is organized as follows. In section 2 the game is described and the upper and lower value maps are introduced. These maps depend on the initial data $(\bar{t}, \bar{x})$ and on the constant $\bar{k} \in [0, K]$ appearing in the integral constraint on c. Hence they turn out to be defined on the domain $[0, T[\times\mathbb{R}^n \times [0, K]$. Moreover, in [24] it has been proved that these maps are continuous and can be continuously extended to $\bar{t} = T$. In section 3, by suitably developing arguments already exploited in the study of slow growth control problems (see [18], [8], [19], [20]), we define the *reparameterized space-time game*. This game turns out to be equivalent to the original one; i.e., it has the same values. However, it enjoys the advantage of involving controls with *bounded values*. Afterwards, we prove a dynamic programming principle (DPP; see section 4) which differs from the usual one in that it involves intervals of integration varying with the controls c. Moreover, this principle takes a special form at the points where $\mathcal{V} > g$ or $\mathcal{U} > g$. This turns out to be essential when one derives the boundary conditions at $\bar{t} = T$. Successively, this DPP is converted into a *reparameterized dynamic programming principle* (RDPP). The latter allows one to establish Isaacs equations with *continuous* Hamiltonians (see section 5). In section

6 we study the boundary conditions of the problem. They are relatively simple, in that they just involve a subsolution relation and a Dirichlet inequality. Unlike the equations, the boundary conditions for the lower value coincide with the ones for the upper value. Though this is true in the conventional case as well, it is not so obvious in the present case. In fact, in a recent paper [5] boundary conditions for the lower value are established which differ from the ones of the upper value.

In section 6 it is also shown that the lower and upper values are the unique solutions of the corresponding boundary value problems. From the uniqueness result one infers an Isaacs condition that has a standard form (i.e., it consists of the identity of two Hamiltonians).

In section 7 we specialize our results to a particular problem, which has already been investigated by Barron, Jensen, and Menaldi [5], with a different approach. In fact, they consider the case where (the state x and) the control c is scalar and belongs to $C \doteq \{c \in \mathbb{R} : c \geq 0\}$. Moreover, they assume that the dynamics and the Lagrangian reduce to

$$(*) \qquad \begin{aligned} f(t,x,a,c) &= f_0(t,x,a) + f_1(t,a)c, \\ h(t,x,a,c) &= h_0(t,x,a) + h_1(t,a)c, \end{aligned}$$

respectively. (Actually, the results in [5] can be easily extended to the case when h_1—but *not* f_1—depends on x; in particular, this allows the authors to show that the classical minimax control problem is a particular case of the game they are investigating.) On the one hand, the fact that f_1 is independent of x is crucial and allows the authors of [5] to adopt a measure-theoretical definition of trajectory. On the other hand, this approach cannot be extended to the larger class of problems investigated here. Moreover, the Hamiltonians utilized in [5] are discontinuous, and their values are calculated by means of minimizations over sets which depend on the gradient of the value functions themselves. On the contrary, the Isaacs equation we derive in section 5 involves continuous Hamiltonians whose values are the results of min-max operations over a *compact, constant* set. Finally, the boundary conditions established here are quite standard, whereas the ones in [5] need the resolution of an auxiliary boundary value problem.

2. The game. We consider the following zero-sum differential game of fixed duration: a dynamics is given of the form

$$(2.1) \qquad \dot{x} = f(t,x,a,c), \qquad x(\bar{t}) = \bar{x}, \qquad t \in [\bar{t}, T],$$

where $0 \leq \bar{t} < T$, $a(\cdot)$ ranges on a compact subset $A \subset \mathbb{R}^q$, and the control $c(\cdot)$ takes values in a closed cone $C \subset \mathbb{R}^m$ (by *cone* we mean a subset of a real vector space closed under multiplication by nonnegative scalars). The admissible controls $a(\cdot)$ are Borel measurable maps from $[\bar{t}, T]$ into A, and we denote this set by $\mathcal{A}(\bar{t})$. These controls may be thought as implemented by a *player*, say, the a-Player. The opponent will be called the c-Player. Let $K > 0$ be an upper bound for the L^1 norm of the control $c(\cdot)$ maneuvered by the c-Player. For every nonnegative $\bar{k} \leq K$ let us define the family $\mathcal{C}(\bar{t}, \bar{k})$ of admissible controls for the c-Player as the set of integrable maps $c(\cdot)$ from $[\bar{t}, T]$ into C satisfying the inequality

$$\int_{\bar{t}}^{T} |c(t)| dt \leq K - \bar{k}.$$

Throughout the paper we shall assume hypotheses (i) and (ii) below, which, for every control pair $(a, c) \in \mathcal{A}(\bar{t}) \times \mathcal{C}(\bar{t}, \bar{k})$, guarantees global existence, uniqueness,

and exponential growth of each solution to (2.1). Let us denote such a solution by $x[\bar{t},\bar{x};a,c]$ or, whenever the initial data are meant from the context, by $x[a,c]$. Let $g:\mathbb{R}^n\to\mathbb{R}$ be a continuous bounded function and let $l=l(t,x,a,c)$ be a scalar function defined on the domain of f. Let us consider the payoff:

$$P[\bar{t},\bar{x};a,c]\doteq g(x[a,c](T))+\int_{\bar{t}}^{T} l(t,x[a,c](t),a(t),c(t))\,dt.$$

Whenever the initial condition is meant by the context we shall write $P[a,c]$ instead of $P[\bar{t},\bar{x};a,c]$.

We shall assume the following hypotheses on the maps l and f.

Hypotheses.

(i) (*Regularity*) The functions f and l are continuous on $[0,T]\times\mathbb{R}^n\times A\times C$. Moreover, for every compact subset $Q\subset\mathbb{R}^n$, there exists a constant $L>0$ such that

$$\begin{aligned}|(f,l)(t',x',a,c)-(f,l)(t,x,a,c)|\le L(1+|c|)|(t',x')-(t,x)|\\ \forall(t',x',a,c),(t,x,a,c)\in[0,T]\times Q\times A\times C.\end{aligned}\tag{2.2}$$

(ii) (*Sublinear growth in c*) There exist continuous maps l^∞ and f^∞, called the *recession function of l and f*, respectively, such that

$$\lim_{r\to+\infty} r^{-1}l(t,x,a,rc)=l^\infty(t,x,a,c),$$
$$\lim_{r\to+\infty} r^{-1}f(t,x,a,rc)=f^\infty(t,x,a,c),$$

uniformly on compact sets of $[0,T]\times\mathbb{R}^n\times A\times C$, with Q compact. Moreover, for suitable constants $M,N\ge 0$, the map l verifies

$$|l(t,x,a,c)|\le M+N|c|$$

for all $(t,x,a,c)\in[0,T]\times\mathbb{R}^n\times A\times C$.

Under hypothesis (i), for every admissible pair $(a,c)\in\mathcal{A}(\bar{t})\times\mathcal{C}(\bar{t},\bar{k})$, the trajectory $x[a,c]$ and the payoff $P[a,c]$ are well defined. Moreover, it is straightforward to check (see [20, Thm. 2.1]) that if $\bar{x}\in Q$, with Q a compact subset, then there exists a compact set $Q'\supset Q$ such that all trajectories starting at $\bar{t}$ from $\bar{x}$ remain inside Q' during the time-interval $[\bar{t},T]$. Obviously, hypothesis (i) can be replaced by any else condition guaranteeing existence, uniqueness, and a priori boundedness of the trajectories.

On the contrary, hypothesis (ii) is not conventional and makes the problem hardly confrontable by means of standard arguments. In fact, on one hand the formal Hamiltonians are highly discontinuous. On the other hand, no a priori bounds can be assumed on the values of the control c.

As in the case with bounded controls, the game consists of the following: at almost each instant $\tilde{t}\in[\bar{t},T]$ the a-Player, with knowledge of the control implemented by the c-Player in the past interval $[\bar{t},\tilde{t}]$, chooses the control a in order to minimize the payoff, while the c-Player chooses c (with knowledge of a on $[\bar{t},\tilde{t}]$) at almost each instant with the goal of maximizing the payoff. To make the notion of game precise we need a notion of *strategy*. This goal will be achieved by formally adapting Roxin, Varaya, Elliot, and Kalton's definition of strategy (see, e.g., [13]).

DEFINITION 2.1. *A strategy for the a-Player relative to the initial data $(\bar{t},\bar{k})$ is a map $\alpha:\mathcal{C}(\bar{t},\bar{k})\to\mathcal{A}(\bar{t})$ which satisfies the following condition: $\forall\tilde{t}\in[\bar{t},T]$ and for any*

pair of controls $c_1, c_2 \in \mathcal{C}(\bar t, \bar k)$,

$$\begin{aligned} &\textit{if} \quad c_1(t) = c_2(t) \quad \textit{for almost every } t \in [\bar t, \tilde t] \\ &\textit{then} \quad \alpha(c_1)(t) = \alpha(c_2)(t) \quad \textit{for almost every} \quad t \in [\bar t, \tilde t]. \end{aligned}$$

The set of strategies for the a-Player relative to the initial data $(\bar t, \bar k)$ *will be denoted by* $\Delta(\bar t, \bar k)$.

A strategy for the c-Player relative to the initial data $(\bar t, \bar k)$ *is a map* $\gamma : \mathcal{A}(\bar t) \to \mathcal{C}(\bar t, \bar k)$ *which satisfies the following condition:* $\forall \tilde t \in [\bar t, T]$ *and for any pair of controls* $a_1, a_2 \in \mathcal{A}(\bar t)$,

$$\begin{aligned} &\textit{if} \quad a_1(t) = a_2(t) \quad \textit{for almost every} \quad t \in [\bar t, \tilde t] \\ &\textit{then} \quad \gamma(v_1)(t) = \gamma(v_2)(t) \quad \textit{for almost every} \quad t \in [\bar t, \tilde t]. \end{aligned}$$

The set of strategies for the c-Player relative to the pair $(\bar t, \bar k)$ *will be denoted by* $\Gamma(\bar t, \bar k)$.

We recall that maps like α and γ are called *nonanticipating.*

DEFINITION 2.2. *For every* $(\bar t, \bar x, \bar k) \in [0, T[\times \mathbb{R}^n \times [0, K]$ *let us set*

$$\mathcal{V}(\bar t, \bar x, \bar k) \doteq \inf_{\alpha \in \Delta(\bar t, \bar k)} \sup_{c \in \mathcal{C}(\bar t, \bar k)} P[\bar t, \bar x, \alpha(c), c]$$

and

$$\mathcal{U}(\bar t, \bar x, \bar k) \doteq \sup_{\gamma \in \Gamma(\bar t, \bar k)} \inf_{a \in \mathcal{A}(\bar t)} P[\bar t, \bar x, a, \gamma(a)].$$

The maps $\mathcal{V}$ *and* $\mathcal{U}$ *are called the lower value and the upper value of the game, respectively.*

Theorem 2.1 below has been proved in [24].

THEOREM 2.1. *The functions* $\mathcal{U}$ *and* $\mathcal{V}$ *are bounded and continuous. Moreover, they can be continuously extended to the closed domain* $[0, T] \times \mathbb{R}^n \times [0, K]$.

3. Reparameterization of the game. A reparameterization technique (extending one already exploited in optimal control problems (see, e.g., [8], [19], [20], [18])) is now introduced in order to obtain an equivalent (space-time) game where only *bounded controls* and *bounded strategies* are implemented. In turn, this will allow us to establish dynamic programming equations for $\mathcal{V}$ and $\mathcal{U}$ which involve *continuous* Hamiltonians. Let us begin by introducing the notions of space-time control and space-time strategy.

DEFINITION 3.1. *The set* $\widetilde{\mathcal{C}}(\bar t, \bar k)$ *of space-time controls for the c-Player consists of all (bounded) measurable maps* (w_0, w) *from an interval* $[0, L]$, $T - \bar t \le L \le T + K - \bar t - \bar k$, *into* $\mathbb{R}^{m+1}$, *enjoying the following properties:* (1) *for almost every* $s \in [0, L]$, $|(w_0, w)(s)| = 1$ *and* $w(s) \in C$; (2) $w_0(s) > 0$ *for almost every* $s \in [0, L]$, *and* $\int_0^L w_0(s)\, ds = T - \bar t$; (3) *the map* w *satisfies the integral bound* $\int_0^L |w(s)|\, ds \le K - \bar k$.

DEFINITION 3.2. *Let* $c \in \mathcal{C}(\bar t, \bar k)$ *and let* $s(t) \doteq \int_{\bar t}^t |(1, c(\tau))|\, d\tau$, $L_c = s(T)$. *Let us define the map* $c^* = (w_0, w) : [0, L_c] \to [0, \infty) \times \mathbb{R}^m$ *by setting, for every* $s \in [0, L_c]$, $c^*(s) = (w_0, w)(s) = \left(\frac{dt}{ds}(s), \frac{dt}{ds}(s) \cdot c \circ t(s)\right)$ *for every* $s \in [0, L_c]$, *where* $t(s)$ *is the inverse of* $s(t)$.

Conversely, for any $(w_0, w) \in \widetilde{\mathcal{C}}(\bar t, \bar k)$, *let* $s(t)$ *be the inverse map of* $t(s) \doteq \bar t + \int_0^s w_0(\sigma)\, d\sigma$ *and let us define the map* $(w_0, w)_* = c : [\bar t, T] \to \mathbb{R}^m$ *by setting* $c(t) \doteq \frac{w}{w_0} \circ s(t)$ *for every* $t \in [\bar t, T]$.

With standard arguments one can check that for every $c \in \mathcal{C}(\bar{t}, \bar{k})$ one has $c^* \in \widetilde{\mathcal{C}}(\bar{t}, \bar{k})$. Conversely, for every $(w_0, w) \in \widetilde{\mathcal{C}}(\bar{t}, \bar{k})$ one has $(w_0, w)_* \in \mathcal{C}(\bar{t}, \bar{k})$. Moreover, it is easy to prove the following.

PROPOSITION 3.1. *For all $c \in \mathcal{C}(\bar{t}, \bar{k})$ and $(w_0, w) \in \widetilde{\mathcal{C}}(\bar{t}, \bar{k})$ one has*

$$(c^*)_* = c, \qquad ((w_0, w)_*)^* = (w_0, w).$$

In other words, the map $c \mapsto c^*$ establishes a one-to-one correspondence between the sets $\mathcal{C}(\bar{t}, \bar{k})$ and $\widetilde{\mathcal{C}}(\bar{t}, \bar{k})$, and the map $(w_0, w) \mapsto (w_0, w)_*$ is its inverse.

We denote the set of Borel measurable maps from $[0, 1]$ into A by $\widetilde{\mathcal{A}}$. It will be called the set of *space-time controls for the a-Player.*

Let $\rho_{\bar{t}}$ denote the unique increasing linear transformation from $[\bar{t}, T]$ onto $[0, 1]$, and let $\tau_{\bar{t}}$ be the inverse of $\rho_{\bar{t}}$. For any $a \in \mathcal{A}(\bar{t})$ and $v \in \tilde{\mathcal{A}}$ let us set

$$a^* \doteq a \circ \tau_{\bar{t}}, \qquad v_* \doteq v \circ \rho_{\bar{t}}.$$

It is trivial to verify that $a \to a^*$ is a bijection and $v \to v_*$ is its inverse.

For every mapping $\nu : \widetilde{\mathcal{C}}(\bar{t}, \bar{k}) \to \widetilde{\mathcal{A}}$ let us define the map $\nu_* : \mathcal{C}(\bar{t}, \bar{k}) \to \mathcal{A}(\bar{t})$ by setting, for every $c \in \mathcal{C}(\bar{t}, \bar{k})$,

$$\nu_*(c) \doteq \big(\nu(c^*)\big)_*.$$

Conversely, if α maps $\mathcal{C}(\bar{t}, \bar{k})$ into $\mathcal{A}(\bar{t})$, let $\alpha^* : \widetilde{\mathcal{C}}(\bar{t}, \bar{k}) \to \widetilde{\mathcal{A}}$ be the function defined by

$$\alpha^*(w_0, w) \doteq \big(\alpha((w_0, w)_*)\big)^*$$

for every $(w_0, w) \in \widetilde{\mathcal{C}}(\bar{t}, \bar{k})$.

DEFINITION 3.3. *A map $\nu : \widetilde{\mathcal{C}}(\bar{t}, \bar{k}) \to \widetilde{\mathcal{A}}$ is called a space-time strategy for the a-Player provided ν_* is a strategy (i.e., ν_* is nonanticipating). We shall denote the set of space-time strategies for the a-Player by $\widetilde{\Delta}(\bar{t}, \bar{k})$.*

PROPOSITION 3.2. *For all strategies $\alpha \in \Delta(\bar{t}, \bar{k})$ and for all space-time strategies $\nu \in \widetilde{\Delta}(\bar{t}, \bar{k})$ one has*

$$(\alpha^*)_* = \alpha, \qquad (\nu_*)^* = \nu.$$

In particular, α^* is a space-time strategy if and only if α is a strategy.

By Proposition 3.2 the map $\alpha \mapsto \alpha^*$ establishes a one-to-one correspondence between the set of strategies $\Delta(\bar{t}, \bar{k})$ and the set of space-time stategies $\widetilde{\Delta}(\bar{t}, \bar{k})$. Moreover, the map $\nu \mapsto \nu_*$ is the inverse of $\alpha \mapsto \alpha^*$.

We now recall from [20] the notion of space-time extension of the pair (l, f).

DEFINITION 3.4 (see [21]). *Let (l, f) be a pair satisfying hypotheses* (i) *and* (ii). *For every $(t, x, a, w_0, w) \in [0, T] \times \mathbb{R}^n \times A \times [0, +\infty) \times C$ we set*

$$(\bar{l}, \overline{f})(t, x, a, w_0, w) \doteq \begin{cases} w_0 \cdot (l, f)(t, x, a, w/w_0) & \text{if } w_0 \neq 0, \\ (l^\infty, f^\infty)(t, x, a, w) & \text{if } w_0 = 0, \end{cases}$$

where l^∞ and f^∞ are the recession functions introduced with hypothesis (i). *The vector field $\overline{f}$ (resp., the map $\bar{l}$) will be called the space-time extension of f (resp., of l).*

Let us introduce the *space-time version of the differential game*. Given a control $(v, w_0, w) \in \widetilde{\mathcal{A}} \times \widetilde{\mathcal{C}}(\bar{t}, \bar{k})$, let us consider the solution (t, z), defined on the domain $[0, L]$ of (w_0, w), of

$$\begin{cases} t'(s) = w_0(s), \\ z'(s) = \overline{f}\big(t(s), z(s), v_* \circ t(s), w_0(s), w(s)\big), \\ (t, z)(0) = (\bar{t}, \bar{x}) \end{cases} \tag{3.1}$$

and the corresponding payoff

$$\overline{P}[\bar{t}, \bar{x}; v, w_0, w] \doteq g(z(L)) + \int_0^L \bar{l}\big(t(s), z(s), v_* \circ t(s), w_0(s), w(s)\big)\, ds.$$

We shall denote the solution of (3.1) by $(t, z)[\bar{t}, \bar{x}; v, w_0, w]$. Whenever the initial condition is known by the context we shall write $(t, z)[v, w_0, w]$ and $\overline{P}[v, w_0, w]$ instead of $(t, z)[\bar{t}, \bar{x}; v, w_0, w]$ and $\overline{P}[\bar{t}, \bar{x}; v, w_0, w]$, respectively.

The relationship occurring between the original game and its space-time version is illustrated by Propositions 3.3 and 3.4 below.

PROPOSITION 3.3. *Let us consider a control $c \in \mathcal{C}(\bar{t}, \bar{k})$ and a strategy $\alpha \in \Delta(\bar{t}, \bar{x})$. Let us set $L_c \doteq \int_{\bar{t}}^T |1, c(t)|\, dt$, $(w_0, w) \doteq c^*$, and $t(s) \doteq \bar{t} + \int_0^s w_0(\sigma)\, d\sigma$. Then, for every $s \in [0, L_c]$, one has*

$$x[\alpha(c), c] \circ t(s) = z[\alpha^*(c^*), c^*](s)$$

and

$$P[\alpha(c), c] = \overline{P}[\alpha^*(c^*), c^*].$$

As a corollary, in view of Propositions 3.1 and 3.2, one also has the converse passage from space-time trajectories to ordinary ones.

PROPOSITION 3.4. *Let us consider a space-time control $(w_0, w) \in \widetilde{\mathcal{C}}(\bar{t}, \bar{x})$ and a space-time strategy $\nu \in \widetilde{\Delta}(\bar{t}, \bar{x})$. Then, for every $t \in [\bar{t}, T]$, one has*

$$z[\nu(w_0, w), (w_0, w)] \circ s(t) = x[\nu_*((w_0, w)_*), (w_0, w)_*](t)$$

and

$$\overline{P}[\nu(w_0, w), (w_0, w)] = P[\nu_*((w_0, w)_*), (w_0, w)_*],$$

where $s(\cdot)$ denotes the inverse of $t(s) \doteq \bar{t} + \int_0^s w_0(\sigma)\, d\sigma$.

Proof of Proposition 3.3. Let us set $x(t) \doteq x[\alpha(c), c](t)$, $z(s) = z[\alpha^*(c^*), c^*](s)$, $(w_0, w) \doteq c^*$. Then, if $s(\cdot)$ denotes the inverse of $t(\cdot)$ one has

$$\begin{aligned} (x \circ t)'(s) &= t'(s) \cdot f\big(t, x(t), \alpha(c)(t), c(t)\big)_{|_{t=t(s)}} \\ &= w_0(s) \cdot f\left(t, x(t), \alpha(c)(t), \frac{w}{w_0} \circ s(t)\right)_{|_{t=t(s)}} \\ &= \overline{f}\left(t(s), x \circ t(s), \big(\alpha^*(c^*)\big)_* \circ t(s), w_0(s), w(s)\right) \\ &= \overline{f}\left(t(s), x \circ t(s), \big(\alpha^*(c^*)\big)_* \circ t(s), c^*(s)\right). \end{aligned}$$

By the uniqueness of the solution to (3.1) we conclude that the former equality of the thesis is verified. The latter equality is a straightforward consequence of the former. □

As a consequence of the previous results we obtain that the value of the original game and of its space-time version do coincide. Indeed, let us set

$$j(\bar t,\bar x,\bar k,\alpha)\doteq\sup_{c\in\mathcal{C}(\bar t,\bar k)}P[\bar t,\bar x,\bar k,\alpha(c),c],$$
$$\tilde j(\bar t,\bar x,\bar k,\nu)\doteq\sup_{(w_0,w)\in\widetilde{\mathcal{C}}(\bar t,\bar k)}\overline{P}[\bar t,\bar x,\bar k,\nu(w_0,w),(w_0,w)],$$
$$\widetilde{\mathcal{V}}(\bar t,\bar x,\bar k)\doteq\inf_{\nu\in\widetilde{\Delta}(\bar t,\bar k)}\tilde j(\bar t,\bar x,\bar k,\nu).$$

The map $\widetilde{\mathcal{V}}$ will be called the *lower value of the space-time version of the game*. Observe that, by definition, $\mathcal{V}(\bar t,\bar x,\bar k)=\inf_{\alpha\in\Delta(\bar t,\bar k)}j(\bar t,\bar x,\bar k,\alpha)$. By Propositions 3.3 and 3.4 it follows that, $\forall\alpha\in\Delta(\bar t,\bar k)$ and $\forall\nu\in\widetilde{\Delta}(\bar t,\bar k)$,

$$\tilde j(\alpha^*)\le j(\alpha),\qquad j(\nu_*)\le\tilde j(\nu).$$

Hence, by $j(\alpha)=j\big((\alpha^*)_*\big)\le\tilde j(\alpha^*)$, one obtains

$$\tilde j(\alpha^*)=j(\alpha),$$

which, by $\nu=(\nu_*)^*$, implies

$$j(\nu_*)=\tilde j(\nu)$$

as well.

From the previous identities and Proposition 3.2 one infers that the lower value of the original game and of the reparameterized one do coincide.

PROPOSITION 3.5. *For every* $(\bar t,\bar x,\bar k)\in[0,T[\times\mathbb{R}^n\times[0,K]$ *one has*

$$\mathcal{V}(\bar t,\bar x,\bar k)=\widetilde{\mathcal{V}}(\bar t,\bar x,\bar k).$$

Similarly, a space-time extension of the notion of strategy for the c-Player can be introduced. We just write down the definitions and the related results, the proofs being quite similar to the ones concerning the lower value case.

Let (ξ_0,ξ) be a mapping from $\widetilde{\mathcal{A}}$ into $\widetilde{\mathcal{C}}(\bar t,\bar k)$ and let us define the function $(\xi_0,\xi)_*:\mathcal{A}(\bar t)\to\mathcal{C}(\bar t,\bar k)$ by setting, for every $a\in\mathcal{A}(\bar t)$,

$$(\xi_0,\xi)_*(a)\doteq[(\xi_0,\xi)(a^*)]_*.$$

DEFINITION 3.5. *The map* (ξ_0,ξ) *is called a space-time strategy for the c-Player provided* $(\xi_0,\xi)_*$ *is a strategy (i.e., it is nonanticipating). The set of space-time strategies for the c-Player will be denoted by* $\tilde\Gamma(\bar t,\bar k)$.

On the other hand, for every map $\gamma:\mathcal{A}(\bar t)\to\mathcal{C}(\bar t,\bar k)$ let us define the function $\gamma^*:\widetilde{\mathcal{A}}\to\widetilde{\mathcal{C}}(\bar t,\bar k)$ by setting, for every $v\in\widetilde{\mathcal{A}}$,

$$\gamma^*(v)\doteq[\gamma(v_*)]^*.$$

PROPOSITION 3.6. *For every strategy* $\gamma\in\Gamma(\bar t,\bar k)$ *and every space-time control* $(\xi_0,\xi)\in\widetilde{\Gamma}(\bar t,\bar x)$, *one has*

$$(\gamma^*)_*=\gamma,\qquad[(\xi_0,\xi)_*]^*=(\xi_0,\xi).$$

In particular γ^* *is a space-time strategy if and only if* γ *is a strategy.*

By Proposition 3.6 we obtain that the map $\gamma \mapsto \gamma^*$ establishes a one-to-one correspondence between the set of strategies $\Gamma(\bar t, \bar k)$ and the set of space-time stategies $\widetilde\Gamma(\bar t, \bar k)$. Moreover, the map $(\xi_0, \xi) \mapsto (\xi_0, \xi)_*$ is the inverse of $\gamma \mapsto \gamma^*$. Propositions 3.7 and 3.8 below are the anologues of Propositions 3.3 and 3.4 for the upper value.

PROPOSITION 3.7. *Let us consider a control $a \in \mathcal{A}(\bar t)$ for the a-Player and a strategy $\gamma \in \Gamma(\bar t, \bar x)$ for the c-Player. Then, for every $s \in [0, L_{\gamma(a)}]$, $L_{\gamma(a)} \doteq \int_{\bar t}^T |1, \gamma(a)(t)|\, dt$, one has*

$$x[a, \gamma(a)] \circ t(s) = z[a^*, \gamma^*(a^*)](s)$$

and

$$P[a, \gamma(a)] = \overline{P}[a^*, \gamma^*(a^*)],$$

where $t(s) \doteq \bar t + \int_0^s w_0(\sigma)\, d\sigma$.

PROPOSITION 3.8. *Let us consider a space-time control $v \in \widetilde{\mathcal{A}}$ and a strategy $(\xi_0, \xi) \in \widetilde\Gamma(\bar t, \bar x)$, and let us set $(w_0, w) \doteq (\xi_0, \xi)(v)$. Then, for every $t \in [\bar t, T]$, one has*

$$z[v, (\xi_0, \xi)(v)] \circ s(t) = x[v_*, (\xi_0, \xi)_*(v_*)](t)$$

and

$$\overline{P}[v, (\xi_0, \xi)(v)] = P[v_*, (\xi_0, \xi)_*(v_*)],$$

where $s(t)$ denotes the inverse of $t(s) \doteq \bar t + \int_0^s w_0(\sigma)\, d\sigma$.

Let us set

$$\begin{aligned} i[\bar t, \bar x, \bar k, \gamma] &\doteq \inf_{a \in \mathcal{A}(\bar t)} P[\bar t, \bar x, \bar k, a, \gamma(a)], \\ \tilde i[\bar t, \bar x, \bar k, (\xi_0, \xi)] &\doteq \inf_{v \in \widetilde{\mathcal{A}}} \overline{P}[\bar t, \bar x, \bar k, v, (\xi_0, \xi)(v)], \\ \widetilde{\mathcal{U}}(\bar t, \bar x, \bar k) &\doteq \sup_{(\xi_0, \xi) \in \widetilde\Gamma(\bar t, \bar k)} \tilde i[\bar t, \bar x, \bar k, (\xi_0, \xi)]. \end{aligned}$$

The map $\widetilde{\mathcal{U}}$ will be called the *upper value of the space-time version of the game.* Just as in the case of the lower value, we have

$$\tilde i[\gamma^*] = i[\gamma], \qquad i[(\xi_0, \xi)_*] = \tilde i[(\xi_0, \xi)],$$

from which the following proposition follows.

PROPOSITION 3.9. *For every $(\bar t, \bar x, \bar k) \in [0, T[\times \mathbb{R}^n \times [0, K]$ one has*

$$\mathcal{U}(\bar t, \bar x, \bar k) = \widetilde{\mathcal{U}}(\bar t, \bar x, \bar k).$$

4. Dynamic programming. Let us set, for every $c \in \mathcal{C}(\bar t, \bar k)$ and $t \in [\bar t, T]$,

$$s_c(t) \doteq \int_{\bar t}^t |1, c(\tau)|\, d\tau,$$

and let $t_c : [0, s_c(T)] \to [\bar t, T]$ be the inverse of s_c. Notice that, for every $c \in \mathcal{C}(\bar t, \bar k)$, one has $[0, T - \bar t] \subseteq [0, s_c(T)]$.

In Theorem 4.1 below and throughout the paper we shall adopt the following convention: whenever an initial condition $(\bar t,\bar x,\bar k)$, a control (a,c), and a pair $(x,k)(\cdot)$ appear in the same formula, it is meant that

$$(x,k)(t) \doteq \left(x[\bar t,\bar x;a,c](t), \bar k + \int_{\bar t}^{t} |c(\tau)|\, d\tau\right).$$

For every $\sigma>0$ let us define the subset $\mathcal{C}_\sigma(\bar t,\bar k)\subset\mathcal{C}(\bar t,\bar k)$ as the set of those controls c such that the domain $[0,L_c]$ of the corresponding c^* and of t_c contains the interval $[0,\sigma]$, i.e.,

$$\mathcal{C}_\sigma(\bar t,\bar k) \doteq \left\{ c\in\mathcal{C}(\bar t,\bar k) \quad : \quad \int_{\bar t}^{T} |(1,c)| > \sigma \right\}.$$

Observe that $\mathcal{C}_\sigma(\bar t,\bar k)\neq\emptyset$ for every $\sigma\in]0,\max\{T-\bar t,K-\bar k\}[$, in that the constant control $c\doteq(K-\bar k)/(T-\bar t)$ belongs to $\mathcal{C}_\sigma(\bar t,\bar k)$. Let us also define the subset of strategies $\Gamma_\sigma(\bar t,\bar k)\subset\Gamma(\bar t,\bar k)$ by setting

$$\Gamma_\sigma(\bar t,\bar k) \doteq \big\{\gamma\in\Gamma(\bar t,\bar k) \quad | \quad \gamma(a)\in\mathcal{C}_\sigma(\bar t,\bar k) \quad \forall a\in\mathcal{A}(\bar t)\big\}.$$

Since the constant mappings from $\mathcal{A}(\bar t)$ into $\mathcal{C}(\bar T,\bar k)$ are strategies, for every $\sigma\in]0,\max\{T-\bar t,K-\bar k\}[$ we have that $\Gamma_\sigma(\bar t,\bar k)\neq\emptyset$ as well. It is also straightforward to check that for every $\sigma<T-\bar t$ one has $\mathcal{C}_\sigma(\bar t,\bar k)=\mathcal{C}(\bar t,\bar k),\Gamma_\sigma(\bar t,\bar k)=\Gamma(\bar t,\bar k)$. Before stating a Dynamic Programming Principle for our problem let us show that as soon as $\mathcal{U}>g$ (resp., $\mathcal{V}>g$) one can replace the set Γ (resp., $\mathcal{C}$) in the definition of $\mathcal{U}$ (resp., $\mathcal{V}$) with a subset Γ_σ (resp., $\mathcal{C}_\sigma$), for a suitable choice of σ.

PROPOSITION 4.1. *Let $Q\subset\mathbb{R}^n$ be a compact subset and let $\eta>0$. Then there exists a constant $\hat\sigma$ such that for every $(\bar t,\bar x,\bar k)\in[0,T[\times Q\times[0,K]$ verifying $\mathcal{U}(\bar t,\bar x,\bar k)-g(\bar x)\geq\eta$ (resp., $\mathcal{V}(\bar t,\bar x,\bar k)-g(\bar x)\geq\eta$) one has*

$$\mathcal{U}(\bar t,\bar x,\bar k) = \sup_{\gamma\in\Gamma_\sigma(\bar t,\bar k)} \inf_{a\in\mathcal{A}(\bar t)} P[\bar t,\bar x;a,\gamma(a)]$$

(resp.,

$$\mathcal{V}(\bar t,\bar x,\bar k) = \inf_{\alpha\in\Delta(\bar t)} \sup_{c\in\mathcal{C}_\sigma(\bar t,\bar k)} P[\bar t,\bar x;\alpha(c),c])$$

for all $\sigma\leq\hat\sigma$.

Proof. The fact that the right-hand sides are less than or equal to the corresponding left-hand sides is trivial. Let us prove that the converse inequalities hold true. We claim that for every $\epsilon>0$ there exists a σ such that the set reachable from $(\bar t,\bar x,\bar k)$ with controls $(a,c)\in\mathcal{A}(\bar t)\times(\mathcal{C}(\bar t,\bar k)\backslash\mathcal{C}_\sigma(\bar t,\bar k))$ is contained in the ball $B[(\bar t,\bar x,\bar k);\epsilon]$ of center $(\bar t,\bar x,\bar k)$ and radius ϵ. Indeed, if $c\in\mathcal{C}(\bar t,\bar k)\backslash\mathcal{C}_\sigma(\bar t,\bar k)$, one has

$$T-\bar t\leq\sigma, \qquad k(T)-\bar k\doteq\int_{\bar t}^{T}|c(t)|dt\leq\sigma.$$

Moreover, for every $(a,c)\in\mathcal{A}(\bar t)\times\mathcal{C}(\bar t,\bar k)$, one has

$$x[\bar t,\bar x;a,c](T)-\bar x = z[\bar t,\bar x;a^*,c^*](s_c(T))-\bar x.$$

Hence the claim follows from a trivial application of Gronwall's lemma to the space-time system (3.1), in that the controls (a^*, c^*) take values in the compact set $A \times \{(w_0, w) \in \mathbb{R}^{m+1} : |(w_0, w)| = 1\}$.

Therefore, in view of the assumptions on l and of the uniform continuity of g on Q, we can infer the existence of a $\hat\sigma$ such that for every $\sigma \le \hat\sigma$ one has

$$|P[\bar t, \bar x; a, c] - g(\bar x)| = |\overline{P}[\bar t, \bar x; a^*, c^*] - g(\bar x)| \le \frac{\eta}{2}$$

for all $(a, c) \in \mathcal{A}(\bar t) \times (\mathcal{C}(\bar t, \bar k) \backslash \mathcal{C}_\sigma(\bar t, \bar k))$. Since the values $\mathcal{U}(\bar t, \bar x, \bar k)$ and $\mathcal{V}(\bar t, \bar x, \bar k)$ belong to the closure of the set $\big\{P[\bar t, \bar x; a, c] \ : \ (a, c) \in \mathcal{A}(\bar t) \times \mathcal{C}(\bar t, \bar k)\big\}$, it follows that

$$\mathcal{U}(\bar t, \bar x, \bar k) > \sup_{\gamma \in (\Gamma(\bar t, \bar k) \backslash \Gamma_\sigma(\bar t, \bar k))} \inf_{a \in \mathcal{A}(\bar t)} P[\bar t, \bar x; a, \gamma(a)]$$

and

$$\mathcal{V}(\bar t, \bar x, \bar k) > \inf_{\alpha \in \Delta(\bar t)} \sup_{c \in (\mathcal{C}(\bar t, \bar k) \backslash \mathcal{C}_\sigma(\bar t, \bar k)} P[\bar t, \bar x; \alpha(c), c],$$

which implies the thesis. □

THEOREM 4.1 (Dynamic Programming Principle (DPP)). *For every* $(\bar t, \bar x, \bar k) \in [0, T[\times \mathbb{R}^n \times [0, K]$ *and* $0 < \sigma < T - \bar t$ *one has*

$$\mathcal{V}(\bar t, \bar x, \bar k) = \inf_{\alpha \in \Delta(\bar t, \bar k)} \sup_{c \in \mathcal{C}(\bar t, \bar x)} \left\{ \int_{\bar t}^{t_c(\sigma)} l\big(t, x(t), \alpha(c)(t), c(t)\big)\, dt + \mathcal{V}\big(t_c(\sigma), x \circ t_c(\sigma), k_c \circ t_c(\sigma)\big) \right\} \tag{4.1}$$

and

$$\mathcal{U}(\bar t, \bar x, \bar k) = \sup_{\gamma \in \Gamma(\bar t, \bar k)} \inf_{a \in \mathcal{A}(\bar t)} \left\{ \int_{\bar t}^{t_{\gamma(a)}(\sigma)} l\big(t, x(t), a(t), \gamma(a)(t)\big)\, dt + \mathcal{U}\big(t_{\gamma(a)}(\sigma), x \circ t_{\gamma(a)}(\sigma), k \circ t_{\gamma(a)}(\sigma)\big) \right\}, \tag{4.2}$$

where, for every $c \in \mathcal{C}(\bar t, \bar k)$, t_c *is defined as in the beginning of this section.*

Moreover, if $Q \subset \mathbb{R}^n$ *is a compact subset and* $(\bar t, \bar x, \bar k) \in [0, T[\times Q \times [0, K[$ *satisfies* $\mathcal{U}(\bar t, \bar x, \bar k) - g(x) \ge \eta > 0$ *(resp.,* $\mathcal{V}(\bar t, \bar x, \bar k) - g(x) \ge \eta > 0$*), there exists a* $\hat\sigma > 0$ *such that*

$$\mathcal{V}(\bar t, \bar x, \bar k) = \inf_{\alpha \in \Delta(\bar t, \bar k)} \sup_{c \in \mathcal{C}_\sigma(\bar t, \bar x)} \left\{ \int_{\bar t}^{t_c(\sigma)} l\big(t, x(t), \alpha(c)(t), c(t)\big)\, dt + \mathcal{V}\big(t_c(\sigma), x \circ t_c(\sigma), k_c \circ t_c(\sigma)\big) \right\} \tag{4.3}$$

(resp.,

$$\mathcal{U}(\bar t, \bar x, \bar k) = \sup_{\gamma \in \Gamma_\sigma(\bar t, \bar k)} \inf_{a \in \mathcal{A}(\bar t)} \left\{ \int_{\bar t}^{t_{\gamma(a)}(\sigma)} l\big(t, x(t), a(t), \gamma(a)(t)\big)\, dt + \mathcal{U}\big(t_{\gamma(a)}(\sigma), x \circ t_{\gamma(a)}(\sigma), k \circ t_{\gamma(a)}(\sigma)\big) \right\} \tag{4.4}$$

for all $\sigma \le \hat\sigma$.

Remark 4.1. The idea of the proof of this theorem is not far from the one exploited in the conventional case (see, e.g., [13]), where only bounded controls are involved. However, unlike the conventional case, the intervals of integration involved in (4.1)–(4.4) are not constant for a fixed σ, and the variable k is constrained to be less than or equal to K. Let us also remark that the validity of (4.3) and (4.4) does not find a counterpart in the conventional case. These equalities mean that as soon as $\mathcal{V}$ or $\mathcal{U}$ is greater than the cost function, one can take a smaller set of controls c in the corresponding DPP. This fact will be *essential* when we establish the boundary conditions for $\mathcal{U}$ and $\mathcal{V}$ on $t = T$ (see Theorem 6.1).

Proof of Theorem 4.1. Denoting the right-hand side of (4.1) and (4.2) by $W(\bar{t}, \bar{x}, \bar{k})$ and $Y(\bar{t}, \bar{x}, \bar{k})$, respectively, we shall limit ourselves to demonstrate that

$$W(\bar{t}, \bar{x}, \bar{k}) \geq \mathcal{V}(\bar{t}, \bar{x}, \bar{k}) \tag{4.5}$$

and

$$Y(\bar{t}, \bar{x}, \bar{k}) \geq \mathcal{U}(\bar{t}, \bar{x}, \bar{k}), \tag{4.6}$$

the proof of the converse inequality of (4.5) (resp., (4.6)) being quite similar to the proof of (4.6) (resp., (4.5)). Moreover, in view of Proposition 4.1, the proofs of (4.3) and (4.4) can be straightforwardly obtained by the ones of (4.1) and (4.2), respectively, by formally replacing $\mathcal{C}(\bar{t}, \bar{k})$ with $\mathcal{C}_\sigma(\bar{t}, \bar{k})$ and $\Gamma(\bar{t}, \bar{k})$ with $\Gamma_\sigma(\bar{t}, \bar{k})$.

Let us fix a $\sigma \in]0, T - \bar{t}[$. For a given $\epsilon > 0$ there exists a strategy $\alpha_1 \in \Delta(\bar{t}, \bar{k})$ such that

$$\begin{aligned} W(\bar{t}, \bar{x}, \bar{k}) \geq \sup_{c \in \mathcal{C}(\bar{t}, \bar{k})} \Bigg\{ & \int_{\bar{t}}^{t_c(\sigma)} l\big(t, x(t), \alpha_1(c)(t), c\big)\, dt \\ & + \mathcal{V}\big(t_c(\sigma), x \circ t_c(\sigma), k \circ t_c(\sigma)\big) \Bigg\} - \epsilon. \end{aligned} \tag{4.7}$$

For every $c \in \mathcal{C}(\bar{t}, \bar{k})$ there is a strategy $\alpha_c \in \Delta\big(t_c(\sigma), k(t_c(\sigma))\big)$ such that

$$\begin{aligned} & \mathcal{V}\big(t_c(\sigma), x \circ t_c(\sigma), k \circ t_c(\sigma)\big) \\ & \quad \geq \sup_{\tilde{c} \in \mathcal{C}\left(t_c(\sigma), k \circ t_c(\sigma)\right)} P[t_c(\sigma), x \circ t_c(\sigma); \alpha_c(\tilde{c}), \tilde{c}] - \epsilon. \end{aligned} \tag{4.8}$$

Let us define a strategy $\alpha_2 \in \mathcal{A}(\bar{t})$ by setting

$$\alpha_2(c)(t) \doteq \begin{cases} \alpha_1(c)(t) & \forall t \in [\bar{t}, t_c(\sigma)[, \\ \alpha_c\left(c_{|_{[t_c(\sigma), T]}}\right)(t), & t \in [t_c(\sigma), T], \end{cases}$$

where, for every $c \in \mathcal{C}(\bar{t}, \bar{k})$, $c_{|_{[t_c(\sigma), T]}}$ denotes the restriction of c to the interval $[t_c(\sigma), T]$. It is not difficult to verify that α_2 is in fact nonanticipating. By (4.7), (4.8), we obtain, for every $c \in \mathcal{C}(\bar{t}, \bar{k})$,

$$W(\bar{t}, \bar{x}, \bar{k}) \geq P[\alpha_2(c), c] - 2\epsilon,$$

which, by the arbitrariness of ϵ, yields (4.5).

In order to prove (4.6), let us select a strategy $\gamma_1 \in \Gamma_\sigma(\bar{t}, \bar{k})$ satisfying

$$\mathcal{U}(\bar{t}, \bar{x}, \bar{k}) \leq \inf_{a \in \mathcal{A}(\bar{t})} P[\bar{t}, \bar{x}; a, \gamma_1(a)] + \epsilon.$$

By the definition of Y one obtains

$$Y(\bar t,\bar x,\bar k)\geq \inf_{a\in\mathcal{A}(\bar t)}\left\{\int_{\bar t}^{t_{\gamma_1(a)}(\sigma)} l\big(t,x(t),a(t),\gamma_1(a)(t)\big)\ dt \right.$$
$$\left. +\,\mathcal{U}\left(t_{\gamma_1(a)}(\sigma),x\circ t_{\gamma_1(a)}(\sigma),k\circ t_{\gamma_1(a)}(\sigma)\right)\right\},$$

thence there exists an $a_1\in\mathcal{A}(\bar t)$ so that

$$Y(\bar t,\bar x,\bar k)+\epsilon\geq \int_{\bar t}^{t_{\gamma_1(a_1)}(\sigma)} l\big(t,x(t),a_1(t),\gamma_1(a_1)(t)\big)\ dt$$
$$+\,\mathcal{U}\left(t_{\gamma_1(a_1)}(\sigma),x\circ t_{\gamma_1(a_1)}(\sigma),k\circ t_{\gamma_1(a_1)}(\sigma)\right).$$

For every $a\in\mathcal{A}(t_{\gamma_1(a_1)}(\sigma))$ let us define a control $\hat a\in\mathcal{A}(\bar t)$ by setting

$$\hat a(t)\doteq\begin{cases} a_1 & \forall t\in[\bar t,t_{\gamma_1(a_1)}(\sigma)],\\ a & \forall t\in]t_{\gamma_1(a_1)}(\sigma),T].\end{cases}$$

The function $\gamma_2:\mathcal{A}(t_{\gamma_1(a_1)}(\sigma))\to\mathcal{C}\big(t_{\gamma_1(a_1)}(\sigma),k\circ t_{\gamma_1(a_1)}(\sigma)\big)$, which maps a control $a\in\mathcal{A}(t_{\gamma_1(a_1)}(\sigma))$ into the control

$$\gamma_2(a)(t)\doteq(\gamma_1(\hat a))(t),\qquad t\in[t_{\gamma_1(a_1)}(\sigma),T],$$

turns out to belong to $\Gamma\big(t_{\gamma_1(a_1)}(\sigma),k\circ t_{\gamma_1(a_1)}(\sigma)\big)$. Indeed, it is nonanticipating, and for every $a\in\mathcal{A}(t_{\gamma_1(a_1)}(\sigma))$, one has

$$\int_{t_{\gamma_1(a_1)}(\sigma)}^{T}|\gamma_2(a)(t)|dt\leq K-\bar k-\left(k\circ t_{\gamma_1(a_1)}(\sigma)-\bar k\right).$$

In particular, it follows that

$$\mathcal{U}\left(t_{\gamma_1(a_1)}(\sigma),x\circ t_{\gamma_1(a_1)}(\sigma),k\circ t_{\gamma_1(a_1)}(\sigma)\right)$$
$$\geq\inf_{a\in\mathcal{A}(t_{\gamma_1(a_1)}(\sigma))}P[a,\gamma_2(a)].$$

Hence, there exists a control $a_2\in\mathcal{A}(t_{\gamma_1(a_1)}(\sigma))$ so that

$$\mathcal{U}\left(t_{\gamma_1(a_1)}(\sigma),x\circ t_{\gamma_1(a_1)}(\sigma),k\circ t_{\gamma_1(a_1)}(\sigma)\right)$$
$$\geq P[a_2,\gamma_2(a_2)]-\epsilon.$$

Finally, the previous inequalities imply that

$$Y(\bar t,\bar x,\bar k)\geq\int_{\bar t}^{t_{\gamma_1(a_1)}(\sigma)} l\left(t,x(t),a_1(t),\gamma_1(a_1)(t)\right)\ dt$$
$$+\int_{t_{\gamma_1(a_1)}(\sigma)}^{T} l\left(t,x(t),a_2(t),\gamma_2(a_2)(t)\right)\ dt+g(x(T))-2\epsilon$$
$$\geq P[\hat a_2,\gamma_1(\hat a_2)]-2\epsilon\geq\inf_{a\in\mathcal{A}(\bar t)}P[a,\gamma_1(a)]-2\epsilon\geq\mathcal{U}(\bar t,\bar x,\bar k)-3\epsilon,$$

from which (4.6) follows, in view of the the arbitrariness of ϵ. □

As a consequence of Theorem 4.1 and of Propositions 3.3–3.5 and 3.7–3.9, we obtain the reparameterized dynamic programming principle (RDPP) below. In turn, the latter allows us to establish boundary value problems for $\mathcal{U}$ and $\mathcal{V}$ involving continuous Hamiltonians and rather simple boundary conditions (see sections 5 and 6). In the statetement of Theorem 4.2 below and in the rest of the paper we adopt the following convention: whenever an initial condition $(\bar t, \bar x, \bar k)$, a control (v, w_0, w), and a triple $(t, z, k)(\cdot)$ appear in the same formula, it is meant that

$$(t,z,k)(s) \doteq \left(\bar t + \int_0^s w_0(\eta)\, d\eta, z[\bar t, \bar x; v, w_0, w](s), \bar k + \int_0^s |w(\eta)|\, d\eta\right)$$

for every $s \in [0,1]$.

Let us define the subsets $\tilde{\mathcal{C}}_\sigma(\bar t, \bar k)$ and $\tilde\Gamma_\sigma(\bar t, \bar k)$ by setting

$$\begin{aligned}\tilde{\mathcal{C}}_\sigma(\bar t, \bar k) &\doteq \big\{(w_0, w) \in \tilde{\mathcal{C}}(\bar t, \bar k) \quad | \quad (w_0, w)_* \in \mathcal{C}_\sigma(\bar t, \bar k)\big\},\\ \tilde\Gamma_\sigma(\bar t, \bar k) &\doteq \big\{(\xi_0, \xi) \in \tilde\Gamma(\bar t, \bar k) \quad | \quad (\xi_0, \xi)_* \in \Gamma_\sigma(\bar t, \bar k)\big\}.\end{aligned}$$

Clearly, for every $\sigma \in]0, \max\{T - \bar t, K - \bar k\}[$ one has $\tilde{\mathcal{C}}_\sigma(\bar t, \bar k) \neq \emptyset$ and $\tilde\Gamma_\sigma(\bar t, \bar k) \neq \emptyset$.

THEOREM 4.2 (reparameterized dynamic programming principle (RDPP)). *For every $(\bar t, \bar x, \bar k) \in [0, T[\times\mathbb{R}^n \times [0, K]$ and $0 < \sigma < T - \bar t$ one has*

$$\mathcal{V}(\bar t, \bar x, \bar k) = \inf_{\nu \in \tilde\Delta(\bar t, \bar k)} \sup_{(w_0, w) \in \tilde{\mathcal{C}}(\bar t, \bar k)} \left\{ \int_0^\sigma \bar l\big(t(s), z(s), (\nu(w_0, w))_* \circ t(s), w_0(s), w(s)\big)\, ds + \mathcal{V}(t(\sigma), z(\sigma), k(\sigma)) \right\}$$

and

$$\mathcal{U}(\bar t, \bar x, \bar k) = \sup_{(\xi_0, \xi) \in \tilde\Gamma(\bar t, \bar k)} \inf_{v \in \tilde{\mathcal{A}}} \left\{ \int_0^\sigma \bar l\big(t(s), z(s), v_* \circ t(s), \xi_0(v)(s), \xi(v)(s)\big)\, ds + \mathcal{U}(t(\sigma), z(\sigma), k(\sigma)) \right\}.$$

Moreover, if $Q \subset \mathbb{R}^n$ is a compact subset and $(\bar t, \bar x, \bar k) \in [0, T[\times Q \times [0, K[$ verifies $\mathcal{U}(\bar t, \bar x, \bar k) - g(x) \geq \eta > 0$ (resp., $\mathcal{V}(\bar t, \bar x, \bar k) - g(x) \geq \eta > 0$), there exists a $\hat\sigma > 0$ such that

$$\mathcal{V}(\bar t, \bar x, \bar k) = \inf_{\nu \in \tilde\Delta(\bar t, \bar k)} \sup_{(w_0, w) \in \tilde{\mathcal{C}}_\sigma(\bar t, \bar k)} \left\{ \int_0^\sigma \bar l\big(t(s), z(s), (\nu(w_0, w))_* \circ t(s), w_0(s), w(s)\big)\, ds + \mathcal{V}(t(\sigma), z(\sigma), k(\sigma)) \right\}$$

(resp.,

$$\mathcal{U}(\bar t, \bar x, \bar k) = \sup_{(\xi_0, \xi) \in \tilde\Gamma_\sigma(\bar t, \bar k)} \inf_{v \in \tilde{\mathcal{A}}} \left\{ \int_0^\sigma \bar l\big(t(s), z(s), v_* \circ t(s), \xi_0(v)(s), \xi(v)(s)\big)\, ds + \mathcal{U}(t(\sigma), z(\sigma), k(\sigma)) \right\}.\Big)$$

5. Hamilton–Jacobi equations for $\mathcal{U}$ and $\mathcal{V}$. In this section we establish Hamilton–Jacobi differential equations for $\mathcal{U}$ and $\mathcal{V}$. These equations are not formal extensions of the ones concerning games with bounded controls. Indeed, this would imply the use of *discontinuous* Hamiltonians. On the contrary, the space-time version of the game presented in section 3 and the consequent RDPP (Theorem 4.2) allow one to obtain equations involving continuous Hamiltonians.

For the reader's convenience let us recall the definition of viscosity solution (see, e.g., [11], [16]).

DEFINITION 5.1. *Let E be any subset of $\mathbb{R}^N$ and let $F : E\times\mathbb{R}\times\mathbb{R}^N \to \mathbb{R}$ be a continuous function. A map $u \in C^0(E)$ is a viscosity subsolution (resp., supersolution) at $y \in E$ of the first-order partial differential equation*

$$F(y, u, \nabla u) = 0 \tag{PDE}$$

if for any $\varphi \in C^\infty(\mathbb{R}^N)$ such that $u - \varphi$ has a local maximum (resp., minimum) at y on E, one has

$$F(y, u, \nabla\varphi) \le 0 \qquad (\textit{resp.,}\ F(y, u, \nabla\varphi) \ge 0),$$

where $\nabla\varphi$ denotes the gradient of φ. A map $u \in C^0(E)$ is called a viscosity solution of (PDE) *at y if it is both a viscosity subsolution and a viscosity supersolution.*

Let us set $\Omega \doteq [0, T[\times\mathbb{R}^n \times [0, K[$. Moreover, let us define the *upper Hamiltonian* H^+ and the *lower Hamiltonian* H^- by setting, for every $(t, x, p_t, p_x, p_k) \in (\mathbb{R}\times\mathbb{R}^n)\times(\mathbb{R}\times\mathbb{R}^n\times\mathbb{R})$,

$$H^+(t, x, p_t, p_x, p_k) \doteq \min_{a\in A} \max_{(w_0,w)\in S^m_+} \big\{p_t w_0 + p_x \cdot \bar f(t, x, a, w_0, w) + \bar l(t, x, a, w_0, w) + p_k|w|\ \big\}$$

and

$$H^-(t, x, p_t, p_x, p_k) \doteq \max_{(w_0,w)\in S^m_+} \min_{a\in A} \big\{p_t w_0 + p_x \cdot \bar f(t, x, a, w_0, w) + \bar l(t, x, a, w_0, w) + p_k|w|\ \big\},$$

respectively, where $S^m_+ \doteq \{(w_0, w) : |(w_0, w)| = 1, w_0 \ge 0\}$.

If Φ is a map defined on Ω, let us denote the gradients of Φ with respect to t, x, and k by $\nabla_t\Phi$, $\nabla_x\Phi$, and $\nabla_k\Phi$, respectively.

THEOREM 5.1. *The maps $\mathcal{V}$ and $\mathcal{U}$ are viscosity solutions on Ω of the lower value equation*

$$-H^-(t, x, \nabla_t\Phi, \nabla_x\Phi, \nabla_k\Phi) = 0 \tag{LVE}$$

and of the upper value equation

$$-H^+(t, x, \nabla_t\Phi, \nabla_x\Phi, \nabla_k\Phi) = 0, \tag{UVE}$$

respectively.

We will just outline the proof of this theorem. Actually, in view of the RDPP, the proof is partially based on some arguments already exploited in the conventional case [13]. However, some changes are needed, due to the fact that within our setting

the integral constraint on c becomes a state constraint, while the finite horizon $t = T$ has to be regarded as a target.

Proof. Let us prove that $\mathcal{V}$ is a subsolution of (LVE). Let $(\bar{t}, \bar{x}, \bar{k}) \in \Omega$ and let φ be a smooth function such that $\mathcal{V} - \varphi$ has a local maximum at $(\bar{t}, \bar{x}, \bar{k})$. Hence, in a neighborhood of $(\bar{t}, \bar{x}, \bar{k})$ one has

$$\mathcal{V}(t, x, k) - \mathcal{V}(\bar{t}, \bar{x}, \bar{k}) \le \varphi(t, x, k) - \varphi(\bar{t}, \bar{x}, \bar{k}). \tag{5.1}$$

If we define

$$\begin{aligned} &Q(t, x, k, a, w_0, w) \\ &\doteq \nabla_t \varphi(t, x, k) \cdot w_0 + \nabla_x \varphi(t, x, k) \cdot \bar{f}(t, x, a, w_0, w) + \bar{l}(t, x, a, w_0, w) + \nabla_k \varphi(t, x, k) \cdot |w|, \end{aligned}$$

the thesis is expressed by the inequality

$$- \max_{(w_0, w) \in S^m_+} \min_{a \in A} Q(\bar{t}, \bar{x}, \bar{k}, a, w_0, w) \le 0.$$

Assume by contradiction that

$$\max_{(w_0, w) \in S^m_+} \min_{a \in A} Q(\bar{t}, \bar{x}, \bar{k}, a, w_0, w) \le -\theta,$$

with $\theta > 0$. By Lemma 5.1 below there exists a strategy $\nu \in \widetilde{\Delta}(\bar{t}, \bar{k})$ and a $\bar{\sigma} \in]0, T - \bar{t}[$ such that, for every $(w_0, w) \in \widetilde{\mathcal{C}}_\sigma(\bar{t}, \bar{k})$ and every $\sigma \le \bar{\sigma}$, one has

$$\int_0^\sigma Q(t(s), z(s), (\nu(w_0, w))_* \circ t(s), w_0(s), w(s)) \le -\frac{\theta \sigma}{2}.$$

By applying the RDPP we obtain a contradiction. Indeed, for every $\sigma \le \min\{\bar{\sigma}, T - \bar{t}\}$ one has

$$\begin{aligned} 0 &= \inf_{\nu \in \widetilde{\Delta}(\bar{t}, \bar{k})} \sup_{(w_0, w) \in \widetilde{\mathcal{C}}(\bar{t}, \bar{k})} \left\{ \int_0^\sigma \bar{l}\big(t(s), z(s), (\nu(w_0, w))_* \circ t(s), w_0(s), w(s)\big)\, ds \right. \\ &\qquad \left. + \mathcal{V}(t(\sigma), z(\sigma), k(\sigma)) - \mathcal{V}(\bar{t}, \bar{x}, \bar{k}) \right\} \\ &\le \inf_{\nu \in \widetilde{\Delta}(\bar{t}, \bar{k})} \sup_{(w_0, w) \in \widetilde{\mathcal{C}}(\bar{t}, \bar{k})} \left\{ \int_0^\sigma \bar{l}\big(t(s), z(s), (\nu(w_0, w))_* \circ t(s), w_0(s), w(s)\big)\, ds \right. \\ &\qquad \left. + \varphi(t(\sigma)), z(\sigma), k(\sigma)) - \varphi(\bar{t}, \bar{x}, \bar{k}) \right\} \\ &= \inf_{\nu \in \widetilde{\Delta}(\bar{t}, \bar{k})} \sup_{(w_0, w) \in \widetilde{\mathcal{C}}(\bar{t}, \bar{k})} \left\{ \int_0^\sigma Q\big(t(s), z(s), (\nu(w_0, w))_* \circ t(s), w_0(s), w(s)\big)\, ds \right\} \\ &\le -\frac{\theta \sigma}{2}. \end{aligned}$$

In view of Lemma 5.1 below, the proofs that $\mathcal{V}$ and $\mathcal{U}$ are a supersolution of (LVE) and a solution of (UVE) on Ω, respectively, proceed similarly. Hence we omit them. □

LEMMA 5.1. *Let Q be defined as in the proof of the previous theorem. Then the following hold.*

(i) *If*

$$\max_{(w_0,w)\in S^m_+}\min_{a\in A} Q(\bar t,\bar x,\bar k,a,w_0,w)\le -\theta \qquad (\textit{resp.},\ \ge\theta),$$

there exist a $\bar\sigma\in]0,T-\bar t[$ *and a strategy* $\nu\in\widetilde\Delta(\bar t,\bar k)$ *(resp., an* $\epsilon>0$ *and a control* $(w_0,w)\in\widetilde{\mathcal C}(\bar t,\bar k)$ $(=\widetilde{\mathcal C}_{\bar\sigma}(\bar t,\bar k))$*, with* $w_0\ge\epsilon$*) such that, for every* $(w_0,w)\in\widetilde{\mathcal C}(\bar t,\bar k)$ $(=\widetilde{\mathcal C}_{\bar\sigma}(\bar t,\bar k))$ *(resp., every strategy* $\nu\in\widetilde\Delta(\bar t,\bar k)$*) and every* $\sigma\le\bar\sigma$*, one has*

$$\int_0^\sigma Q(t(s),z(s),k(s),(\nu(w_0,w))_*\circ t(s),w_0(s),w(s))\le -\frac{\theta\sigma}{2}\qquad\left(\textit{resp.},\ \ge\frac{\theta\sigma}{2}\right).$$

(ii) *If*

$$\min_{a\in A}\max_{(w_0,w)\in S^m_+} Q(\bar t,\bar x,\bar k,a,w_0,w)\ge\theta\qquad(\text{resp.},\ \le-\theta),$$

there exist a $\bar\sigma\in]0,T-\bar t[$*, an* $\epsilon>0$*, and a strategy* $(\xi_0,\xi)\in\widetilde\Gamma(\bar t,\bar k)$ $(=\widetilde\Gamma_{\bar\sigma}(\bar t,\bar k))$ *verifying* $\xi_0(v)\ge\epsilon\ \forall v\in\widetilde{\mathcal A}$ *(resp., a control* $v\in\widetilde{\mathcal A}$*) such that, for every* $v\in\widetilde{\mathcal A}$ *(resp., every strategy* $(\xi_0,\xi)\in\widetilde\Gamma(\bar t,\bar k)$ $(=\widetilde\Gamma_{\bar\sigma}(\bar t,\bar k)))$ *and every* $\sigma\le\bar\sigma$*, one has*

$$\int_0^\sigma Q(t(s),z(s),k(s),v_*\circ t(s),\xi_0(v)(s),\xi(v)(s))\ge\frac{\theta\sigma}{2}\qquad\left(\textit{resp.},\le-\frac{\theta\sigma}{2}\right).$$

Proof. We shall limit ourselves to prove only the second of the statements in (i) and the first of the statements in (ii), the proofs of the remaining two statements being similar.

Let us assume

$$\max_{(w_0,w)\in S^m_+}\min_{a\in A} Q(\bar t,\bar x,\bar k,a,w_0,w)\ge\theta.$$

Hence there exists an $\epsilon>0$ and a $(\bar w_0,\bar w)\in S^m_+$, with $\bar w_0>\epsilon$, such that

$$Q(\bar t,\bar x,\bar k,a,\bar w_0,\bar w)\ge\frac34\theta$$

for every $a\in A$. Since A is a compact subset, by the hypotheses on f and l, there exists a σ such that for every $s\le\sigma$ and $v\in\widetilde{\mathcal A}(\bar t)$ one has

$$|(t(s),z(s),k(s))-(\bar t,\bar x,\bar k)|\le Ms,$$

where we have set $\big(t(s),z(s),k(s)\big)\doteq((t,z)[v,\bar w_0,\bar w](s),\bar k+\int_0^s|\bar w|(\eta)d\eta\big)$. Hence, since Q is uniformly continuous, by choosing a suitable $\bar\sigma>0$ one has

$$Q(t(s),z(s),k(s),(\nu(\bar w_0,\bar w))_*\circ t(s),\bar w_0,\bar w)\ge\frac12\theta$$

for every $s\in[0,\bar\sigma]$ and every $\nu\in\widetilde\Delta(\bar t,\bar x)$, from which the thesis follows, with $(w_0,w)(s)\doteq(\bar w_0,\bar w)$.

Now let us assume that

$$\min_{a\in A}\max_{(w_0,w)\in S^m_+} Q(\bar t,\bar x,\bar k,a,w_0,w)\ge\theta.$$

By the continuity of Q and the compactness of A this implies that there exists an $\epsilon > 0$ such that

$$\max_{(w_0,w)\in S^m_+, w_0\geq\epsilon} Q(\bar t,\bar x,\bar k,a,w_0,w)\geq \frac{5}{6}\theta$$

for every $a\in A$. Since the multivalued map

$$W(a)\doteq \operatorname{argmax}\big\{Q(\bar t,\bar x,\bar k,a,w_0,w),\ (w_0,w)\in S^m_+, w_0\geq\epsilon\big\}$$

is upper semicontinuous, it admits a Borel measurable selection, say, $(w_0^a,w^a)\in W(a)$ (see, e.g., [1]). The map $\tilde c: A\to C$ defined by $\tilde c(a)\doteq \frac{w^a}{w_0^a}$ turns out to be Borel measurable. Then it is easy to check that the functional γ which maps a control $a\in\mathcal{A}(\bar t)$ into the control

$$\gamma(a)(t)\doteq\begin{cases}\tilde c\big(a(t)\big) & \forall t\in[\bar t,t_a],\\ 0 & \forall t\in]t_a,T],\end{cases}$$

with

$$t_a\doteq\sup\left\{t\in[\bar t,T]:\int_{\bar t}^t|\tilde c\circ a(\tau)|\,d\tau\leq K-\bar k\right\},$$

is a nonanticipating map from $\mathcal{A}(\bar t)$ into $\mathcal{C}(\bar t,\bar k)$; i.e., $\gamma\in\Gamma(\bar t,\bar k)$. Let us consider the strategy $(\xi_0,\xi)\doteq\gamma^*\in\widetilde\Gamma(\bar t,\bar k)$. As in the first part of the proof, we can find a $\bar\sigma\in]0,T-\bar t[$ such that, for every $v\in\widetilde{\mathcal{A}}(\bar t)$ and $s\in[0,\bar\sigma]$, one has

$$|Q\big(t(s),z(s),k(s),v_*\circ t(s),(\xi_0,\xi)(v)(s)\big)-Q\big(\bar t,\bar x,\bar k,v_*\circ t(s),(\xi_0,\xi)(v)(s)\big)\leq\frac{2}{6}\theta,$$

where

$$(t,z,k)(s)\doteq\left((t,z)[\bar t,\bar x;v,(\xi_0,\xi)(v)](s),\bar k+\int_0^s|\xi(v)(\eta)|d\eta\right).$$

Since $(\xi_0,\xi)(v)(s)=\big(w_0^{v_*(t(s))},w^{v_*(t(s))}\big)$ we obtain that, for every $s\in[0,\bar\sigma]$,

$$Q\big(t(s),z(s),k(s),v_*\circ t(s),(\xi_0,\xi)(v)(s)\big)\geq\frac{5}{6}\theta-\frac{2}{6}\theta=\frac{1}{2}\theta,$$

which yields the thesis. □

6. Boundary value problems for $\mathcal{U}$ and $\mathcal{V}$. In this section we establish the LVBVP and the UVBVP below, whose unique bounded continuous solutions coincide with $\mathcal{V}$ and $\mathcal{U}$, respectively.

On the basis of Theorem 2.1 the maps $\mathcal{U}$ $(=\widetilde{\mathcal{U}})$ and $\mathcal{V}$ $(=\widetilde{\mathcal{V}})$ can be continuously extended on the closure of $[0,T[\times\mathbb{R}^n\times[0,K]$.

Let $\partial\Omega$ denote the boundary of Ω and let us set

$$\partial_T\Omega\doteq\{T\}\times\mathbb{R}^n\times[0,K],\qquad \partial'\Omega\doteq\partial\Omega\backslash\partial_T\Omega.$$

DEFINITION 6.1. *We say that a continuous map Φ is a solution of the LVBVP (resp., UVBVP) if the following hold:* (1) Φ *is a viscosity solution of* (LVE) (*resp.,* (UVE)) *in* Ω; (2) $\Phi(T,x,k)\geq g(x)$ *for every* $(T,x,k)\in\partial_T\Omega$; (3) Φ *is a viscosity*

subsolution of (LVE) (*resp., of* (UVE)) *on* $\partial'\Omega$ *and at any point* $(T,\bar x,\bar k)\in\partial_T\Omega$ *such that* $\Phi(T,\bar x,\bar k)>g(\bar x)$.

THEOREM 6.1. *The maps* $\mathcal{V}$ *and* $\mathcal{U}$ *are the unique bounded solutions of the LVBVP and the UVBVP, respectively.*

Proof. The fact that $\mathcal{V}$ and $\mathcal{U}$ satisfy (1) in Definition 6.1 coincides with the statement of Theorem 5.1.

In order to prove that $\mathcal{V}$ and $\mathcal{U}$ verify (2) in Definition 6.1, let us consider the constant space-time control $(\hat w_0,\hat w)(s)=(1,0)$, $s\in[0,\frac1n]$, which belongs to $\widetilde{\mathcal{C}}(T-\frac1n,k)$.

For each control $v:[0,1]\to A$ the trajectory

$$(t(s),z(s))\doteq(t,z)[T-\frac1n,\bar x,k;v,\hat w_0,\hat w](s)$$

verifies

$$\left|z\left(\frac1n\right)-\bar x\right|\le M\frac1n, \tag{6.1}$$

$$\left|\int_0^{\frac1n}\bar l(t(s),z(s),v_*\circ t(s),\hat w_0(s),\hat w(s))\,ds\right|\le M\frac1n$$

for a suitable $M>0$. For every strategy ν one has

$$\int_0^{\frac1n}\bar l(t(s),z(s),(\nu(\hat w_0,\hat w)_*)\circ t(s),\hat w_0(s),\hat w(s))\,ds+g\left(z\left(\frac1n\right)\right)$$

$$\le\sup_{(w_0,w)\in\widetilde{\mathcal{C}}(T-\frac1n,k)}\left\{\int_0^L\bar l(t(s),z(s),(\nu(w_0,w)_*)\circ t(s),w_0(s),w(s))\,ds+g(z(L))\right\},$$

where the L in the expression in brackets denotes the right endpoint of the domain of (w_0,w). As a consequence, one has

$$\mathcal{V}\left(T-\frac1n,\bar x,k\right)$$

$$\ge\inf_{\nu\in\widetilde\Delta(T-\frac1n,k)}\left\{\int_0^{\frac1n}\bar l(t(s),z(s),(\nu(\hat w_0,\hat w))_*\circ t(s),\hat w_0(s),\hat w(s))\,ds+g\left(z\left(\frac1n\right)\right)\right\}$$

for every natural number n. Together with 6.1, this implies

$$\mathcal{V}(T,\bar x,k)\ge g(\bar x).$$

On the other hand, for every $n\in\mathbb{N}$, there exists a control $\hat v\in\widetilde{\mathcal{A}}$ such that

$$\int_0^{\frac1n}\bar l(t(s),z(s),\hat v_*\circ t(s),\hat w_0(s),\hat w(s))\,ds+g\left(z\left(\frac1n\right)\right)$$

$$\le\inf_{v\in\widetilde{\mathcal{A}}}\left\{\int_0^{\frac1n}\bar l(t(s),z(s),v_*\circ t(s),\hat w_0(s),\hat w(s))\,ds+g\left(z\left(\frac1n\right)\right)\right\}+\epsilon$$

$$\le\mathcal{U}\left(T-\frac1n,\bar x,k\right)+\epsilon,$$

which, in view of (6.1), implies

$$g(\bar x)\le\mathcal{U}(T,\bar x,k).$$

Let us prove that $\mathcal{V}$ and $\mathcal{U}$ verify condition (3) of Definition 6.1. The part of the proof of Theorem 5.1 which concerns the subsolution properties of $\mathcal{V}$ and $\mathcal{U}$ can be easily extended in order to include the points of $[0,T[\times\mathbb{R}^n\times\{K\}$. Hence, in order to verify (3) in Definition 6.1, it remains to prove that $\mathcal{V}$ (resp., $\mathcal{U}$) is a viscosity subsolution of (LVE) (resp., of (UVE)) at any point of $\partial_T\Omega$ where $\mathcal{V} > g$ (resp., $\mathcal{U} > g$).

Let $(T,\bar{x},\bar{k})$ verify $\bar{k} < K$ and $\mathcal{V}(T,\bar{x},\bar{k}) > g(\bar{x})$, and let φ be a smooth function such that $\mathcal{V}-\varphi$ has a maximum at $(T,\bar{x},\bar{k}) \in \{T\}\times\mathbb{R}^n\times[0,K[$. Hence, in a neighborhood of $(T,\bar{x},\bar{k})$, one has

$$\mathcal{V}(t,x,k) - \mathcal{V}(T,\bar{x},\bar{k}) \le \varphi(t,x,k) - \varphi(T,\bar{x},\bar{k}).$$

Setting, as in the proof of Theorem 5.1,

$$\begin{aligned} Q(t,x,k,a,w_0,w) \doteq \nabla_t\varphi(t,x,k)w_0 + &\nabla_x\varphi(t,x,k)\cdot\bar{f}(t,x,a,w_0,w) \\ &+ \bar{l}(t,x,a,w_0,w) + \nabla_k\varphi(t,x,k)|w|, \end{aligned}$$

we have to prove that

$$\max_{(w_0,w)\in S^m_+}\min_{a\in A} Q(T,\bar{x},\bar{k},a,w_0,w) \ge 0.$$

Assume, by contradiction, that there exists a positive θ such that

$$\min_{a\in A} Q(T,\bar{x},\bar{k},a,w_0,w) \le -\theta$$

for all $(w_0,w)\in S^m_+$. Rephrasing the second part of the proof of Lemma 5.1 one can find a Borel measurable map $v(w_0,w)$ such that

$$Q(T,\bar{x},\bar{k},v(w_0,w),w_0,w) \le -\theta$$

for every $(w_0,w)\in S^m_+$. Let $\{\bar{t}_n\}$ be a sequence converging to T from the left, and let us define the strategies $\nu_n \in \widetilde{\Gamma}(\bar{t}_n,\bar{k})$ as follows. If $(w_0,w)\in\widetilde{\mathcal{C}}(\bar{t}_n,\bar{k})$, set $t_n(s) \doteq \bar{t}_n + \int_0^s w_0(\eta)\,d\eta$ and let $s_n(\cdot)$ denote the inverse of $t_n(\cdot)$. Then it is not difficult to check that the maps $\nu_n : \widetilde{\mathcal{C}}(\bar{t}_n,\bar{k}) \to \widetilde{\mathcal{A}}$ defined by

$$\nu_n(w_0,w) \doteq (v\circ(w_0,w)\circ s_n)^*$$

are strategies (see Definition 3.1). From the inequality

$$Q(T,\bar{x},\bar{k},(\nu_n(w_0,w))_*\circ t_n(s),w_0(s),w(s)) \le -\theta,$$

valid for every control $(w_0,w)\in\widetilde{\mathcal{C}}(\bar{t}_n,\bar{k})$ and every $s\in[0,s_n(T)]$, it follows that there exist $\epsilon,\bar{\sigma}>0$ such that, if $\bar{t}_n\in[T-\epsilon,T]$ and $0\le s\le\bar{\sigma}$, one has

$$Q\big(t_n(s),z(s),k(s),(\nu_n(w_0,w))_*\circ t_n(s),w_0(s),w(s)\big) \le -\frac{\theta}{2},$$

where $(t_n,z)(s) \doteq (t,z)[\bar{t}_n,\bar{x},\nu_n(w_0,w),w_0,w](s)$, $k(s)\doteq\bar{k}+\int_0^s|w(\xi)|d\xi$. Integrating the terms in the last inequality, for every $\sigma\le\bar{\sigma}$ and every $(w_0,w)\in\widetilde{\mathcal{C}}_\sigma(\bar{t}_n,\bar{k})$, one has

$$\begin{aligned} \int_0^\sigma \bar{l}\big(t_n(s),z(s),(\nu_n(w_0,w))_*\circ t_n(s),w_0(s),w(s)\big)\,ds \\ + \varphi(t_n(\sigma),z(\sigma),k(\sigma)) - \varphi(\bar{t}_n,\bar{x},\bar{k}) \le \frac{-\theta\sigma}{2}. \end{aligned} \tag{6.2}$$

On the other hand, in a neighborhood of $(T,\bar x,\bar k)$ one has $\mathcal{V}-g>\eta>0$. Hence, on the basis of the RDPP (Theorem 4.2), there exists a $\hat\sigma>0$ such that for every $\sigma<\hat\sigma$

$$0\le \sup_{(w_0,w)\in\widetilde{\mathcal{C}}_\sigma(\bar t_n,\bar k)}\Big\{\int_0^\sigma \bar l\big(t_n(s),z(s),(\nu_n(w_0,w))_*\circ t_n(s),w_0(s),w(s)\big)\,ds \\ +\mathcal{V}\big(t_n(\sigma),z(\sigma),k(\sigma)\big)-\mathcal{V}(\bar t_n,\bar x,\bar k)\Big\}.$$

In particular, for every $\sigma<\hat\sigma$ and every n, there exists a control $(w_{0\,n}^\sigma,{w^\sigma}_n)\in\widetilde{\mathcal{C}}_\sigma(\bar t_n,\bar k)$ such that

$$\int_0^\sigma \bar l(t_n^\sigma(s),z_n^\sigma(s),(\nu_n(w_{0\,n}^\sigma,{w^\sigma}_n))_*\circ t_n^\sigma(s),w_{0\,n}^\sigma(s),{w^\sigma}_n(s))\,ds \\ +\mathcal{V}\big(t_n^\sigma(\sigma),z_n^\sigma(\sigma),k_n^\sigma(\sigma)\big)-\mathcal{V}(\bar t_n,\bar x,\bar k)>-\frac{\theta\sigma}{4}, \tag{6.3}$$

where we have set

$$z_n^\sigma(s)\doteq z[\bar t_n,\bar x;\nu_n(w_{0\,n}^\sigma,{w^\sigma}_n),w_{0\,n}^\sigma,{w^\sigma}_n](s),$$
$$k_n^\sigma(s)\doteq\bar k+\int_0^s|{w^\sigma}_n(\mu)|\,d\mu.$$

Setting

$$p_n^\sigma\doteq\int_0^\sigma \bar l\,(t_n^\sigma(s),z_n^\sigma(s),(\nu_n(w_{0\,n}^\sigma,{w^\sigma}_n))_*\circ t_n^\sigma(s),w_{0\,n}^\sigma(s),{w^\sigma}_n(s))\;ds,$$

by Ascoli–Arzelà's theorem we obtain the existence of a subsequence $\{z_{n'}^\sigma,k_{n'}^\sigma,p_{n'}^\sigma\}$ such that $\{z_{n'}^\sigma\}$ and $\{k_{n'}^\sigma\}$ converge to continuous maps z_∞^σ and k_∞^σ, respectively, uniformly on $[0,\sigma]$, and $\{p_{n'}^\sigma\}$ converge to a constant p_∞^σ.

For any $\sigma\in]0,\min\{\bar\sigma,\hat\sigma\}[$, letting n tend to infinity in (6.2)–(6.3), one obtains

$$-\frac{\theta\sigma}{4}\le p_\infty^\sigma+\mathcal{V}(T,z_\infty^\sigma(\sigma),k_\infty^\sigma(\sigma))-\mathcal{V}(T,\bar x,\bar k)$$
$$\le p_\infty^\sigma+\varphi(T,z_\infty^\sigma(\sigma),k_\infty^\sigma(\sigma))-\varphi(T,\bar x,\bar k)\le-\frac{\theta\sigma}{2},$$

a contradiction for every $\sigma>0$. Let us prove now that $\mathcal{U}$ is a subsolution of (UVE) at a point $(T,\bar x,\bar k)$ such that $\bar k<K$ and $\mathcal{U}(T,\bar x,\bar k)>g(\bar x)$. Let φ be a smooth map such that

$$\mathcal{U}(t,x,k)-\mathcal{U}(T,\bar x,\bar k)\le\varphi(t,x,k)-\varphi(T,\bar x,\bar k)$$

for every (t,x,k) in a neighborhood of $(T,\bar x,\bar k)$. Let Q be defined as above. The thesis amounts to proving that

$$\min_{a\in A}\max_{(w_0,w)\in S_+^m}Q(T,\bar x,\bar t,a,w_0,w)\ge0.$$

Assume, by contradiction, that there is a $\theta>0$ such that

$$\min_{a\in A}\max_{(w_0,w)\in S_+^m}Q(T,\bar x,\bar t,a,w_0,w)\le-\theta.$$

Then there exists an $\bar{a} \in A$ such that

$$\max_{(w_0,w)\in S^m_+} Q(T, \bar{x}, \bar{t}, \bar{a}, w_0, w) \leq -\theta.$$

It follows that, setting $v(s) = \bar{a}\ \forall s \in [0,1]$, for a $\bar{\sigma}$ sufficiently small and $\bar{t}$ sufficiently close to T one has

$$Q(t(s), z(s), k(s), v_* \circ t(s), \xi_0(a)(s), \xi(a)(s)) \leq -\frac{\theta}{2}$$

for every $s \in [0, \bar{\sigma}]$ and every strategy $(\xi_0, \xi) \in \widetilde{\Gamma}(\bar{t}, \bar{k})$. Integrating both terms, one obtains

$$\int_0^\sigma \bar{l}(t(s), z(s), v_* \circ t(s), \xi_0(v)(s), \xi(v)(s))\, ds + \varphi\big(t(\sigma), z(\sigma), k(\sigma)\big) - \varphi(\bar{t}, \bar{x}, \bar{k}) \leq -\frac{\theta\sigma}{2}$$

for every $\sigma \leq \bar{\sigma}$ and every $(\xi_0, \xi) \in \widetilde{\Gamma}_\sigma(\bar{t}, \bar{k})$. An argument analogous to the one exploited for $\mathcal{V}$ now leads to a contradiction with the RDPP, according to which there exists a $\hat{\sigma}$ such that for every sequence $\{\bar{t}_n\}$ converging to T and every $\sigma < \hat{\sigma}$ one can find strategies $({\xi_0}^\sigma{}_n, {\xi}^\sigma{}_n) \in \widetilde{\Gamma}(\bar{t}_n, \bar{k})$ verifying

$$\int_0^\sigma \bar{l}\big(t_n^\sigma(s), z_n^\sigma(s), v_* \circ t_n^\sigma(s), {\xi_0}^\sigma{}_n(v)(s), {\xi}^\sigma{}_n(v)(s)\big) + \mathcal{U}\big(t_n^\sigma(\sigma), z_n^\sigma(\sigma) k_n^\sigma(\sigma)\big) - \mathcal{U}\big(\bar{t}_n, \bar{x}, \bar{k}\big) > -\frac{\theta\sigma}{4},$$

with obvious meanings of t_n^σ, z_n^σ, and k_n^σ. This concludes the proof of condition (3) of Definition 6.1.

The uniqueness property stated in the thesis is a straighforward consequence of the comparison criterion below, which, in turn, is deduced—via suitable changes of variable—from Theorem 1.1 of [3]. □

Remark 6.1. In the proof of the boundary conditions at $\bar{t} = T$, an essential role is played by the fact that at the points of $\{T\} \times Q \times [0, K]$, where $\mathcal{V} - g \geq \eta$ (resp., $\mathcal{U} - g \geq \eta$), the DPP for $\mathcal{V}$ (resp., $\mathcal{U}$) (and its reparameterized version) holds for every $\sigma \in]0, \hat{\sigma}[$, with $\hat{\sigma}$ depending just on η (and Q). This fact reflects the essentially *impulsive* nature of the problem.

PROPOSITION 6.1 (comparison). *Let $\mathcal{G}_i : \overline{\Omega} \to \mathbb{R}, i = 1, 2$, be continuous and bounded and assume that $\mathcal{G}_1$ is a viscosity subsolution of* (LVE) *and $\mathcal{G}_2$ is a viscosity supersolution of* (LVE) *in Ω. Moreover, let us assume that in $\partial\Omega$ either $\mathcal{G}_1 \leq \mathcal{G}_2$ or $\mathcal{G}_1$ is a subsolution of* (LVE). *Then $\mathcal{G}_1 \leq \mathcal{G}_2$ in $\overline{\Omega}$. The same statement holds true if we replace equation* (LVE) *with* (UVE).

Proof. Let G be a lower bound for both $\mathcal{G}_1$ and $\mathcal{G}_2$ and, for some $q > 0$ to be fixed later, consider the maps $\mathcal{F}_i$ defined by

$$\mathcal{F}_i(t, x, k) \doteq \frac{1}{q(1 + t + k)} \log[\mathcal{G}_i(T - t, x, K - k) + G + 1], \qquad i = 1, 2.$$

Then, setting

$$\tilde{H}^-(t,x,k,r,p_t,p_x,p_k)$$
$$\doteq \min_{(w_0,w)\in S^m_+}\max_{a\in A}\frac{1+t+k}{w_0+|w|}\left\{qp_tw_0 - qp_x\cdot\bar{f}(T-t,x,a,w_0,w) + qp_k|w| \right.$$
$$\left. -\frac{\bar{l}(T-t,x,a,w_0,w)}{(1+t+k)\exp(qr)} + (q-1)r(w_0+|w|)\right\},$$

one can easily check that the maps $\mathcal{F}_1$ and $\mathcal{F}_2$ are a viscosity subsolution and supersolution of

$$\mathcal{F} + \tilde{H}^-(t,x,k,\mathcal{F},\nabla\mathcal{F}) = 0 \tag{LVE$'$}$$

in $]0,T]\times\mathbb{R}^n\times]0,K]$, respectively. Moreover, on $\partial\Omega$, either $\mathcal{F}_1\le\mathcal{F}_2$ or $\mathcal{F}_1$ is a subsolution of (LVE$'$).

Let us observe that, in view of the hypotheses assumed on l, the map $\bar{l}$ is bounded. This implies that for a q sufficiently large the function

$$r\mapsto\tilde{H}^-(t,x,k,r,p_t,p_x,p_k)$$

is nondecreasing for every $(t,x,k,p_t,p_x,p_k)\in[0,T]\times\overline{\Theta}\times[0,K]\times\mathbb{R}^{1+n+1}$.

Hence it is straightforward to check that the Hamiltonian $\tilde{H}^-$ and the domain $\overline{\Theta}$ verify the hypotheses of Theorem 1.1 in [3], from which it follows that $\mathcal{F}_1\le\mathcal{F}_2$ on $\overline{\Theta}$.

By inverting the considered change of (dependent and independent) coordinates, one obtains the assertion concerning (LVE).

The results concerning (UVE) can be straightforwardly proved by just replacing $\tilde{H}^-$ with the Hamiltonian

$$\tilde{H}^+(t,x,k,r,p_t,p_x,p_k)$$
$$\doteq \max_{a\in A}\min_{(w_0,w)\in S^m_+}\frac{1+t+k}{w_0+|w|}\left\{qp_tw_0 - qp_x\cdot\bar{f}(T-t,x,a,w_0,w) + qp_k|w| \right.$$
$$\left. -\frac{\bar{l}(T-t,x,a,w_0,w)}{(1+t+k)\exp(qr)} + (q-1)r(w_0+|w|)\right\}. \qquad \square$$

As a corollary of Theorem 6.1 we obtain an *Isaacs condition*, i.e., a sufficient condition for the game to have a value.

COROLLARY 6.1 (Isaacs condition). *If*

$$H^+ = H^-, \tag{IC}$$

then the game has a value; i.e., $\mathcal{V}=\mathcal{U}$.

Remark 6.2. The case where both the dynamics f and the Lagrangian l are affine in the unbounded control, which includes the problem investigated in [5] (see section 7), i.e.

$$f = f_0(t,x,v) + \sum_{i=1}^m f_i(t,x,v)c_i,$$
$$l = l_0(t,x,v) + \sum_{i=1}^m l_i(t,x,v)c_i,$$

is particularly interesting: indeed, as soon as $f_0,\dots,f_m,l_0,\dots,l_m$ are convex in v the min-max theorem (see e.g., [2]) implies that (IC) is satisfied, so the game has a value.

7. On a former result concerning a particular case. Barron, Jensen, and Menaldi [5] have studied the particular case where c is scalar-valued and positive, i.e. $c \in C \doteq \mathbb{R}^+$, and

$$f(t,x,v,c) = f_0(t,x,v) + f_1(t,v)c, \qquad l(t,x,v,c) = l_0(t,x,v) + l_1(t,v)c. \tag{7.1}$$

The fact that f_1 is independent of the state x is crucial. Indeed, this allows the authors to give a notion of (possibly discontinuous) trajectory even when c is a measure.

Remark 7.1. Incidentally, let us remark that such a trajectory depends strongly on the values of v at the atoms of the measure c. However, this does not trouble the dynamic programming approach pursued in [5], for each of these trajectories can be interpreted as a *space-time solution*, according to [20]. And this allows one to interpret these trajectories as elements of the closure—in a suitable topology—of the set of trajectories corresponding to ordinary controls.

Let us begin by observing that the approach presented in [5] cannot be extended to some cases of interest, which, instead, are included in our investigation. Indeed, the arguments of [5] do not apply as soon as f_1 depends on x. The main reason for that relies on the fact that, unless very restrictive commutative conditions are in force [9], no robust notion of t-parameterized solution can be given (see, e.g., [8], [14], [18], [19], [25]).

Second, the approach proposed in the present paper, once specialized to the case under consideration, yields a boundary value problem much more regular than the one established in [5].

In order to be more precise, let us record the main result in [5] and afterwards let us specialize Theorem 6.1 to the present case.

THEOREM 7.1 (see [5]). *Let us assume that f and l are as in* (7.1)*, and let $K = 1$, $x \in \mathbb{R}$.*

Define the upper Hamiltonian $\hat{H}^+$ as

$$\hat{H}^+(t,x,p_x,p_k) \doteq \min_{a \in A(t,p_x,p_k)} \{p_x \cdot f_0(t,x,a) + l_0(t,x,a)\},$$

where

$$A(t,p_k,p_x) \doteq \{a \in A : \quad p_x \cdot f_1(t,a) + l_1(t,a) + p_k \le 0\}.$$

If $A(t,p_x,p_k) = \emptyset$ then one sets $\hat{H}^+ = +\infty$.

Moreover, define the lower Hamiltonian $\hat{H}^-$ as

$$\hat{H}^-(t,x,p_x,p_k) \doteq \min\{p_x \cdot \beta_0 + \delta_0 : (\beta_0,\delta_0,\beta_1,\delta_1) \in \mathcal{R}(t,x,p_x,p_k)\},$$

where

$$\mathcal{F}(t,x) \doteq \overline{co}\left\{\left(f_0(t,x,a), l_0(t,x,a), f_1(t,a), l_1(t,a)\right) : a \in A\right\},$$
$$\mathcal{R}(t,x,p_x,p_k) \doteq \{(\beta_0,\delta_0,\beta_1,\delta_1) \in \mathcal{F}(t,x) : p_x\beta_1 + \delta_1 + p_k \le 0\}.$$

The upper value $\mathcal{U}$ (resp., the lower value $\mathcal{V}$) is the unique continuous (viscosity) solution in $]0,T[\times\mathbb{R}\times]0,1[$ of

$$-\nabla_t u - \hat{H}^+(t,x,\nabla_x u, \nabla_k u) = 0 \tag{ÛV̂Ê}$$

(resp.,

$$(\hat{\mathrm{L}}\hat{\mathrm{V}}\hat{\mathrm{E}}) \qquad -\nabla_t u - \hat{H}^-(t,x,\nabla_x u,\nabla_k u) = 0)$$

in $]0,T[\times\mathbb{R}\times]0,1[$ *that satisfies the boundary conditions*

$$\text{(TC)} \quad \begin{gathered} u(T,x,k) = \min_{a\in A}\max_{k\le\chi\le 1}\big\{g\big(x+f_1(T,a)(\chi-k)\big)+l_1(T,a)(\chi-k)\big\} \\ \Big(\textit{resp.},\quad u(T,x,k) = \max_{k\le\chi\le 1}\min_{a\in A}\big\{g\big(x+f_1(T,a)(\chi-k)\big)+l_1(T,a)(\chi-k)\big\}\Big) \end{gathered}$$

and

$$\text{(BC)} \qquad u(t,x,1) = r(t,x),$$

where $r(t,x)$ *is the value function of the minimum problem*

$$\underset{a(\cdot)\in\mathcal{A}(t)}{\textit{minimize}} \int_t^T l_0(\tau,x(\tau),a(\tau))\,d\tau, \qquad \frac{dx}{d\tau} = f_0(\tau,x(\tau),a(\tau)), \quad x(t)=x.$$

Remark 7.2. In the previous theorem an extension (due to Ishii [30]) of the original notion of viscosity solution is adopted, in that the Hamiltonians $\hat{H}^+$ and $\hat{H}^-$ are discontinuous.

On the other hand, by specializing Theorem 6.1 to the problem considered in [5] we obtain the following theorem.

THEOREM 7.2. *The upper value* $\mathcal{U}$ *(resp., the lower value* $\mathcal{V}$*) is the unique viscosity solution in* $\Omega \doteq [0,T[\times\mathbb{R}\times[0,1[$ *of the equation*

$$\text{(UVE)} \qquad -H^+(t,x,\nabla_t,\nabla_x u,\nabla_k u) = 0$$

(resp.,

$$\text{(LVE)} \qquad -H^-(t,x,\nabla_t,\nabla_x u,\nabla_k u) = 0),$$

which satisfies the following boundary conditions: (1) *for every* $(x,k)\in\mathbb{R}\times[0,K]$, *one has*

$$\mathcal{U}(T,x,k) \ge g(x) \quad (\textit{resp.},\ \mathcal{V}(T,x,k)\ge g(x)].$$

(2) $\mathcal{U}$ *(resp.,* $\mathcal{V}$*) is a viscosity subsolution of* (UVE) (*resp., of* (LVE)) *on* $\partial'\Omega$ *and at the points of* $\partial_T\Omega$ *where* $\mathcal{U}-g>0$ (*resp.,* $\mathcal{V}-g>0$).

Four main facts can be pointed out from comparing the two results.

1. Unlike the Hamiltonians $\hat{H}^+$ and $\hat{H}^-$ exploited in [5], the Hamiltonians H^+ and H^- here considered are continuous.
2. The domains of minimization involved in the definitions of $\hat{H}^+$ and $\hat{H}^-$ *depend on the gradient of the solution* (this happens also in the minimax control problem [4], which in fact can be considered (see [5]) as a particular case of this problem); on the contrary, the domain involved in the definitions of H^+ and H^- is a *constant, compact* set.
3. The boundary conditions of (BC) need the resolution of an auxiliary boundary value problem and, moreover, condition (TC) for the lower value does not coincide with condition (TC) for the upper value; on the contrary, the boundary conditions (1) and (2) established here, besides being common to $\mathcal{V}$ and $\mathcal{U}$, involve just a subsolution relation and a Dirichlet inequality.
4. The Isaacs condition established in [5] involves the final cost function g, whereas the Isaacs condition found here (Corollary 6.1) has a standard form; i.e., it reduces to the equality of the lower Hamiltonian and the upper one.

REFERENCES

[1] J. P. Aubin and A. Cellina, *Differential Inclusion*, Springer-Verlag, Berlin, 1984.

[2] J. P. Aubin and I. Ekeland, *Applied Nonlinear Analysis*, John Wiley, New York, 1984.

[3] M. Bardi and P. Soravia, *A comparison result for Hamilton–Jacobi equations and applications to some differential games lacking controllability*, Funkcial. Ekvac., 37 (1994), pp. 19–43.

[4] E. N. Barron and H. Ishi, *The Bellman equation for minimizing the maximum cost*, Nonlinear Anal., 13 (1989), pp. 1067–1090.

[5] E. N. Barron, R. Jensen, and J. L. Menaldi, *Optimal control and differential games with measures*, Nonlinear Anal., 21 (1993), pp. 241–268.

[6] T. Başar and P. Bernhard, $\mathcal{H}_\infty$ *Optimal Control and Related Minimax Design Problems*, Birkhäuser, Boston, 1990.

[7] A. Bensoussans and J. L. Lions, *Impulse Control and Quasi-Variational Inequalities*, Bordes, Paris, 1983.

[8] A. Bressan and F. Rampazzo, *On differential systems with vector-valued impulsive controls*, Boll. Un. Mat. Ital. (7), 2 (1988), pp. 641–656.

[9] A. Bressan and F. Rampazzo, *Impulsive control systems with commutative vector fields*, J. Optim. Theory Appl., 71 (1991), pp. 67–83.

[10] C. W. Clark, F. H. Clarke, and G. R. Munro, *The optimal exploitation of renewable resource stocks*, Econometrica, 48 (1979), pp. 25–47.

[11] M. G. Crandall, L. C. Evans, and P. L. Lions, *Some properties of viscosity solutions of Hamilton–Jacobi equations*, Trans. Amer. Math. Soc., 282 (1984), pp. 487–502.

[12] J. R. Dorroh and G. Ferreyra, *A multi-state, multi-control problem with unbounded controls*, SIAM J. Control Optim., 34 (1994), pp. 25–47.

[13] L. C. Evans and P. E. Souganidis, *Differential games and representation formulas of Hamilton–Jacobi equations*, Indiana Univ. Math. J., 33 (1984), pp. 773–797.

[14] O. Hájek, *Book review of S. G. Pandit, S. G. Deo*, Differential Systems Involving Impulses, Lecture Notes in Math. 954, Springer-Verlag, Berlin-New York, 1982, Bull. Amer. Math. Soc., 12 (1985), pp. 272–279.

[15] D. F. Lawden, *Optimal Trajectories for Space Navigation*, Butterworth, London, 1963.

[16] P. L. Lions, *Generalized Solutions of Hamilton–Jacobi equations*, Pitman, London, 1982.

[17] J. P. Marec, *Optimal Space Trajectories*, Elsevier, Amsterdam, Oxford, 1979.

[18] B. M. Miller, *The generalized solutions of ordinary differential equations in the impulse control problems*, J. Math. Systems Estim. Control, 4 (1994), pp. 385–388.

[19] M. Motta and F. Rampazzo, *Space-time trajectories of nonlinear systems driven by ordinary and impulsive controls*, Differential Integral Equations, 8 (1995), pp. 269–288.

[20] M. Motta and F. Rampazzo, *Nonlinear systems with unbounded controls and state constraints: A problem of proper extension*, Nonlinear Differential Equations Appl., 3 (1996), pp. 191–216.

[21] M. Motta and F. Rampazzo, *Dynamic programming for nonlinear systems driven by ordinary and impulsive controls*, SIAM J. Control Optim., 34 (1996), pp. 199–225.

[22] M. Motta and F. Rampazzo, *The value function of a slow growth control problem with state constraints* (*short version*), J. Math. Systems Estim. Control, 7 (1997), pp. 375–378.

[23] L. W. Neustadt, *A general theory of minimum-fuel space trajectories*, SIAM J. Control, 3 (1965), pp. 317–356.

[24] F. Rampazzo, *Continuity of the Upper and Lower Value of Slow Growth Differential Games*, J. Math. Anal. Appl., 213 (1997), pp. 15–31.

[25] A. Sesekin, *Nonlinear differential equations in the class of functions of bounded variation*, Automat. Remote Control, 1990, pp. 1356–1361.

[26] S. P. Sethi, *Dynamic optimal control problems in advertising: A survey*, SIAM Rev., 19 (1977), pp. 685–725.

[27] P. Soravia, $\mathcal{H}_\infty$ *control of nonlinear systems: Differential games and viscosity solutions*, SIAM J. Control Optim., 34 (1996), pp. 1071–1097.

[28] A. J. Van der Shaft, L_2 *gain analysis for nonlinear systems and nonlinear* $\mathcal{H}_\infty$ *control*, IEEE Trans. Automat. Control, 37 (1992), pp. 770–784.

[29] J. Yong, *Zero-sum differential games involving impulse controls*, Appl. Math. Optim., 29 (1994), pp. 243–261.

[30] H. Ishii, *Perron's method for Hamilton-Jacobi equations*, Duke Math. J., 55 (1987), pp. 369–384.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 3, pp. 840–851, May 1998

003

ASYNCHRONOUS STOCHASTIC APPROXIMATIONS*

VIVEK S. BORKAR†

Abstract. The asymptotic behavior of a distributed, asynchronous stochastic approximation scheme is analyzed in terms of a limiting nonautonomous differential equation. The relation between the latter and the relative values of suitably rescaled relative frequencies of updates of different components is underscored.

Key words. distributed algorithms, asynchronous algorithms, communication delays, stochastic approximation, ODE limit

AMS subject classifications. 62L20, 93E25

PII. S0363012995282784

1. Introduction.

There has been a resurgence of interest in stochastic approximation algorithms, particularly as mechanisms for learning systems. They can, for example, be a learning algorithm for neural networks [13] or a model of learning by boundedly rational agents in a macroeconomic system [20], in addition to their traditional applications in adaptive engineering systems [2]. These applications call for a distributed, asynchronous implementation of stochastic approximation schemes. In engineering applications, this is a natural consequence of dealing with large interconnected systems. In macroeconomics, it is simply the reality of life. It is not, however, apparent that the traditional analysis of these schemes, extensively dealt with in [2], automatically holds ground in the new scenario. Prompted by these and similar concerns, there have been studies of distributed implementations of these algorithms [17, 18, 21, 22]. (See [3] for an extensive account of parallel distributed algorithms in general). The present work is in the same spirit, but with some crucial differences.

1. Our model of asynchronism postulates a set-valued random process that marks the indices to be updated at each iteration. This clumping of indices into sets can be an artifice as long as causal relationships are not violated; thus the set-up is very general indeed. We impose on this process a natural condition that requires all indices to be updated comparably often in a precise sense.

2. In addition, we allow random, possibly nonstationary and unbounded delays that are required to satisfy a mild conditional moment condition.

3. The analysis depends on proving that the algorithm asymptotically tracks a nonautonomous ODE, in contrast to the traditional autonomous "ODE limit." In particular, it gives a handle on situations when the latter may not be feasible.

4. The ODE in question differs from the traditional one in that each component of the driving vector field is now weighted by a time-varying nonnegative scalar. These scalars add to 1 and may be interpreted as relative frequencies of updates of different components after suitable time-scaling. This clearly brings out the desired relationships between update schemes for different components and paves the way for analyzing situations where they are not desirable (see remark 4 of the conclusion).

*Received by the editors March 10, 1995; accepted for publication (in revised form) February 19, 1997. This research was supported by the Homi Bhabha Fellowship.
http://www.siam.org/journals/sicon/36-3/28278.html
†Department of Computer Science and Automation, Indian Institute of Science, Bangalore 560012, India (borkar@csa.iisc.ernet.in).

The paper is organized as follows. The remainder of this section describes the problem framework. The next section states the key assumptions and their immediate consequences. The third section provides the convergence analysis. The final section highlights some further possibilities.

Let TS (for tapering stepsize) denote the set of sequences $\{a(n)\}$ in (0,1) satisfying

$$\sum_n a(n) = \infty, \qquad \sum_n a(n)^2 < \infty. \tag{1.1}$$

The standard stochastic approximation algorithm is the recursive scheme in R^d, $d \geq 1$, described by

$$X(n+1) = X(n) + a(n)F(X(n), \xi(n)), \tag{1.2}$$

where $\{a(n)\} \in \text{TS}$, $X(n) = [X_1(n), \ldots, X_d(n)]^T \in R^d$ with a prescribed $X(0)$, $F(\cdot,\cdot) : R^d \times R^k \to R^d$, and $\{\xi(n)\}$ is a stationary random process in R^k. For simplicity, we take $\{\xi(n)\}$ to be independently and identically distributed (i.i.d.) with a common law ψ (say). The ith row of this vector iteration reads

$$X_i(n+1) = X_i(n) + a(n)F_i(X_1(n), \ldots, X_d(n), \xi(n)). \tag{1.3}$$

A distributed but synchronous version of (1.2) could be as follows. Let $I = \{1, 2, \ldots, d\}$ and S be a collection of nonempty subsets of I that cover I. Let $\{Y_n\}$ be an S-valued random process that selects the coordinates to be updated at time n, and for each n, let $\tau_{ij}(n)$, $i \neq j \in I$, be random variables taking values in $\{0, 1, \ldots, n\}$ that represent communication delays. We set $\tau_{ii}(n) = 0\ \forall i, n$. The synchronous distributed version of (1.3) is then

$$X_i(n+1) = X_i(n) + a(n)F_i(X_1(n - \tau_{i1}(n)), \ldots, X_d(n - \tau_{id}(n)), \xi(n))I\{i \in Y_n\} \tag{1.4}$$

for $i \in I$, $n \geq 0$. Special instances of (1.4) were studied in [5, 6]. The reason this is a synchronous version is that the decision to use stepsize $a(n)$ at time n by the processor updating the ith component (say) presupposes the availability of a global clock to all processors. This is not reasonable in an asynchronous environment. The asynchronous version we propose is as follows. Let $\{a(n,i)\} \in \text{TS}$, $i \in I$, and define

$$\nu(n,i) = \sum_{m=0}^{n} I\{i \in Y_m\}, \quad i \in I, \quad n \geq 0,$$
$$\overline{a}(n,i) = a(\nu(n,i), i), \quad i \in I, \quad n \geq 0.$$

The first of these is the total number of times the ith component was updated up until time n. Assuming that each component of the iteration is assigned to one and only one processor once and for all, $\overline{a}(n,i)$ is a random variable known to the ith processor at time n. The proposed algorithm is

$$X_i(n+1) = X_i(n) + \overline{a}(n,i)F_i(X_1(n - \tau_{i1}(n)), \ldots, X_d(n - \tau_{id}(n)), \xi(n))I\{i \in Y_n\} \tag{1.5}$$

for $i \in I$, $n \geq 0$. This is the algorithm analyzed in this paper, under the assumptions stipulated in the next section. We conclude this section with the remark that even the implicit presence of an unobserved global clock in the background in (1.5) is not really needed. The clumping of updated coordinates into Y_m's could be a complete artifice as long as causal relationships are not violated and the additional assumptions of the next section (notably (A3)) remain valid.

2. Preliminaries. The additional assumptions and their consequences that we present in this section concern, respectively, the stepsize routines $\{a(n,i)\}$, the sampling process $\{Y_n\}$, the communication delays $\{\tau_{ij}(n)\}$, and the function F. We proceed in that order. These assumptions, (A1)–(A5), are enforced throughout the paper without further mention.

Let ITS (for "ideal tapering stepsize") denote the subset of TS consisting of $\{a(n)\}$ satisfying:

(i) $a(n+1) \leq a(n)$ from some n onwards;

(ii) there exists $r \in (0,1)$ such that

$$\sum_n a(n)^{1+q} < \infty, \qquad q \geq r; \tag{2.1}$$

(iii) for $x \in (0,1)$,

$$\sup_n a([xn])/a(n) < \infty, \tag{2.2}$$

where $[\cdots]$ stands for the integer part of "$\cdots$";

(iv) for $x \in (0,1)$ and $A(n) \triangleq \sum_{i=0}^{n} a(i)$,

$$A([yn])/A(n) \to 1 \tag{2.3}$$

uniformly in $y \in [x,1]$.

By (i), (2.2) may be strengthened to

$$\sup_n \sup_{y \in [x,1]} a([yn])/a(n) < \infty. \tag{2.4}$$

It is easy to construct examples of $\{a(n)\}$ in TS which violate (2.2). Condition (iv) can be given an alternative formulation. Let $h : R^+ \to R^+$ be an eventually nonincreasing function satisfying $h(n) = a(n)$, $n \geq 0$. Then (2.3) is equivalent to

$$\lim_{t \to \infty} \frac{\int_0^{yt} h(s)ds}{\int_0^t h(s)ds} = 1, \tag{2.5}$$

which, by l'Hôpital's rule, reduces to

$$\lim_{t \to \infty} yh(yt)/h(t) = 1.$$

One needs this to hold uniformly in $y \in [x,1]$. One sufficient condition for this would be that the derivative of the left-hand side of (2.5) in y, which is $th(yt)/\int_0^t h(s)ds$, be bounded uniformly in y, t, ensuring the equicontinuity in y for the ratio in (2.5). It is not clear whether (iv) is implied by (i)–(iii). Examples of $\{a(n)\}$ satisfying (i)–(iv) are $\{1/n\}$, $\{1/n\log n\}$, and $\{\log n/n\}$, with suitable modification for $n = 0, 1$ where needed.

One property of $\{a(n)\} \in$ TS that we shall need later is the following.

LEMMA 2.1. *For $s \in (0,1)$, $a(n)^{-s}/n \to 0$.*

Proof. It suffices to prove that $(a(n)n^x)^{-1} \to 0$ for $x = 1/s > 1$, or equivalently, that $(a(n) + n^{-x})/a(n) \to 1$. Let $h_1, h_2 : R^+ \to R^+$ be continuous functions linearly interpolated from $h_1(n) = a(n) + n^{-x}$, $h_2(n) = a(n)$, $n \geq 0$. Since $\int_0^t h_1(y)dy \to \infty$ as $t \to \infty$ and $\int_1^\infty t^{-x}dt < \infty$, we have

$$\lim_{t \to \infty} \frac{\int_0^t h_2(y)dy}{\int_0^t h_1(y)dy} = 1.$$

The claim now follows from l'Hôpital's rule. ∎

Our first assumption then is:

(A1) $\{a(n,i)\} \in$ ITS for $i \in I$.

Next, introduce for $n \geq 0$ the σ-fields $\mathcal{F}_n = \sigma(X(m), Y(m), m \leq n, \tau_{ij}(m), \xi(m), m < n, i, j \in I)$ and $\mathcal{G}_n = \sigma(X(m), Y(m), \tau_{ij}(m), \xi(m), m \leq n, i, j \in I)$. Our assumption concerning $\{Y_n\}$ is as follows.

(A2) There exists a $\delta > 0$ such that for any A, $B \in S$, the quantity

$$P(Y_{n+1} = B / Y_n = A, \mathcal{G}_n) \tag{2.6}$$

is either always zero almost surely (a.s.) or always exceeds δ a.s. That is, having picked A at time n, picking B at time $n+1$ is either improbable or probable with a conditional probability of at least δ, regardless of n and the "history" $\mathcal{G}_n$. Furthermore, if we draw a directed graph with node set S and an edge from A to B whenever (2.6) exceeds δ a.s., the graph is irreducible; i.e., there is a directed path from any $A \in S$ to any $B \in S$. (As will become apparent later, this may be replaced by the weaker requirement that every communicating class of the directed graph comprises sets that together cover I.)

This has the following important consequence. Let $\mathcal{P}(\cdots)$ denote the space of probability vectors on "$\cdots$."

LEMMA 2.2. *There exists a deterministic constant $\Delta > 0$ such that for any $A \in S$,*

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{m=0}^{n-1} I\{Y_m = A\} \geq \Delta \quad a.s. \tag{2.7}$$

Proof. For $A \in S$, let $D_A = \{B \in S |$ (2.6) exceeds δ a.s.$\}$ and $V_A = \{u \in \mathcal{P}(D_A) | u(B) \geq \delta \ \forall B \in D_A\}$, $V = \prod_A V_A$. Define $p : S \times S \times V \to [0,1]$ by $p(A,B,u) = u_A(B)$, where u_A is the Ath component of u. Define V-valued random variables $\{Z^n\}$ by

$$Z_A^n(B) = P(Y_{n+1} = B / \mathcal{G}_n) I\{Y_n = A\} + \psi_A I\{Y_n \neq A\},$$

where ψ_A is a fixed element of V_A for $A \in S$. Then (2.6) equals $p(A, B, Z^n)$ and $\{Y_n\}$ may be viewed as an S-valued controlled Markov chain with action space V and transition probability function p. (This is a pure artifice for the sake of the proof. It is in no way implied that $\{Z^n\}$ is an actual control process.) In particular, this allows us to conceive of a stationary policy π associated with a map π: $S \to V$. (A1) implies, in particular, that $\{Y_n\}$ will be an ergodic Markov chain under a stationary policy π with a corresponding stationary distribution $\nu_\pi \in \mathcal{P}(S)$. Then the left-hand side of (2.7) a.s. exceeds $\min_\pi \nu_\pi(A) > 0$ by Lemmas 1.2 and 2.1 of [4, pp. 56, 60]. □

For the communication delays, we assume the following. (Recall the r in (2.1).)

(A3) $\tau_{ij}(n) \in \{0, 1, \ldots, n\}, \tau_{ii}(n) = 0 \ \forall i, a$. There exist $b > r/(1-r), C > 0$ such that

$$E[(\tau_{ij}(n))^b / \mathcal{F}_n] \leq C \quad \text{a.s.} \quad \forall i, j, n. \tag{2.8}$$

(In particular, we do not require the delays to be either bounded or stationary.) Also, $\{\xi(n)\}$ is i.i.d. and independent of $\{X_0, \xi(m), \tau_{ij}(m), m < n\}$ for all n.

Next come the conditions on F.

(A4) F is assumed to be measurable and uniformly Lipschitz in the first argument; i.e., for some $K > 0$,

$$\|F(x,z) - F(y,z)\| \leq K\|x - y\| \quad \forall\ x, y, z.$$

Other conditions on F will be given in terms of the function $f : R^d \to R^d$ defined by

$$f(x) = \int F(x,y)\psi(dy). \tag{2.9}$$

Under our conditions on F, f is Lipschitz with Lipschitz constant K. The traditional analysis of (1.2) [2] proceeds by showing that it asymptotically tracks the ODE

$$\dot{x}(t) = f(x(t)), \tag{2.10}$$

which in turn has trajectories converging to $J = \{x | f(x) = 0\}$.

(A5) J is assumed to be compact and nonempty.

We shall also have reason to consider a related nonautonomous ODE. Let $\mathcal{D}$ denote the set of diagonal $d \times d$ matrices with nonnegative diagonal entries that add to 1. For $a > 0$, say that $M = \text{diag}(m_1, \ldots, m_d)$ is a-thick if $m_i \geq a \ \forall i$. The ODE in question is

$$\dot{x}(t) = M(t)f(x(t)), \tag{2.11}$$

where $t \to M(t)$ is a $\mathcal{D}$-valued measurable process.

We consider two scenarios.

Case 1: *Strict Liapunov systems.* A continuously differentiable function $V : R^d \to R^+$ is said to be a strict Liapunov function for (2.10) if $\nabla V.\, f < 0$ outside J. Call (2.10) a strict Liapunov system if it has bounded trajectories and a strict Liapunov function V exists. The latter implies the former if $V(x) \to \infty$ as $\|x\| \to \infty$, which we assume to hold. (Call this assumption (A6).) Examples of such systems can be found among gradient systems and their variants, certain systems arising in neural networks [14], and analog fixed point algorithms wherein $f(x) = g(x) - x$ and g is either a contraction under a $\|.\|_p$-norm for $p \in [1, \infty]$ or nonexpansive under a $\|.\|_p$- norm for $p \in (1, \infty)$. (Here $V(.) = \|. - x^*\|_p$, where $x^* \in J$, will do. For $p = \infty$, this is not continuously differentiable, but this does not pose any problems for contractions [7].)

Finally, a strict Liapunov system as above will be said to be a-robust for some $a > 0$ if $\nabla V.\, Mf < 0$ outside J for any a-thick $M \in \mathcal{D}$.

Given $T, \delta > 0$, a (T, δ)-perturbation of (2.10) (resp., (2.11)) is a function $y : R^+ \to R^d$ such that there exist $0 = T_0 < T_1 < \cdots < T_n \uparrow \infty$ and solutions $x^j(t)$, $t \in [T_j, T_{j+1}]$, $j \geq 0$, of (2.10) (resp., (2.11)) such that $T_{j+1} - T_j \geq T$ for $j \geq 0$ and

$$\|y(t) - x^j(t)\| < \delta, \quad T_j \leq t \leq T_{j+1}, j \geq 0.$$

For $\epsilon > 0$, let $J^\epsilon = \{x \in R^d | \, \|x - y\| < \epsilon \text{ for some } y \in J\}$.

LEMMA 2.3. *Under* (A6), *we have:* (a) *For any* $T, \epsilon > 0$, *there exists a* $\delta_0 = \delta_0(T, \epsilon) > 0$ *such that for* $\delta \in (0, \delta_0)$, *any* (T, δ)*-perturbation of* (2.10) *converges to* J^ϵ. *(In particular, solutions of* (2.10) *converge to* J.*)*

(b) *Suppose that* (2.10) *is a-robust for some* $a > 0$ *and* $M(t)$ *in* (2.11) *is a-thick for almost every* t. *Then, for any* $T, \epsilon > 0$, *there exists a* $\delta_0 = \delta_0(T, \epsilon, a) > 0$ *such that for* $\delta \in (0, \delta_0)$, *any* (T, δ)*-perturbation of* (2.11) *converges to* J^ϵ. *(In particular, the solutions of* (2.11) *converge to* J.*)*

These are straightforward adaptations of Theorem 1 of [14, p. 339].

Case 2: ∞*-nonexpansive maps.* In this case $f(x) = g(x) - x$, where g is ∞-nonexpansive, i.e., $\|g(x) - g(y)\|_\infty \leq \|x - y\|_\infty, x, y \in R^d$. Thus J is the set of fixed points of g. This case is important in dynamic programming applications [3, 7].

For this case, we have the following analog of Lemma 2.3.

LEMMA 2.4. *The conclusions of Lemma* 2.3(a) *continue to hold. Those of Lemma* 2.3(b) *hold if* $M(t)$ *is* a*-thick for almost every* t*, for some* $a > 0$.

This is proved in Theorem 2.1 and Corollary 2.2 of [5].

3. Convergence. We start by establishing a link between (1.4) and (2.11). Our first observation is that we may equivalently consider the recursion

$$X_i(n+1) = X_i(n)+\overline{a}(n,i)F_i(X_i(n-\tau_{ij}(n)),\dots,X_d(n-\tau_{id}(n)),\tilde{\xi}(n))I\{\varphi_n = i\}, \tag{3.1}$$

where $\{\varphi_n\}$ is an I-valued random process satisfying the following statement. There exists a deterministic constant $\eta > 0$ such that

$$\liminf_{n\to\infty} \frac{1}{n}\sum_{m=0}^{n-1} I\{\varphi_n = i\} \geq \eta \quad \text{a.s.} \quad \forall i \in I \tag{3.2}$$

and $\tilde{\xi}(n) = \xi(k(n))$ for a nondecreasing map $n \to k(n)$, satisfying $k(n+1) - k(n) \in \{0,1\}$.

This is achieved simply by unfolding each iteration as follows.

Let $Y_n = \{i_1,\dots,i_{c(n)}\}$ (say) with the elements arranged in ascending order. Replace the iteration (1.4) by $c(n)$ distinct iterations such that the jth iteration among them updates only the i_jth component in accordance with (1.4). Next, relabel the iteration index and the delays to obtain a correspondence with (3.1). Then (3.2) is an immediate consequence of Lemma 2.2. Note that this blows up the delays at most d fold, thus still retaining (A3). Note also that for $m > n$, $\varphi_m = \varphi_n$ implies $k(m) > k(n)$. With these considerations, we proceed to analyze (3.1). We start with some preliminaries.

Let U be the space of $\mathcal{P}(I)$-valued trajectories $\overline{\mu} = \{\mu_t, t \geq 0\}$ with the coarsest topology that renders continuous the maps $\overline{\mu} \to \int_0^T h(t)\mu_t(i)dt$ for $T \geq 0$, $i \in I$, $h \in L_2[0,T]$. U is compact metrizable. Say that $\mu \in \mathcal{P}(I)$ is α-thick for some $\alpha > 0$ if $\mu(i) \geq \alpha\ \forall i$. Say that $\overline{\mu} \in U$ is α-thick, $\alpha > 0$, if μ_t is so for almost every t. Say that μ (resp., $\overline{\mu}$) is thick if it is α-thick for some $\alpha > 0$.

LEMMA 3.1. (a) *For* $\alpha > 0$, $\{\overline{\mu}|\overline{\mu}$ *is* α*-thick*$\}$ *is compact in* U. (b) *The map* $(\overline{\mu}, x) \to x(\,.\,) : U \times R^d \to C([0,\infty); R^d)$ *defined via*

$$\dot{x}(t) = M^{\overline{\mu}}(t)f(x(t)), \qquad x(0) = x, \tag{3.3}$$

with $M^{\overline{\mu}}(t) = \mathrm{diag}(\mu_t(1),\dots,\mu_t(d))$, *is continuous.*

Proof. (i) For $i \in I, t > s, n \geq 1$, and any α-thick $\overline{\mu}$, $\alpha > 0$,

$$\int_s^t \mu_y(i)dy \geq \alpha(t-s).$$

The relation is preserved under limits in U, implying the claim.

(ii) Let $(\overline{\mu}^n, x_n) \to (\overline{\mu}^\infty, x_\infty)$. For $n \geq 1$, let $x^n(\,.\,)$ satisfy

$$\dot{x}^n(t) = M^{\overline{\mu}^n}(t)f(x^n(t)), \qquad x^n(0) = x_n. \tag{3.4}$$

Using the Gronwall lemma and the Arzela–Ascoli theorem, one verifies that $\{x^n(\,.\,)\}$ is relatively compact in $C([0,\infty); R^d)$, and a straightforward limiting argument (keeping in mind our topology on U) shows that any limit $x^\infty(\,.\,)$ thereof must satisfy (3.4) with $n = \infty$. The claim follows. □

Let $\tilde{a}(n) = \overline{a}(n, \varphi_n)$ and rewrite (3.1) as

$$X(n+1) = X(n) + \tilde{a}(n)W(n)$$

for appropriately defined $W(n) = [W_1(n), \ldots, W_d(n)]^T$. Redefine $\mathcal{F}_n$ by $\mathcal{F}_n = \sigma(X(m), m \leq k^{-1}(n), \xi(m), m < k^{-1}(n), \tau_{ij}(m), m < n, \varphi_m, m \leq n)$, where $k^{-1}(n) = \min\{j | k(j) = n\}$. Set $\hat{W}(n) = E[W(n)/\mathcal{F}_n], n \geq 0$, the conditioning bring componentwise. Write $\hat{W}(n) = [\hat{W}_1(n), \ldots, \hat{W}_d(n)]^T$. Define $f^i : R^d \to R^d$ by $f^i_j(x) = f_i(x)\delta_{ij}$, $i, j \in I$, δ_{ij} being the Kronecker delta. Let $b(n) = \max_i \overline{a}(n, i), n \geq 0$. Let Q denote the set of sample points for which $\hat{X} \triangleq \sup_n \|X(n)\| < \infty$.

LEMMA 3.2. *$\{b(n)\}$ satisfies $\sum_n b(n)^{1+r} < \infty$ a.s., and for $a \in (0, 1]$,*

$$\sup_n \sup_{\alpha \in [a,1]} b([\alpha n])/b(n) < \infty \quad a.s.$$

Proof. By (2.4) and (3.2), $\sup_n \overline{a}(n, i)/a(n, i) < \infty$ a.s., $i \in I$. Combining this with property (2.1) for $\{a(n, i)\}$, we have $\sum \overline{a}(n, i)^{1+r} < \infty$ a.s. The first claim follows. The second follows easily from (2.4) applied to $\{a(n, i)\}$. □

LEMMA 3.3. *Almost surely on Q, there exist $K_1 > 0$, $N \geq 1$ (random) such that for $n \geq N$,*

$$\|f^{\varphi_n}(X(n)) - \hat{W}(n)\| < K_1 b(n)^r.$$

Proof. Consider $\omega \in Q$. Let K_2 be an upper bound on $\{\|f(x)\|_\infty \mid \|x\| \leq \hat{X}\}$. Let $\tilde{W}_i(n) = f_i^{\varphi_n}(X_1(n - \tau_{i1}(n)), \ldots, X_d(n - \tau_{id}(n)))$. Let $c = 1 - r$. For $i \in I$, we have

(3.5)

$$|f_i^{\varphi_n}(X(n)) - \hat{W}_i(n)| \leq E[|f_i^{\varphi_n}(X(n)) - \tilde{W}_i(n)| I\{\tau_{ij}(n) \leq b(n)^{-c} \ \forall i, j\}/\mathcal{F}_n]$$
$$+ E[|f_i^{\varphi_n}(X(n)) - \tilde{W}_i(n)| I\{\tau_{ij}(n) > b(n)^{-c} \text{ for some } i, j\}/\mathcal{F}_n] \quad \text{a.s.}$$

By (A3) and the conditional Chebyshev inequality, the second term is a.s. bounded by $2K_2 C d^2 b(n)^{bc}$. Let $\overline{n} = [b(n)^{-c}]$. By Lemma 2.1, $\overline{n}$ is $o(1)$ a.s. as $n \to \infty$, and outside a zero probability set, we may pick n large enough so that $n > \overline{n}$. Then for $m \leq \overline{n}$,

$$\|X(n) - X(n-m)\| \leq 2K_2 d \sum_{j=n-\overline{n}}^{n} b(j) \leq K_3 b(n)^{1-c}$$

for a suitable (random) $K_3 > 0$, by the above lemma. Thus the first term in (3.5) is bounded by $K_4 b(n)^r$ for a suitable (random) $K_4 > 0$. Since $b > r/(1-r)$, the claim follows. □

Let $T > 0$. Define $t_0 = T_0 = 0, t_n = \sum_{m=0}^{n-1} \tilde{a}(m)$, $n \geq 1$, and $T_n = \min\{t_m | t_m \geq T_{n-1} + T\}$, $n \geq 1$. Then $T_n = t_{m(n)}$ for a strictly increasing sequence $\{m(n)\}$. Let $I_n = [T_n, T_{n+1}]$, $n \geq 0$. Define $\overline{x}^n(t)$, $t \in I_n$, by $\overline{x}^n(T_n) = X(m(n))$ and

$$\overline{x}^n(t_{m(n)+k+1}) = \overline{x}^n(t_{m(n)+k}) + \tilde{a}(m(n)+k) f^{\varphi_{m(n)+k}}(\overline{x}^n(t_{m(n)+k})),$$

with linear interpolation on each interval $[t_{m(n)+k}, t_{m(n)+k+1}]$. Define $x(t)$, $t \geq 0$, by $x(t_n) = X(n)$ with linear interpolation on each interval $[t_n, t_{n+1}]$.

LEMMA 3.4. $\lim_{n\to\infty} \sup_{t\in I_n} \|x(t) - \overline{x}^n(t)\| = 0$ *a.s. on* Q.

Proof. Let $n \geq 1$. For $i \geq m(n)$, we have

$$x(t_{i+1}) = x(t_i) + \tilde{a}(i)f^{\varphi_i}(x(t_i)) + \tilde{a}(i)(\hat{W}(i) - f^{\varphi_i}(x(t_i))) + \tilde{a}(i)(W(i) - \hat{W}(i)).$$

Let $\overline{M}_i = \sum_{j=0}^{i} \tilde{a}(j)(W(j) - \hat{W}(j))$ and $\lambda_i = \overline{M}_i - \overline{M}_{m(n)}, i \geq m(n)$. Also, let $M_i^k = \sum_{j=0}^{k} \tilde{a}(j)(W(j) - \hat{W}(j))I\{\varphi_j = k\}, 1 \leq k \leq d$. Recall that $m > n$ and $\varphi_m = \varphi_n$ implies $k(m) > k(n)$. Then, for each k, $\{M_i^k, \mathcal{F}_i\}$ is a zero mean–bounded increment vector martingale, and the quadratic variation process of each of its component martingales is a.s. convergent on Q. By Proposition VII-3-(c) of [19, pp. 149–150], each $\{M_i^k\}$ and hence $\{\overline{M}_i\}$ converges a.s. on Q. Fix a sample point for which this convergence holds and let $\epsilon > 0$. Then $\sup_{i\geq m(n)} \|\lambda_i\| < \epsilon/2$ for sufficiently large n. Let $\hat{x}_{i+1} = x(t_{i+1}) - \lambda_i, i \geq m(n)$, with $\hat{x}_{m(n)} = X(m(n))$. Then, for $i \geq m(n)$, we have

$$\hat{x}_{i+1} = \hat{x}_i + \tilde{a}(i)f^{\varphi_i}(\hat{x}_i) + \tilde{a}(i)(f^{\varphi_i}(\hat{x}_i + \lambda_{i-1}) - f^{\varphi_i}(\hat{x}_i)) + \tilde{a}(i)(\hat{W}(i) - f^{\varphi_i}(x(t_i))).$$

Also,

$$\overline{x}^n(t_{i+1}) = \overline{x}^n(t_i) + \tilde{a}(i)f^{\varphi_i}(\overline{x}^n(t_i)).$$

Fix $\omega \in Q$, where the foregoing and Lemma 3.3 hold. Subtracting and using Lemma 3.3, we have, for n sufficiently large,

$$\|\hat{x}_{i+1} - \overline{x}^n(t_{i+1})\| \leq (1 + K\tilde{a}(i))\|\hat{x}_i - \overline{x}^n(t_i)\| + \tilde{a}(i)\|\lambda_{i-1}\|K + K_1\tilde{a}(i)b(i)^{1+r}.$$

By increasing n if necessary, we may suppose that

$$\sum_{i\geq n} b(i)^{1+r} < \epsilon/2.$$

Then using the inequality $1 + x \leq \exp(x)$ and iterating, we have

$$\sup_{m(n)\leq i\leq m(n+1)} \|\hat{x}_i - \overline{x}^n(t_i)\| \leq e^{K(T+1)}(K_1 + K(T+1))\epsilon$$

for sufficiently large n. Since $\|\hat{x}_i - x(t_i)\| < \epsilon/2$, $i \geq m(n)$ for sufficiently large n, $\sup_{m(n)\leq i\leq m(n+1)} \|x(t_i) - \overline{x}^n(t_i)\| \leq \tilde{K}\epsilon$ for a suitable $\tilde{K} > 0$. Since $\epsilon > 0$ was arbitrary, the claim follows on noting that both $x(\,.\,)$ and $\overline{x}^n(\,.\,)$ are linearly interpolated from their values at $\{t_i\}$. □

Next, define $\overline{\mu} \in U$ by $\mu_t =$ the Dirac measure at φ_n for $t \in [t_n, t_{n+1}), n \geq 0$. Define $\tilde{x}^n(t)$, $t \in I_n$, by $\tilde{x}^n(t_{m(n)}) = x(t_{m(n)})$ and

$$\dot{\tilde{x}}^n(t) = M^{\overline{\mu}}(t)f(\tilde{x}^n(t)), \qquad t \in I_n. \tag{3.6}$$

LEMMA 3.5. $\lim_{n\to\infty} \sup_{t\in I_n} \|\tilde{x}^n(t) - \overline{x}^n(t)\| = 0$ *a.s.*

Proof. This follows easily from the Gronwall inequality. □

For $\overline{\mu}$ as above, define $\overline{\mu}^t = \{\mu_{t+s}, i \geq 0\} \in U$ for $t \geq 0$. Combining the foregoing with Lemmas 2.3 and 2.4, we have the following theorem.

THEOREM 3.1. (a) *Suppose there exists an* $a > 0$ *such that* (2.10) *is an* a*-robust strict Liapunov system,* (A6) *applies, and all limit points of* $\overline{\mu}^t$ *in* U *as* $t \to \infty$ *are* a*-thick a.s. Then the algorithm converges to* J *a.s. on* Q.

(b) *For the ∞-nonexpansive case (Case 2), suppose all limit points of $\overline{\mu}^t$ in U as $t \to \infty$ are a-thick for some $a > 0$, a.s. Then the algorithm converges to J a.s. on Q.*

Remark. For Case 1 without the a-robustness hypothesis, the above analysis still gives some clue about the convergence of the algorithm: if all limit points of $\overline{\mu}^t$ are a-thick a.s., the algorithm will converge to the smallest closed set outside which $\nabla V.\, Mf < 0$ for a-thick M.

Clearly, one would like $P(Q) = 1 = P\ (\hat{X} < \infty)$. One observes that the boundedness of $\hat{X}$ is used twice: in Lemma 3.3 and to prove almost sure convergence on Q of $\{\overline{M}_n\}$. In either case, it is unnecessary if f (or, in Case 2, g) is bounded. If not, the problem of establishing $P(Q) = 1$ remains. This is so even for the traditional "centralized" algorithm, and it is not unusual to find results that state convergence if the iterates remain bounded or visit a neighborhood of the desired attractor infinitely often, a.s. There is no general scheme for showing $P(Q) = 1$. There are, however, problem-specific techniques for special problem classes. We list a few recent ones below without details, referring the reader to the original works for those.

(i) *Martingale methods.* These usually take the form of establishing the "almost supermartingale" property [19, p. 33] for $\{V(X(n))\}$, where $V : R^d \to R^+$ is a continuously differentiable "stochastic Liapunov function" satisfying $V(x) \to \infty$ as $\|x\| \to \infty$. This leads to the almost sure boundedness of $\{V(X(n))\}$, hence of $\{X(n)\}$. For strict Liapunov systems, the Liapunov function therein will itself suffice in most cases. The adaptation of this approach to the asynchronous case, however, is rendered difficult by the presence of delays. Specific instances of it have been worked out, a good example being the stochastic gradient schemes discussed in [3, section 7.8]. For the "centralized" case without delays, see [2, p. 239].

(ii) *Projection and related schemes.* One way to escape the boundedness issue is to alter the algorithm by projecting the iterates back onto a prescribed, large bounded set whenever they exit from the same. The trade-off is that the limiting ODE becomes more complicated. It is now confined to the said set and thus involves a "reflection at the boundary" of the same in an appropriate sense. The analysis of such schemes for the centralized case is by now standard, and an excellent exposition appears in [16, pp. 191–194]. It seems possible to extend it to the present case. (See [1] for a specific instance.)

In an ingenious boundedness proof for the case when $F(x, y)$ is homogeneous of degree 1 in its first argument (important in certain "learning" algorithms), Jaakola, Jordan, and Singh [15] use the almost sure convergence of the algorithm with rescaling to deduce the almost sure boundedness of the one without. See [1] for some extensions of this idea and application to a specific asynchronous situation.

In a somewhat similar spirit, but using different techniques, Chen [8] discusses stabilization of the (centralized) algorithm by truncating the iterates while slowly increasing the truncation bounds.

(iii) *Tsitsiklis conditions.* For Case 2 (nonexpansive maps) studied above, Tsitsiklis [21] gives a remarkable set of conditions for almost sure boundedness when additional structure is available, such as an appropriate monotonicity property of the map or contraction property under a suitable norm. These are very useful for applications arising from dynamic programming.

In some special cases (e.g., when $F(\cdot\,, y)$ has a common fixed point), one may adapt the conditions of [3, p. 433], for deterministic algorithms to prove almost sure boundedness. See [5] for an instance of this.

In [5], almost sure a-thickness of the limit points of $\{\overline{\mu}^t\}$ for a suitable $a > 0$ is established for the synchronous case. That argument does not follow for the asynchronous case. In fact, it will soon become clear that such a result need not hold in general, and whether it does depends crucially on the relationships between the sequences $\{a(n,i)\} \in \text{ITS}$, $i \in I$. We now consider an important special case where things work out.

Say that the family $\{a(n,i)\}, i \in I$, is balanced if there exist $a_{ij} > 0, i, j \in I$, such that

$$\lim_{n\to\infty} \frac{\sum_{m=0}^{n} a(m,j)}{\sum_{m=0}^{n} a(m,i)} = a_{ij}.$$

Equivalently, if h_i, h_j are continuous, eventually nonincreasing functions $R^+ \to R^d$ that restrict to $\{a(n,i)\}$, $\{a(n,j)\}$, respectively, at integer values of their arguments, then

$$\lim_{t\to\infty} \frac{\int_0^t h_j(s)ds}{\int_0^t h_i(s)ds} = a_{ij}. \tag{3.7}$$

Certain relations between a_{ij}'s are obvious: $a_{ii} = 1$, $a_{ik} = a_{ij}a_{jk}$, $a_{ji} = 1/a_{ij}$. An important special case is $a_{ij} = 1\ \forall i, j$, which would be true, e.g., when all $\{a(n,i)\}, i \in I$, are identical. Let $\beta(i) = a_{1i}/a_{11}$ and $\overline{\beta}(i) = \beta(i)/\sum_j \beta(j)$. Then $\overline{\beta}(i) \in (0,1)\ \forall i$ and $\sum_i \overline{\beta}(i) = 1$. Also $a_{ij} = \overline{\beta}(j)/\overline{\beta}(i)\ \forall i, j$. Set $\overline{a} = \min \overline{\beta}(i)$.

THEOREM 3.2. *If $\{a(n,i)\}, i \in I$, are balanced, the conclusions of Theorem* 3.1(b) *hold. Those of Theorem* 3.1(a) *hold if, in addition, $a \le \overline{a}$.*

Proof. For $i \in I$, let $q(i,n) = \sum_{m=0}^{n} I\{\varphi_m = i\}$. By (3.2),

$$\liminf_{n\to\infty} q(i,n)/n \ge \eta \quad \text{a.s.}, \quad i \in I. \tag{3.8}$$

Fix $i, j \in I$. Then, for $z > 0$,

$$\begin{aligned}
\lim_{t\to\infty} \frac{\int_z^t \mu_s(j)ds}{\int_z^t \mu_s(i)ds} &= \lim_{n\to\infty} \frac{\sum_{m=0}^{q(j,n)} a(m,j)}{\sum_{m=0}^{q(i,n)} a(m,i)} \\
&= \lim_{n\to\infty} \frac{\int_0^{q(j,n)} h_j(s)ds}{\int_0^{q(i,n)} h_i(s)ds} \\
&= \lim_{n\to\infty} \frac{\int_0^{q(j,n)} h_j(s)ds}{\int_0^n h_j(s)ds} \cdot \frac{\int_0^n h_j(s)ds}{\int_0^n h_i(s)ds} \cdot \frac{\int_0^n h_i(s)ds}{\int_0^{q(i,n)} h_i(s)ds} \\
&= a_{ij} \quad \text{a.s.}
\end{aligned}$$

uniformly in z in a compact interval, by (2.5) and (3.8).

Thus, for $x > 0$,

$$\lim_{t\to\infty} \frac{\int_0^x \int_0^t \mu_{s+y}(j)dsdy}{\int_0^x \int_0^t \mu_{s+y}(i)dsdy} = \lim_{t\to\infty} \frac{\int_0^t \int_0^x \mu_{s+y}(j)dsdy}{\int_0^t \int_0^x \mu_{s+y}(i)dsdy} = a_{ij} \quad \text{a.s.}$$

By l'Hôpital's rule,

$$\lim_{t\to\infty} \frac{\int_0^x \mu_{t+y}(j)dy}{\int_0^x \mu_{t+y}(i)dy} = a_{ij} \quad \text{a.s.}$$

It follows that, a.s., any limit point $\overline{\mu}^*$ of $\{\overline{\mu}^t\}$ in U as $t \to \infty$ must satisfy $\int_0^x \mu_t^*(j)dt / \int_0^x \mu_t^*(i)dt = a_{ij}$. Then so will $\overline{\mu}^{*t}$, $t \geq 0$. Since $x > 0$ was arbitrary, we have $\mu_t^*(j)/\mu_t^*(i) = a_{ij}$ for almost every t, where we may drop the "almost every t" by taking a suitable modification. Then we must have $\mu_t^*(i) = \overline{\beta}(i)\ \forall i,\ t$, and the matrix $M^{\overline{\mu}^*}(t)$ is the constant diagonal matrix $M^* = \text{diag}(\overline{\beta}(1), \ldots, \overline{\beta}(d))$. The rest is easy. □

Remark. In the latter case, one may in fact replace the a-robustness condition and the condition $a \leq \overline{a}$ by the simpler condition $\nabla V.\, M^* f < 0$ outside J.

In particular, if $\{a(n,i)\}$ are identical, $M^{\overline{\mu}^*}$ is $1/d$ times the identity matrix, implying that the rescaled time axis is apportioned equally to all components. One may dub this the "asymptotic equipartition of time."

4. Conclusions. The foregoing analysis raises several interesting issues, which are listed below.

1. We have not presented any results on the convergence rate. For the ODE, the rate of convergence to J^ϵ for $\epsilon > 0$ could be gleaned from the Liapunov function and would be eventually mimicked by the interpolated algorithm $x(\cdot)$. There are two catches here. One is that "eventually" could be a long way into the future. Second, the passage from $\{X(n)\}$ to $x(\cdot)$ involves a time-scaling $n \to t(n)$, which has to be inverted to obtain the actual convergence rate of $\{X(n)\}$. These aspects need further study.

2. It seems plausible that one could retain the above results if (3.7) were replaced by the weaker requirement that the corresponding liminf be bounded away from zero. One cannot then expect $\{\overline{\mu}^t\}$ to a.s. converge to a fixed element, but it is conjectured that one will still retain the property that all limit points of $\{\overline{\mu}^t\}$ in U are a-thick for some $a > 0$.

3. In engineering applications, $\{a(n,i)\}$ are design parameters and can be chosen to be balanced. This may not, however, be so in "emergent" computations or when (1.4) is merely a computational paradigm for a natural process such as a macroeconomic learning system. An interesting problem, then, is to let each agent (processor) "learn" its stepsize scheme in real time based on observations of stepsizes used by, say, "neighboring" agents. One may then try to show that under reasonable conditions, this leads to balanced schemes.

4. If we had allowed some of the a_{ij}'s to be zero, it is clear that the corresponding diagonal elements of M^* will be zero and M^* is no longer thick. This reflects different time scales in the speed of adjustment of different learners. It would be interesting to analyze this situation using the theory of singularly perturbed differential equations.

5. If (1.2) had no extraneous randomness, i.e., $F(X(n), \xi(n)) = H(X(n))\ \forall n$ for a suitable H, the foregoing shows that a stepsize scheme from ITS suppresses the effects of communication delays in deterministic recursions under a mild conditional moment condition (A3). This is in contrast to the usual role of ITS as a pure noise-suppressing mechanism. Compare this with the fact that even linear recursions with constant stepsize show very complex behavior in the presence of communication delays [12].

6. Yet another possibility to explore is the use of the Wentzell–Freidlin theory of small noise asymptotics [10] to get a dynamic picture of the behavior of the algorithm in the vicinity of J, in particular, to see if it favors certain points in J. This is in the spirit of some recent work on annealing algorithms [11] and equilibrium selection in evolutionary games [9].

Acknowledgments. This work has benefited significantly from discussions with Profs. Vinod Sharma and Daniel Ocone, and by the comments of the referees and the associate editor.

REFERENCES

[1] J. Abounadi, D. Bertsekas, and V.S. Borkar, *O.D.E. Analysis for Q-Learning Algorithms*, preprint, Laboratory for Information and Decision Systems, M.I.T., Cambridge, MA, 1996.

[2] A. Benveniste, M. Metivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximations*, Springer-Verlag, Berlin, Heidelberg, 1990.

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

[4] V.S. Borkar, *Topics in Controlled Markov Chains*, Pitman Res. Notes in Math. Ser. 240, Longman Scientific and Technical, Harlow, UK, 1991.

[5] V.S. Borkar, *Distributed computation of fixed points of ∞-nonexpansive maps*, Proc. Indian Acad. Sci. Math. Sci., 106 (1996), pp. 289–300.

[6] V.S. Borkar and V.V. Phansalkar, *Managing interprocessor delays in distributed recursive algorithms*, Sādhanā, 19 (1994), pp. 995–1003.

[7] V.S. Borkar and K. Soumyanath, *A new analog parallel scheme for fixed point computation*, Part I: Theory, IEEE Trans. Circuits Systems I Fund. Theory Appl., 44 (1997), pp. 509–522.

[8] H.F. Chen, *Stochastic approximation and its new applications*, in Proc. 1994 Hong Kong Internat. Workshop on New Directions in Control and Manufacturing, 1994, pp. 2–12.

[9] D. Foster and P. Young, *Stochastic evolutionary game dynamics*, Theoret. Population Biol., 38 (1990), pp. 229–232.

[10] M. Freidlin and A. Wentzell, *Random Perturbations of Dynamical Systems*, Springer-Verlag, Berlin, New York, 1984.

[11] S.B. Gelfand and S.K. Mitter, *Recursive stochastic algorithms for global optimization in R^d*, SIAM J. Control Optim., 29 (1992), pp. 999–1018.

[12] R. Gharavi and V. Anantharam, *A Structure Theorem for Partially Asynchronous Relaxations with Random Delays*, ERL Memo. No. M92/143, Electronics Research Laboratory, University of California, Berkeley, 1993.

[13] S. Haykin, *Neural Networks: A Comprehensive Foundation*, McMillan, New York, 1994.

[14] M. Hirsch, *Convergent activation dynamics in continuous time networks*, Neural Networks, 2 (1987), pp. 331–349.

[15] T. Jaakola, M. Jordan, and S.P. Singh, *On the convergence of stochastic iterative dynamic programming algorithms*, Neural Computation, 6 (1994), pp. 1185–1201.

[16] H. Kushner and D. Clark, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Springer-Verlag, New York, 1978.

[17] H. Kushner and G. Yin, *Stochastic approximation algorithms for parallel and distributed processing*, Stochastics, 22 (1987), pp. 219–250.

[18] H. Kushner and G. Yin, *Asymptotic properties of distributed and communicating stochastic approximation algorithms*, SIAM J. Control Optim., 25 (1987), pp. 1266–1290.

[19] J. Neveu, *Discrete Parameter Martingales*, North-Holland, Amsterdam, 1975.

[20] T. Sargent, *Bounded Rationality in Macroeconomics*, Clarendon Press, Oxford, UK, 1993.

[21] J. Tsitsiklis, *Asynchronous stochastic approximation and Q-learning*, Machine Learning, 16 (1994), pp. 185–202.

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

SIAM J. CONTROL OPTIM.
Vol. 36, No. 3, pp. 852–894, May 1998

004

BOUNDARY VALUE PROBLEMS AND OPTIMAL BOUNDARY CONTROL FOR THE NAVIER–STOKES SYSTEM: THE TWO-DIMENSIONAL CASE*

A. V. FURSIKOV†, M. D. GUNZBURGER‡, AND L. S. HOU§

Abstract. We study optimal boundary control problems for the two-dimensional Navier–Stokes equations in an unbounded domain. Control is effected through the Dirichlet boundary condition and is sought in a subset of the trace space of velocity fields with minimal regularity satisfying the energy estimates. An objective of interest is the drag functional. We first establish three important results for inhomogeneous boundary value problems for the Navier–Stokes equations; namely, we identify the trace space for the velocity fields possessing finite energy, we prove the existence of a solution for the Navier–Stokes equations with boundary data belonging to the trace space, and we identify the space in which the stress vector (along the boundary) of admissible solutions is well defined. Then, we prove the existence of an optimal solution over the control set. Finally, we justify the use of Lagrange multiplier principles, derive an optimality system of equations in the weak sense from which optimal states and controls may be determined, and prove that the optimality system of equations satisfies in appropriate senses a system of partial differential equations with boundary values.

Key words. optimal control, Navier–Stokes equations, boundary value problem, drag minimization

AMS subject classifications. 76D05, 49J20, 49K20, 35K50

PII. S0363012994273374

1. Introduction.

Optimal control problems for fluid flows have been a subject of interest to experimenters and designers since at least the time of Prandtl. In more recent times, they have also become of substantial interest to mathematicians and computational scientists. For the steady state Navier–Stokes system, complete and systematic mathematical and numerical analyses of optimal control problems of different types (e.g., having Dirichlet, Neumann, and distributed controls and also finite-dimensional controls) were given in [15, 16, 17, 18]. Mathematical treatments of optimal control problems for the time-dependent Navier–Stokes system were given in [2], [6, 7, 8, 9, 10, 11, 12, 13], [20], and [24, 25, 26, 27]. In [6], free convection problems with boundary heat flux controls were considered; the existence of optimal solutions was proved and necessary conditions that characterize optimal controls and states were derived. In [11, 12, 13], the existence of optimal distributed controls was shown, an optimality system of equations was derived, and the question of the uniqueness

*Received by the editors August 24, 1994; accepted for publication (in revised form) February 24, 1997.

http://www.siam.org/journals/sicon/36-3/27337.html

†Department of Mechanics and Mathematics, Moscow State University, Moscow 119899, Russia (fursikov@dial01.msu.ru). The research of this author was supported by Air Force Office of Scientific Research grant AFOSR-93-1-0280 while he was visiting Virginia Tech, and by Natural Science and Engineering Research Council of Canada grant OGP-0137436 while he was visiting York University.

‡Interdisciplinary Center for Applied Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061-0531. Current address: Department of Mathematics, Iowa State University, Ames, IA 50011-2066 (gunzburg@iastate.edu). The research of this author was supported by Air Force Office of Scientific Research grant AFOSR-93-1-0280 and Office of Naval Research grant N00014-91-J-1493.

§Department of Mathematics and Statistics, York University, North York, ON M3J 1P3, Canada (hou@mathstat.yorku.ca). The research of this author was supported by the Natural Science and Engineering Research Council of Canada grant OGP-0137436.

of optimal solutions was resolved. Distributed controls were also considered in [20]. Various optimal control problems involving both distributed and boundary controls were considered in [2], although detailed proofs were provided only for the case of distributed controls. In [7, 8, 9, 10] and [24, 25, 26, 27] extensive studies of optimal control problems were given for Dirichlet controls in a special case, namely, when the control is of the separation-of-variable type.

In this paper we consider general Dirichlet controls for the time-dependent, two-dimensional Navier–Stokes system in the exterior of a bounded domain. Our eventual goal is to derive an optimality system from which optimal controls and states may be determined. A feature of the Dirichlet boundary control problem is as follows: one can derive an optimality system only in spaces of sufficiently smooth functions for which the nonlinear terms of the Navier–Stokes system are subordinate to the linear terms. (In the case of distributed control the situation is different; see [13].) In the two-dimensional case, the space of minimal smoothness possessing this property is the space of functions with "finite energy." Therefore, we first identify the space of boundary values which allows us to obtain finite energy solutions for the Navier–Stokes equations. Then, we prove the existence of an optimal solution in the finite energy space. Note that it would be easier to prove the solvability of an optimal control problem in a certain space of nonsmooth functions, but such a result is useless for the derivation of an optimality system. Finally, we use Lagrange multiplier techniques to derive a boundary value problem that the optimal states and control must satisfy. This boundary value problem is called the optimality system. We rigorously justify the boundary conditions for this system by means of techniques for elliptic boundary value problems in spaces of distributions and a theory, given below, about stress regularity for solutions of the Navier–Stokes equations with nonhomogeneous Dirichlet boundary conditions. In contrast to parabolic boundary value problems, it is here necessary to fulfill the compatibility conditions for boundary and initial values even in the case of nonsmooth solutions.

2. Formulation of the problem.

2.1. Derivation of the cost functional. We will formally derive the drag functional for flows surrounding a finite body. We consider the motion of an incompressible fluid in an unbounded domain that is described by the system

$$\rho\partial_t \mathbf{v} - \mu\Delta\mathbf{v} + \rho\mathbf{v}\cdot\nabla\mathbf{v} + \nabla p = \mathbf{0} \qquad \text{and} \qquad \nabla\cdot\mathbf{v} = 0 \qquad \text{in } (0,T)\times\Omega\,, \tag{2.1}$$

$$\mathbf{v}|_{t=0} = \mathbf{v}_0 \quad \text{for } \mathbf{x}\in\Omega\,, \qquad \mathbf{v}|_{\partial\Omega} = \mathbf{g} \quad \text{for } t\in(0,T)\,, \tag{2.2}$$

and

$$\mathbf{v} \to \mathbf{v}_\infty \quad \text{as } |\mathbf{x}|\to\infty\,. \tag{2.3}$$

Here, $\partial_t = \partial/\partial t$, Ω is the region exterior to a bounded body $B \subset \mathbb{R}^2$, and $\partial\Omega$ is its boundary. For simplicity we assume $\partial\Omega$ is of class C^∞ and is a connected closed curve without self-intersections. Also, the density ρ is a constant and $\mathbf{v}_\infty$ is a constant vector; the exact nature of the behavior at infinity will be discussed later. Later on we will add a condition on p so that, for given $\mathbf{v}_0$, $\mathbf{g}$, and $\mathbf{v}_\infty$, the problem (2.1)–(2.3) has a unique solution. When $\mathbf{g} = \mathbf{0}$, (2.1)–(2.3) is the problem of a fluid moving around the body B with uniform velocity $\mathbf{v}_\infty$ at infinity.

Denote by $\partial\Omega_\epsilon$ a smooth closed curve in a neighborhood of $\partial\Omega$, surrounding $\partial\Omega$ and lying inside Ω; Ω_ϵ is the part of Ω bounded by $\partial\Omega_\epsilon$ and containing the point at

infinity. Let $\mathcal{T} = -p\mathcal{I} + 2\mu\mathcal{D}$ be the stress tensor; here, $\mathcal{D} = \mathcal{D}(\mathbf{v}) = \frac{1}{2}(\nabla\mathbf{v} + \nabla\mathbf{v}^T)$ is the rate of deformation tensor for the flow. Then, for $\mathbf{x} \in \partial\Omega_\epsilon$, $(\mathcal{T}\mathbf{n})(t, \mathbf{x})$ is the force at a point $\mathbf{x}$ on $\partial\Omega_\epsilon$ which acts on the fluid in Ω_ϵ at the time t; here, $\mathbf{n}$ denotes the unit normal to the curve $\partial\Omega_\epsilon$ which is outer with respect to Ω_ϵ. Thus,

$$\int_0^T dt \int_{\partial\Omega_\epsilon} (\mathbf{v} - \mathbf{v}_\infty) \cdot (\mathcal{T}\mathbf{n})\, ds$$

is the work needed to overcome the drag exerted on the "body" $B_\epsilon = \mathbb{R}^2 \backslash \Omega_\epsilon$ over the time interval $(0, T)$. After passing to the limit as $\epsilon \to 0$, we obtain the work needed to overcome the drag exerted on the given body $B = \mathbb{R}^2 \backslash \Omega$:

$$\mathcal{W} = \int_0^T dt \int_{\partial\Omega} (\mathbf{v} - \mathbf{v}_\infty) \cdot (\mathcal{T}\mathbf{n})\, ds\,.$$

Using the definitions of $\mathcal{T}$ and D, and taking into account that $\mathbf{v}_\infty$ is a constant vector, we have that

$$\begin{aligned}\mathcal{W} &= \int_0^T dt \int_{\partial\Omega} (\mathbf{v} - \mathbf{v}_\infty) \cdot \{-p\mathbf{n} + \mu(\nabla\mathbf{v} + \nabla\mathbf{v}^T)\mathbf{n}\}\, ds \\ &= \int_0^T dt \int_{\partial\Omega} (\mathbf{v} - \mathbf{v}_\infty) \cdot \big\{-p\mathbf{n} + \mu\big(\nabla(\mathbf{v} - \mathbf{v}_\infty) + \nabla(\mathbf{v} - \mathbf{v}_\infty)^T\big)\mathbf{n}\big\}\, ds\,.\end{aligned}$$

Upon setting $\mathbf{w} = \mathbf{v} - \mathbf{v}_\infty$,

$$\mathcal{W} = \int_0^T \int_{\partial\Omega} \mathbf{w} \cdot \{-p\mathbf{n} + 2\mu\mathcal{D}(\mathbf{w})\mathbf{n}\}\, ds\, dt\,. \tag{2.4}$$

Let $\Omega_R = \Omega \cap \{\mathbf{x} \in \mathbb{R}^2 : |\mathbf{x}| < R\}$ and $\Gamma_R = \partial\Omega_R \setminus \partial\Omega$ for sufficiently large R such that the circle of radius R centered at the origin contains Ω. Using Green's formula we obtain

$$\begin{aligned}&\int_{\Omega_R} \mathbf{w} \cdot \big(\nabla \cdot \mathcal{D}(\mathbf{w})\big)\, d\mathbf{x} \\ &\qquad = \int_{\partial\Omega} \mathbf{w} \cdot \mathcal{D}(\mathbf{w})\mathbf{n}\, ds + \int_{\Gamma_R} \mathbf{w} \cdot \mathcal{D}(\mathbf{w})\mathbf{n}\, ds - \int_{\Omega_R} \mathcal{D}(\mathbf{w}) : \nabla\mathbf{w}\, d\mathbf{x}\,,\end{aligned} \tag{2.5}$$

where $\partial_j = \partial/\partial x_j$; i.e., ∂_j denotes the partial derivative with respect to the jth coordinate, $\nabla \cdot \mathcal{S}$ for a two-tensor $\mathcal{S} = \{S_{ij}\}$ is defined as the vector $(\partial_j S_{1j}, \partial_j S_{2j})^T$, and the colon notation denotes the scalar product operation on two two-tensors; i.e., for two-tensors $\mathcal{T} = \{T_{ij}\}$ and $\mathcal{S} = \{S_{ij}\}$, $\mathcal{T} : \mathcal{S} = T_{ij}S_{ij}$. Also, we have employed the convention that repeated indices imply summation. From (2.1) we have that $\mathbf{w} = \mathbf{v} - \mathbf{v}_\infty$ satisfies $\nabla \cdot \mathbf{w} = 0$ so that (2.5) and the identity

$$2\nabla \cdot \mathcal{D}(\mathbf{w}) = \Delta\mathbf{w} + \nabla(\nabla \cdot \mathbf{w})$$

yield

$$\int_{\partial\Omega} \mathbf{w} \cdot \mathcal{D}(\mathbf{w})\mathbf{n}\, ds = \frac{1}{2}\int_{\Omega_R} \mathbf{w} \cdot \Delta\mathbf{w}\, d\mathbf{x} + \int_{\Omega_R} \mathcal{D}(\mathbf{w}) : \nabla\mathbf{w}\, d\mathbf{x} - \int_{\Gamma_R} \mathbf{w} \cdot \mathcal{D}(\mathbf{w})\mathbf{n}\, ds\,.$$

The symmetry of the tensor $\mathcal{D}(\mathbf{w})$ yields

$$\int_{\Omega_R} \mathcal{D}(\mathbf{w}) : \nabla\mathbf{w}\, d\mathbf{x} = \int_{\Omega_R} \mathcal{D}(\mathbf{w}) : \mathcal{D}(\mathbf{w})\, d\mathbf{x}$$

so that

$$\int_{\partial\Omega} \mathbf{w}\cdot\mathcal{D}(\mathbf{w})\mathbf{n}\,ds = \frac{1}{2}\int_{\Omega_R} \mathbf{w}\cdot\Delta\mathbf{w}\,d\mathbf{x} + \int_{\Omega_R} \mathcal{D}(\mathbf{w}):\mathcal{D}(\mathbf{w})\,d\mathbf{x} - \int_{\Gamma_R} \mathbf{w}\cdot\mathcal{D}(\mathbf{w})\mathbf{n}\,ds\,.$$

Since $\mathcal{D}(\mathbf{w}) = \mathcal{D}(\mathbf{v})$, the substitution of the last equation into (2.4) yields

$$\begin{aligned}\mathcal{W} = \int_0^T\int_{\Omega_R} \mathbf{w}\cdot(\mu\Delta\mathbf{w} - \nabla p)\,d\mathbf{x}\,dt + 2\mu\int_0^T\int_{\Omega_R} \mathcal{D}(\mathbf{v}):\mathcal{D}(\mathbf{v})\,d\mathbf{x}\,dt\\ - 2\mu\int_0^T\int_{\Gamma_R} \mathbf{w}\cdot\mathcal{D}(\mathbf{w})\mathbf{n}\,ds\,dt + \int_0^T\int_{\Gamma_R} p\mathbf{w}\cdot\mathbf{n}\,ds\,dt\end{aligned}$$

so that by taking the limit $R\to\infty$ we obtain

$$\mathcal{W} = \int_0^T\int_{\Omega} \mathbf{w}\cdot(\mu\Delta\mathbf{w} - \nabla p)\,d\mathbf{x}\,dt + 2\mu\int_0^T\int_{\Omega} \mathcal{D}(\mathbf{v}):\mathcal{D}(\mathbf{v})\,d\mathbf{x}\,dt\,. \tag{2.6}$$

From (2.1) we have that $\mathbf{w} = \mathbf{v} - \mathbf{v}_\infty$ satisfies

$$\rho\partial_t\mathbf{w} - \mu\Delta\mathbf{w} + \rho\mathbf{v}\cdot\nabla\mathbf{w} + \nabla p = \mathbf{0}\,.$$

Combining (2.6) and the last equation yields

$$\begin{aligned}\mathcal{W} &= 2\mu\int_0^T dt\int_\Omega \mathcal{D}(\mathbf{v}):\mathcal{D}(\mathbf{v})\,d\mathbf{x} + \frac{\rho}{2}\int_0^T dt\int_\Omega \partial_t|\mathbf{w}|^2\,d\mathbf{x}\\ &\qquad + \rho\int_0^T dt\int_\Omega (\mathbf{v}\cdot\nabla\mathbf{w})\cdot\mathbf{w}\,d\mathbf{x}\\ &= 2\mu\int_0^T dt\int_\Omega \mathcal{D}(\mathbf{v}):\mathcal{D}(\mathbf{v})\,d\mathbf{x} + \frac{\rho}{2}\int_0^T dt\int_{\partial\Omega} |\mathbf{w}|^2\mathbf{v}\cdot\mathbf{n}\,ds\\ &\qquad + \frac{\rho}{2}\int_\Omega |\mathbf{w}(T,\mathbf{x})|^2\,d\mathbf{x} - \frac{\rho}{2}\int_\Omega |\mathbf{w}(0,\mathbf{x})|^2\,d\mathbf{x}\,.\end{aligned} \tag{2.7}$$

The integral $\frac{\rho}{2}\int_\Omega |\mathbf{w}(t,\mathbf{x})|^2\,d\mathbf{x}$ is the (finite) kinetic energy of the difference flow $\mathbf{w} = \mathbf{v} - \mathbf{v}_\infty$. (Note that the kinetic energy of the flow $\frac{\rho}{2}\int_\Omega |\mathbf{v}(t,\mathbf{x})|^2\,d\mathbf{x} = \infty$.) We can rewrite (2.7) as the energy equality

$$\begin{aligned}&\frac{\rho}{2}\int_\Omega |\mathbf{w}(0,\mathbf{x})|^2\,d\mathbf{x} + \int_0^T dt\int_{\partial\Omega} \mathbf{w}\cdot\mathcal{T}\mathbf{n}\,ds\\ &\qquad = \frac{\rho}{2}\int_\Omega |\mathbf{w}(T,\mathbf{x})|^2\,d\mathbf{x} + 2\mu\int_0^T dt\int_\Omega \mathcal{D}(\mathbf{v}):\mathcal{D}(\mathbf{v})\,d\mathbf{x} + \frac{\rho}{2}\int_0^T dt\int_{\partial\Omega} |\mathbf{w}|^2\mathbf{v}\cdot\mathbf{n}\,ds\,.\end{aligned}$$

This relation may be interpreted as follows: the initial kinetic energy of the difference flow plus the work due to drag is equal to the final, i.e., at $t = T$, value of the kinetic energy of the difference flow plus the energy dissipated due to friction plus the work done by the boundary control. Whenever the control is absent, i.e., whenever $\mathbf{v}|_{\partial\Omega} = \mathbf{g} = \mathbf{0}$, the third integral on the right-hand side of the last equation vanishes. Since the initial kinetic energy of the difference flow is given, it is quite natural to take the right-hand side of the last equation as the cost functional (for convenience, we introduce a factor of one-half):

$$\begin{aligned}\mathcal{J}(\mathbf{w}) = \frac{\rho}{4}\int_\Omega |\mathbf{w}(T,\mathbf{x})|^2\,d\mathbf{x} + \mu\int_0^T dt\int_\Omega \mathcal{D}(\mathbf{v}):\mathcal{D}(\mathbf{v})\,d\mathbf{x}\\ + \frac{\rho}{4}\int_0^T dt\int_{\partial\Omega} |\mathbf{w}|^2\mathbf{v}\cdot\mathbf{n}\,ds\,.\end{aligned} \tag{2.8}$$

2.2. Constraints on the control. For both physical and mathematical reasons, the size of the control should be constrained. Physically, one cannot realize controls of arbitrary size. Moreover, the cost of effecting control should be accounted for in the optimization process; e.g., one would not usually want to reduce the drag by a small amount if the cost of doing so is prohibitive. Limits on the size of the control are also needed in order to obtain a mathematically meaningful problem, e.g., to guarantee the existence of an optimal solution in a certain function class. Of course, the physical and mathematical needs for limiting the size of the control are not unrelated.

It is simpler to explain the ideas concerning constraining the size of the control in the steady state context in which we have the governing system

$$-\mu\Delta\mathbf{v} + \rho\mathbf{v}\cdot\nabla\mathbf{v} + \nabla p = \mathbf{0} \qquad \text{and} \qquad \nabla\cdot\mathbf{v} = 0 \qquad \text{in } \Omega\,, \tag{2.9}$$

$$\mathbf{v}|_{\partial\Omega} = \mathbf{g}\,, \qquad \text{and} \qquad \mathbf{v} \to \mathbf{v}_\infty \quad \text{as } |\mathbf{x}| \to \infty \tag{2.10}$$

and the cost functional

$$\mathcal{J}_s(\mathbf{v}) = \int_{\partial\Omega} \mathbf{w}\cdot\mathcal{T}\,\mathbf{n}\,ds = \mu\int_\Omega \mathcal{D}(\mathbf{v}) : \mathcal{D}(\mathbf{v})\,d\mathbf{x} + \frac{\rho}{4}\int_{\partial\Omega} |\mathbf{w}|^2\mathbf{v}\cdot\mathbf{n}\,ds\,, \tag{2.11}$$

where $\mathbf{w} = \mathbf{v} - \mathbf{v}_\infty$, as noted previously. In all physically interesting situations one would want to minimize the drag functional (2.11). If there are no constraints on the control, i.e., on $\mathbf{v}$ along the boundary $\partial\Omega$, then it is easy to find a trivial control such that $\mathcal{J}_s(\mathbf{v}) = 0$. Indeed, if we take $\mathbf{v}|_{\partial\Omega} = \mathbf{v}_\infty$, then the solution of (2.9)–(2.10) is given by $\mathbf{v}(\mathbf{x}) = \mathbf{v}_\infty$ and $\nabla p = \mathbf{0}$, and then, clearly, $\mathcal{J}_s(\mathbf{v}) = \mathcal{J}_s(\mathbf{v}_\infty) = 0$. This implies that $\mathcal{J}_s(\mathbf{v})$ can possibly be negative, thereby the object occupying the region B is being propelled rather than being dragged, exactly the opposite of what we want to study. Thus, constraining the control is not only natural from the physical point of view of conserving resources, but is necessary for the minimization problem to model properly the desired physical objectives. (Note that in the time-dependent case, we cannot choose $\mathbf{v} = \mathbf{v}_\infty$ due to the initial condition of (2.2); however, we still want to limit the size of the control for the same reasons as in the steady state case.)

There are two common ways of constraining the control. The first one is to impose an explicit bound on the control. In the steady state case, we can impose

$$\int_{\partial\Omega} |\mathbf{v}|^k\,ds \le M \text{ for some } k \ge 3 \tag{2.12}$$

or

$$|\mathbf{v}(\mathbf{x})| \le M \quad \forall\,\mathbf{x} \in \partial\Omega\,, \tag{2.13}$$

where M is a prescribed positive constant. The constraint (2.12) allows the control to concentrate on small portions of the boundary and is therefore more useful in providing information about the locations where the control is most effective. Such information will be helpful in the study of "local controls," i.e., the application of control at a number of chosen locations on the boundary. (We will study local control problems elsewhere.) For this reason we will not pursue constraints of the type (2.13) any further in this paper. The second way of constraining the control is to add some norm of the control to the cost functional; e.g., instead of (2.11), we consider the functional

$$\mu\int_\Omega \mathcal{D}(\mathbf{v}) : \mathcal{D}(\mathbf{v})\,d\mathbf{x} + \frac{\rho}{4}\int_{\partial\Omega} |\mathbf{w}|^2\mathbf{v}\cdot\mathbf{n}\,ds + \rho N\int_{\partial\Omega} |\mathbf{v}|^k\,ds \tag{2.14}$$

for some $k \geq 3$ and $N > 0$. If $k = 3$, we will need $N > \frac{1}{4}$. In both ways of constraining the control, one can use different norms to measure the control. The physical problem does not tell us which norm to use, although desirable physical properties, e.g., having no sharp peaks in the control along the boundary, should influence the choice. The choice of norm is also influenced by the need to establish the well-posedness of the problem, e.g., the existence of an optimal solution in some function class, the regularity of the optimal solution, etc. For example, the constraints on the value of k are motivated by the need to have the cost of control, i.e., the last term in (2.14), dominate (in an appropriate sense that will be made clear later in this paper) the second term.

Our interest in this paper is in the time-dependent problem, and we now discuss how we can choose a convenient norm for measuring the control. Our starting point is the requirement that solutions of the Navier–Stokes system have energy estimates that will be needed later in this paper in studying the optimal control problems, and particularly in the derivation of the optimality system of equations. The minimum level of smoothness for the velocity field $\mathbf{v}$ at which the energy estimates are valid is $\mathbf{v} - \mathbf{v}_\infty \in L^2\big(0,T;\mathbf{H}^1(\Omega)\big)$ and $\partial_t \mathbf{v} \in L^2\big(0,T;\mathbf{H}^{-1}(\Omega)\big)$. Note that these inclusions imply a certain behavior at infinity. (The Sobolev space notation used here is established in section 3.1.) The boundary control should belong to a subset of the trace space on $\partial\Omega$ of the space for the vector field $\mathbf{v}$. The norm on the trace space will be shown to be

$$\begin{aligned}&\|\mathbf{v}\cdot\mathbf{n}\|_{L^2(0,T;H^{1/2}(\partial\Omega))} + \|\mathbf{v}\cdot\mathbf{n}\|_{H^{3/4}(0,T;H^{-1}(\partial\Omega))}\\ &+ \|\mathbf{v}\cdot\boldsymbol{\tau}\|_{L^2(0,T;H^{1/2}(\partial\Omega))} + \|\mathbf{v}\cdot\boldsymbol{\tau}\|_{H^{1/4}(0,T;L^2(\partial\Omega))}\,,\end{aligned}$$

where $\boldsymbol{\tau}$ denotes the counterclockwise unit tangent vector to $\partial\Omega$. Naturally, the control should be measured in a norm that is not weaker than the norm for the desired trace space. For computational convenience, we will strengthen the fractional time derivative to the first derivative ∂_t in the functional. Also, the particular form of the functional (2.8), i.e., the term $\int_0^T\!\int_\Omega \mathcal{D}(\mathbf{v}) : \mathcal{D}(\mathbf{v})\, d\mathbf{x}dt$, implies that in order for $\mathbf{v}$ to belong to the desired trace space, it is sufficient to use the norm

$$\int_0^T \int_{\partial\Omega} |\partial_t \mathbf{v}|^2 \, ds\, dt$$

for the controls. Also, as in (2.12), we have to include the constraints connected with the term $\int_0^T\!\int_{\partial\Omega} |\mathbf{v}|^k \, ds\, dt$ for some $k \geq 3$.

Hence, the two approaches of constraining the control in the time-dependent case can now be described as follows. The first approach, i.e., imposing an explicit bound on the control, requires that, for some constant $M > 0$,

$$\int_0^T \int_{\partial\Omega} |\mathbf{v}|^k \, ds\, dt + \int_0^T \int_{\partial\Omega} |\partial_t \mathbf{v}|^2 \, ds\, dt \leq M\,, \tag{2.15}$$

where $k \geq 3$. The second approach, i.e., adding a norm of the control to the functional, uses the functional

(2.16)

$$\begin{aligned}\mathcal{J}_N(\mathbf{v}) = \mu \int_0^T \int_\Omega \mathcal{D}(\mathbf{v}) : \mathcal{D}(\mathbf{v})\, d\mathbf{x}dt + \frac{\rho}{4}\int_0^T \int_{\partial\Omega} |\mathbf{w}|^2 \mathbf{v}\cdot\mathbf{n}\, ds\, dt\\ + \frac{\rho}{4}\int_\Omega |\mathbf{w}(T,\mathbf{x})|^2\, d\mathbf{x} + \rho N \left(\int_0^T \int_{\partial\Omega} |\mathbf{v}|^k \, ds\, dt + \int_0^T \int_{\partial\Omega} |\partial_t \mathbf{v}|^2 \, ds\, dt \right),\end{aligned}$$

where $\mathbf{w} = \mathbf{v} - \mathbf{v}_\infty$, $k \geq 3$, and $N > 0$ ($N > \frac{1}{4}$ if $k = 3$).

3. Precise statement of extremal problems. We will use the standard notations for the Lebesgue function space $L^r(\Omega)$ and the Sobolev spaces $W^{m,r}(\Omega)$, $H^m(\Omega)$, $W^{l,s}(\partial\Omega)$, and $H^l(\partial\Omega)$ for real numbers r, m, l, s, where m, l are the smoothness indices and r, s are the integrability indices. Also, $H^m(\Omega) = W^{m,2}(\Omega)$ and $H^l(\partial\Omega) = W^{l,2}(\partial\Omega)$. For $m \geq 0$, we introduce the subspaces of the Sobolev spaces $W^{m,r}(\Omega)$:

$$W_0^{m,r}(\Omega) = \text{the closure of } C_0^\infty(\Omega) \text{ in } W^{m,r}(\Omega)$$

and the dual spaces

$$W^{-m,r}(\Omega) = \big(W_0^{m,r'}(\Omega)\big)^*, \quad \text{where } \frac{1}{r} + \frac{1}{r'} = 1,\ 1 < r, r' < \infty\,.$$

Also, $H_0^m(\Omega) = W_0^{m,2}(\Omega)$ and $H^{-m}(\Omega) = W^{-m,2}(\Omega)$. The vector counterparts of these spaces are denoted by $\mathbf{L}^r(\Omega)$, $\mathbf{W}^{m,r}(\Omega)$, $\mathbf{H}^m(\Omega)$, $\mathbf{W}^{l,s}(\partial\Omega)$, $\mathbf{H}^l(\partial\Omega)$, $\mathbf{W}_0^{m,r}(\Omega)$, and $\mathbf{H}_0^m(\Omega)$. For details, see [1] and [14]. We will also use the solenoidal spaces

$$\mathbf{V}^m(\Omega) = \left\{\mathbf{u} \in \mathbf{H}^m(\Omega)\ :\ \nabla\cdot\mathbf{u} = 0,\ \int_{\partial\Omega} \mathbf{u}\cdot\mathbf{n}\, ds = 0\right\} \qquad \text{for } m \geq 0$$

and

$$\mathbf{V}_0^m(\Omega) = \text{the closure of } \mathbf{C}_0^\infty(\Omega) \cap \mathbf{V}^0(\Omega) \text{ in the } \mathbf{H}^m(\Omega)\text{-norm} \qquad \text{for } m \geq 0\,,$$

where when $m = 0$, $\int_{\partial\Omega} \mathbf{u}\cdot\mathbf{n}\, ds$ is understood as the $H^{-1/2}(\partial\Omega)$–$H^{1/2}(\partial\Omega)$ duality pairing between the function $(\mathbf{u}\cdot\mathbf{n}) \in H^{-1/2}(\partial\Omega)$ and the constant scalar function $1 \in H^{1/2}(\partial\Omega)$. Note that in the definition of $\mathbf{V}^m(\Omega)$ (Ω being unbounded), the condition $\int_{\partial\Omega} \mathbf{u}\cdot\mathbf{n}\, ds = 0$ does not follow from $\operatorname{div}\mathbf{u} = 0$ unless some additional assumptions are made on $\mathbf{u}$ at ∞. Note also that for simplicity, we have assumed $\partial\Omega$ is a connected curve; otherwise, we need to require $\int_{\Gamma_i} \mathbf{u}\cdot\mathbf{n}\, ds = 0$ on each connected component Γ_i of $\partial\Omega$. Identifying $\big(\mathbf{V}^0(\Omega)\big)^*$ with $\mathbf{V}^0(\Omega)$ we introduce the dual spaces

$$\mathbf{V}^{-m}(\Omega) = [\mathbf{V}_0^m(\Omega)]^* \qquad \text{for } m \geq 1\,.$$

The norms on $\mathbf{V}^m(\Omega)$ and $\mathbf{V}_0^m(\Omega)$ are chosen to be that of $\mathbf{H}^m(\Omega)$. We also introduce the temporal-spatial function space, defined on $Q = \mathbb{R}\times\Omega$,

$$\mathcal{H}^{(s)}(Q) = \{f \in L^2\big(\mathbb{R}; H^s(\Omega)\big)\ :\ \partial_t f \in L^2\big(\mathbb{R}; H^{s-2}(\Omega)\big)\}$$

with norm

$$\|f\|^2_{\mathcal{H}^{(s)}(Q)} = \|f\|^2_{L^2(\mathbb{R};H^s(\Omega))} + \|\partial_t f\|^2_{L^2(\mathbb{R};H^{s-2}(\Omega))}\,,$$

and the corresponding solenoidal function space

$$\mathcal{V}^{(s)}(Q) = \{\mathbf{v} \in L^2\big(\mathbb{R}; \mathbf{V}^s(\Omega)\big) : \partial_t\mathbf{v} \in L^2\big(\mathbb{R}; \mathbf{V}^{s-2}(\Omega)\big)\}$$

with norm

$$\|\mathbf{v}\|^2_{\mathcal{V}^{(s)}(Q)} = \|\mathbf{v}\|^2_{L^2(\mathbb{R};\mathbf{V}^s(\Omega))} + \|\partial_t\mathbf{v}\|^2_{L^2(\mathbb{R};\mathbf{V}^{s-2}(\Omega))}\,.$$

Analogously, we may define the function spaces $\mathcal{H}^{(s)}(Q_T)$ and $\mathcal{V}^{(s)}(Q_T)$ defined on $Q_T = (0,T)\times\Omega$.

With the help of the spaces defined above, we may define the solution for the Navier–Stokes equations (2.1)–(2.3). We first quote a useful lemma.

LEMMA 3.1. *The space* $\mathcal{V}^{(1)}(Q_T)$ *is continuously imbedded into* $C\big([0,T];\mathbf{V}^0(\Omega)\big)$.

Proof. Solving in Q_T the equations $\partial F(t,x)/\partial x_2 = u_1$, $\partial F(t,x)/\partial x_1 = -u_2$, and $\int_\Omega F(t,x)dx = 0$ almost everywhere (a.e.) $t \in (0,T)$ for an arbitrary $u = (u_1,u_2) \in \mathcal{V}^{(1)}(Q_T)$, we reduce the proof of the lemma to a proof of the continuity of the embedding $\mathcal{H}^{(2)}(Q_T) \subset C([0,T];H^1(\Omega))$. The last assertion is proved in [4] or [21]. (For an alternate proof, see [5] or [28].) ∎

Below, for the sake of simplicity, we set the constant density $\rho = 1$ or, more precisely, we introduce nondimensionalized variables so that now μ is the inverse of the Reynolds number.

DEFINITION 3.2. $\mathbf{v}$ *is said to be a solution of* (2.1)–(2.3) *if* $\mathbf{v} = \mathbf{w} + \mathbf{v}_\infty$, *where* $\mathbf{w} \in \mathcal{V}^{(1)}(Q_T)$ *satisfies*

$$\begin{aligned}\langle \partial_t \mathbf{w}(t), \mathbf{z}\rangle + 2\mu\int_\Omega \mathcal{D}(\mathbf{w}(t)) : \mathcal{D}(\mathbf{z})\,d\mathbf{x} + \int_\Omega \big(\mathbf{w}(t)\cdot\nabla\big)\mathbf{w}(t)\cdot\mathbf{z}\,d\mathbf{x} \\ + \int_\Omega \big(\mathbf{v}_\infty\cdot\nabla\big)\mathbf{w}(t)\cdot\mathbf{z}\,d\mathbf{x} = 0 \qquad \forall\,\mathbf{z}\in\mathbf{V}_0^1(\Omega),\quad \textit{a.e. } t\in(0,T)\,,\end{aligned} \tag{3.1}$$

$$\mathbf{w} = \mathbf{b} \equiv \mathbf{g} - \mathbf{v}_\infty \qquad \textit{in } L^2\big(0,T;\mathbf{H}^{1/2}(\partial\Omega)\big)\,, \tag{3.2}$$

and

$$\mathbf{w}|_{t=0} = \mathbf{w}_0 \equiv \mathbf{v}_0 - \mathbf{v}_\infty \qquad \textit{in } \mathbf{V}^0(\Omega)\,. \qquad ∎$$

Note that the initial condition in Definition 3.2 makes sense because of Lemma 3.1. Here and elsewhere in this paper, $\langle\cdot,\cdot\rangle$ denotes the duality pairing between a Banach space and its dual space; the underlying Banach space may vary depending on the context. In particular, $\langle\cdot,\cdot\rangle$ in (3.1) denotes the duality pairing between $\mathbf{V}^{-1}(\Omega)$ and $\mathbf{V}_0^1(\Omega)$. Also, note that we have used the identity

$$2\int_\Omega \mathcal{D}(\mathbf{w}) : \mathcal{D}(\mathbf{z})\,d\mathbf{x} = \int_\Omega \nabla\mathbf{w} : \nabla\mathbf{z}\,d\mathbf{x} \qquad \forall\,\mathbf{w}\in\mathbf{V}^1(\Omega),\,\mathbf{z}\in\mathbf{H}_0^1(\Omega)\,.$$

The extremal problems we study involve the objective of drag minimization. Based on the two ways of constraining the control, we have the two functionals (2.8) or (2.16) so that we state two extremal problems. It is more convenient to use the variable $\mathbf{w} = \mathbf{v} - \mathbf{v}_\infty$. Also, we will simply use $\mathbf{w}|_{\partial\Omega}$ to denote the Dirichlet control and, thus, we will not introduce a separate notation to denote the control variable and the boundary condition (3.2) will not be explicitly imposed as a constraint. Extremal solutions are sought in the space

$$\begin{aligned}Y = \Bigg\{ \mathbf{w}\in\mathcal{V}^{(1)}(Q_T)\ :\ (\partial_t\mathbf{w})|_{\partial\Omega}\in L^2\big(0,T;\mathbf{L}^2(\partial\Omega)\big), \\ \int_{\partial\Omega}\partial_t\mathbf{w}\cdot\mathbf{n}\,ds = 0\,,\quad \mathbf{w}|_{\partial\Omega}\in\mathbf{L}^k\big((0,T)\times\partial\Omega\big)\Bigg\}\end{aligned}$$

equipped with the norm

$$\|\mathbf{w}\|_Y = \|\mathbf{w}\|_{\mathcal{V}^{(1)}(Q_T)} + \|\partial_t\mathbf{w}\|_{L^2(0,T;\mathbf{L}^2(\partial\Omega))} + \|\mathbf{w}\|_{\mathbf{L}^k((0,T)\times\partial\Omega)}\,,$$

where $k \geq 3$ and $\mathbf{n}$ is the outward normal on $\partial\Omega$.

We also introduce the space

$$\mathbf{W} = \{\mathbf{w} \in \mathbf{V}^0(\Omega) : (\mathbf{w}\cdot\mathbf{n})|_{\partial\Omega} \in H^{1/4}(\partial\Omega) \cap L^{1+k/2}(\partial\Omega)\}\,.$$

Since the trace $\gamma_{n,\partial\Omega}(\mathbf{w}) \equiv (\mathbf{w}\cdot\mathbf{n})|_{\partial\Omega}$ is well defined and belongs to $H^{-1/2}(\partial\Omega)$ (see [28]), the definition of $\mathbf{W}$ makes sense. Note that the restriction operator

$$\gamma_0 : Y \to \mathbf{W}$$

defined by $\gamma_0\mathbf{w} = \mathbf{w}|_{t=0}$ is continuous. Indeed, we denote

$$Y_\delta = \left\{ \mathbf{w} \in L^2\big(0,T;\mathbf{H}^{1/2}(\partial\Omega)\big) \cap \mathbf{L}^k\big((0,T)\times\partial\Omega\big) : \right.$$
$$\left. \partial_t\mathbf{w} \in \mathbf{L}^2\big((0,T)\times\partial\Omega\big), \int_{\partial\Omega} \partial_t\mathbf{w}\, ds\, ds = 0 \right\}.$$

Then, since the trace operators $\gamma_{n,\partial\Omega} : Y \to Y_\delta$ and $\gamma_0 : Y_\delta \to H^{1/4}(\partial\Omega)\cap L^{1+k/2}(\partial\Omega)$ are continuous, the restriction $\gamma_0\mathbf{w}$ for an arbitrary $\mathbf{w} \in Y$ possesses the property

$$\gamma_{n,\partial\Omega}(\gamma_0\mathbf{w}) = \gamma_0(\gamma_{n,\partial\Omega}\mathbf{w}) \in H^{1/4}(\partial\Omega) \cap L^{1+k/2}(\partial\Omega)\,.$$

This proves that the imbedding $\gamma_0 Y \subset \mathbf{W}$ is continuous. We intend to look for an extremal solution in the space Y. Thus, we are compelled to replace the initial condition in Definition 3.2 by

$$\mathbf{w}|_{t=0} = \mathbf{w}_0 \equiv \mathbf{v}_0 - \mathbf{v}_\infty \in \mathbf{W}. \tag{3.3}$$

Problem I. Suppose that $\mathbf{w}_0 \equiv \mathbf{v}_0 - \mathbf{v}_\infty \in \mathbf{W}$. Seek a $\mathbf{w} \in Y$ such that the functional

$$\begin{aligned}\mathcal{J}_N(\mathbf{w}) &= \mu\int_0^T\int_\Omega \mathcal{D}(\mathbf{w}) : \mathcal{D}(\mathbf{w})\, d\mathbf{x}dt \\ &\quad + \frac{1}{4}\int_0^T\int_{\partial\Omega} |\mathbf{w}|^2(\mathbf{w}+\mathbf{v}_\infty)\cdot\mathbf{n}\, ds\, dt + \frac{1}{4}\int_\Omega |\mathbf{w}(T,\mathbf{x})|^2\, d\mathbf{x} \\ &\quad + N\int_0^T\int_{\partial\Omega}\big(|\mathbf{w}+\mathbf{v}_\infty|^k + |\partial_t\mathbf{w}|^2\big)\, ds\, dt\end{aligned} \tag{3.4}$$

is minimized subject to the constraints (3.1) and (3.3), where $k \geq 3$ and $N > 0$ with $N > \frac{1}{4}$ if $k = 3$.

Problem II. Suppose that $\mathbf{w}_0 \equiv \mathbf{v}_0 - \mathbf{v}_\infty \in \mathbf{W}$. Seek a $\mathbf{w} \in Y$ such that the functional

$$\begin{aligned}\mathcal{J}(\mathbf{w}) = \mu\int_0^T\int_\Omega \mathcal{D}(\mathbf{w}) : \mathcal{D}(\mathbf{w})\, d\mathbf{x}dt + \frac{1}{4}\int_0^T\int_{\partial\Omega} |\mathbf{w}|^2(\mathbf{w}+\mathbf{v}_\infty)\cdot\mathbf{n}\, ds\, dt \\ + \frac{1}{4}\int_\Omega |\mathbf{w}(T,\mathbf{x})|^2\, d\mathbf{x}\end{aligned} \tag{3.5}$$

is minimized subject to the constraints (3.1), (3.3), and

$$\int_0^T\int_{\partial\Omega}\big(|\mathbf{w}+\mathbf{v}_\infty|^k + |\partial_t\mathbf{w}|^2\big)\, ds\, dt \leq M\,, \tag{3.6}$$

where $k \geq 3$ and $M > 0$.

Note that Lemma 3.1 ensures that the functionals (3.4) and (3.5) are well defined on Y. We now give definitions for an admissible element and for a solution of Problem I or II.

DEFINITION 3.3. *An element* $\mathbf{w} \in Y$ *is called admissible if it satisfies* (3.1) *and* (3.3) *in the case of Problem* I *and satisfies* (3.1), (3.3), *and* (3.6) *in the case of Problem* II. *The set of admissible elements is denoted by* $\mathcal{V}_{ad}$. □

DEFINITION 3.4. *An element* $\widehat{\mathbf{w}} \in \mathcal{V}_{ad}$ *is called a solution of Problem* I *if*

$$\mathcal{J}_N(\widehat{\mathbf{w}}) = \inf_{\mathbf{w}\in\mathcal{V}_{ad}} \mathcal{J}_N(\mathbf{w})\,,$$

where $\mathcal{J}_N$ *is defined by* (3.4). *An element* $\widehat{\mathbf{w}} \in \mathcal{V}_{ad}$ *is called a solution of Problem* II *if*

$$\mathcal{J}(\widehat{\mathbf{w}}) = \inf_{\mathbf{w}\in\mathcal{V}_{ad}} \mathcal{J}(\mathbf{w})\,,$$

where $\mathcal{J}$ *is defined by* (3.5). □

4. An extension theorem, solutions of the Navier–Stokes equations, and the stress vector on the boundary. Our aim is to prove the existence of optimal solutions for Problems I and II and to obtain optimality systems of partial differential equations that optimal solutions must satisfy. To this end, we first prove three results that are of considerable interest in their own right in the study of Dirichlet boundary value problems for the Navier–Stokes equations.

The first result (section 4.1) is the identification of the trace space of $\mathcal{V}^{(1)}(Q_T) = L^2\big(0,T;\mathbf{V}^1(\Omega)\big)\cap H^1\big(0,T;\mathbf{V}^{-1}(\Omega)\big)$, i.e., the collection of velocity boundary data that can be extended into functions belonging to $\mathcal{V}^{(1)}(Q_T)$. The second result (section 4.2) is the existence of a solution of the Navier–Stokes equations with boundary values in these trace spaces along with a priori estimates for the solution. The third result (section 4.3) is the identification of the space in which the trace of the stress vector (on the boundary) of admissible solutions is well defined.

4.1. An extension theorem for boundary data. We prove some results concerning the extension of functions from the lateral surface of the time-space cylinder to the entire cylinder, i.e., from $(0,T)\times\partial\Omega$ to $(0,T)\times\Omega$.

We set $Q = \mathbb{R}\times\Omega$ (the infinite time-space cylinder) and $S = \mathbb{R}\times\partial\Omega$ (the lateral surface of the infinite time-space cylinder). The problem we want to consider is to describe the space of vector fields defined on S which can be extended to solenoidal vector fields defined on Q which belong to the space $\mathcal{V}^{(1)}(Q)$, where we recall, from section 3, the definition

$$\mathcal{V}^{(s)}(Q) = \{\mathbf{v}\in L^2\big(\mathbb{R};\mathbf{V}^{(s)}(\Omega)\big)\ :\ \partial_t\mathbf{v}\in L^2\big(\mathbb{R};\mathbf{V}^{(s-2)}(\Omega)\big)\}\,. \tag{4.1}$$

Alternatively, the task here is to characterize the trace space of $\mathcal{V}^{(1)}(Q)$. We will see that it is necessary to examine the normal trace and tangential trace separately, as they belong to different function spaces.

We denote by $\boldsymbol{\tau} = (\tau_1,\tau_2)^T$ and $\mathbf{n} = (n_1,n_2)^T$ the unit counterclockwise tangent and outward normal vectors, respectively, along $\partial\Omega$. We have the following relations:

$$\tau_1 = n_2 \quad\text{and}\quad \tau_2 = -n_1\,.$$

Given a boundary vector field

$$\mathbf{b}(t,\mathbf{x}) = b_n(t,\mathbf{x})\mathbf{n}(\mathbf{x}) + b_\tau(t,\mathbf{x})\boldsymbol{\tau}(\mathbf{x}) \qquad \text{a.e. } (t,\mathbf{x})\in S \tag{4.2}$$

satisfying

$$\int_{\partial\Omega} \mathbf{b}\cdot\mathbf{n}\, ds = 0 \qquad \text{a.e. } t \in \mathbb{R}\,, \tag{4.3}$$

where $b_n = \mathbf{b}\cdot\mathbf{n}$ and $b_\tau = \mathbf{b}\cdot\boldsymbol{\tau}$, we seek a solenoidal extension $\mathbf{u} = (u_1, u_2)^T \in \mathcal{V}^{(1)}(Q)$ of the form (see [19])

$$u_1 = \partial_2 F \qquad \text{and} \qquad u_2 = -\partial_1 F\,, \tag{4.4}$$

where F is the streamfunction for $\mathbf{u}$ and $\partial_i F = \partial F/\partial x_i$. In other words, given a boundary vector field $\mathbf{b}$ satisfying (4.3), we seek an F such that

$$b_n = -(\nabla F\cdot\boldsymbol{\tau})|_S \equiv -\partial_\tau F|_S \tag{4.5}$$

and

$$b_\tau = (\nabla F\cdot\mathbf{n})|_S \equiv \partial_n F|_S\,. \tag{4.6}$$

Note that the assumption (4.3) is necessary since we are seeking a solenoidal extension of the boundary data $\mathbf{b}$.

With the assumption (4.3), the relation (4.5) is equivalent to

$$F|_S = h \equiv -\int_{\mathbf{x}_0}^{\mathbf{x}} b_n(t, \mathbf{x}(s))\, ds\,, \tag{4.7}$$

where the line integral is taken along $\partial\Omega$ in the counterclockwise direction starting from a fixed point $\mathbf{x}_0 \in \partial\Omega$. Thus, for each given pair (b_τ, h) defined on S, we want to construct an $F \in \mathcal{H}^{(2)}(Q)$ satisfying (4.6) and (4.7), where

$$\mathcal{H}^{(s)}(\mathbb{R}\times\Theta) = \{u \in L^2\big(\mathbb{R}; H^s(\Theta)\big) \;:\; \partial_t u \in L^2\big(\mathbb{R}; H^{s-2}(\Theta)\big)\}\,.$$

Here $s \in \mathbb{R}$, Θ is any spatial domain, and the norm on $\mathcal{H}^{(s)}(\mathbb{R}\times\Theta)$ is defined by

$$\|F\|^2_{\mathcal{H}^{(s)}(\mathbb{R}\times\Theta)} = \|F\|^2_{L^2(\mathbb{R};H^s(\Theta))} + \|\partial_t F\|^2_{L^2(\mathbb{R};H^{s-2}(\Theta))} \quad \forall\, F \in \mathcal{H}^{(s)}(\mathbb{R}\times\Theta)\,.$$

We now prove that such an extension F exists provided that the boundary data (b_τ, h) belongs to an appropriate function space.

PROPOSITION 4.1. *A pair of functions (b_τ, h) defined on S possess an extension $F \in \mathcal{H}^{(2)}(Q)$ satisfying* (4.6), (4.7),

$$\begin{aligned}\|F\|^2_{\mathcal{H}^{(2)}(Q)} \le C\Big\{ &\|b_\tau\|^2_{L^2(\mathbb{R};H^{1/2}(\partial\Omega))} + \|b_\tau\|^2_{H^{1/4}(\mathbb{R};L^2(\partial\Omega))} \\ &+ \|h\|^2_{L^2(\mathbb{R};H^{3/2}(\partial\Omega))} + \|h\|^2_{H^{3/4}(\mathbb{R};L^2(\partial\Omega))}\Big\}\,,\end{aligned} \tag{4.8}$$

and

$$F \text{ vanishes outside a neighborhood of } S = \mathbb{R}\times\partial\Omega\,, \tag{4.9}$$

where C is a constant independent of b_τ, h, and F, if and only if

$$b_\tau \in L^2\big(\mathbb{R}; H^{1/2}(\partial\Omega)\big) \cap H^{1/4}\big(\mathbb{R}; L^2(\partial\Omega)\big) \tag{4.10}$$

and

$$h \in L^2\big(\mathbb{R}; H^{3/2}(\partial\Omega)\big) \cap H^{3/4}\big(\mathbb{R}; L^2(\partial\Omega)\big)\,. \tag{4.11}$$

Proof. Given b_τ and h satisfying (4.10)–(4.11), we construct an extension F satisfying (4.6)–(4.9); the converse result is easily proved as well. $\partial\Omega$ being of class C^∞, we may choose a neighborhood U of $\partial\Omega$ and a coordinate system $(x_1', x_2')^T$ such that $U = \big\{\mathbf{x} = (x_1', x_2')^T : (x_1', 0)^T \in \partial\Omega, x_2' \in [0, \epsilon]\big\}$ for some $\epsilon > 0$. The space $\mathcal{H}^{(2)}(\mathbb{R} \times U)$ can be rewritten in the form

$$\begin{aligned}
&\mathcal{H}^{(2)}(\mathbb{R} \times U)\\
&\quad = \big\{F(x_2', t, x_1') \in L^2\big(0, \epsilon; L^2(\mathbb{R}; H^2(\partial\Omega))\big) \cap L^2\big(0, \epsilon; H^1(\mathbb{R}; L^2(\partial\Omega))\big) :\\
&\qquad\qquad \partial_{x_2'x_2'} F \in L^2\big(0, \epsilon; L^2(\mathbb{R}; L^2(\partial\Omega))\big)\big\}\,.
\end{aligned}$$

By virtue of a trace theorem of [21], we have that the mappings $\gamma_0 : F \mapsto F|_{x_2'=0}$ and $\gamma_1 : F \mapsto \partial_{x_2'} F|_{x_2'=0}$ are well defined on $\mathcal{H}^{(2)}(\mathbb{R} \times U)$; furthermore, the mapping

$$\begin{aligned}
&F \mapsto (\gamma_0 F, \gamma_1 F) :\\
&\quad \mathcal{H}^{(2)}(\mathbb{R} \times U) \to \big[L^2\big(\mathbb{R}; H^2(\partial\Omega)\big) \cap H^1\big(\mathbb{R}; L^2(\partial\Omega)\big), L^2\big(\mathbb{R}; L^2(\partial\Omega)\big)\big]_{3/4}\\
&\qquad\qquad \times \big[L^2\big(\mathbb{R}; H^2(\partial\Omega)\big) \cap H^1\big(\mathbb{R}; L^2(\partial\Omega)\big), L^2\big(\mathbb{R}; L^2(\partial\Omega)\big)\big]_{1/4}
\end{aligned}$$

is continuous and surjective. Here we have used the intermediate spaces $[\mathcal{X}, \mathcal{Y}]_\alpha$, $\alpha \in [0, 1]$, of the Hilbert spaces $\mathcal{X}$ and $\mathcal{Y}$ as defined in [21]. Using the definition of these intermediate spaces (see [21]), we obtain

$$\begin{aligned}
&\big[L^2\big(\mathbb{R}; H^2(\partial\Omega)\big) \cap H^1\big(\mathbb{R}; L^2(\partial\Omega)\big), L^2\big(\mathbb{R}; L^2(\partial\Omega)\big)\big]_{3/4}\\
&\qquad = L^2\big(\mathbb{R}; H^{3/2}(\partial\Omega)\big) \cap H^{3/4}\big(\mathbb{R}; L^2(\partial\Omega)\big)
\end{aligned}$$

and

$$\begin{aligned}
&\big[L^2\big(\mathbb{R}; H^2(\partial\Omega)\big) \cap H^1\big(\mathbb{R}; L^2(\partial\Omega)\big), L^2\big(\mathbb{R}; L^2(\partial\Omega)\big)\big]_{1/4}\\
&\qquad = L^2\big(\mathbb{R}; H^{1/2}(\partial\Omega)\big) \cap H^{1/4}\big(\mathbb{R}; L^2(\partial\Omega)\big)\,.
\end{aligned}$$

Hence the mapping $F \mapsto (\gamma_0 F, \gamma_1 F)$ is continuous and surjective from $\mathcal{H}^{(2)}(\mathbb{R} \times U)$ to $\big[L^2\big(\mathbb{R}; H^{3/2}(\partial\Omega)\big) \cap H^{3/4}\big(\mathbb{R}; L^2(\partial\Omega)\big)\big] \times \big[L^2\big(\mathbb{R}; \mathbf{H}^{1/2}(\partial\Omega)\big) \cap H^{1/4}\big(\mathbb{R}; \mathbf{L}^2(\partial\Omega)\big)\big]$. Finally, we may choose another neighborhood $\widetilde{U}$ of $(0, \varepsilon) \times \partial\Omega$ such that the closure of U is contained in $\widetilde{U}$. Well-known extension results allow us to extend continuously the space $\mathcal{H}^{(2)}(\mathbb{R} \times U)$ into the space $\big\{F \in \mathcal{H}^{(2)}(\mathbb{R} \times \Omega) : F \text{ vanishes outside } \widetilde{U}\big\}$. ∎

We are now in a position to prove the main extension result. We denote the finite time-space cylinder by $Q_T = (0, T) \times \Omega$ and its lateral surface by $S_T = (0, T) \times \partial\Omega$.

THEOREM 4.2. *Assume that b_n and b_τ satisfy*

$$\int_{\partial\Omega} b_n \, ds = 0 \quad \text{a.e. } t \in [0, T]\,, \tag{4.12}$$

$$b_n \in L^2\big(0, T; H^{1/2}(\partial\Omega)\big) \cap H^{3/4}\big(0, T; H^{-1}(\partial\Omega)\big)\,, \tag{4.13}$$

and

$$b_\tau \in L^2\big(0,T;H^{1/2}(\partial\Omega)\big) \cap H^{1/4}\big(0,T;L^2(\partial\Omega)\big)\,. \tag{4.14}$$

Then, there exists a $\mathbf{u} \in \mathcal{V}^{(1)}(Q_T)$ *satisfying*

$$\mathbf{u}|_{S_T} = \mathbf{b} \equiv b_n\mathbf{n} + b_\tau\boldsymbol{\tau} \tag{4.15}$$

and the estimate

$$\begin{aligned}\|\mathbf{u}\|^2_{\mathcal{V}^{(1)}(Q_T)} \le C\Big\{ &\|b_n\|^2_{L^2(0,T;H^{1/2}(\partial\Omega))} + \|b_n\|^2_{H^{3/4}(0,T;H^{-1}(\partial\Omega))} \\ &+ \|b_\tau\|^2_{L^2(0,T;H^{1/2}(\partial\Omega))} + \|b_\tau\|^2_{H^{1/4}(0,T;L^2(\partial\Omega))}\Big\}\,,\end{aligned} \tag{4.16}$$

where C *is a constant independent of* b_n *and* b_τ*, and such that* $\mathbf{u}$ *vanishes outside a neighborhood of* $(0,T)\times\partial\Omega$*.*

Proof. By definition, the space $H^r\big(0,T;H^s(\partial\Omega)\big)$ with fractional indices r and s is the restriction to $(0,T)\times\partial\Omega$ of $H^r\big(\mathbb{R};H^s(\partial\Omega)\big)$. Thus, we may extend the data in time; i.e., there exists a $\widetilde{b}_n \in L^2\big(\mathbb{R};H^{1/2}(\partial\Omega)\big)\cap H^{3/4}\big(\mathbb{R};H^{-1}(\partial\Omega)\big)$ and $\widetilde{b}_\tau \in L^2\big(\mathbb{R};H^{1/2}(\partial\Omega)\big)\cap H^{1/4}\big(\mathbb{R};L^2(\partial\Omega)\big)$ such that

$$\widetilde{b}_n = b_n \quad\text{and}\quad \widetilde{b}_\tau = b_\tau \quad \text{on } (0,T)\times\partial\Omega\,,$$

$$\begin{aligned}&\|\widetilde{b}_n\|^2_{L^2(\mathbb{R};H^{1/2}(\partial\Omega))} + \|\widetilde{b}_n\|^2_{H^{3/4}(\mathbb{R};H^{-1}(\partial\Omega))} \\ &\qquad\le C\big\{\|b_n\|^2_{L^2(0,T;H^{1/2}(\partial\Omega))} + \|b_n\|^2_{H^{3/4}(0,T;H^{-1}(\partial\Omega))}\big\}\end{aligned}$$

and

$$\begin{aligned}&\|\widetilde{b}_\tau\|^2_{L^2(\mathbb{R};H^{1/2}(\partial\Omega))} + \|\widetilde{b}_\tau\|^2_{H^{1/4}(\mathbb{R};L^2(\partial\Omega))} \\ &\qquad\le C\big\{\|b_\tau\|^2_{L^2(0,T;H^{1/2}(\partial\Omega))} + \|b_\tau\|^2_{H^{1/4}(0,T;L^2(\partial\Omega))}\big\}\,.\end{aligned}$$

Furthermore, we may assume, without loss of generality, that

$$\int_{\partial\Omega} \widetilde{b}_n\, ds = 0 \qquad \text{a.e. } t \in \mathbb{R}\,.$$

Indeed, we can reset $\widetilde{\widetilde{b}}_n = \widetilde{b}_n - \big(\int_{\partial\Omega}\widetilde{b}_n\, ds \Big/ \int_{\partial\Omega} ds\big)$, if necessary. We define

$$\widetilde{h}(t,\mathbf{x}) = -\int_{\mathbf{x}_0}^{\mathbf{x}} \widetilde{b}_n(t,\mathbf{x}(s))\, ds \qquad \forall\,\mathbf{x}\in\partial\Omega\,,$$

where the line integral on the right-hand side is taken counterclockwise along $\partial\Omega$, starting from a given point $\mathbf{x}_0 \in \partial\Omega$. Evidently, $\widetilde{h} \in L^2\big(\mathbb{R};H^{3/2}(\partial\Omega)\big)\cap H^{3/4}\big(\mathbb{R};L^2(\partial\Omega)\big)$. Then, Proposition 4.1 implies that there exists an $F \in \mathcal{H}^{(2)}(Q)$ which vanishes outside a neighborhood of $\mathbb{R}\times\partial\Omega$ such that

$$F|_S = \widetilde{h} \qquad\text{and}\qquad \partial_n F|_S = \widetilde{b}_\tau\,.$$

By setting

$$\mathbf{u} = \mathbf{curl}\, F = \begin{pmatrix} \partial_2 F \\ -\partial_1 F \end{pmatrix}$$

we see that

$$\mathbf{u} \in \mathcal{V}^{(1)}(Q)\,,$$

$$(\mathbf{u}\cdot\mathbf{n})|_S = \mathbf{curl}\, F\cdot\mathbf{n}|_S = -\nabla F\cdot\boldsymbol{\tau} = -\partial_\tau F = -\partial_\tau \widetilde{h} = \widetilde{b}_n\,,$$

and

$$(\mathbf{u}\cdot\boldsymbol{\tau})|_S = \mathbf{curl}\, F\cdot\boldsymbol{\tau}|_S = \nabla F\cdot\mathbf{n} = \partial_n F = \widetilde{b}_\tau\,.$$

Hence, $\mathbf{u}|_{S_T} = \widetilde{\mathbf{b}} \equiv \widetilde{b}_n\mathbf{n} + \widetilde{b}_\tau\boldsymbol{\tau}$; i.e., $\mathbf{u}$ satisfies (4.15). The estimate (4.16) follows from Proposition 4.1. □

REMARK. We see from the proofs of Proposition 4.1 and Theorem 4.2 that the restriction operator $\mathbf{u}\mapsto(\mathbf{u}\cdot\mathbf{n})|_{\partial\Omega}$ is continuous from $\mathcal{V}^{(1)}(Q_T)$ to $H^{3/4}\big(0,T;H^{-1}(\partial\Omega)\big)\cap L^2\big(0,T;H^{1/2}(\partial\Omega)\big)$. Also, the trace operator $b_n\mapsto b_n|_{t=0}$ is continuous from the space $H^{3/4}\big(0,T;H^{-1}(\partial\Omega)\big)\cap L^2\big(0,T;H^{1/2}(\partial\Omega)\big)$ to $H^{-1/2}(\partial\Omega)$ (see [21]). Hence, the composition of these two operators, i.e., the operator $\mathbf{u}\mapsto[(\mathbf{u}\cdot\mathbf{n})|_{\partial\Omega}]|_{t=0}$, is continuous from $\mathcal{V}^{(1)}(Q_T)$ to $H^{-1/2}(\partial\Omega)$. On the other hand, the composition of the operators $\mathbf{u}\mapsto\mathbf{u}|_{t=0}$ and $\mathbf{u}|_{t=0}\mapsto(\mathbf{u}|_{t=0}\cdot\mathbf{n})|_{\partial\Omega}$ is continuous from $\mathcal{V}^{(1)}(Q_T)$ to $H^{-1/2}(\partial\Omega)$ (see Lemma 3.1 and [28]). Hence, using the denseness of $\mathbf{C}^\infty(Q_T)\cap\mathcal{V}^{(1)}(Q_T)$ in $\mathcal{V}^{(1)}(Q_T)$ we obtain the following compatibility condition for the extension $\mathbf{u}$ of Theorem 4.2:

$$(\mathbf{u}|_{t=0}\cdot\mathbf{n})\big|_{\partial\Omega} = \big((\mathbf{u}\cdot\mathbf{n})|_{\partial\Omega}\big)\big|_{t=0} \qquad \forall\,\mathbf{u}\in\mathcal{V}^{(1)}(Q_T)\,. \qquad \square$$

4.2. Estimates for the solutions of the Navier–Stokes equations with nonhomogeneous Dirichlet boundary data. We now consider the boundary value problem for the Navier–Stokes equation in the form introduced in Definition 3.2. The boundary data $\mathbf{b}$ is assumed to satisfy the compatibility condition (4.12). Our goal here is, with the help of the extension theorem of section 4.1, to establish the existence of a solution for (3.1)–(3.3) and derive estimates for the solutions in the space of *critical smoothness* in terms of the data $\mathbf{w}_0$ and $\mathbf{b}$.

Let b_n and b_τ be the normal and tangential components of the boundary value $\mathbf{b}$. We assume that b_n and b_τ satisfy (4.12)–(4.14) and that

$$\mathbf{w}_0 \in \mathbf{V}^0(\Omega)\,. \tag{4.17}$$

We also assume the compatibility condition

$$(\mathbf{w}_0\cdot\mathbf{n})\big|_{\partial\Omega} = b_n\big|_{t=0} \tag{4.18}$$

(see the remark at the end of section 4.1). We express the solution $\mathbf{w}$ of (3.1)–(3.3) in the form

$$\mathbf{w} = \mathbf{u} + \boldsymbol{\eta}\,,$$

where $\mathbf{u}\in\mathcal{V}^{(1)}(Q_T)$ is the vector field constructed in Theorem 4.2 satisfying (4.15) and (4.16). Note that the fact that $\mathbf{u}\in\mathcal{V}^{(1)}(Q)$ implies that $\mathbf{u}|_{t=0}\in\mathbf{L}^2(\Omega)$; see Lemma 3.1. Substituting $\mathbf{w}=\mathbf{u}+\boldsymbol{\eta}$ into (3.1)–(3.3), we obtain for $\boldsymbol{\eta}$

$$\begin{aligned}\langle\partial_t\boldsymbol{\eta}(t),\mathbf{z}\rangle + \mu\int_\Omega \nabla\boldsymbol{\eta}(t):\nabla\mathbf{z}\,d\mathbf{x} + \int_\Omega\big((\boldsymbol{\eta}(t)+\mathbf{u}(t)+\mathbf{v}_\infty)\cdot\nabla\big)\boldsymbol{\eta}(t)\cdot\mathbf{z}\,d\mathbf{x}\\ + \int_\Omega\big(\boldsymbol{\eta}(t)\cdot\nabla\big)\mathbf{u}(t)\cdot\mathbf{z}\,d\mathbf{x} = \langle\mathbf{f}(t),\mathbf{z}\rangle \qquad \forall\,\mathbf{z}\in\mathbf{V}^1_0(\Omega),\ \text{ a.e. } t\in(0,T)\,,\end{aligned} \tag{4.19}$$

$$\boldsymbol{\eta}|_{S_T} = \mathbf{0} \qquad \text{in } L^2\big(0,T;\mathbf{H}^{1/2}(\partial\Omega)\big)\,, \tag{4.20}$$

and

$$\boldsymbol{\eta}|_{t=0} = \boldsymbol{\eta}_0 \equiv \mathbf{w}_0 - \mathbf{u}|_{t=0} \qquad \text{in } \mathbf{V}_0^0(\Omega)\,, \tag{4.21}$$

where

$$\langle \mathbf{f}(t), \mathbf{z}\rangle = -\mu\int_\Omega \nabla\mathbf{u}(t) : \nabla\mathbf{z}\,d\mathbf{x} - \langle \partial_t\mathbf{u}(t), \mathbf{z}\rangle - \int_\Omega [(\mathbf{u}(t)+\mathbf{v}_\infty)\cdot\nabla]\mathbf{u}(t)\cdot\mathbf{z}\,d\mathbf{x}\,. \tag{4.22}$$

LEMMA 4.3. *Assume that the hypotheses of Theorem* 4.2 *hold. Let* $\mathbf{u}$ *be the vector field constructed in Theorem* 4.2. *Assume also that the compatibility condition* (4.18) *holds. Then, there exists a unique solution* $\boldsymbol{\eta} \in \mathcal{V}^{(1)}(Q_T)$ *of system* (4.19)–(4.21). *Moreover,* $\boldsymbol{\eta}$ *satisfies the estimate*

$$\begin{aligned}&\|\partial_t\boldsymbol{\eta}\|^2_{L^2(0,T;\mathbf{V}^{-1}(\Omega))} + \|\boldsymbol{\eta}\|^2_{L^\infty(0,T;\mathbf{V}^0(\Omega))} + \|\boldsymbol{\eta}\|^2_{L^2(0,T;\mathbf{V}^1(\Omega))}\\ &\quad\le A\left(\|\mathbf{f}\|_{L^2(0,T;\mathbf{V}^{-1}(\Omega))}\,,\ \|\mathbf{u}\|_{\mathcal{V}^{(1)}(Q_T)}\,,\ \|\boldsymbol{\eta}_0\|_{\mathbf{V}^0(\Omega)}\,,\ |\mathbf{v}_\infty|\right)\,,\end{aligned}$$

where $A(\cdot,\cdot,\cdot,\cdot)$ *is a continuous positive function defined on* $\mathbb{R}\times\mathbb{R}\times\mathbb{R}\times\mathbb{R}$ *and* $A(\lambda_1,\lambda_2,\lambda_3,|\mathbf{v}_\infty|)\to 0$ *as* $\boldsymbol{\lambda} = (\lambda_1,\lambda_2,\lambda_3)\to(0,0,0)$.

Proof. The existence and uniqueness of the solution $\boldsymbol{\eta}\in\mathcal{V}^{(1)}(Q_T)$ for (4.19)–(4.21) can be proved in exactly the same way as that for the two-dimensional Navier–Stokes equations with homogeneous boundary conditions in exterior domains; see, e.g., [19] or [28]. We only need to prove the estimate. (Note that $\boldsymbol{\eta}_0\in\mathbf{V}_0^0(\Omega)$.)

Letting $\mathbf{z} = \boldsymbol{\eta}(t,\cdot)$ in (4.19) yields

$$\begin{aligned}\frac{1}{2}\partial_t\|\boldsymbol{\eta}(t,\cdot)\|^2_{\mathbf{L}^2(\Omega)} + \mu\|\nabla\boldsymbol{\eta}(t,\cdot)\|^2_{\mathbf{L}^2(\Omega)} &= \langle\mathbf{f}(t,\cdot),\boldsymbol{\eta}(t,\cdot)\rangle - \int_\Omega(\boldsymbol{\eta}\cdot\nabla)\mathbf{u}\cdot\boldsymbol{\eta}\,d\mathbf{x}\\ &\le \frac{1}{\mu}\|\mathbf{f}(t,\cdot)\|^2_{\mathbf{V}^{-1}(\Omega)} + \frac{\mu}{4}\left(\|\nabla\boldsymbol{\eta}(t,\cdot)\|^2_{\mathbf{L}^2(\Omega)} + \|\boldsymbol{\eta}(t,\cdot)\|^2_{\mathbf{L}^2(\Omega)}\right)\\ &\qquad + \|\nabla\mathbf{u}(t,\cdot)\|_{\mathbf{L}^2(\Omega)}\,\|\boldsymbol{\eta}(t,\cdot)\|^2_{\mathbf{L}^4(\Omega)}\,.\end{aligned}$$

Applying to the last term the Ladyzhenskaya inequality (see [19, Lemma 1])

$$\|\eta\|^2_{L^4(\Omega)} \le \sqrt{2}\,\|\eta\|_{L^2(\Omega)}\,\|\nabla\eta\|_{L^2(\Omega)} \quad \forall\,\eta\in H^1(\Omega)$$

and then integrating with respect to t, we obtain

$$\begin{aligned}&\|\boldsymbol{\eta}(t,\cdot)\|^2_{\mathbf{L}^2(\Omega)} + \mu\int_0^t\|\nabla\boldsymbol{\eta}(t,\cdot)\|^2_{\mathbf{L}^2(\Omega)}\,dt \le \|\boldsymbol{\eta}_0\|^2_{\mathbf{L}^2(\Omega)}\\ &\quad + \frac{2}{\mu}\int_0^t\|\mathbf{f}(\tau,\cdot)\|^2_{\mathbf{V}^{-1}(\Omega)}\,d\tau + \int_0^t\left(\frac{\mu}{2} + \frac{4}{\mu}\|\nabla\mathbf{u}(\tau,\cdot)\|^2_{\mathbf{L}^2(\Omega)}\right)\|\boldsymbol{\eta}(\tau,\cdot)\|^2_{\mathbf{L}^2(\Omega)}\,d\tau\,.\end{aligned}$$

Then, the Gronwall inequality yields the estimate

$$\begin{aligned}&\|\boldsymbol{\eta}\|^2_{L^\infty(0,T;\mathbf{L}^2(\Omega))} + \|\boldsymbol{\eta}\|^2_{L^2(0,T;\mathbf{H}^1(\Omega))}\\ &\quad\le A_1\left(\|\mathbf{f}\|_{L^2(0,T;\mathbf{V}^{-1}(\Omega))}\,,\ \|\mathbf{u}\|_{\mathbf{L}^2(0,T;\mathbf{H}^1(\Omega))}\,,\ \|\boldsymbol{\eta}_0\|_{\mathbf{L}^2(\Omega)}\right)\,,\end{aligned}$$

where $A_1(\cdot,\cdot,\cdot)$ is a continuous positive function defined on $\mathbb{R}\times\mathbb{R}\times\mathbb{R}$. Evidently, this last estimate implies

$$\begin{aligned}&\|\boldsymbol{\eta}\|^2_{L^\infty(0,T;\mathbf{V}^0(\Omega))} + \|\boldsymbol{\eta}\|^2_{L^2(0,T;\mathbf{V}^1(\Omega))}\\ &\quad\le A_1\left(\|\mathbf{f}\|_{L^2(0,T;\mathbf{V}^{-1}(\Omega))}\,,\ \|\mathbf{u}\|_{\mathcal{V}^{(1)}(Q_T)}\,,\ \|\boldsymbol{\eta}_0\|_{\mathbf{V}^0(\Omega)}\right)\,.\end{aligned} \tag{4.23}$$

Now, taking the supremum of (4.19) with respect to $\mathbf{z} \in \mathbf{V}_0^1(\Omega)$ with $\|\mathbf{z}\|_{\mathbf{V}_0^1(\Omega)} = 1$ and again applying the Ladyzhenskaya inequality we obtain

$$\begin{aligned}\|\partial_t \boldsymbol{\eta}(t,\cdot)\|_{\mathbf{V}^{-1}(\Omega)} \le{}& \mu\|\boldsymbol{\eta}(t,\cdot)\|_{\mathbf{H}^1(\Omega)} + C\|\boldsymbol{\eta}(t,\cdot)\|_{\mathbf{H}^1(\Omega)}\,\|\boldsymbol{\eta}(t,\cdot)\|_{\mathbf{L}^2(\Omega)}\\ &+ C\|\mathbf{u}(t,\cdot)\|_{\mathbf{H}^1(\Omega)}^{1/2}\,\|\mathbf{u}(t,\cdot)\|_{\mathbf{L}^2(\Omega)}^{1/2}\,\|\boldsymbol{\eta}(t,\cdot)\|_{\mathbf{H}^1(\Omega)}^{1/2}\,\|\boldsymbol{\eta}(t,\cdot)\|_{\mathbf{L}^2(\Omega)}^{1/2}\\ &+ C|\mathbf{v}_\infty|\|\boldsymbol{\eta}(t,\cdot)\|_{\mathbf{L}^2(\Omega)} + \|\mathbf{f}(t,\cdot)\|_{\mathbf{V}^{-1}(\Omega)}\end{aligned}$$

so that

$$\begin{aligned}&\|\partial_t \boldsymbol{\eta}(t,\cdot)\|^2_{L^2(0,T;\mathbf{V}^{-1}(\Omega))}\\ &\le C\|\boldsymbol{\eta}(t,\cdot)\|^2_{L^2(0,T;\mathbf{H}^1(\Omega))} + C\|\boldsymbol{\eta}(t,\cdot)\|^2_{L^\infty(0,T;\mathbf{L}^2(\Omega))}\,\|\boldsymbol{\eta}(t,\cdot)\|^2_{L^2(0,T;\mathbf{H}^1(\Omega))}\\ &\quad + C\,\|\mathbf{u}(t,\cdot)\|_{L^\infty(0,T;\mathbf{L}^2(\Omega))}\,\|\boldsymbol{\eta}(t,\cdot)\|_{L^\infty(0,T;\mathbf{L}^2(\Omega))}\\ &\qquad\quad \cdot \|\mathbf{u}(t,\cdot)\|_{L^2(0,T;\mathbf{H}^1(\Omega))}\,\|\boldsymbol{\eta}(t,\cdot)\|_{L^2(0,T;\mathbf{H}^1(\Omega))}\\ &\quad + C|\mathbf{v}_\infty|^2\,\|\boldsymbol{\eta}(t,\cdot)\|^2_{L^2(0,T;\mathbf{L}^2(\Omega))} + C\|\mathbf{f}(t,\cdot)\|^2_{L^2(0,T;\mathbf{V}^{-1}(\Omega))}\,.\end{aligned}$$

Hence, using (4.23) and Lemma 3.1, we obtain the desired estimate. ∎

Lemma 4.3 and Theorem 4.2 lead to the following result.

THEOREM 4.4. *Let* $\mathbf{b}$ *and* $\mathbf{w}_0$ *satisfy* (4.12)–(4.14) *and* (4.17)–(4.18). *Then, there exists a unique solution* $\mathbf{w} \in \mathcal{V}^{(1)}(Q_T)$ *for the problem* (3.1)–(3.3). *Moreover, the solution satisfies the estimate*

$$\begin{aligned}&\|\mathbf{w}\|^2_{\mathcal{V}^{(1)}(Q_T)}\\ &\qquad \le B\Big(\|\mathbf{w}_0\|_{\mathbf{L}^2(\Omega)}\,,\ \|b_n\|_{L^2(0,T;H^{1/2}(\partial\Omega))} + \|b_n\|_{H^{3/4}(0,T;H^{-1}(\partial\Omega))}\,,\\ &\qquad\qquad \|b_\tau\|_{L^2(0,T;H^{1/2}(\partial\Omega))} + \|b_\tau\|_{H^{1/4}(0,T;L^2(\partial\Omega))}\,,\ |\mathbf{v}_\infty|\Big)\,,\end{aligned} \tag{4.24}$$

where $B(\cdot,\cdot,\cdot,\cdot)$ *is a continuous positive function defined on* $\mathbb{R}\times\mathbb{R}\times\mathbb{R}\times\mathbb{R}$.

Proof. Let $\mathbf{u} \in \mathcal{V}^{(1)}(Q_T)$ be the extension of the data $\mathbf{b}$ into Q_T constructed in Theorem 4.2 and let $\boldsymbol{\eta}$ be the solution of (4.19)–(4.21) with $\mathbf{f}$ defined by (4.22). The existence and uniqueness of such an $\boldsymbol{\eta}$ is guaranteed by Lemma 4.3. Set $\mathbf{w} = \mathbf{u} + \boldsymbol{\eta}$; then $\mathbf{w}$ is clearly the unique solution of (3.1)–(3.3). Thus, it only remains to prove the estimate (4.24).

From (4.22) and the fact (see Theorem 4.2) that $\mathbf{u}$ has bounded support, we have that

$$\begin{aligned}\|\mathbf{f}(t,\cdot)\|_{\mathbf{V}^{-1}(\Omega)} \le{}& \mu\|\mathbf{u}(t,\cdot)\|_{\mathbf{H}^1(\Omega)} + \|\partial_t\mathbf{u}(t,\cdot)\|_{\mathbf{V}^{-1}(\Omega)}\\ &+ \|\mathbf{u}(t,\cdot)\|_{\mathbf{L}^2(\Omega)}\|\mathbf{u}(t,\cdot)\|_{\mathbf{H}^1(\Omega)} + |\mathbf{v}_\infty|\,\|\mathbf{u}(t,\cdot)\|_{\mathbf{H}^1(\Omega)}\,.\end{aligned} \tag{4.25}$$

Also, from (4.21), we have that

$$\|\boldsymbol{\eta}_0\|_{\mathbf{L}^2(\Omega)} \le \|\mathbf{w}_0\|_{\mathbf{L}^2(\Omega)} + \|\mathbf{u}(0,\cdot)\|_{\mathbf{L}^2(\Omega)}\,. \tag{4.26}$$

Hence, (4.24) follows from Theorem 4.2, Lemmas 3.1 and 4.3, (4.25), and (4.26). ∎

REMARK. We stress that the normal and tangential components of the boundary condition for the Navier–Stokes equations have different smoothness. This is a feature that is not exhibited in boundary value problems for general second-order parabolic systems. ∎

4.3. The stress vector (on the boundary) of admissible solutions. We now show that the stress vector on the boundary $\big(-pI+\mu(\nabla\mathbf{w}+(\nabla\mathbf{w})^T)\big)\cdot\mathbf{n}|_{(0,T)\times\partial\Omega}$, where $\mathbf{w}\in Y$ is an admissible solution in the sense of Definition 3.3 and p is an associated pressure field, is well defined in a certain function space. This result will be needed in section 6.4 in order to derive the optimality system in the form of a boundary value problem for a system of partial differential equations. Note that the requirement $\mathbf{w}\in\mathcal{V}_{ad}$ is stronger than $\mathbf{w}\in\mathcal{V}^{(1)}(Q_T)$ merely being a solution of (3.1) and (3.3). (We will actually show that each of $(p\mathbf{n})|_{(0,T)\times\partial\Omega}$, $(\nabla\mathbf{w}\cdot\mathbf{n})|_{(0,T)\times\partial\Omega}$, and $((\nabla\mathbf{w})^T\cdot\mathbf{n})|_{(0,T)\times\partial\Omega}$ is well defined.)

Let $\mathbf{w}\in Y$ be an admissible element; then $\mathbf{w}$ satisfies (3.1) and (3.3). From the definition of Y we see that $\mathbf{w}|_{(0,T)\times\partial\Omega}$ is well defined and

$$\mathbf{w}|_{(0,T)\times\partial\Omega}\in H^1\big(0,T;\mathbf{L}^2(\partial\Omega)\big)\cap L^2\big(0,T;\mathbf{H}^{1/2}(\partial\Omega)\big)\cap\mathbf{L}^k\big((0,T)\times\partial\Omega\big)\,.$$

By de Rham's lemma (see [14] and [28]), there exists a $p\in L^2\big(0,T;L^2_{\rm loc}(\Omega)\big)$ such that $\nabla p\in L^2\big(0,T;\mathbf{H}^{-1}(\Omega)\big)$ and

$$\partial_t\mathbf{w}-\mu\Delta\mathbf{w}+[(\mathbf{w}+\mathbf{v}_\infty)\cdot\nabla]\mathbf{w}+\nabla p=\mathbf{0} \tag{4.27}$$

in the sense of distributions on Q_T. To study the normal stress on $\partial\Omega$, the behavior of $\mathbf{w}$ and p at infinity is irrelevant and we can restrict our attention to a bounded domain whose boundary contains $\partial\Omega$. To this end, we let $\Theta\subset\Omega$ be a bounded domain with C^∞ boundary $\partial\Theta$ such that $\partial\Theta\cap\partial\Omega=\partial\Omega$. We denote by γ the restriction operator on $\partial\Theta$. Let F be a streamfunction of $\mathbf{w}$ which can constructed as in section 4.1, satisfying on Θ

$$w_1=\partial_2 F\qquad\text{and}\qquad w_2=-\partial_1 F\,. \tag{4.28}$$

Since $\mathbf{w}\in Y(Q_\Theta)\subset\mathcal{V}^{(1)}(Q_\Theta)$, where $Q_\Theta=(0,T)\times\Theta$, we have $F\in\mathcal{H}^{(2)}(Q_\Theta)$. The restriction of (4.27) and the divergence-free condition for $\mathbf{w}$ on Q_Θ yields

$$\partial_t\mathbf{w}-\mu\Delta\mathbf{w}+[(\mathbf{w}+\mathbf{v}_\infty)\cdot\nabla]\mathbf{w}+\nabla p=\mathbf{0} \tag{4.29}$$

and

$$\operatorname{div}\mathbf{w}=0\,, \tag{4.30}$$

where the derivatives are understood in the sense of distributions in Q_Θ. Applying the curl operator to (4.29) and taking into account (4.28) we obtain

$$\partial_t\Delta F-\mu\Delta^2F=G\,, \tag{4.31}$$

where

$$G=-\operatorname{curl}\{[(\mathbf{w}+\mathbf{v}_\infty)\cdot\nabla]\mathbf{w}\}=-(w_1+v_{\infty,1})\Delta w_2+(w_2+v_{\infty,2})\Delta w_1\,. \tag{4.32}$$

LEMMA 4.5. *Assume $\mathbf{w}\in Y$ is a solution of (3.1) and G is defined by (4.32). Then, $G\in L^1\big(0,T;W^{-1,\alpha}(\Theta)\big)$ for every $\alpha\in(1,2)$.*

Proof. Let α' and β be defined by

$$\frac{1}{\alpha'}+\frac{1}{\alpha}=1\qquad\text{and}\qquad\frac{1}{\beta}+\frac{1}{\alpha'}=\frac{1}{2}\,.$$

Let $\phi \in W_0^{1,\alpha'}(\Theta)$ be given. By integration by parts and Hölder's inequality, we have that a.e. $t \in (0,T)$,

$$\begin{aligned}
&\Big| \int_\Theta (w_1 + v_{\infty,1}) \Delta w_2 \phi \, d\mathbf{x} \Big| \\
&\qquad = \Big| \int_\Theta \Big(\phi \nabla w_1 \cdot \nabla w_2 + (w_1 + v_{\infty,1}) \nabla w_2 \cdot \nabla \phi \Big) \, d\mathbf{x} \Big| \\
&\qquad\qquad \le \|\nabla w_1\|_{\mathbf{L}^2(\Theta)} \, \|\nabla w_2\|_{\mathbf{L}^2(\Theta)} \, \|\phi\|_{L^\infty(\Theta)} \\
&\qquad\qquad\quad + \|w_1 + v_{\infty,1}\|_{L^\beta(\Theta)} \, \|\nabla w_2\|_{\mathbf{L}^2(\Theta)} \, \|\nabla \phi\|_{\mathbf{L}^{\alpha'}(\Theta)} \, .
\end{aligned} \tag{4.33}$$

Since $\alpha' \in (2,\infty)$ and $\beta \in (2,\infty)$, Sobolev imbedding theorems imply

$$\|\phi\|_{L^\infty(\Theta)} \le C \, \|\phi\|_{W^{1,\alpha'}(\Theta)} \qquad \text{and} \qquad \|w_1 + v_{\infty,1}\|_{L^\beta(\Theta)} \le C \, \|w_1 + v_{\infty,1}\|_{H^1(\Theta)}$$

so that, from (4.33),

$$\|(w_1 + v_{\infty,1}) \Delta w_2\|_{L^1(0,T;W^{-1,\alpha}(\Theta))} \le C\big(\|\mathbf{w}\|^2_{\mathcal{V}^{(1)}(Q_\Theta)} + |\mathbf{v}_\infty|^2 \big) \, .$$

Similarly, we can show

$$\|(w_2 + v_{\infty,2}) \Delta w_1\|_{L^1(0,T;W^{-1,\alpha}(\Theta))} \le C\big(\|\mathbf{w}\|^2_{\mathcal{V}^{(1)}(Q_\Theta)} + |\mathbf{v}_\infty|^2 \big) \, .$$

It follows from the last two inequalities and (4.32) that $G \in L^1\big(0,T;W^{-1,\alpha}(\Theta)\big)$. □

Since $F \in \mathcal{H}^{(2)}(Q_\Theta)$, we have $\Delta F \in L^2\big(0,T;L^2(\Theta)\big)$. From (4.31), (4.32), and Lemma 4.5, we see that

$$\Delta(\partial_t F - \mu \, \Delta F) \in L^1\big(0,T;W^{-1,\alpha}(\Theta)\big) \, .$$

We now introduce the space

$$X_\alpha = \big\{ f \in L^2(\Theta) : \Delta f \in W^{-1,\alpha}(\Theta) \big\}$$

equipped with the norm

$$\|f\|_{X_\alpha} = \|f\|_{L^2(\Theta)} + \|\Delta f\|_{W^{-1,\alpha}(\Theta)} \quad \forall \, f \in X_\alpha \, .$$

It is easy to verify that X_α is a Banach space. We will establish a trace theorem for X_α. To this end, we first prove two lemmas.

LEMMA 4.6. *Every bounded linear functional L on X_α has the representation*

$$Lf = (f,\phi) + \langle \Delta f, \psi \rangle \quad \forall \, f \in X_\alpha \, , \tag{4.34}$$

where $\phi \in L^2(\Theta)$, $\psi \in W_0^{1,\alpha'}(\Theta)$, $(\cdot,\cdot)$ denotes the $L^2(\Theta)$-inner product, and $\langle \cdot,\cdot \rangle$ denotes the duality pairing between $W^{-1,\alpha}(\Theta)$ and $W_0^{1,\alpha'}(\Theta)$.

Proof. In $L^2(\Theta) \times W^{-1,\alpha}(\Theta)$, we consider the subspace $\Pi = \big\{ (f, \Delta f) : f \in X_\alpha \big\}$. Clearly, Π is closed under the Cartesian norm for $L^2(\Theta) \times W^{-1,\alpha}(\Theta)$ and the mapping $\pi : f \mapsto (f, \Delta f)$ establishes an isomorphism between X_α and Π. Let an arbitrary bounded linear functional L on X_α be given. Then, there exists a unique functional K on Π such that $Lf = K(f, \Delta f)$. Using the Hahn–Banach theorem we can extend the functional K defined on Π into a functional $\widetilde{K}$ defined on the entire space $L^2(\Theta) \times W^{-1,\alpha}(\Theta)$ with the functional norm preserved, i.e., with $\|\widetilde{K}\| = \|K\|$.

Since $L^2(\Theta) \times W^{-1,\alpha}(\Theta)$ is reflexive, there exist $\phi \in L^2(\Theta)$ and $\psi \in W_0^{1,\alpha'}(\Theta)$ such that

$$\widetilde{K}(f,g) = (f,\phi) + \langle g,\psi\rangle \quad \forall\, (f,g) \in L^2(\Theta) \times W^{-1,\alpha}(\Theta)\,,$$

so that on the subspace Π,

$$K(f,\Delta f) = (f,\phi) + \langle \Delta f,\psi\rangle \quad \forall\, (f,\Delta f) \in \Pi\,.$$

The last relation is equivalent to (4.34). ∎

LEMMA 4.7. *$C^\infty(\overline{\Theta})$ is dense in X_α.*

Proof. We need only show that if a bounded linear functional L on X_α satisfies $Lf = 0$ for all $f \in C^\infty(\overline{\Theta})$, then $L = 0$. We assume that L is a bounded linear functional on X_α satisfying $Lf = 0$ for all $f \in C^\infty(\overline{\Theta})$. By Lemma 4.6, there exist $\phi \in L^2(\Theta)$ and $\psi \in W_0^{1,\alpha'}(\Theta)$ such that

$$Lf = (f,\phi) + \langle \Delta f,\psi\rangle = 0 \quad \forall\, f \in C^\infty(\overline{\Theta})\,.$$

This implies that, in the sense of distributions,

$$\Delta\psi = -\phi\,.$$

As $\psi \in W_0^{1,\alpha'}(\Theta)$ and $\phi \in L^2(\Theta)$, we deduce from elliptic regularity that $\psi \in W_0^{1,\alpha'}(\Theta) \cap H^2(\Theta)$ and $\Delta\psi = -\phi$ in $L^2(\Theta)$, which in turn implies $\partial_n\psi \in H^{1/2}(\partial\Omega)$. For each $f \in C^\infty(\overline{\Theta})$, we are justified in using integration by parts to obtain

$$\begin{aligned} 0 = (f,\phi) + \langle \Delta f,\psi\rangle &= (f,\phi) + (\Delta f,\psi) \\ &= (f,\phi) + (f,\Delta\psi) - \langle \partial_n\psi, f\rangle \\ &= (f,\phi) + (f,-\phi) - \langle \partial_n\psi, f\rangle = -\langle \partial_n\psi, f\rangle \end{aligned}$$

so that $\partial_n\psi\big|_{\partial\Omega} = 0$ and $\psi \in H_0^2(\Theta)$. Using the denseness of $C_0^\infty(\Theta)$ in $H_0^2(\Theta)$ we may choose a sequence $\{\psi_n\} \subset C_0^\infty(\Theta)$ such that $\psi_n \to \psi$ in $H^2(\overline{\Theta})$. Then, for each $f \in X_\alpha$ we have

$$\begin{aligned} (f,\phi) + \langle \Delta f,\psi\rangle &= (f,\phi) + \lim_{n\to\infty} \langle \Delta f,\psi_n\rangle \\ &= (f,\phi) + \lim_{n\to\infty} (f,\Delta\psi_n) = (f,\phi) + (f,\Delta\psi) = (f,\phi) + (f,-\phi) = 0; \end{aligned}$$

i.e., we have shown that

$$Lf = 0 \quad \forall\, f \in X_\alpha\,.$$

Hence, $L = 0$. ∎

In the sequel, we will make use of Besov spaces $B^{s,q}(\partial\Theta)$, where s is the smoothness index and q is the integrability index. For the definition of Besov spaces, see [4] and [29], where the Besov spaces $B^{s,q}(\partial\Theta)$ are denoted by $B^s_{q,q}(\partial\Theta)$. One can also consult [1] for the definition of Besov spaces and the relations between Besov spaces and Sobolev spaces. One important feature of Besov spaces is that they coincide with the traces of Sobolev spaces. In particular, we have the following precise result: if we denote by γ the mapping $\gamma f = f|_{\partial\Theta}$ for functions defined in Θ, then the mapping

$$(\gamma, \gamma\partial_n) : W^{2,\alpha'}(\Theta) \to B^{2-1/\alpha',\alpha'}(\partial\Theta) \times B^{1-1/\alpha',\alpha'}(\partial\Theta) \tag{4.35}$$

is continuous and establishes an epimorphism; see [4] and [29].

PROPOSITION 4.8. *Assume that* $1 < \alpha < 2$. *Then, the operator* γ, *defined on* $C^\infty(\overline{\Theta})$ *by* $\gamma f = f|_{\partial\Theta}$, *can be extended continuously into the trace operator*

$$\gamma \in \mathcal{L}\big(X_\alpha; B^{-1/\alpha,\alpha}(\partial\Theta)\big)\,. \tag{4.36}$$

Proof. By (4.35), we can choose a continuous linear operator

$$K : B^{1-1/\alpha',\alpha'}(\partial\Theta) \to W^{2,\alpha'}(\Theta) \tag{4.37}$$

such that

$$\gamma K\phi = 0 \qquad \text{and} \qquad \gamma\partial_n K\phi = \phi \qquad \forall\,\phi \in B^{1-1/\alpha',\alpha'}(\partial\Theta)\,. \tag{4.38}$$

Let $f \in X_\alpha$. We define a linear functional Z on $B^{1-1/\alpha',\alpha'}(\partial\Theta)$ by

$$Z\phi = Z_K(\phi) = (f, \Delta K\phi) - \langle \Delta f, K\phi\rangle \qquad \forall\,\phi \in B^{1-1/\alpha',\alpha'}(\partial\Theta)\,.$$

We claim that Z does not depend on the choice of K. Indeed, let K_1 and K_2 be two continuous linear operators satisfying (4.37)–(4.38). Then, by (4.38),

$$\gamma(K_1 - K_2)\phi = 0 \qquad \text{and} \qquad \gamma\partial_n(K_1 - K_2)\phi = 0 \qquad \forall\,\phi \in B^{1-1/\alpha',\alpha'}(\partial\Theta)$$

so that if $f \in C^\infty(\overline{\Theta})$, then integration by parts yields

$$Z_{K_1}(\phi) - Z_{K_2}(\phi) = (f, \Delta(K_1 - K_2)\phi) - \langle \Delta f, (K_1 - K_2)\phi\rangle = 0\,.$$

By virtue of Lemma 4.7, this equality is true for an arbitrary $f \in X_\alpha$. Hence, we have shown that $Z_{K_1} = Z_{K_2}$, i.e., that the operator Z is well defined. Evidently, Z_K is bounded on $B^{1-1/\alpha',\alpha'}(\partial\Theta)$. Hence, by the Riesz theorem, there exists an element $Rf \in B^{-1/\alpha,\alpha}(\partial\Theta)$ such that

$$(Rf, \phi) = Z(\phi) = (f, \Delta K\phi) - \langle \Delta f, K\phi\rangle \quad \forall\,\phi \in B^{1-1/\alpha',\alpha'}(\partial\Theta)\,,$$

where R is the Riesz map. If $f \in C^\infty(\overline{\Theta})$, then using Green's formula in the last equation we obtain $Rf = \gamma f$. By virtue of Lemma 4.7 and the boundedness of the operator in (4.37), we can extend the operator γ continuously into the mapping of (4.36). □

We introduce the set

$$\Upsilon = \{\mathbf{w} \in Y : \mathbf{w} \text{ satisfies (3.1)}\}$$

equipped with the topology generated by the norm of Y.

THEOREM 4.9. *Let* $\mathbf{w} \in Y$ *be a solution of* (3.1) *and* $F \in \mathcal{H}^{(2)}(Q_\Theta)$ *be defined by* (4.28). *Let* $G \in L^1\big(0,T; W^{-1,\alpha}(\Theta)\big)$, $\alpha \in (1,2)$, *be defined by* (4.32). *Then,* $\gamma\big(\nabla\mathbf{w}\cdot\mathbf{n}\big) \in L^1\big(0,T;\mathbf{B}^{-1/\alpha,\alpha}(\partial\Theta)\big)$ *and* $\gamma\big((\nabla\mathbf{w})^T\cdot\mathbf{n}\big) \in L^1\big(0,T;\mathbf{B}^{-1/\alpha,\alpha}(\partial\Theta)\big)$. *Moreover, the mappings* $\mathbf{w} \mapsto \gamma(\nabla\mathbf{w}\cdot\mathbf{n})$ *and* $\mathbf{w} \mapsto \gamma\big((\nabla\mathbf{w})^T\cdot\mathbf{n}\big)$ *are continuous from the topological space* Υ *to* $L^1\big(0,T;\mathbf{B}^{-1/\alpha,\alpha}(\partial\Theta)\big)$.

Proof. From the assumptions on F and G, we easily deduce that $\partial_t F - \mu\,\Delta F \in L^2\big(0,T;L^2(\Theta)\big)$ and $\Delta(\partial_t F - \mu\,\Delta F) \in L^1\big(0,T;W^{-1,\alpha}(\Theta)\big)$. Hence, Proposition 4.8 implies that for almost every $t \in (0,T)$, the restriction $\gamma\big(\partial_t F(t,\cdot) - \mu\,\Delta F(t,\cdot)\big)$ is well defined and

$$\gamma(\partial_t F - \mu\,\Delta F) \in L^1\big(0,T; B^{-1/\alpha,\alpha}(\partial\Theta)\big)\,, \tag{4.39}$$

where $1 < \alpha < 2$. Since $F \in \mathcal{H}^{(2)}(Q_\Theta)$, we have that $\nabla\partial_t F \in H^1\big(\Theta; H^{-1}(0,T)\big)$. Therefore, the restriction $F \mapsto \gamma(\partial_t \nabla F)$ on $(0,T)\times\partial\Theta$ is well defined on the space $H^{1/2}\big(\partial\Theta; H^{-1}(0,T)\big)$. Moreover, the fact that $\mathbf{w} = (\partial_2 F, -\partial_1 F) \in Y$ implies $\gamma\partial_t\nabla F \in \mathbf{L}^2\big((0,T)\times\partial\Theta\big)$. Hence

$$\gamma\partial_t\partial_n F \in L^2\big((0,T)\times\partial\Theta\big) \quad \text{and} \quad \gamma\partial_t\partial_\tau F \in L^2\big((0,T)\times\partial\Theta\big)\,. \tag{4.40}$$

Relation (4.40) implies that

$$\gamma\partial_t F \in L^2\big(0,T; H^1(\partial\Theta)\big)\,. \tag{4.41}$$

By (4.39) and (4.41), we have that

$$\gamma\Delta F \in L^1\big(0,T; B^{-1/\alpha,\alpha}(\partial\Theta)\big) \tag{4.42}$$

for $1 < \alpha < 2$. Since

$$F \in \mathcal{H}^{(2)}(Q_\Theta) \subset L^2\big(0,T; H^2(\Theta)\big)\,, \tag{4.43}$$

we see that

$$\gamma F \in L^2\big(0,T; H^{3/2}(\partial\Theta)\big)\,, \quad \gamma\partial_\tau F \in L^2\big(0,T; H^{1/2}(\partial\Theta)\big) \tag{4.44}$$

and

$$\gamma\partial_n F \in L^2\big(0,T; H^{1/2}(\partial\Theta)\big)\,. \tag{4.45}$$

We claim that

$$\gamma\partial_{n\tau} F \in L^2\big(0,T; H^{-1/2}(\partial\Theta)\big) \quad \text{and} \quad \gamma\partial_{\tau\tau} F \in L^2\big(0,T; H^{-1/2}(\partial\Theta)\big)\,. \tag{4.46}$$

To prove this claim, we proceed as follows. We multiply F by a cut-off function with support in a neighborhood of $\partial\Theta$. We assume without loss of generality that Θ coincides with the half-plane $\mathbb{R}^2_+ = \{(x_1,x_2) : x_2 \ge 0\}$ and $\partial\Theta$ coincides with $\{(x_1,x_2) : x_2 = 0\}$. We set $F_1 = \partial_\tau F$ and $F_2 = \partial_{\tau\tau} F$. From (4.43), we easily deduce that

$$F_2 \in L^2\big((0,T)\times\mathbb{R}^2_+\big)\,, \quad \partial_{\tau\tau}F_2 \in L^2\big((0,T)\times\mathbb{R}_+; H^{-2}(\mathbb{R})\big) \tag{4.47}$$

and

$$\Delta F_2 = \partial_{\tau\tau}(\partial_{nn} + \partial_{\tau\tau})F \in L^2\big((0,T)\times\mathbb{R}_+; H^{-2}(\mathbb{R})\big)\,. \tag{4.48}$$

These in turn imply

$$\partial_{nn}F_2 = \Delta F_2 - \partial_{\tau\tau}F_2 \in L^2\big((0,T)\times\mathbb{R}_+; H^{-2}(\mathbb{R})\big)\,. \tag{4.49}$$

Relations (4.47)–(4.49) and the trace theorem [21, Chapter 5, section 3] yield the second relation in our claim (4.46). We can similarly prove the first relation in (4.46).

By denoting the unit normal vector by $\mathbf{n} = (n_1, n_2)$ and the unit tangential vector by $\boldsymbol{\tau} = (n_2, -n_1)$, we obtain that

$$\partial_1 F = n_1\partial_n F + n_2\partial_\tau F, \qquad \partial_2 F = n_2\partial_n F - n_1\partial_\tau F, \tag{4.50}$$

and

$$\Delta F = \partial_{nn}F + \partial_{\tau\tau}F + \rho_1\partial_n F + \rho_2\partial_\tau F\,, \tag{4.51}$$

where ρ_1 and ρ_2 are smooth functions. Expressing $\partial_{nn}F$ by ΔF, $\partial_{\tau\tau}F$, $\partial_n F$, and $\partial_\tau F$ in (4.51) and taking into account (4.42), (4.44)–(4.46), and the imbedding

$$L^2\big(0,T;H^{-1/2}(\partial\Theta)\big) \subset L^1\big(0,T;B^{-1/\alpha,\alpha}(\partial\Theta)\big)\,, \qquad 1<\alpha<2\,,$$

we deduce that $\partial_{nn}F$ possesses a trace on $\partial\Theta$ and that the trace satisfies

$$\gamma\partial_{nn}F \in L^1\big(0,T;B^{-1/\alpha,\alpha}(\partial\Theta)\big)\,. \tag{4.52}$$

Equations (4.28) and (4.50) yield

$$\nabla w_1\cdot\mathbf{n} = n_2\partial_{nn}F - n_1\partial_{n\tau}F + \beta_1\partial_n F + \beta_2\partial_\tau F$$

and

$$\nabla w_2\cdot\mathbf{n} = -n_1\partial_{nn}F - n_2\partial_{n\tau}F + \delta_1\partial_n F + \delta_2\partial_\tau F\,,$$

where β_1, β_2, δ_1, and δ_2 are smooth functions. These two relations give us the expression for $\nabla\mathbf{w}\cdot\mathbf{n}$ in terms of $\partial_{nn}F$, $\partial_{n\tau}F$, $\partial_n F$, and $\partial_\tau F$. Similarly, we obtain the expression for $(\nabla\mathbf{w})^T\cdot\mathbf{n}$ in terms of $\partial_{nn}F$, $\partial_{n\tau}F$, $\partial_n F$, and $\partial_\tau F$:

$$(\partial_1\mathbf{w})\cdot\mathbf{n} = -n_2\partial_{\tau\tau}F - n_1\partial_{n\tau}F + b_1\partial_n F + b_2\partial_\tau F$$

and

$$(\partial_2\mathbf{w})\cdot\mathbf{n} = n_1\partial_{\tau\tau}F - n_2\partial_{n\tau}F + d_1\partial_n F + d_2\partial_\tau F\,,$$

where b_1, b_2, d_1, and d_2 are smooth functions. These relations together with (4.44)–(4.46) and (4.52) imply the assertions of the theorem. □

Now we prove a trace result for the pressure field p that satisfies (4.27).

THEOREM 4.10. *Assume* $\mathbf{w}$ *satisfies the hypotheses of* Theorem 4.9 *and let* p *be a scalar field such that* $p\in L^2\big(0,T;L^2_{\rm loc}(\Omega)\big)$, $\nabla p\in L^2\big(0,T;\mathbf{H}^{-1}(\Omega)\big)$, *and* (4.27) *holds. Then* $p\in L^1\big(0,T;X_\alpha(\Theta)\big)$ *and the restriction mapping* $\gamma: p\mapsto (p\mathbf{n})\big|_{\partial\Omega}$ *belongs to* $\mathcal{L}\Big(L^1\big(0,T;X_\alpha(\Theta)\big), L^1\big(0,T;\mathbf{B}^{-1/\alpha,\alpha}(\partial\Omega)\big)\Big)$, *where* $1<\alpha<2$.

Proof. Taking the divergence of (4.27) and using the divergence-free condition (4.30) for $\mathbf{w}$, we obtain

$$\Delta p = E, \tag{4.53}$$

where $E=-2[(\partial_1 w_1)^2+(\partial_1 w_2)(\partial_2 w_1)]$. Let α' be the reciprocal conjugate of α, i.e., $(1/\alpha)+(1/\alpha')=1$. Since $\alpha'>2$, the imbedding $W^{1,\alpha'}(\Theta)\hookrightarrow C(\overline{\Theta})$ is continuous so that

$$\begin{aligned}\int_0^T\int_\Theta E(t,\mathbf{x})\phi(\mathbf{x})\,d\mathbf{x}\,dt &\le 2\int_0^T \|\mathbf{w}\|^2_{\mathbf{V}^1(\Theta)}\,dt\,\|\phi\|_{C(\overline{\Theta})}\\ &\le C\,\|\mathbf{w}\|^2_{L^2(0,T;\mathbf{V}^1(\Theta))}\,\|\phi\|_{W^{1,\alpha'}(\Theta)} \qquad \forall\,\phi\in W^{1,\alpha'}(\Theta)\,.\end{aligned}$$

Hence $\Delta p\in L^1\big(0,T;W^{-1,\alpha}(\Theta)\big)$. Also, $p\in L^2\big(0,T;L^2(\Theta)\big)$. Hence we conclude that $p\in L^1\big(0,T;X_\alpha\big)$ so that the desired result about the trace of p follows from Proposition 4.8 and the fact that $\partial\Omega$ is of class C^∞. □

Combining Theorems 4.9 and 4.10, we obtain the following result for the stress vector on the boundary corresponding to admissible solutions.

COROLLARY 4.11. *Assume that* $\mathbf{w}$ *and* p *satisfy the hypotheses of Theorems* 4.9 *and* 4.10. *Then, the stress vector* $\big(-p\mathbf{n}+\mu(\nabla\mathbf{w}+(\nabla\mathbf{w})^T\big)\cdot\mathbf{n})|_{(0,T)\times\partial\Omega}$ *on the boundary belongs to* $L^1\big(0,T;\mathbf{B}^{-1/\alpha,\alpha}(\partial\Omega)\big)$.

5. The existence of an optimal solution. In this section we prove the existence of an optimal solution for both Problem I and Problem II. We first establish a useful lemma.

LEMMA 5.1. *Let* $R>0$ *be a constant such that* $\partial\Omega\subset\big\{\mathbf{x}:|\mathbf{x}|<R\big\}$ *and define* $\Omega_R=\Omega\cap\big\{\mathbf{x}:|\mathbf{x}|<R\big\}$. *Then there exists a positive constant* C *depending only on* R *such that*

$$\|\mathbf{u}\|_{\mathbf{H}^1(\Omega_R)}\le C\Big(\int_{\Omega_R}|D(\mathbf{u})|^2\,d\mathbf{x}+\int_{\partial\Omega}|\mathbf{u}|^2\,ds\Big)\qquad\forall\,\mathbf{u}\in\mathbf{H}^1(\Omega_R).$$

Proof. Assume the lemma is false; then we may choose a sequence $\{\mathbf{u}_n\}\subset\mathbf{H}^1(\Omega_R)$ such that $\|\mathbf{u}_n\|_{\mathbf{H}^1(\Omega_R)}=1$ and

$$1>n\Big(\int_{\Omega_R}|D(\mathbf{u}_n)|^2\,d\mathbf{x}+\int_{\partial\Omega}|\mathbf{u}_n|^2\,ds\Big)$$

so that

$$\int_{\Omega_R}|D(\mathbf{u}_n)|^2\,d\mathbf{x}\to0\qquad\text{and}\qquad\int_{\partial\Omega}|\mathbf{u}_n|^2\,ds\to0\qquad\text{as } n\to\infty;\tag{5.1}$$

i.e., $D(\mathbf{u}_n)\to\mathbf{O}$ in $\mathbf{L}^2(\Omega_R)$ (where $\mathbf{O}$ is the zero tensor) and $\mathbf{u}_n\to\mathbf{u}$ in $\mathbf{L}^2(\partial\Omega)$. The fact that $\|\mathbf{u}_n\|_{\mathbf{H}^1(\Omega_R)}=1$ implies that there exists a subsequence (still denoted by $\{\mathbf{u}_n\}$) such that as $n\to\infty$,

$$\mathbf{u}_n\rightharpoonup\mathbf{u}\text{ in }\mathbf{H}^1(\Omega_R),\qquad\mathbf{u}_n\to\mathbf{u}\text{ in }\mathbf{L}^2(\Omega_R),\qquad\text{and}\qquad\mathbf{u}_n\to\mathbf{u}\text{ in }\mathbf{L}^2(\partial\Omega)\tag{5.2}$$

for some $\mathbf{u}\in\mathbf{H}^1(\Omega_R)$, which in turn implies $D(\mathbf{u}_n)\rightharpoonup D(\mathbf{u})$ in $\mathbf{L}^2(\Omega_R)$. Hence we have $D(\mathbf{u})=\mathbf{O}$ in $\mathbf{L}^2(\Omega_R)$ and $\mathbf{u}=\mathbf{0}$ in $\mathbf{L}^2(\partial\Omega)$, i.e., $D(\mathbf{u})\equiv\mathbf{O}$ in Ω_R and $\mathbf{u}\equiv\mathbf{0}$ on $\partial\Omega$. Hence $\mathbf{u}$ is a rigid-body motion which can be expressed in the form $\mathbf{u}=\mathbf{a}+\mathbf{b}\times\mathbf{x}$ for all $\mathbf{x}\in\Omega_R$, where $\mathbf{a}$ and $\mathbf{b}$ are constant vectors (see [22] or [23]). $\mathbf{u}$ being a linear function and $\mathbf{u}\equiv\mathbf{0}$ on $\partial\Omega$ easily leads us to $\mathbf{a}=\mathbf{0}$ and $\mathbf{b}=\mathbf{0}$, i.e., $\mathbf{u}\equiv\mathbf{0}$ in Ω_R. On the other hand, we deduce from (5.1) and (5.2) that $D(\mathbf{u}_n)\to D(\mathbf{u})$ in $\mathbf{L}^2(\Omega_R)$ and $\mathbf{u}_n\to\mathbf{u}$ in $\mathbf{L}^2(\Omega_R)$. Then, using Korn's second inequality (see, e.g., [22, p. 31]),

$$\int_{\Omega_R}\Big(|D(\mathbf{z})|^2+|\mathbf{z}|^2\Big)\,d\mathbf{x}\ge C\|\mathbf{z}\|^2_{\mathbf{H}^1(\Omega_R)}\qquad\forall\,\mathbf{z}\in\mathbf{H}^1(\Omega_R),$$

we conclude $\mathbf{u}_n\to\mathbf{u}$ in $\mathbf{H}^1(\Omega_R)$ so that $\|\mathbf{u}\|_{\mathbf{H}^1(\Omega_R)}=\lim_{n\to\infty}\|\mathbf{u}_n\|_{\mathbf{H}^1(\Omega_R)}=1$, i.e., $\mathbf{u}\ne\mathbf{0}$. This gives a contradiction. Hence the lemma is proved. ∎

As a consequence of Lemma 5.1, we obtain the following.

COROLLARY 5.2. *There exists a constant* $C>0$ *depending only on* Ω *such that*

$$\|\mathbf{u}\|_{\mathbf{H}^{1/2}(\partial\Omega)}\le C\Big(\int_{\Omega}|D(\mathbf{u})|^2\,d\mathbf{x}+\int_{\partial\Omega}|\mathbf{u}|^2\,ds\Big)\qquad\forall\,\mathbf{u}\in\mathbf{H}^1(\Omega).$$

Proof. We fix an $R>0$ such that $\partial\Omega\subset\{\mathbf{x}:|\mathbf{x}|<R\}$ (R is determined, albeit not uniquely, by Ω). Then by Lemma 5.1 and the trace theorem for $\mathbf{H}^1(\Omega_R)$, there

exists a constant $C > 0$ (depending only on R and Ω) such that

$$\|\mathbf{u}\|_{\mathbf{H}^{1/2}(\partial\Omega)} \le C\Big(\int_{\Omega_R} |D(\mathbf{u})|^2 \, d\mathbf{x} + \int_{\partial\Omega} |\mathbf{u}|^2 \, ds \Big) \qquad \forall\, \mathbf{u} \in \mathbf{H}^1(\Omega_R).$$

Thus the desired estimate follows from the last inequality and the fact that $\Omega_R \subset \Omega$. □

THEOREM 5.3. *There exists a solution* $\mathbf{w} \in Y$ *for Problem* I*; there exists a solution* $\mathbf{w} \in Y$ *for Problem* II.

Proof. The proofs for Problem I and Problem II are essentially the same, and we will only consider Problem I. Theorem 4.4 guarantees that the admissible set $\mathcal{V}_{ad}$ is nonempty; indeed, we choose a boundary data in $\mathbf{C}^\infty\big([0,T] \times \partial\Omega\big)$, and then by Theorem 4.4, there exists a solution in $\mathcal{V}^{(1)}(Q_T)$ for the Navier–Stokes equations satisfying this chosen smooth boundary data. This solution clearly belongs to $\mathcal{V}_{ad}$; i.e., $\mathcal{V}_{ad}$; is nonempty. It is easy to verify that $\mathcal{J}_N(\cdot)$ is bounded from below in Y. Thus, we may choose a sequence $\{\mathbf{w}_n\} \subset \mathcal{V}_{ad}$ such that

$$\lim_{n\to\infty} \mathcal{J}_N(\mathbf{w}_n) = \inf_{\mathbf{w}\in\mathcal{V}_{ad}} \mathcal{J}_N(\mathbf{w}) \,,$$

$$\begin{aligned} \langle \partial_t \mathbf{w}_n, \mathbf{z} \rangle + \mu \int_\Omega \nabla \mathbf{w}_n : \nabla \mathbf{z} \, d\mathbf{x} + \int_\Omega \big([\mathbf{w}_n + \mathbf{v}_\infty] \cdot \nabla \big) \mathbf{w}_n \cdot \mathbf{z} \, d\mathbf{x} \\ = 0 \qquad \forall\, \mathbf{z} \in \mathbf{V}_0^1(\Omega), \quad \text{a.e. } t \in (0,T) \,, \end{aligned} \tag{5.3}$$

and

$$\mathbf{w}_n|_{t=0} = \mathbf{w}_0 \qquad \text{in } \mathbf{V}^0(\Omega) \,.$$

Using (3.4) and the conditions on k and N given in the definition of Problem I, we obtain that

$$\mu \int_0^T \int_\Omega \mathcal{D}(\mathbf{w}_n) : \mathcal{D}(\mathbf{w}_n) \, d\mathbf{x}dt + \int_0^T \int_{\partial\Omega} \big(|\partial_t \mathbf{w}_n|^2 + |\mathbf{w}_n|^k \big) \, ds \, dt \le C \,. \tag{5.4}$$

The last inequality and Corollary 5.2 imply that

$$\|\mathbf{w}_n\|_{L^k((0,T)\times\partial\Omega)} + \|\mathbf{w}_n\|_{H^1(0,T;\mathbf{L}^2(\partial\Omega))} + \|\mathbf{w}_n\|_{L^2(0,T;\mathbf{H}^{1/2}(\partial\Omega))} \le C \,. \tag{5.5}$$

Thus, the estimate of Theorem 4.4 with (5.5) gives us the bound

$$\|\partial_t \mathbf{w}_n\|_{L^2(0,T;\mathbf{V}^{-1}(\Omega))} + \|\mathbf{w}_n\|_{L^2(0,T;\mathbf{V}^1(\Omega))} \le C,$$

which allows us to choose a weakly convergent subsequence

$$\mathbf{w}_n \rightharpoonup \widehat{\mathbf{w}} \quad \text{in } L^2\big(0,T;\mathbf{V}^1(\Omega)\big) \tag{5.6}$$

and

$$\partial_t \mathbf{w}_n \rightharpoonup \partial_t \widehat{\mathbf{w}} \quad \text{in } L^2\big(0,T;\mathbf{V}^{-1}(\Omega)\big) \tag{5.7}$$

for some $\widehat{\mathbf{w}} \in \mathcal{V}^{(1)}(Q_T)$. For each $R > 0$ we let $B_R = \{\mathbf{x} \in \mathbb{R}^2 : |\mathbf{x}| < R\}$. Since the space

$$\mathcal{V}_R^{(1)}(\Omega \cap B_R) \equiv \big\{ \mathbf{u} \in L^2\big(0,T;\mathbf{V}^1(\Omega \cap B_R)\big) : \partial_t \mathbf{u} \in L^2\big(0,T;\mathbf{V}^{-1}(\Omega \cap B_R)\big) \big\}$$

equipped with the norm $\|\mathbf{u}\|^2_{\mathcal{V}^{(1)}_R(\Omega\cap B_R)} = \|\mathbf{u}\|^2_{L^2(0,T;\mathbf{V}^1(\Omega\cap B_R))}+\|\partial_t\mathbf{u}\|^2_{L^2(0,T;\mathbf{V}^{-1}(\Omega\cap B_R))}$ is compactly imbedded into $\mathbf{L}^2\big((0,T)\times(\Omega\cap B_R)\big)$, we may use (5.6) and (5.7) to conclude that

$$\mathbf{w}_n \to \widehat{\mathbf{w}} \quad \text{in } \mathbf{L}^2\big((0,T)\times(\Omega\cap B_R)\big)\,. \tag{5.8}$$

For each arbitrarily given $\mathbf{z}\in\mathbf{C}^\infty_0(\Omega)\cap\mathbf{V}^1_0(\Omega)$, relations (5.6)–(5.8) allow us to pass to the limit in (5.3) to deduce that

$$\langle\partial_t\widehat{\mathbf{w}}(t),\mathbf{z}\rangle + \mu\int_\Omega \nabla\widehat{\mathbf{w}}(t):\nabla\mathbf{z}\,d\mathbf{x} + \int_\Omega \big([\widehat{\mathbf{w}}(t)+\mathbf{v}_\infty]\cdot\nabla\big)\widehat{\mathbf{w}}(t)\cdot\mathbf{z}\,d\mathbf{x} = 0 \qquad \text{a.e. } t\in(0,T)\,.$$

Then, using the denseness of $\mathbf{C}^\infty_0(\Omega)\cap\mathbf{V}^1_0(\Omega)$ in $\mathbf{V}^1_0(\Omega)$, we obtain

$$\langle\partial_t\widehat{\mathbf{w}}(t),\mathbf{z}\rangle + \mu\int_\Omega \nabla\widehat{\mathbf{w}}(t):\nabla\mathbf{z}\,d\mathbf{x} + \int_\Omega \big([\widehat{\mathbf{w}}(t)+\mathbf{v}_\infty]\cdot\nabla\big)\widehat{\mathbf{w}}(t)\cdot\mathbf{z}\,d\mathbf{x} = 0 \qquad \forall\,\mathbf{z}\in\mathbf{V}^1_0(\Omega),\ \text{ a.e. } t\in(0,T); \tag{5.9}$$

i.e., $\widehat{\mathbf{w}}$ satisfies the weak form of the Navier–Stokes equations.

Relations (5.6) and trace theorems imply

$$\mathbf{w}_n \rightharpoonup \widehat{\mathbf{w}} \quad \text{in } L^2\big(0,T;\mathbf{H}^{1/2}(\partial\Omega)\big)$$

so that

$$\mathbf{w}_n \rightharpoonup \widehat{\mathbf{w}} \quad \text{in } \mathbf{L}^2\big((0,T)\times\partial\Omega\big)\,.$$

The estimate (5.5) implies

$$\mathbf{w}_n \rightharpoonup \mathbf{h} \quad \text{in } \mathbf{L}^k\big((0,T)\times\partial\Omega\big)\cap H^1\big(0,T;\mathbf{L}^2(\partial\Omega)\big)\cap L^2\big(0,T;\mathbf{H}^{1/2}(\partial\Omega)\big)$$

for some $\mathbf{h}\in\mathbf{L}^k\big((0,T)\times\partial\Omega\big)\cap H^1\big(0,T;\mathbf{L}^2(\partial\Omega)\big)\cap L^2\big(0,T;\mathbf{H}^{1/2}(\partial\Omega)\big)$. Hence we deduce that $\mathbf{h}=\widehat{\mathbf{w}}$ on $(0,T)\times\partial\Omega$ so that

$$\mathbf{w}_n \rightharpoonup \widehat{\mathbf{w}} \quad \text{in } \mathbf{L}^k\big((0,T)\times\partial\Omega\big)\cap H^1\big(0,T;\mathbf{L}^2(\partial\Omega)\big)\cap L^2\big(0,T;\mathbf{H}^{1/2}(\partial\Omega)\big)\,. \tag{5.10}$$

Thus, we have shown that $\widehat{\mathbf{w}}\in Y$. The continuous imbedding of $\mathcal{V}^{(1)}(Q_T)$ into $\mathbf{C}\big([0,T];\mathbf{V^0}\big)$ (see Lemma 3.1) yields that for each $\tau\in[0,T]$, the trace operator $\mathbf{w}\mapsto\mathbf{w}|_{t=\tau}$ is bounded from $\mathcal{V}^{(1)}(Q_T)$ into $\mathbf{V}^0(\Omega)$. Hence, using the weak convergence $\mathbf{w}_n\rightharpoonup\widehat{\mathbf{w}}$ in $\mathcal{V}^{(1)}(Q_T)$ and the fact that bounded linear operators preserve weak convergence we obtain

$$\mathbf{w}_0=\mathbf{w}_n|_{t=0}\rightharpoonup\widehat{\mathbf{w}}|_{t=0} \quad \text{in } \mathbf{V}^0(\Omega)$$

and

$$\mathbf{w}_n|_{t=T}\rightharpoonup\widehat{\mathbf{w}}|_{t=T} \quad \text{in } \mathbf{V}^0(\Omega)\,.$$

Now, we pass to the limit in the functional $\mathcal{J}_N$. We first examine the term $\int_0^T\int_{\partial\Omega}|\mathbf{w}|^2\mathbf{w}\cdot\mathbf{n}\,ds\,dt$ in the functional. By the compact imbedding result (see [1])

$$\begin{aligned}&\mathbf{L}^k\big((0,T)\times\partial\Omega\big)\cap H^1\big(0,T;\mathbf{L}^2(\partial\Omega)\big)\cap L^2\big(0,T;\mathbf{H}^{1/2}(\partial\Omega)\big)\\ &\hookrightarrow \mathbf{H}^{1/2}\big((0,T)\times\partial\Omega\big)\hookrightarrow\hookrightarrow\mathbf{L}^3\big((0,T)\times\partial\Omega\big)\,,\end{aligned}$$

we obtain from (5.10) that

$$\mathbf{w}_n|_{\partial\Omega} \to \widehat{\mathbf{w}}|_{\partial\Omega} \quad \text{in } \mathbf{L}^3\big((0,T)\times\partial\Omega\big)$$

so that

$$\lim_{n\to\infty}\int_0^T\int_{\partial\Omega}|\mathbf{w}_n|^2\mathbf{w}_n\cdot\mathbf{n}\,ds\,dt=\int_0^T\int_{\partial\Omega}|\widehat{\mathbf{w}}|^2\hat{\mathbf{w}}\cdot\mathbf{n}\,ds\,dt\,.$$

All the remaining terms in the functional $\mathcal{J}_N$ are sequentially weakly lower semicontinuous; thus, using the weak convergence results obtained earlier, we have that

$$\mathcal{J}_N(\widehat{\mathbf{w}})\le\liminf_{n\to\infty}\mathcal{J}_N(\mathbf{w}_n)\,.$$

Hence, we have shown that $\widehat{\mathbf{w}}\in Y$ is indeed a solution to Problem I. □

REMARK. The proof of Theorem 5.2 for Problem II can proceed first by substituting $w=w_n$ into (3.5) and (3.6) to obtain the estimate (5.4) and then passing to the limit as $n\to\infty$. □

REMARK. Since the optimal solution is sought in the space Y whose boundary values are more regular than the trace of $\mathcal{V}^{(1)}(Q_T)$, we expect the optimal solution to be more regular than merely in $\mathcal{V}^{(1)}(Q_T)$. □

REMARK. The result also holds for many other cost functionals such as the L^2-norm of the vorticity functional used in [2] or the velocity matching functional

$$\mathcal{K}(\mathbf{w})=\frac{1}{2}\int_0^T\int_\Omega|\mathbf{w}-\mathbf{w}_0|^2\,d\mathbf{x}dt+\frac{1}{2}\int_0^T\int_{\partial\Omega}\big(|\partial_t\mathbf{w}|^2+|\mathbf{w}|^2+|\nabla_s\mathbf{w}|^2\big)\,ds\,dt\,,$$

where ∇_s denotes the surface gradient on $\partial\Omega$. Using similar arguments we may, for example, conclude that there exists a $\widehat{\mathbf{w}}\in\mathcal{V}_{ad}$ such that $\mathcal{K}(\widehat{\mathbf{w}})=\inf_{\mathbf{w}\in\mathcal{V}_{ad}}\mathcal{K}(\mathbf{w})$. □

6. The optimality system. Having proved that an optimal solution $\mathbf{w}$ exists, we now use Lagrange multiplier principles to characterize the optimal solution; i.e., we obtain an optimality system of partial differential equations that the optimal solution $\mathbf{w}$ and Lagrange multipliers must satisfy. This optimality system can serve as the basis for computing approximations to optimal solutions numerically.

6.1. Abstract Lagrange multiplier principles. We consider an abstract minimization problem. Let X_1 and X_2 be two Banach spaces. Let $f:X_1\to\mathbb{R}$ and $g_j:X_1\to\mathbb{R}$ be functionals and $F:X_1\to X_2$ be a mapping. We seek a $w\in X_1$ such that

$$f(w)=\inf_{u\in\mathcal{W}_{ad}}f(u)\,, \tag{6.1}$$

where

$$\mathcal{W}_{ad}=\big\{u\in X_1:F(u)=0,\ \text{and}\ \ g_j(u)\le 0\ \text{for}\ j=1,\dots,m\big\}\,.$$

The Lagrange functional for the minimization problem (6.1) is defined by

$$\mathcal{L}(w,\boldsymbol{\lambda},q)=\lambda_0 f(w)+\langle F(w),q\rangle+\sum_{i=1}^m\lambda_j g_j(w)$$

for all $w \in X_1$, $\boldsymbol{\lambda} = (\lambda_0, \lambda_1, \ldots, \lambda_m)^T \in \mathbb{R}^{m+1}$, and $q \in X_2^*$. We quote a standard abstract Lagrange principle in the following particular form (see [3]).

THEOREM 6.1. *Let w be a solution of (6.1). Assume that the mappings f, g_j, and F are continuously differentiable and that the image of the operator $F'(w) : X_1 \to X_2$ is closed. Then there exists a $q \in X_2^*$ and a $\boldsymbol{\lambda} = (\lambda_0, \lambda_1, \ldots, \lambda_m)^T \in \mathbb{R}^{m+1}$ such that the pair $(q, \boldsymbol{\lambda}) \neq (0, \mathbf{0})$,*

$$\langle \mathcal{L}_w(w, \boldsymbol{\lambda}, q), h \rangle = 0 \qquad \forall\, h \in X_1\,, \tag{6.2}$$

$$\lambda_j \geq 0, \quad j = 0, 1, \ldots, m\,, \qquad \text{and} \qquad \lambda_j g_j(w) = 0, \quad j = 1, \ldots, m\,, \tag{6.3}$$

where $\mathcal{L}_w(\cdot, \cdot, \cdot)$ denotes the Fréchet derivative of $\mathcal{L}$ with respect to the first argument. Furthermore, if $F'(w) : X_1 \to X_2$ is an epimorphism and the constraints $g_i(w) \leq 0$ are absent in problem (6.1), then $\lambda_0 \neq 0$ and λ_0 can be taken as 1. ∎

6.2. The weak form of an optimality system. Now we apply the abstract Lagrange principle to Problem I and Problem II to obtain an optimality system of equations for each case. We first examine Problem I. We first derive the adjoint equation, in the weak form, for the optimal control problem.

THEOREM 6.2. *Assume $\mathbf{w} \in \mathcal{V}^{(1)}(Q_T)$ is a solution for Problem* I. *Then there exists a $\mathbf{q} \in \mathcal{V}^{(1)}(Q_T) \cap L^2\big(0, T; \mathbf{V}_0^1(\Omega)\big)$ such that*

$$\begin{aligned}
&2\mu \int_0^T \int_\Omega \mathcal{D}(\mathbf{w}) : \mathcal{D}(\mathbf{h})\, d\mathbf{x}dt + 2\mu \int_0^T \int_\Omega \mathcal{D}(\mathbf{h}) : \mathcal{D}(\mathbf{q})\, d\mathbf{x}dt \\
&\quad + \int_0^T \int_\Omega \big\{(\mathbf{h} \cdot \nabla)\mathbf{w} + (\mathbf{w} \cdot \nabla)\mathbf{h} + (\mathbf{v}_\infty \cdot \nabla)\mathbf{h}\big\} \cdot \mathbf{q}\, d\mathbf{x}dt \\
&\quad + \int_0^T \langle \partial_t \mathbf{h}(t, \cdot)\,,\, \mathbf{q}(t, \cdot) \rangle\, dt \\
&\quad + N\Big(\int_0^T \int_{\partial\Omega} 2\partial_t \mathbf{w} \cdot \partial_t \mathbf{h}\, ds\, dt + k \int_0^T \int_{\partial\Omega} |\mathbf{w} + \mathbf{v}_\infty|^{k-2} (\mathbf{w} + \mathbf{v}_\infty) \cdot \mathbf{h}\, ds\, dt \Big) \\
&\quad + \frac{1}{2} \int_0^T \int_{\partial\Omega} \left\{ (\mathbf{w} + \mathbf{v}_\infty) \cdot \mathbf{n} (\mathbf{w} \cdot \mathbf{h}) + \frac{1}{2} (\mathbf{h} \cdot \mathbf{n}) |\mathbf{w}|^2 \right\} ds\, dt \\
&\quad + \frac{1}{2} \int_\Omega \mathbf{w}(T, \mathbf{x}) \cdot \mathbf{h}(T, \mathbf{x})\, d\mathbf{x} = 0 \qquad \forall\, \mathbf{h} \in Y_0,
\end{aligned} \tag{6.4}$$

where $Y_0 \equiv \{\mathbf{y} \in Y : \mathbf{y}|_{t=0} = \mathbf{0}\}$.

Proof. We use the Lagrange multiplier principle (Theorem 6.1) to prove the desired result. We set $X_1 = Y_0$ and $X_2 = L^2\big(0, T; \mathbf{V}^{-1}(\Omega)\big)$. We define the mappings $f : X_1 \to \mathbb{R}$ and $F : X_1 \to X_2$ as follows:

$$f(\mathbf{y}) = \mathcal{J}_N(\mathbf{w} + \mathbf{y})$$

and

$$F(\mathbf{y}) = \partial_t(\mathbf{w} + \mathbf{y}) - \mu P \Delta(\mathbf{w} + \mathbf{y}) + P\big[\big((\mathbf{w} + \mathbf{y} + \mathbf{v}_\infty) \cdot \nabla\big)(\mathbf{w} + \mathbf{y})\big],$$

where $P : \mathbf{H}^{-1}(\Omega) \to \mathbf{V}^{-1}(\Omega)$ is the projection operator. Constraints $g_i \leq 0$ are absent in Problem I. Then $\mathbf{y} = \mathbf{0}$ is the solution of the corresponding extremal problem and $F'(\mathbf{0}) : X_1 \to X_2$ is defined by

$$\langle F'(\mathbf{0}), \mathbf{y} \rangle = \partial_t \mathbf{y} - \mu P \Delta \mathbf{y} + P\big[(\mathbf{y} \cdot \nabla)\mathbf{w} + \big((\mathbf{w} + \mathbf{v}_\infty) \cdot \nabla\big)\mathbf{y}\big].$$

To show that $F'(\mathbf{0})$ is an epimorphism, we first observe that this operator is continuous. Next we need to show that for each $\mathbf{f} \in L^2\big(0,T;\mathbf{V}^{-1}(\Omega)\big)$ the system

(6.5)
$$\begin{aligned}&\langle \partial_t \mathbf{y}(t), \mathbf{z}\rangle + \mu\int_\Omega \nabla\mathbf{y}(t) : \nabla\mathbf{z}\,d\mathbf{x} + \int_\Omega \big((\mathbf{w}(t)+\mathbf{v}_\infty)\cdot\nabla\big)\mathbf{y}(t)\cdot\mathbf{z}\,d\mathbf{x}\\ &\quad + \int_\Omega \big(\mathbf{y}(t)\cdot\nabla\big)\mathbf{w}(t)\cdot\mathbf{z}\,d\mathbf{x} = \int_\Omega \mathbf{f}(t)\cdot\mathbf{z}\,d\mathbf{x} \qquad \forall\,\mathbf{z}\in\mathbf{V}_0^1(\Omega),\ \text{ a.e. } t\in(0,T)\,,\end{aligned}$$

and

$$\mathbf{y}|_{t=0} = \mathbf{0} \qquad \text{in } \mathbf{V}^0(\Omega) \tag{6.6}$$

has a solution $\mathbf{y}\in Y_0$. We supplement this system with the boundary condition

$$\mathbf{y}|_{(0,T)\times\partial\Omega} = \mathbf{0}\,. \tag{6.7}$$

Using the techniques in the proof of Theorem 4.4 we see that (6.5)–(6.7) indeed has a (unique) solution $\mathbf{y}\in\mathcal{V}^{(1)}(Q_T)$. (The situation now is even simpler, as the system (6.5)–(6.7) is linear.) Clearly, $\mathbf{y}\in Y_0$. Hence, we have verified all the assumptions in Theorem 6.1 and we conclude that there exists a $\mathbf{q}\in X_2^* = L^2\big(0,T;\mathbf{V}_0^1(\Omega)\big)$ such that

$$\big\langle \mathcal{L}_{\mathbf{y}}\big(\mathbf{y},\mathbf{q}\big),\mathbf{h}\big\rangle|_{y=0} = 0 \qquad \forall\,\mathbf{h}\in Y_0\,, \tag{6.8}$$

where the Lagrange functional for Problem I is defined by

$$\begin{aligned}\mathcal{L}\big(\mathbf{y},\mathbf{q}\big) &= \mu\int_0^T\!\!\int_\Omega |\mathcal{D}(\mathbf{w}+\mathbf{y})|^2\,d\mathbf{x}dt + \frac14\int_\Omega |\mathbf{w}(T,\mathbf{x})+\mathbf{y}(T,\mathbf{x})|^2\,d\mathbf{x}\\ &\quad + \frac14\int_0^T\!\!\int_{\partial\Omega} (\mathbf{w}+\mathbf{y}+\mathbf{v}_\infty)\cdot\mathbf{n}|\mathbf{w}+\mathbf{y}|^2\,ds\,dt\\ &\quad + \int_0^T\!\!\int_\Omega \partial_t(\mathbf{w}+\mathbf{y})\cdot\mathbf{q}\,d\mathbf{x}dt + 2\mu\int_0^T\!\!\int_\Omega \mathcal{D}(\mathbf{w}+\mathbf{y}) : \mathcal{D}(\mathbf{q})\,d\mathbf{x}dt\\ &\quad + \int_0^T\!\!\int_\Omega \big\{[(\mathbf{w}+\mathbf{y})\cdot\nabla](\mathbf{w}+\mathbf{y}) + (\mathbf{v}_\infty\cdot\nabla)(\mathbf{w}+\mathbf{y})\big\}\cdot\mathbf{q}\,d\mathbf{x}dt\\ &\quad + N\Big(\int_0^T\!\!\int_{\partial\Omega} |\partial_t\mathbf{w}+\partial_t\mathbf{y}|^2\,ds\,dt + \int_0^T\!\!\int_{\partial\Omega} |\mathbf{w}+\mathbf{y}+\mathbf{v}_\infty|^k\,ds\,dt\Big)\end{aligned} \tag{6.9}$$

for all $\mathbf{y}\in X_1$ and $\mathbf{q}\in X_2^* = L^2\big(0,T;\mathbf{V}_0^1(\Omega)\big)$. (Note that we have chosen $\lambda_0 = 1$ in the definition (6.9); this is justified by Theorem 6.1 and the fact that $F'(\mathbf{0})$ is an epimorphism.) Substituting (6.9) into (6.8) we obtain (6.4). By varying $\mathbf{h}$ in $\mathcal{E} = \big\{\mathbf{v}\in\mathbf{C}_0^\infty\big((0,T)\times\Omega\big) : \operatorname{div}\mathbf{v} = 0\big\}\subset Y_0$, we obtain in the sense of distributions defined on solenoidal vector fields:

$$-\partial_t\mathbf{q} - \mu\Delta\mathbf{q} + \mathbf{q}\cdot(\nabla\mathbf{w})^T - (\mathbf{w}\cdot\nabla)\mathbf{q} - (\mathbf{v}_\infty\cdot\nabla)\mathbf{q} = \mu\Delta\mathbf{w} \qquad \text{in } \mathcal{E}'\,, \tag{6.10}$$

or equivalently,

$$-\partial_t\mathbf{q} = \mu\Delta\mathbf{q} - \mathbf{q}\cdot(\nabla\mathbf{w})^T + (\mathbf{w}\cdot\nabla)\mathbf{q} + (\mathbf{v}_\infty\cdot\nabla)\mathbf{q} + \mu\Delta\mathbf{w} \qquad \text{in } \mathcal{E}'\,.$$

From the fact that $\mathbf{w}\in\mathcal{V}^{(1)}(Q_T)$ and $\mathbf{q}\in L^2\big(0,T;\mathbf{V}_0^1(\Omega)\big)$, we easily deduce $\partial_t\mathbf{q}\in L^2\big(0,T;\mathbf{V}^{-1}(\Omega)\big)$. Hence, we have proved $\mathbf{q}\in\mathcal{V}^{(1)}(Q_T)$. □

6.3. Green's formulae. To interpret the weak optimality system (6.4) as a system of partial differential equations with boundary conditions, we will need some Green's formulae for the optimal solution $\mathbf{w}$, the Lagrange multiplier $\mathbf{q}$, and their associated pressure fields p and r, respectively.

We note that if $\mathbf{q}$ is a solution of (6.4) or (6.10), then by De Rham's lemma (see [14] and [28]), there exists an $\widetilde{r} \in L^2\big(0,T;L^2_{\rm loc}(\Omega)\big)$ such that $\nabla\widetilde{r} \in L^2\big(0,T;\mathbf{H}^{-1}(\Omega)\big)$ and

$$-\partial_t\mathbf{q} - \mu\Delta\mathbf{q} + \mathbf{q}\cdot(\nabla\mathbf{w})^T - (\mathbf{w}\cdot\nabla)\mathbf{q} - (\mathbf{v}_\infty\cdot\nabla)\mathbf{q} + \nabla\widetilde{r} = \mu\Delta\mathbf{w} \tag{6.11}$$

in the sense of distributions. Through the change of variable $r = \widetilde{r} + p$, where p satisfies (4.27), we see that $r \in L^2\big(0,T;L^2_{\rm loc}(\Omega)\big)$, $\nabla r \in L^2\big(0,T;\mathbf{H}^{-1}(\Omega)\big)$, and

$$-\partial_t\mathbf{q} - \mu\Delta\mathbf{q} + \mathbf{q}\cdot(\nabla\mathbf{w})^T - (\mathbf{w}\cdot\nabla)\mathbf{q} - (\mathbf{v}_\infty\cdot\nabla)\mathbf{q} + \nabla r = \mu\Delta\mathbf{w} - \nabla p \tag{6.12}$$

in the sense of distributions. We now prove the trace theorems for $\big((\nabla\mathbf{q}) + (\nabla\mathbf{q})^T\big)\cdot\mathbf{n}$ and r as was done for $\big((\nabla\mathbf{w}) + (\nabla\mathbf{w})^T\big)\cdot\mathbf{n}$ and p in section 4.3; we will also derive some Green's formulae that are useful in interpreting the weak optimality system as a boundary value problem for a system of partial differential equations.

LEMMA 6.3. *Assume* $\mathbf{w}$ *is a solution for Problem* I *and let* $\mathbf{q} \in \mathcal{V}^{(1)}(Q_T) \cap L^2\big(0,T;\mathbf{V}^1_0(\Omega)\big)$ *be a solution of* (6.4). *Let* $r \in L^2\big(0,T;L^2_{\rm loc}(\Omega)\big)$ *satisfy* (6.12) *and* $\nabla r \in L^2\big(0,T;\mathbf{H}^{-1}(\Omega)\big)$. *Then*

$$\gamma\big[(\nabla\mathbf{q})\cdot\mathbf{n}\big] \in L^1\big(0,T;\mathbf{B}^{-1/\alpha,\alpha}(\partial\Omega)\big)\,,\ \gamma\big[(\nabla\mathbf{q})^T\cdot\mathbf{n}\big] \in L^1\big(0,T;\mathbf{B}^{-1/\alpha,\alpha}(\partial\Omega)\big)\,, \tag{6.13}$$

$$\gamma(r\mathbf{n}) \in L^1\big(0,T;\mathbf{B}^{-1/\alpha,\alpha}(\partial\Omega)\big)\,, \tag{6.14}$$

and therefore

$$\gamma\big[\big(-rI + (\nabla\mathbf{q}) + (\nabla\mathbf{q})^T\big)\cdot\mathbf{n} \in L^1\big(0,T;\mathbf{B}^{-1/\alpha,\alpha}(\partial\Omega)\big)\big]\,, \tag{6.15}$$

where $1 < \alpha < 2$.

Proof. As in section 4.3, we introduce the bounded subdomain $\Theta \subset \Omega$ such that $\partial\Omega \subset \partial\Theta$. We introduce on Θ the streamfunction E for $\mathbf{q} = (q_1, q_2)$ such that

$$q_1 = \partial_2 E \qquad \text{and} \qquad q_2 = -\partial_1 E\,.$$

Then, by applying the curl operator to (6.11), we obtain that

$$\Delta(\partial_t E + \mu E + \mu\Delta F) = G_1, \tag{6.16}$$

where F is the streamfunction for $\mathbf{w}$ introduced in (4.28) and

$$G_1 = 2(\partial_1 w_1)(\partial_2 q_1 + \partial_1 q_2) + 2(\partial_2 q_2)(\partial_1 w_2 + \partial_2 w_1) + (w_1 + v_{\infty,1})\Delta q_2 - (w_2 + v_{\infty,2})\Delta q_1\,.$$

Also, the fact that $\mathbf{q}|_{S_T} = \mathbf{0}$ allows us to choose E to satisfy $E|_{S_T} = 0$. Analogous to the proof of Lemma 4.5, we obtain

$$G_1 \in L^1\big(0,T;W^{-1,\alpha}(\Theta)\big), \qquad 1 < \alpha < 2\,. \tag{6.17}$$

Since $\mathbf{q} \in \mathcal{V}^{(1)}(Q_T)$ we have $E \in \mathcal{H}^{(2)}(Q_T)$. Thus, (6.16) and (6.17) yield

$$\partial_t E + \mu\Delta E + \mu\Delta F \in L^1(0,T;X_\alpha), \qquad 1 < \alpha < 2\,.$$

By virtue of Proposition 4.8 we obtain

$$\gamma(\partial_t E + \mu\Delta E + \mu\Delta F) \in L^1\big(0,T;B^{-1/\alpha,\alpha}(\partial\Omega)\big)\,. \tag{6.18}$$

The fact that $E|_{S_T} = 0$ implies $\gamma\partial_t E|_{S_T} = 0$. Recall from (4.42) that $\gamma\Delta F \in L^1\big(0,T;B^{-1/\alpha,\alpha}(\partial\Omega)\big)$ so that from (6.18),

$$\gamma\Delta E \in L^1\big(0,T;B^{-1/\alpha,\alpha}(\partial\Omega)\big), \qquad 1<\alpha<2\,.$$

Repeating the arguments in the proof of Theorem 4.9 we obtain (6.13). To prove the trace result for r, we proceed in the same way as in the proof of Theorem 4.10 for p. Taking the divergence of (6.12) we obtain

$$-\Delta r = G_2 + \Delta p,$$

where

$$G_2 = \partial_1 q_2(\partial_1 w_2 - \partial_2 w_1) + \partial_2 q_1(\partial_2 w_1 - \partial_1 w_2) + q_1\Delta w_1 + q_2\Delta w_2\,.$$

Analogous to the proof of Lemma 4.5 we have

$$G_2 \in L^1\big(0,T;W^{-1,\alpha}(\Theta)\big), \qquad 1<\alpha<2\,.$$

Hence, the last three relations and the fact that $r \in L^2\big(0,T;L^2(\Theta)\big)$ and $\Delta p \in L^1\big(0,T;W^{-1,\alpha}(\Theta)\big)$ yield

$$r \in L^1\big(0,T;X_\alpha\big),$$

so that from Proposition 4.8, $\gamma r \in L^1\big(0,T;B^{-1/\alpha,\alpha}(\partial\Omega)\big)$ for $1<\alpha<2$; i.e., (6.14) holds. Finally, (6.15) follows trivially from (6.13)–(6.14). □

We now establish some Green's formulae.

LEMMA 6.4. *Let $r \in X_\alpha(\Theta)$, $1<\alpha<2$. Then the distribution ∇r can be extended continuously into the functional defined by*

$$\langle \nabla r, \mathbf{h}\rangle = \langle r\mathbf{n}, \mathbf{h}\rangle_{\partial\Omega} - \int_\Theta r \operatorname{div}\mathbf{h}\, d\mathbf{x} \tag{6.19}$$

for every $\mathbf{h} \in \mathbf{C}^\infty(\overline{\Theta})$ which vanishes near $(0,T)\times(\partial\Theta\setminus\partial\Omega)$. Here $\langle\nabla r,\cdot\rangle$ denotes the defined functional and $\langle\cdot,\cdot\rangle_{\partial\Omega}$ denotes the duality pairing between $\mathbf{B}^{1/\alpha,\alpha'}(\partial\Theta)$ and $\mathbf{B}^{-1/\alpha,\alpha}(\partial\Theta)$.

Proof. By Lemma 4.7 we may choose a sequence $\{r_n\} \subset C^\infty(\overline{\Theta})$ such that $r_n \to r$ in X_α. Formula (6.19) holds for $r = r_n$ by the classical Stokes theorem. Since the right side of (6.19) with $r = r_n$ converges as $n\to\infty$ to the same expression with r, formula (6.19) defines the desired functional for $r \in X_\alpha(\Theta)$. □

REMARK. On $C^\infty(\overline{\Theta})$, the definition of the operator ∇ found in Lemma 6.4 coincides with the classical definition. □

LEMMA 6.5. *Let $\mathbf{w} \in Y$ be a solution of* (3.1)–(3.3). *Then there exists a sequence of solutions of* (3.1), *$\{\mathbf{w}^{(k)}\} \subset Y \cap L^\infty\big(0,T;\mathbf{V}^2(\Theta)\big)$, such that $\mathbf{w}^{(k)} \to \mathbf{w}$ in Y.*

Proof. Let $U \subset \Theta$ be a (bounded) neighborhood of $\partial\Omega$ such that the extension $\mathbf{u}$ of the boundary data $\mathbf{b}$ constructed in Theorem 4.2 has support in U and we can choose a coordinate system $(x_1,x_2)^T$ such that $U = \{(x_1,x_2)^T \in \mathbb{R}^2 : 0<x_2<d\}$ and $\partial\Omega = \{(x_1,x_2)^T \in \mathbb{R}^2 : x_2 = 0\}$. We consider the sequence $\{\mathbf{u}^{(k)}\}$ defined by

$$\mathbf{u}^{(k)}(t,\mathbf{x}) \equiv \int_0^T\int_U k^3\phi\big(k(t-s),k(x_1-y_1),k(x_2-y_2)\big)\mathbf{u}(s,y_1,y_2)\,dy_1\,dy_2\,ds\,, \tag{6.20}$$

where $\phi \in C_0^\infty(\mathbb{R}^3)$, $\int_{\mathbb{R}^3} \phi \, dt \, d\mathbf{x} = 1$, and $\operatorname{supp} \phi \in \{(t, x_1, x_2)^T \in \mathbb{R}^3 : -1 < t < 1, -1 < x_1 < 1, -1 < x_2 < 0\}$. Evidently, each $\mathbf{u}^{(k)} \in \mathbf{C}^\infty\big((0,T) \times \Theta\big)$ and

$$\|\mathbf{u} - \mathbf{u}^{(k)}\|_{\mathcal{V}^{(1)}(Q_T)} \to 0 \quad \text{as } k \to \infty \,. \tag{6.21}$$

We set

$$\mathbf{b}^{(k)} = \mathbf{u}^{(k)}|_{(0,T)\times\partial\Omega} \,.$$

We choose a sequence $\{\mathbf{w}_0^{(k)}\} \subset \mathbf{V}^2(\Omega)$ such that

$$\mathbf{w}_0^{(k)} \to \mathbf{w}_0 \quad \text{in } \mathbf{V}^0(\Omega) \tag{6.22}$$

and each $\mathbf{w}_0^{(k)}$ satisfies the compatibility condition

$$(\mathbf{w}_0^{(k)} \cdot \mathbf{n})|_{\partial\Omega} = (\mathbf{b}_0^{(k)} \cdot \mathbf{n})|_{t=0} \,.$$

Moreover, we choose $\mathbf{w}_0^{(k)}$ such that $\mathbf{w}_0^{(k)} - \mathbf{u}^{(k)}|_{t=0} \in \mathbf{V}_0^1(\Omega)$. We then consider (3.1) with the boundary and initial conditions

$$\mathbf{w}|_{(0,T)\times\partial\Omega} = \mathbf{b}^{(k)} \qquad \text{and} \qquad \mathbf{w}|_{t=0} = \mathbf{w}_0^{(k)} \,. \tag{6.23}$$

Let $b_n^{(k)}$ and $b_\tau^{(k)}$ be the normal and tangential components of $\mathbf{b}^{(k)}$, respectively:

$$\mathbf{b}^{(k)} = b_n^{(k)} \mathbf{n} + b_\tau^{(k)} \boldsymbol{\tau} \,.$$

Clearly, $b_n^{(k)}$, $b_\tau^{(k)}$, and $\mathbf{w}_0^{(k)}$ satisfy conditions (4.12)–(4.14) and (4.18) of Theorem 4.4. Hence, by Theorem 4.4, there exists a solution $\mathbf{w}^{(k)} \in Y$ for (3.1) and (6.23). We now show that $\mathbf{w}^{(k)} \to \mathbf{w}$ in Y. We obviously have

$$(\mathbf{w}^{(k)} - \mathbf{w})|_{(0,T)\times\partial\Omega} = \mathbf{b}^{(k)} - \mathbf{b} \to \mathbf{0}$$

in $H^1\big(0,T;\mathbf{L}^2(\partial\Omega)\big) \cap L^2\big(0,T;\mathbf{H}^{1/2}(\partial\Omega)\big) \cap \mathbf{L}^k\big((0,T)\times\partial\Omega\big)$. Thus we only need to show $\mathbf{w}^{(k)} \to \mathbf{w}$ in $\mathcal{V}^{(1)}(Q_T)$. We rewrite $\mathbf{w} - \mathbf{w}^{(k)}$ in the form

$$\mathbf{w} - \mathbf{w}^{(k)} = (\mathbf{u} - \mathbf{u}^{(k)}) + \boldsymbol{\eta}^{(k)}, \tag{6.24}$$

where $\boldsymbol{\eta}^{(k)}$ satisfies (4.20), (4.21) with $\boldsymbol{\eta}_0^{(k)} = (\mathbf{w}_0 - \mathbf{w}_0^{(k)}) - (\mathbf{u} - \mathbf{u}^{(k)})|_{t=0}$, and the following analog of (4.19):

$$\begin{aligned}
&\langle \partial_t \boldsymbol{\eta}^{(k)}(t), \mathbf{z}\rangle + \mu \int_\Omega \nabla \boldsymbol{\eta}^{(k)}(t) : \nabla \mathbf{z} \, d\mathbf{x} + \int_\Omega (\boldsymbol{\eta}^{(k)}(t) \cdot \nabla)\mathbf{w}(t) \cdot \mathbf{z} \, d\mathbf{x} \\
&+ \int_\Omega \big((\mathbf{w}^{(k)}(t) + \mathbf{v}_\infty) \cdot \nabla\big)\boldsymbol{\eta}^{(k)}(t) \cdot \mathbf{z} \, d\mathbf{x} = \langle \mathbf{f}^{(k)}(t), \mathbf{z}\rangle \qquad \forall\, \mathbf{z} \in \mathbf{V}^1(\Omega), \text{ a.e. } t \in (0,T),
\end{aligned}$$

where $\mathbf{f}^{(k)}$ is defined by the following analog of (4.22):

$$\begin{aligned}
\langle \mathbf{f}^{(k)}(t), \mathbf{z}\rangle = &-\mu \int_\Omega \nabla(\mathbf{u}(t) - \mathbf{u}^{(k)}(t)) : \nabla \mathbf{z} \, d\mathbf{x} - \langle \partial_t(\mathbf{u}(t) - \mathbf{u}^{(k)}(t)), \mathbf{z}\rangle \\
&- \int_\Omega \big((\mathbf{u}(t) - \mathbf{u}^{(k)}(t)) \cdot \nabla\big)\mathbf{w}(t) \cdot \mathbf{z} \, d\mathbf{x} \\
&- \int_\Omega \big((\mathbf{w}^{(k)}(t) + \mathbf{v}_\infty) \cdot \nabla\big)(\mathbf{u}(t) - \mathbf{u}^{(k)}(t)) \cdot \mathbf{z} \, d\mathbf{x}.
\end{aligned}$$

We can estimate $\boldsymbol{\eta}^{(k)}$ and $\mathbf{f}^{(k)}$ as in the proof of Lemma 4.3 and (4.25) to obtain

$$\begin{aligned}&\|\boldsymbol{\eta}^{(k)}\|_{\mathcal{V}^{(1)}(Q_T)}\\&\le A\big(\|\mathbf{f}^{(k)}\|_{L^2(0,T;\mathbf{V}^{-1}(\Omega))}, \|\mathbf{u}-\mathbf{u}^{(k)}\|_{\mathcal{V}^{(1)}(Q_T)}, \|\boldsymbol{\eta}_0^{(k)}\|_{\mathbf{V}^0(\Omega)}, \|\mathbf{w}\|_{\mathcal{V}^{(1)}(Q_T)}, |\mathbf{v}_\infty|\big)\end{aligned}$$

and

$$\begin{aligned}\|\mathbf{f}^{(k)}\|_{L^2(0,T;\mathbf{V}^{-1}(\Omega))} \le C\Big(&\|\mathbf{u}-\mathbf{u}^{(k)}\|_{\mathcal{V}^{(1)}(Q_T)} + \|\mathbf{u}-\mathbf{u}^{(k)}\|_{\mathcal{V}^{(1)}(Q_T)}\,\|\mathbf{w}\|_{\mathcal{V}^{(1)}(Q_T)}\\&+ \|\mathbf{u}-\mathbf{u}^{(k)}\|_{\mathcal{V}^{(1)}(Q_T)}\left(\|\mathbf{w}^{(k)}\|_{\mathcal{V}^{(1)}(Q_T)} + |\mathbf{v}_\infty|\right)\Big),\end{aligned}$$

where $A(\cdot)$ is a continuous positive function on $\mathbb{R}^5$ and $A(\lambda_1,\lambda_2,\lambda_3,\lambda_4,\lambda_5)\to 0$ as $(\lambda_1,\lambda_2,\lambda_3)\to(0,0,0)$ for fixed λ_4,λ_5. Taking into account (6.21)–(6.22) and the boundedness of $\|\mathbf{w}^{(k)}\|_{\mathcal{V}^{(1)}(Q_T)}$ (which follows from Theorem 4.4), we obtain

$$\|\boldsymbol{\eta}^{(k)}\|_{\mathcal{V}^{(1)}(Q_T)} \to 0 \quad \text{as } k\to\infty\,. \tag{6.25}$$

Relations (6.21) and (6.25) imply the convergence $\mathbf{w}^{(k)}\to\mathbf{w}$ in $\mathcal{V}^{(1)}(Q_T)$. Finally, we prove that $\mathbf{w}^{(k)}\in L^\infty\big(0,T;\mathbf{V}^2(\Theta)\big)$. To this end, we write $\mathbf{w}^{(k)}$ in the form

$$\mathbf{w}^{(k)} = \mathbf{u}^{(k)} + \boldsymbol{\xi}^{(k)}\,. \tag{6.26}$$

Then $\boldsymbol{\xi}^{(k)}$ is the solution of problem (4.19)–(4.21) with $\mathbf{u}$ replaced by $\mathbf{u}^{(k)}$ and $\mathbf{f}$ defined by (4.22), wherein $\mathbf{f}$ is replaced by $\mathbf{f}^{(k)}$ and $\mathbf{u}$ is replaced by $\mathbf{u}^{(k)}$. Note that by (6.20) the inclusion $\mathbf{u}^{(k)}\in\mathbf{C}^\infty(\overline{\Theta})$ holds. By (4.22) we have

$$\mathbf{f}^{(k)} = P\Big(\mu\Delta\mathbf{u}^{(k)} - \partial_t\mathbf{u}^{(k)} + [(\mathbf{u}^{(k)}+\mathbf{v}_\infty)\cdot\nabla]\mathbf{u}^{(k)}\Big), \tag{6.27}$$

where P is the orthogonal projection from $\mathbf{L}^2(\Theta)$ onto $\mathbf{V}_0^0(\Theta)$. Since $\mathbf{u}^{(k)}\in\mathbf{C}^\infty(\overline{\Theta})$, we have $\mathbf{f}^{(k)}\in L^\infty\big(0,T;\mathbf{V}_0^0(\Theta)\big)$, $\partial_t\mathbf{f}^{(k)}\in L^2\big(0,T;\widehat{\mathbf{V}}^{-1}(\Theta)\big)$, and $\mathbf{f}^{(k)}|_{t=0}\in\mathbf{V}_0^0(\Theta)$, where $\widehat{\mathbf{V}}^{-1}$ is the completion of $\mathbf{V}_0^0$ under the norm $\sup_{\|\phi\|_{\mathbf{V}_0^1=1}}\int_\Theta f\cdot\phi\,d\mathbf{x}$. By (6.26) and by properties of $(\mathbf{w}^{(k)}-\mathbf{u}^{(k)})|_{t=0}$ we have that $\boldsymbol{\xi}^{(k)}|_{t=0}\in\mathbf{V}^2(\Omega)\cap\mathbf{V}_0^1(\Omega)$. Hence, by a result in [28, pp. 299–302], we have $\boldsymbol{\xi}^{(k)}\in L^\infty\big(0,T;\mathbf{V}^2(\Theta)\big)$ so that $\mathbf{w}^{(k)}\in L^\infty\big(0,T;\mathbf{V}^2(\Theta)\big)$. □

We may prove a similar result for the solution $\mathbf{q}$ for the adjoint equation (6.12). We consider the boundary value problem

$$-\partial_t\mathbf{q}^{(k)} - \mu\Delta\mathbf{q}^{(k)} + \mathbf{q}^{(k)}\cdot(\nabla\mathbf{w}^{(k)})^T - [(\mathbf{w}^{(k)}+\mathbf{v}_\infty)\cdot\nabla]\mathbf{q}^{(k)} + \nabla\widetilde{r}^{(k)} = \mu\Delta\mathbf{w}^{(k)}, \tag{6.28}$$

$$\operatorname{div}\mathbf{q}^{(k)} = 0, \tag{6.29}$$

and

$$\mathbf{q}^{(k)}|_{t=T} = \mathbf{q}_0^{(k)}\in\mathbf{V}_0^0(\Theta), \qquad \mathbf{q}^{(k)}|_{(0,T)\times\partial\Omega} = \mathbf{0}\,. \tag{6.30}$$

The existence and uniqueness of the solution $\mathbf{q}^{(k)}\in\mathcal{V}^{(1)}(Q_T)$ for (6.28)–(6.30) can be shown by the standard techniques (see [19], [28]).

LEMMA 6.6. *Let* $\mathbf{q}\in\mathcal{V}^{(1)}(Q_T)\cap L^2\big(0,T;\mathbf{V}_0^1(\Omega)\big)$ *and* $\widetilde{r}\in L^2\big(0,T;L^2_{\mathrm{loc}}(\Omega)\big)$ *be the solution of* (6.11), $\mathbf{w}^{(k)}$ *be the solution of* (3.1) *and* (6.22), *and* $\mathbf{q}^{(k)}$ *be the solution*

of (6.28)–(6.30)*, where* $\mathbf{q}_0^{(k)} \in \mathbf{V}_0^1(\Omega)$ *and* $\|\mathbf{q}_0^{(k)} - \mathbf{q}(T,\cdot)\|_{\mathbf{V}_0^0(\Omega)} \to 0$ *as* $k \to \infty$*. Then for every* k*,* $\mathbf{q}^{(k)} \in L^2\big(0,T;\mathbf{V}^2(\Theta)\big)$ *and* $\|\mathbf{q}^{(k)} - \mathbf{q}\|_{\mathcal{V}^{(1)}(Q_T)} \to 0$ *as* $k \to \infty$*.*

Proof. By Lemma 6.5, $\|\mathbf{w}^{(k)} - \mathbf{w}\|_{\mathcal{V}^{(1)}(Q_T)} \to 0$ as $k \to \infty$. Subtracting (6.11) from (6.28) and doing estimation as in Lemma 4.3 we obtain

$$\|\mathbf{q}^{(k)} - \mathbf{q}\|_{\mathcal{V}^{(1)}(Q_T)} \to 0 \quad \text{as } k \to \infty\,.$$

By Lemma 6.5, $\mathbf{w}^{(k)} \in \mathcal{V}^{(1)}(Q_T) \cap L^\infty\big(0,T;\mathbf{V}^2(\Theta)\big)$. Also, $\mathbf{q}^{(k)} \in \mathcal{V}^{(1)}(Q_T)$. Thus we have that

$$\mathbf{g} \equiv \mu\Delta\mathbf{w}^{(k)} - \mathbf{q}^{(k)} \cdot (\nabla\mathbf{w}^{(k)})^T - [(\mathbf{w}^{(k)} + \mathbf{v}_\infty)\cdot\nabla]\mathbf{q}^{(k)} \in \mathbf{L}^2\big((0,T)\times\Theta\big)\,.$$

We rewrite (6.28) in the form

$$-\partial_t\mathbf{q}^{(k)} - \mu\Delta\mathbf{q}^{(k)} + \nabla\widetilde{r}^{(k)} = \mathbf{g}\,.$$

Applying to this last equation the regularity result for the Stokes equations (see [19]) we obtain $\mathbf{q}^{(k)} \in L^2\big(0,T;\mathbf{V}^2(\Theta)\big)$. ∎

LEMMA 6.7. *Let* $\mathbf{w} \in Y$ *be a solution of problem* (3.1)–(3.3) *and* $\mathbf{q} \in \mathcal{V}^{(1)}(Q_T) \cap L^2\big(0,T;\mathbf{V}_0^1(\Omega)\big)$ *be a solution of the adjoint equation* (6.11)*. Then the distribution* $\Delta\mathbf{w}$ *defined on* $\mathbf{C}_0^\infty(\Omega)\cap\mathbf{V}^0(\Omega)$ *can be extended continuously into a functional defined by*

$$\int_\Theta \Delta\mathbf{w}\cdot\mathbf{h}\,d\mathbf{x} = \int_{\partial\Omega}\big((\nabla\mathbf{w}) + (\nabla\mathbf{w})^T\big)\mathbf{n}\cdot\mathbf{h}\,ds - 2\int_\Theta \mathcal{D}(\mathbf{w}) : \mathcal{D}(\mathbf{h})\,d\mathbf{x} \tag{6.31}$$

for every $\mathbf{h} \in \mathbf{C}^\infty(\overline{\Theta})$ *which vanishes near* $(0,T)\times(\partial\Theta\setminus\partial\Omega)$*. Furthermore, the time-dependent version of* (6.31) *also holds; i.e.,*

$$\begin{aligned}&\int_0^T\int_\Theta \Delta\mathbf{w}\cdot\mathbf{h}\,d\mathbf{x}\,dt\\ &\quad= \int_0^T\int_{\partial\Omega}\big((\nabla\mathbf{w}) + (\nabla\mathbf{w})^T\big)\mathbf{n}\cdot\mathbf{h}\,ds\,dt - 2\int_0^T\int_\Theta \mathcal{D}(\mathbf{w}) : \mathcal{D}(\mathbf{h})\,d\mathbf{x}\,dt\end{aligned} \tag{6.32}$$

for every $\mathbf{h} \in \mathbf{C}^\infty((0,T)\times\overline{\Theta})$ *which vanishes near* $(0,T)\times(\partial\Theta\setminus\partial\Omega)$*. Formulae* (6.31) *and* (6.32) *also hold if* $\mathbf{w}$ *is replaced by* $\mathbf{q}$*.*

Proof. If $\mathbf{w} = \mathbf{w}^{(k)} \in L^\infty\big(0,T;\mathbf{V}^2(\Theta)\big)$, (6.32) is the well-known Green's formula (see (2.5) and the ensuing formulae). We substitute into (6.32) the solution $\mathbf{w}^{(k)}$ for the problem (3.1) and (6.23) as constructed in Lemma 6.5. By this lemma we have $\mathbf{w}^{(k)} \to \mathbf{w}$ in $\mathcal{V}^{(1)}(Q_T)$, and therefore,

$$\int_0^T\int_\Theta \mathcal{D}(\mathbf{w}^{(k)}) : \mathcal{D}(\mathbf{h})\,d\mathbf{x}\,dt \to \int_0^T\int_\Theta \mathcal{D}(\mathbf{w}) : \mathcal{D}(\mathbf{h})\,d\mathbf{x}\,dt$$

as $k \to \infty$. Theorem 4.9 yields that the operator $\mathbf{w} \mapsto \gamma\big((\nabla\mathbf{w}) + (\nabla\mathbf{w})^T\big)\mathbf{n}$ is continuous from the set $\{\mathbf{w} \in Y : \mathbf{w} \text{ satisfies (3.1)}\}$ to the space $L^1\big(0,T;\mathbf{B}^{-1/\alpha,\alpha}(\partial\Omega)\big)$, $1 < \alpha < 2$. Hence,

$$\int_0^T\int_{\partial\Omega}\big((\nabla\mathbf{w}^{(k)}) + (\nabla\mathbf{w}^{(k)})^T\big)\mathbf{n}\cdot\mathbf{h}\,ds\,dt \to \int_0^T\int_{\partial\Omega}\big((\nabla\mathbf{w}) + (\nabla\mathbf{w})^T\big)\mathbf{n}\cdot\mathbf{h}\,ds\,dt$$

as $k \to \infty$. Substituting $\mathbf{w} = \mathbf{w}^{(k)}$ into (6.32) and passing to the limit as $k \to \infty$ in the right-hand side of this formula we obtain the desired extension of the distribution $\Delta\mathbf{w}$ which is defined by (6.32). The steady state formula (6.31) can be similarly proved. The case of the distribution $\Delta\mathbf{q}$ can be proved analogously. ∎

6.4. The optimality system in the form of a boundary value problem for a system of partial differential equations. We now interpret the optimality system (3.1), (3.3), and (6.4) as a system of partial differential equations with boundary conditions on the entire boundary of the cylinder $Q_T = (0,T)\times\Omega$.

We first recall from section 3 that

$$\mathbf{V}^0(\Omega) = \{\mathbf{v}\in\mathbf{L}^2(\Omega) : \operatorname{div}\mathbf{v} = 0\}$$

and

$$\mathbf{V}_0^0(\Omega) = \text{the closure of } \mathbf{C}_0^\infty(\Omega)\cap\mathbf{V}^0(\Omega) \text{ in the } \mathbf{L}^2(\Omega)\text{-norm}\,.$$

We have the well-known Weyl decomposition (see [14] and [19])

$$\mathbf{L}^2(\Omega) = \mathbf{V}_0^0(\Omega)\oplus\big(\nabla H^1(\Omega)\big),$$

where $\nabla H^1(\Omega) = \{\nabla g : g\in H^1(\Omega)\}$. We claim that

$$\mathbf{V}^0(\Omega) = \mathbf{V}_0^0(\Omega)\oplus\big(\nabla H_\pi\big),$$

where $\nabla H_\pi = \{\nabla g : g\in H^1(\Omega),\ \Delta g = 0\}$. Indeed, since for each $\mathbf{w}\in\mathbf{V}^0(\Omega)$ we have $\mathbf{w} = \mathbf{w}_\sigma + \nabla w_\pi$ from the Weyl decomposition with $\mathbf{w}_\sigma\in\mathbf{V}_0^0(\Omega)$ and $w_\pi\in H^1(\Omega)$, we obtain by taking the divergence of $\mathbf{w}$ that $\Delta w_\pi = \operatorname{div}\mathbf{w} - \operatorname{div}\mathbf{w}_\sigma = 0$.

We are now prepared to interpret the optimality system in the weak form as a system of partial differential equations with boundary conditions on the entire boundary of the cylinder $Q_T = (0,T)\times\Omega$.

THEOREM 6.8. *Assume* $\mathbf{w}\in Y$ *is a solution for Problem* I *and* $\mathbf{q}\in\mathcal{V}^{(1)}(Q_T)$ *is as defined in Theorem* 6.2. *Then there exist a* $p\in L^2\big(0,T;L^2_{\mathrm{loc}}(\Omega)\big)$ *and an* $r\in L^2\big(0,T;L^2_{\mathrm{loc}}(\Omega)\big)$ *such that the quadruplet* $(\mathbf{w},p,\mathbf{q},r)$ *satisfies the partial differential equations (in the sense of distributions)*

$$\partial_t\mathbf{w} - \mu\Delta\mathbf{w} + ((\mathbf{w}+\mathbf{v}_\infty)\cdot\nabla)\mathbf{w} + \nabla p = \mathbf{0} \qquad \text{in } (0,T)\times\Omega\,, \tag{6.33}$$

$$\nabla\cdot\mathbf{w} = 0 \qquad \text{in } (0,T)\times\Omega\,, \tag{6.34}$$

$$-\partial_t\mathbf{q} - \mu\Delta\mathbf{q} + \mathbf{q}\cdot(\nabla\mathbf{w})^T - (\mathbf{w}\cdot\nabla)\mathbf{q} - (\mathbf{v}_\infty\cdot\nabla)\mathbf{q} + \nabla r = \mu\Delta\mathbf{w} - \nabla p\,, \tag{6.35}$$

and

$$\nabla\cdot\mathbf{q} = 0 \qquad \text{in } (0,T)\times\Omega\,, \tag{6.36}$$

the initial and terminal conditions

$$\mathbf{w}(0,\cdot) = \mathbf{w}_0(\cdot) \quad \text{in } \mathbf{V}^0(\Omega)\,, \tag{6.37}$$

and

$$\mathbf{q}(T,\cdot) + \frac{1}{2}\mathbf{w}_\sigma(T,\cdot) = 0 \quad \text{in } \mathbf{V}_0^0(\Omega), \tag{6.38}$$

and the (lateral) boundary condition

$$\mathbf{q}|_{S_T} = \mathbf{0}\,, \tag{6.39}$$

where $\mathbf{w}_\sigma(T,\cdot)$ *is the projection of* $\mathbf{w}(T,\cdot)$ *onto* $\mathbf{V}_0^0(\Omega)$. *Moreover,*

$$\partial_{tt}(\gamma\mathbf{w}) \in L^1\big(0,T;\mathbf{B}^{-1/\alpha,\alpha}(\partial\Omega)\big)\,, \tag{6.40}$$

$$\gamma\big[\big((\nabla\mathbf{w})+(\nabla\mathbf{w})^T\big)\cdot\mathbf{n}\big] \in L^1\big(0,T;\mathbf{B}^{-1/\alpha,\alpha}(\partial\Omega)\big)\,, \tag{6.41}$$

$$\gamma p \in L^1\big(0,T;B^{-1/\alpha,\alpha}(\partial\Omega)\big)\,, \quad \int_{\partial\Omega} p\,ds = 0, \tag{6.42}$$

$$\gamma\big[\big((\nabla\mathbf{q})+(\nabla\mathbf{q})^T\big)\cdot\mathbf{n}\big] \in L^1\big(0,T;\mathbf{B}^{-1/\alpha,\alpha}(\partial\Omega)\big)\,, \tag{6.43}$$

and

$$\gamma r \in L^1\big(0,T;B^{-1/\alpha,\alpha}(\partial\Omega)\big)\,, \quad \int_{\partial\Omega} r\,ds = 0, \tag{6.44}$$

where $1<\alpha<2$ *and the following boundary conditions hold:*

$$2N\partial_{tt}^2(\gamma\mathbf{w}) - \mathcal{A}(\mathbf{w}) - \mathcal{T}(\mathbf{w},p)\mathbf{n} - \mathcal{T}(\mathbf{q},r)\mathbf{n} = \eta(t)\mathbf{n}, \tag{6.45}$$

where

$$\mathcal{T}(\mathbf{w},p) = -pI + 2\mu\mathcal{D}(\mathbf{w}) \quad \textit{and} \quad \mathcal{T}(\mathbf{q},r) = -rI + 2\mu\mathcal{D}(\mathbf{q})\,, \tag{6.46}$$

$$\mathcal{A}(\mathbf{w}) = \gamma\Big(kN|\mathbf{w}+\mathbf{v}_\infty|^{k-2}(\mathbf{w}+\mathbf{v}_\infty) + \frac{1}{2}\big((\mathbf{w}+\mathbf{v}_\infty)\cdot\mathbf{n}\big)\mathbf{w} + \frac{|\mathbf{w}|^2}{4}\mathbf{n}\Big) \tag{6.47}$$

and

$$\eta(t) = -\int_{\partial\Omega}\mathcal{A}(\mathbf{w})\cdot\mathbf{n}\,ds \Big/ \int_{\partial\Omega} ds\,. \tag{6.48}$$

Furthermore, the following compatibility conditions hold:

$$\big[(\gamma\mathbf{w})\cdot\mathbf{n}\big]\big|_{t=0} = (\gamma\mathbf{w}_0)\cdot\mathbf{n}\,, \tag{6.49}$$

$$(\partial_t\gamma\mathbf{w})\cdot\boldsymbol{\tau}|_{t=T} = 0 \qquad \textit{and} \qquad \left(\frac{1}{2}\gamma w_\pi + 2N\partial_t\gamma\mathbf{w}\cdot\mathbf{n}\right)|_{t=T} = 0\,, \tag{6.50}$$

where $\boldsymbol{\tau}$ *is the unit tangential along* $\partial\Omega$ *and* $w_\pi(t,\cdot)$ *is the primitive function of the projection of* $\mathbf{w}(t,\cdot)$ *onto* ∇H_π *determined by the condition*

$$\int_{\partial\Omega} w_\pi(t,\cdot)\,ds = 0\,. \tag{6.51}$$

Proof. $\mathbf{w}\in\mathcal{V}^{(1)}(Q_T)$ as a solution of Problem I satisfies (3.1) and (6.37). By the De Rham lemma, or recall from (4.27), there exists a $p\in L^2\big(0,T;L^2_{\rm loc}(\Omega)\big)$ such that (6.33) holds. Relation (6.49) follows from the inclusion $\mathbf{w}\in Y$ and the remark at the end of section 4.1. For the Lagrange multiplier $\mathbf{q}\in\mathcal{V}^{(1)}(Q_T)\cap L^2\big(0,T;\mathbf{V}_0^1\big)$ we recall from (6.12) that there exists an $r\in L^2\big(0,T;L^2_{\rm loc}(\Omega)\big)$ such that (6.35) holds. (6.34) and (6.36) simply follow from the fact that $\mathbf{w}\in\mathcal{V}^{(1)}(Q_T)$ and $\mathbf{q}\in\mathcal{V}^{(1)}(Q_T)$; (6.39)

follows from the fact that $\mathbf{q} \in L^2\big(0,T;\mathbf{V}_0^1(\Omega)\big)$. From Theorem 4.10 and Lemma 6.3 we see that the traces of p and r live in the space $L^1\big(0,T;B^{-1/\alpha,\alpha}(\partial\Omega)\big)$ for $1<\alpha<2$. Also, note that in (6.33) and (6.35), we can add an arbitrary constant to p and r so that, in particular, we can choose p and r satisfying $\int_{\partial\Omega} p\,ds = 0$ and $\int_{\partial\Omega} r\,ds = 0$, respectively, where the integrals are understood as the duality pairings $\langle p, 1\rangle$ and $\langle r, 1\rangle$, respectively. This eliminates the arbitrary constant from p and r and also will facilitate some later discussion. Hence (6.42) and (6.44) are verified. Relations (6.41)–(6.44) follow from Theorems 4.9 and 4.10 and Lemma 6.3.

We now examine (6.45). By taking $\mathbf{h} \in \mathbf{C}^\infty$ in (6.4) with $\operatorname{div}\mathbf{h} = 0$, $\mathbf{h}|_{t=0} = \mathbf{0}$, $\mathbf{h}|_{t=T} = \mathbf{0}$ and integrating by parts (which is justified by Lemma 6.7), we obtain

$$\begin{aligned}(6.52)\qquad &\int_0^T\!\!\int_\Omega \Big(-\partial_t\mathbf{q} - \mu\Delta\mathbf{q} + \mathbf{q}\cdot(\nabla\mathbf{w})^T - [(\mathbf{w}+\mathbf{v}_\infty)\cdot\nabla]\mathbf{q} - \mu\Delta\mathbf{w}\Big)\cdot\mathbf{h}\,d\mathbf{x}dt\\ &\quad + \int_0^T\!\!\int_{\partial\Omega}\Big(2N\partial_t\mathbf{w}\cdot\partial_t\mathbf{h} + \big(\mathcal{A}(\mathbf{w}) + 2\mu\mathcal{D}(\mathbf{w})\mathbf{n} + 2\mu\mathcal{D}(\mathbf{q})\mathbf{n}\big)\cdot\mathbf{h}\Big)\,ds\,dt = 0\,,\end{aligned}$$

where $\mathcal{A}(\mathbf{w})$ is defined by (6.47). Also, the integrals are understood as duality pairings if necessary. Equations (6.35) and (6.52) yield

$$\begin{aligned}(6.53)\qquad &-\int_0^T\!\!\int_\Omega (\nabla r + \nabla p)\cdot\mathbf{h}\,d\mathbf{x}dt\\ &\quad + \int_0^T\!\!\int_{\partial\Omega}\Big(2N\partial_t\mathbf{w}\cdot\partial_t\mathbf{h} + \big(\mathcal{A}(\mathbf{w}) + 2\mu\mathcal{D}(\mathbf{w})\mathbf{n} + 2\mu\mathcal{D}(\mathbf{q})\mathbf{n}\big)\cdot\mathbf{h}\Big)\,ds\,dt = 0\,.\end{aligned}$$

Using Lemma 6.4 and the last equation, we obtain

$$(6.54)\qquad \int_0^T\!\!\int_{\partial\Omega}\Big(2N\partial_t\mathbf{w}\cdot\partial_t\mathbf{h} + \big(\mathcal{A}(\mathbf{w}) + \mathcal{T}(p,\mathbf{w})\mathbf{n} + \mathcal{T}(r,\mathbf{q})\mathbf{n}\big)\cdot\mathbf{h}\Big)\,ds\,dt = 0\,,$$

where $\mathcal{T}$ is the stress tensor defined by (6.46). Since $1<\alpha<2$, we have the continuous imbeddings $B^{1/\alpha,\alpha'}(\partial\Omega) \hookrightarrow L^\infty(\partial\Omega)$, where $\alpha' = \alpha/(\alpha-1)$ so that $L^1(\partial\Omega) \hookrightarrow B^{-1/\alpha,\alpha}(\partial\Omega)$. Hence,

$$\gamma\Big(2kN|\mathbf{w}+\mathbf{v}_\infty|^{k-2}(\mathbf{w}+\mathbf{v}_\infty) + \frac{1}{2}\big((\mathbf{w}+\mathbf{v}_\infty)\cdot\mathbf{n}\big)\mathbf{w} + \frac{|\mathbf{w}|^2}{4}\mathbf{n}\Big) \in \mathbf{L}^1(\partial\Omega) \hookrightarrow \mathbf{B}^{-1/\alpha,\alpha}(\partial\Omega)\,.$$

Using Theorems 4.9 and 4.10, Lemma 6.3, and the last relation, we see that

$$(6.55)\qquad \mathcal{A}(\mathbf{w}) + \mathcal{T}(p,\mathbf{w})\mathbf{n} + \mathcal{T}(r,\mathbf{q})\mathbf{n} \in L^1\big(0,T;\mathbf{B}^{-1/\alpha,\alpha}(\partial\Omega)\big).$$

Since $\mathbf{h}$ and $\partial_t\mathbf{w}$ are solenoidal (from the definition of the spaces $\mathcal{V}^{(1)}(Q_T)$ and Y), we have

$$\int_{\partial\Omega}\mathbf{h}\cdot\mathbf{n}\,ds = 0 \qquad \text{and} \qquad \int_{\partial\Omega}\partial_t\mathbf{w}\cdot\mathbf{n}\,ds = 0\,.$$

Thus, from (6.54)–(6.55), we deduce (6.40) and

$$(6.56)\qquad \int_0^T\!\!\int_{\partial\Omega}\Big(2N\partial_{tt}\mathbf{w} - \mathcal{A}(\mathbf{w}) - \mathcal{T}(p,\mathbf{w})\mathbf{n} - \mathcal{T}(r,\mathbf{q})\mathbf{n}\Big)\cdot\mathbf{h}\,ds\,dt = 0\,.$$

Hence, (6.45) follows from (6.55) and (6.56) with $\eta(t)$ defined by

$$(6.57)\qquad \eta(t) = \int_{\partial\Omega}\mathbf{n}\cdot\Big(2N\partial_{tt}\mathbf{w} - \mathcal{A}(\mathbf{w}) - \mathcal{T}(p,\mathbf{w})\mathbf{n} - \mathcal{T}(r,\mathbf{q})\mathbf{n}\Big)\,ds \Big/ \int_{\partial\Omega} ds\,.$$

Note that for every $\phi(t) \in C_0^\infty(0,T)$,

$$\int_0^T \int_{\partial\Omega} \partial_{tt}\mathbf{w}\cdot\mathbf{n}\,ds\,\phi(t)\,dt = \int_0^T \int_{\partial\Omega} \mathbf{w}\cdot\mathbf{n}\,ds\,\partial_{tt}\phi(t)\,dt = 0$$

as $\int_{\partial\Omega} \mathbf{w}\cdot\mathbf{n}\,ds = 0$ for every divergence-free function $\mathbf{w}$. Thus, for almost all $t \in (0,T)$,

$$\int_{\partial\Omega} \partial_{tt}\mathbf{w}\cdot\mathbf{n}\,ds = 0\,. \tag{6.58}$$

Taking into account (6.42) we obtain for each $\phi(t) \in C_0^\infty(0,T)$,

$$\begin{aligned} &\int_0^T \int_{\partial\Omega} \mathbf{n}\cdot\mathcal{T}(p,\mathbf{w})\mathbf{n}\,ds\,\phi(t)\,dt \\ &= \int_0^T \int_{\partial\Omega} \Big(\mathbf{n}\cdot\mu\big((\nabla\mathbf{w}) + (\nabla\mathbf{w})^T\big)\mathbf{n} - p\Big)\,ds\,\phi(t)\,dt \\ &= \int_0^T \int_{\partial\Omega} \Big(\mathbf{n}\cdot\mu\big((\nabla\mathbf{w}) + (\nabla\mathbf{w})^T\big)\mathbf{n}\Big)\,ds\,\phi(t)\,dt\,. \end{aligned} \tag{6.59}$$

Let $\epsilon > 0$ be given. For each $s \in \partial\Omega$, we consider the normal $\widetilde{\mathbf{n}}(s)$ to $\partial\Omega$ which is directed into Ω. We choose the point $K(\epsilon, s)$ along $\widetilde{\mathbf{n}}(s)$ such that the distance between $K(\epsilon, s)$ and $\partial\Omega$ equals ϵ. If ϵ is sufficiently small, then the set $\{K(\epsilon, s) : s \in \partial\Omega\}$ is a C^∞-manifold which we denote by $\partial\Omega_\epsilon$. Since $\mathbf{w} \in \mathbf{V}^0(\Omega)$,

$$\int_{\partial\Omega} \mathbf{w}\cdot\widetilde{\mathbf{n}}\,ds = 0 \quad \text{and} \quad \int_{\partial\Omega_\epsilon} \mathbf{w}\cdot\widetilde{\mathbf{n}}_\epsilon\,ds = 0,$$

where $\widetilde{\mathbf{n}}_\epsilon$ is the outward normal to $\partial\Omega_\epsilon$. Hence,

$$\begin{aligned} &\int_0^T \int_{\partial\Omega} \Big(\mathbf{n}\cdot\big((\nabla\mathbf{w}) + (\nabla\mathbf{w})^T\big)\mathbf{n}\Big)\,ds\,\phi(t)\,dt \\ &= 2\int_0^T \int_{\partial\Omega} \partial_n\mathbf{w}\cdot\mathbf{n}\,ds\,\phi(t)\,dt \\ &= \lim_{\epsilon\to 0} \frac{1}{\epsilon} \int_0^T \Big(\int_{\partial\Omega} \mathbf{w}\cdot\mathbf{n}\,ds - \int_{\partial\Omega_\epsilon} \mathbf{w}\cdot\widetilde{\mathbf{n}}_\epsilon\,ds\Big)\,\phi(t)\,dt = 0\,. \end{aligned} \tag{6.60}$$

Thus, (6.59) and (6.60) yield

$$\int_{\partial\Omega} \mathbf{n}\cdot\mathcal{T}(p,\mathbf{w})\mathbf{n}\,ds = 0 \quad \text{for almost every } t \in (0,T)\,. \tag{6.61}$$

Similarly, we have

$$\int_{\partial\Omega} \mathbf{n}\cdot\mathcal{T}(r,\mathbf{q})\mathbf{n}\,ds = 0 \quad \text{for almost every } t \in (0,T)\,. \tag{6.62}$$

From relations (6.58) and (6.61)–(6.62) we conclude that the function $\eta(t)$ defined in (6.57) equals the function defined in (6.48).

Now we choose $\mathbf{h} \in \mathbf{C}^\infty(Q_T)$ in (6.4) with $\operatorname{div} \mathbf{h} = 0$ and $\mathbf{h}|_{t=0} = \mathbf{0}$. Integration by parts (which again is justified by Lemmas 6.4 and 6.7) yields

$$\begin{aligned}
&\int_0^T \int_\Omega \Big(- \partial_t \mathbf{q} - \mu \Delta \mathbf{q} + \mathbf{q} \cdot (\nabla \mathbf{w})^T - [(\mathbf{w} + \mathbf{v}_\infty) \cdot] \mathbf{q} - \mu \Delta \mathbf{w} \Big) \cdot \mathbf{h} \, d\mathbf{x} dt \\
&+ \int_0^T \int_{\partial\Omega} \Big(- 2N \partial_{tt} \mathbf{w} \cdot \mathbf{h} + \big(\mathcal{A}(\mathbf{w}) + 2\mu \mathcal{D}(\mathbf{w}) \mathbf{n} + 2\mu \mathcal{D}(\mathbf{q}) \mathbf{n} \big) \cdot \mathbf{h} \Big) \, ds \, dt \\
&+ \int_\Omega \left[\mathbf{q}(T, \mathbf{x}) + \frac{1}{2} \mathbf{w}(T, \mathbf{x}) \right] \cdot \mathbf{h}(T, \mathbf{x}) \, d\mathbf{x} + \int_{\partial\Omega} 2N \partial_t \mathbf{w}(T, \mathbf{x}) \cdot \mathbf{h}(T, \mathbf{x}) \, ds = 0 \,.
\end{aligned} \tag{6.63}$$

Note that (6.53) and (6.56) hold for the present $\mathbf{h}$ so that using (6.35), (6.53), (6.56), and (6.63), we are led to

$$\int_\Omega \left[\mathbf{q}(T, \mathbf{x}) + \frac{1}{2} \mathbf{w}(T, \mathbf{x}) \right] \cdot \mathbf{h}(T, \mathbf{x}) \, d\mathbf{x} + \int_{\partial\Omega} 2N \partial_t \mathbf{w}(T, \mathbf{x}) \cdot \mathbf{h}(T, \mathbf{x}) \, ds = 0 \,. \tag{6.64}$$

Using the fact that $\mathbf{q}(T, \cdot) \in \mathbf{V}_0^0(\Omega)$ and $\mathbf{w}(T, \cdot) \in \mathbf{V}^0(\Omega)$ we obtain (6.38). Substituting (6.38) into (6.64) we obtain by integration by parts that

$$\int_{\partial\Omega} \Big(\frac{1}{2} w_\pi(T, \mathbf{x}) \mathbf{n}(\mathbf{x}) \cdot \mathbf{h}(T, \mathbf{x}) + 2N \partial_t \mathbf{w}(T, \mathbf{x}) \cdot \mathbf{h}(T, \mathbf{x}) \Big) \, ds = 0,$$

which implies (6.50) with (6.51). □

6.5. The case of Problem II. We derive now the optimality system for Problem II.

THEOREM 6.9. *Assume that* $\mathbf{v}_0 \equiv \mathbf{w}_0 + \mathbf{v}_\infty \in \mathbf{V}_0^1(\Omega)$ *and* $\mathbf{w} \in Y$ *is a solution of Problem* II. *Then there exists a triplet* $(\mathbf{q}, r, \lambda) \in \mathcal{V}^{(1)}(Q_T) \times L^2_{\mathrm{loc}}(Q_T) \times \mathbb{R}_+$ *and* $p \in L^2_{\mathrm{loc}}(Q_T)$ *such that* $(\mathbf{q}, r, \lambda) \neq (\mathbf{0}, 0, 0)$ *and the collection* $(\mathbf{w}, \mathbf{q}, p, r, \lambda)$ *satisfies* (6.33)–(6.44) *and the boundary conditions*

$$2\lambda \partial_{tt}^2 \gamma \mathbf{w} - \widetilde{\mathcal{A}}(\mathbf{w}) - \mathcal{T}(\mathbf{w}, p)\mathbf{n} - \mathcal{T}(\mathbf{q}, r)\mathbf{n} = \widetilde{\eta}(t)\mathbf{n}, \tag{6.65}$$

where $\mathcal{T}(\mathbf{w}, p)$ *and* $\mathcal{T}(\mathbf{q}, r)$ *are defined by* (6.46),

$$\widetilde{\mathcal{A}}(\mathbf{w}) = \gamma \Big(k\lambda |\mathbf{w} + \mathbf{v}_\infty|^{k-2} (\mathbf{w} + \mathbf{v}_\infty) + \frac{1}{2} \big((\mathbf{w} + \mathbf{v}_\infty) \cdot \mathbf{n} \big) \mathbf{w} + \frac{|\mathbf{w}|^2}{4} \mathbf{n} \Big), \tag{6.66}$$

and

$$\widetilde{\eta}(t) = - \int_{\partial\Omega} \widetilde{\mathcal{A}}(\mathbf{w}) \cdot \mathbf{n} \, ds \Big/ \int_{\partial\Omega} ds \,. \tag{6.67}$$

Moreover, the following compatibility conditions hold:

$$(\gamma \mathbf{w}) \cdot \mathbf{n}|_{t=0} = (\gamma \mathbf{w}_0) \cdot \mathbf{n} \,, \tag{6.68}$$

$$\lambda \big[(\partial_t \gamma \mathbf{w}) \cdot \boldsymbol{\tau} \big] \big|_{t=T} = 0 \,, \qquad \textit{and} \qquad \Big(\frac{1}{2} \gamma w_\pi + 2\lambda \partial_t \gamma \mathbf{w} \cdot \mathbf{n} \Big) \Big|_{t=T} = 0 \,, \tag{6.69}$$

where $\boldsymbol{\tau}$ *is the unit tangential along* $\partial\Omega$ *and* $w_\pi(t, \cdot)$ *is the primitive function of the projection of* $\mathbf{w}(t, \cdot)$ *onto* ∇H_π *determined by the condition* (6.51). *Furthermore, the conditions of nonnegativeness and complementary slackness are valid; i.e.,*

$$\lambda \geq 0 \tag{6.70}$$

and

$$\lambda\Big(\int_0^T\int_{\partial\Omega}\big(|\mathbf{w}+\mathbf{v}_\infty|^k+|\partial_t\mathbf{w}|^2\big)\,ds\,dt-M\Big)=0\,. \tag{6.71}$$

Proof. Let $\mathbf{w}\in Y$ be a solution of Problem II. We fit Problem II into the framework of Theorem 6.1. We set $X_1=Y_0$ and $X_2=L^2\big(0,T;\mathbf{V}^{-1}(\Omega)\big)$. The Lagrange functional for Problem II is defined by

$$\begin{aligned}\mathcal{L}(\mathbf{y},\mathbf{q})=\lambda_0\Big(&\mu\int_0^T\int_\Omega|\mathcal{D}(\mathbf{w}+\mathbf{y})|^2\,d\mathbf{x}dt+\frac14\int_\Omega|\mathbf{w}(T,\mathbf{x})+\mathbf{y}(T,\mathbf{x})|^2\,d\mathbf{x}\\&+\frac14\int_0^T\int_{\partial\Omega}|\mathbf{w}+\mathbf{y}|^2(\mathbf{w}+\mathbf{y}+\mathbf{v}_\infty)\cdot\mathbf{n}\,ds\,dt\Big)\\&+\lambda\int_0^T\int_{\partial\Omega}\Big(|\mathbf{w}+\mathbf{y}+\mathbf{v}_\infty|^k+|\partial_t\mathbf{w}+\partial_t\mathbf{y}|^2-M\Big)\,ds\,dt\\&+\int_0^T\int_\Omega(\partial_t\mathbf{w}+\partial_t\mathbf{y})\cdot\mathbf{q}\,d\mathbf{x}dt+2\mu\int_0^T\int_\Omega\mathcal{D}(\mathbf{w}+\mathbf{y}):\mathcal{D}(\mathbf{q})\,d\mathbf{x}dt\\&+\int_0^T\int_\Omega\big\{[(\mathbf{w}+\mathbf{y})\cdot\nabla](\mathbf{w}+\mathbf{y})+(\mathbf{v}_\infty\cdot\nabla)(\mathbf{w}+\mathbf{y})\big\}\cdot\mathbf{q}\,d\mathbf{x}dt\end{aligned} \tag{6.72}$$

for all $\mathbf{y}\in X_1$ and $\mathbf{q}\in X_2^*=L^2\big(0,T;\mathbf{V}_0^1(\Omega)\big)$. Note that (6.72) differs from (6.9) in that λ_0 has to be included in the Lagrangian functional and λ_0 can be zero. We define the functionals

$$f(\mathbf{y})=\mathcal{J}(\mathbf{w}+\mathbf{y})$$

and

$$g_1(\mathbf{y})=\int_0^T\int_{\partial\Omega}\big(|\mathbf{w}+\mathbf{y}+\mathbf{v}_\infty|^k+|\partial_t\mathbf{w}+\partial_t\mathbf{y}|^2\big)\,ds\,dt-M\,;$$

see (3.5)–(3.6). We define the mapping $F:X_1\to X_2$ as in the proof of Theorem 6.2. Analogous to the proof of Theorem 6.2, we can verify that $F'(\mathbf{0})$ is an epimorphism, and therefore the image of $F'(\mathbf{0})$ is closed. Hence, we have verified all the assumptions in Theorem 6.1 and we conclude that there exist a $\mathbf{q}\in X_2^*=L^2\big(0,T;\mathbf{V}_0^1(\Omega)\big)$ and a $(\lambda_0,\lambda)\in\mathbb{R}^2$ such that $(\mathbf{q},\lambda_0,\lambda)\neq(\mathbf{0},0,0)$,

$$\big\langle\mathcal{L}_{\mathbf{y}}(\mathbf{y},\mathbf{q}),\mathbf{h}\big\rangle\Big|_{\mathbf{y}=0}=0\qquad\forall\,\mathbf{h}\in Y\,, \tag{6.73}$$

$$\lambda_0\geq0\,,\qquad\lambda\geq0\,, \tag{6.74}$$

and

$$\lambda\Big(\int_0^T\int_{\partial\Omega}\big(|\partial_t\mathbf{w}|^2+|\mathbf{w}+\mathbf{v}_\infty|^k\big)\,ds\,dt-M\Big)=0\,. \tag{6.75}$$

As in Theorems 6.2 and 6.8, we derive from (6.73) relations (6.33)–(6.34), (6.36)–(6.37), (6.39)–(6.44),

$$-\partial_t\mathbf{q}-\mu\Delta\mathbf{q}+\mathbf{q}\cdot(\nabla\mathbf{w})^T-(\mathbf{w}\cdot\nabla)\mathbf{q}-(\mathbf{v}_\infty\cdot\nabla)\mathbf{q}+\nabla r=\lambda_0(\mu\Delta\mathbf{w}-\nabla p) \tag{6.76}$$

and

$$\mathbf{q}(T,\cdot)+\frac{\lambda_0}{2}\mathbf{w}_\sigma(T,\cdot)=0 \quad \text{in } \mathbf{V}_0^0(\Omega)\,, \tag{6.77}$$

where $\mathbf{w}_\sigma(T,\cdot)$ is the projection of $\mathbf{w}(T,\cdot)$ onto $\mathbf{V}_0^0(\Omega)$. Moreover, we derive the following boundary condition:

$$2\lambda\partial_{tt}^2\gamma\mathbf{w}-\widetilde{\mathcal{A}}(\mathbf{w})-\mathcal{T}(\mathbf{w},p)\mathbf{n}-\mathcal{T}(\mathbf{q},r)\mathbf{n}=\widetilde{\eta}(t)\mathbf{n}\,, \tag{6.78}$$

where

$$\widetilde{\mathcal{A}}(\mathbf{w})=\gamma\Big\{k\lambda|\mathbf{w}+\mathbf{v}_\infty|^{k-2}(\mathbf{w}+\mathbf{v}_\infty)+\lambda_0\Big(\frac{1}{2}\big((\mathbf{w}+\mathbf{v}_\infty)\cdot\mathbf{n}\big)\mathbf{w}+\frac{|\mathbf{w}|^2}{4}\mathbf{n}\Big)\Big\}$$

and

$$\widetilde{\eta}(t)=-\int_{\partial\Omega}\widetilde{\mathcal{A}}(\mathbf{w})\cdot\mathbf{n}\,ds\Big/\int_{\partial\Omega}ds\,.$$

Furthermore, the following compatibility conditions hold:

$$\big[(\gamma\mathbf{w})\cdot\mathbf{n}\big]\big|_{t=0}=(\gamma\mathbf{w}_0)\cdot\mathbf{n}\,,$$

$$\Big(\frac{1}{2}\lambda_0\gamma w_\pi+2\lambda\partial_t\gamma\mathbf{w}\cdot\mathbf{n}\Big)|_{t=T}=0\,,\quad (\lambda\partial_t\gamma\mathbf{w})\cdot\boldsymbol{\tau}|_{t=T}=0\,, \tag{6.79}$$

where w_π is the primitive for the projection ∇w_σ of $\mathbf{w}$ onto ∇H_π determined by (6.51).

To complete the proof, it remains to show that $\lambda_0\neq 0$ so that we can choose $\lambda_0=1$. Assume $\lambda_0=0$. Then (6.76), (6.77), (6.36), and (6.39) yield $\mathbf{q}\equiv\mathbf{0}$ by standard techniques of energy estimates and the Gronwall inequality. Also, equation (6.76) and the condition $\int_{\partial\Omega}r\,ds=0$ imply $r=0$. We note that $\lambda\neq 0$ because $(\mathbf{q},\lambda_0,\lambda,r)=(\mathbf{0},0,\lambda,0)\neq(\mathbf{0},0,0,0)$. By (6.70), $\lambda>0$. Then, by virtue of (6.71), $\mathbf{w}$ is also a solution of the following modified minimization problem: seek a $\mathbf{w}\in Y$ such that the functional (3.5) is minimized subject to the *equality* constraints (3.1), (3.3), and

$$\int_0^T\int_{\partial\Omega}(|\mathbf{w}+\mathbf{v}_\infty|^k+|\partial_t\mathbf{w}|^2)\,ds\,dt=M\,. \tag{6.80}$$

We now show that this minimization problem satisfies the conditions of Theorem 6.1. Indeed, we set $X_1=Y_0$ and $X_2=L^2\big(0,T;\mathbf{V}^{-1}(\Omega)\big)\times\mathbb{R}$. We define the mapping f by $f(\mathbf{y})=\mathcal{J}(\mathbf{w}+\mathbf{y})$, where $\mathcal{J}$ is the functional (3.5) and define

$$\widetilde{F}(\mathbf{y})=\begin{pmatrix}\partial_t(\mathbf{w}+\mathbf{y})-\mu P\Delta(\mathbf{w}+\mathbf{y})+P\big[\big((\mathbf{w}+\mathbf{y}+\mathbf{v}_\infty)\cdot\nabla\big)(\mathbf{w}+\mathbf{y})\big]\\ \int_0^T\int_{\partial\Omega}(|\mathbf{w}+\mathbf{y}+\mathbf{v}_\infty|^k+|\partial_t\mathbf{w}+\partial_t\mathbf{y}|^2)\,ds\,dt-M\end{pmatrix},$$

where $P:\mathbf{H}^{-1}(\Omega)\to\mathbf{V}^{-1}(\Omega)$ is the orthogonal projection. Then $\widetilde{F}'(\mathbf{0}):X_1\to X_2$ is defined by

$$\langle\widetilde{F}'(\mathbf{0}),\mathbf{y}\rangle=\begin{pmatrix}\partial_t\mathbf{y}-\mu P\Delta\mathbf{y}+P(\mathbf{y}\cdot\nabla)\mathbf{w}+P\big((\mathbf{w}+\mathbf{v}_\infty)\cdot\nabla\big)\mathbf{y}\\ \int_0^T\int_{\partial\Omega}(k|\mathbf{w}+\mathbf{v}_\infty|^{k-2}(\mathbf{w}+\mathbf{v}_\infty)\cdot\mathbf{y}+2\partial_t\mathbf{w}\cdot\partial_t\mathbf{y})\,ds\,dt\end{pmatrix}.$$

To show that $\widetilde{F}'(\mathbf{0})$ is an epimorphism, we first observe that this operator is continuous. Next we need to show that for each $\mathbf{f} \in L^2\big(0,T;\mathbf{V}^{-1}(\Omega)\big)$ and $\zeta \in \mathbb{R}$, the system

$$\begin{aligned}&\langle \partial_t \mathbf{y}(t), \mathbf{z}\rangle + \mu \int_\Omega \nabla \mathbf{y}(t) : \nabla \mathbf{z}\, d\mathbf{x} + \int_\Omega \big((\mathbf{w}(t)+\mathbf{v}_\infty)\cdot\nabla\big)\mathbf{y}(t)\cdot\mathbf{z}\, d\mathbf{x} \\ &+ \int_\Omega \big(\mathbf{y}(t)\cdot\nabla\big)\mathbf{w}(t)\cdot\mathbf{z}\, d\mathbf{x} = \int_\Omega \mathbf{f}(t)\cdot\mathbf{z}\, d\mathbf{x} \qquad \forall\, \mathbf{z}\in \mathbf{V}_0^1(\Omega),\ \text{a.e. } t\in(0,T)\,,\end{aligned} \tag{6.81}$$

$$\mathbf{y}|_{t=0} = \mathbf{0} \qquad \text{in } \mathbf{V}^0(\Omega), \tag{6.82}$$

and

$$\int_0^T \int_{\partial\Omega} \Big(k|\mathbf{w}+\mathbf{v}_\infty|^{k-2}(\mathbf{w}+\mathbf{v}_\infty)\cdot\mathbf{y} + 2\partial_t\mathbf{w}\cdot\partial_t\mathbf{y}\Big)\, ds\, dt = \zeta \tag{6.83}$$

has a solution $\mathbf{y} \in Y_0$.

To this end we first look for a $\mathbf{y} \in \gamma_{S_T} Y_0$ satisfying (6.83). ($\gamma_{S_T} Y_0$ is the space of functions belonging to Y_0 restricted to S_T.) It suffices to show that there exists a $\mathbf{y} \in \gamma_{S_T} Y_0$ for which the left side of (6.83) is not zero, for then we can obtain (6.83) by multiplying $\mathbf{y}$ by a suitable constant. Suppose that for every $\mathbf{y} \in \gamma_{S_T} Y_0$ the equality

$$\int_0^T \int_{\partial\Omega} \Big(k|\mathbf{w}+\mathbf{v}_\infty|^{k-2}(\mathbf{w}+\mathbf{v}_\infty)\cdot\mathbf{y} + 2\partial_t(\mathbf{w}+\mathbf{v}_\infty)\cdot\partial_t\mathbf{y}\Big)\, ds\, dt = 0$$

holds. This equality and (6.79) with $\lambda_0 = 0$ imply that $\mathbf{w}+\mathbf{v}_\infty$ satisfies the relations

$$-2\partial_{tt}(\mathbf{w}+\mathbf{v}_\infty) + k|\mathbf{w}+\mathbf{v}_\infty|^{k-2}(\mathbf{w}+\mathbf{v}_\infty) = \mathbf{0} \qquad \text{on } (0,T)\times\partial\Omega \tag{6.84}$$

and

$$\partial_t(\mathbf{w}+\mathbf{v}_\infty)\big|_{t=T} = \mathbf{0} \qquad \text{on } \partial\Omega\,. \tag{6.85}$$

Note also that

$$(\mathbf{w}+\mathbf{v}_\infty)\big|_{t=0,\mathbf{x}\in\partial\Omega} = \mathbf{v}_0\big|_{\partial\Omega} \equiv \mathbf{0}\,. \tag{6.86}$$

We multiply (6.84) by $\partial_t\big(\mathbf{w}+\mathbf{v}_\infty\big)$ and obtain

$$-\partial_t|\partial_t(\mathbf{w}+\mathbf{v}_\infty)|^2 + \partial_t|\mathbf{w}+\mathbf{v}_\infty|^k = 0 \qquad \text{on } (0,T)\times\partial\Omega\,,$$

which together with (6.85) implies

$$-|\partial_t(\mathbf{w}+\mathbf{v}_\infty)|^2 + |\mathbf{w}+\mathbf{v}_\infty|^k = |\mathbf{w}(T,\cdot)+\mathbf{v}_\infty|^k \qquad \text{on } (0,T)\times\partial\Omega\,.$$

This equality taken at $t=0$ yields

$$-\big|\partial_t\big(\mathbf{w}(0,\mathbf{x})+\mathbf{v}_\infty\big)\big|^2 = |\mathbf{w}(T,\mathbf{x})+\mathbf{v}_\infty|^k \qquad \text{on } \partial\Omega\,,$$

which implies

$$\big|\partial_t\big(\mathbf{w}(0,\mathbf{x})+\mathbf{v}_\infty\big)\big|^2 = 0 \qquad \text{and} \qquad |\mathbf{w}(T,\mathbf{x})+\mathbf{v}_\infty|^k = 0 \qquad \text{on } \partial\Omega\,,$$

or we rewrite the last relation as

$$\mathbf{w}(T,\mathbf{x}) + \mathbf{v}_\infty = \mathbf{0} \qquad \text{on } (0,T)\times\partial\Omega\,. \tag{6.87}$$

We deduce from the differential equation (6.84) and boundary conditions (6.85) and (6.87) that $(\mathbf{w}+\mathbf{v}_\infty)=\mathbf{0}$ on $(0,T)\times\partial\Omega$. This contradicts (6.80). Therefore, there exists a $\mathbf{z}\in\gamma_{S_T}Y_0$ satisfying (6.83), where $\mathbf{y}$ is replaced by $\mathbf{z}$.

We supplement the system (6.81)–(6.82) with the boundary condition

$$\mathbf{y}|_{(0,T)\times\partial\Omega} = \mathbf{z}\,. \tag{6.88}$$

Using the techniques in the proof of Theorem 4.4 we see that (6.81)–(6.82) and (6.88) indeed has a (unique) solution $\mathbf{y}\in\mathcal{V}^{(1)}(Q_T)$ (the situation now is even simpler, as the system (6.81)–(6.82) is linear). Note that substituting $\mathbf{z}$ from (6.88) into (6.83) in place of $\mathbf{y}$ makes (6.83) valid. Clearly, $\mathbf{y}\in Y_0$. Hence we have proved that $\widetilde{F}'(0)$ is an epimorphism, so that we have verified all the assumptions in Theorem 6.1. By virtue of Theorem 6.1, every Lagrange multiplier triplet $(\widetilde{\mathbf{q}},\widetilde{\lambda}_0,\widetilde{\lambda})$ such that (6.73) holds where $\mathcal{L}$ is defined by (6.72) satisfies $\widetilde{\lambda}_0\neq 0$; in particular, $(\mathbf{q},\lambda_0,\lambda)$ is such a triplet, and therefore $\lambda_0\neq 0$. This contradicts the assumption $\lambda_0=0$. Hence $\lambda_0\neq 0$. □

REMARK. Note that, since we have not employed a separate variable for the control, the boundary condition (3.2) does not appear in the optimality systems of sections 6.3–6.5. In fact, in order to satisfy (3.2) one merely has to choose, once $\mathbf{w}$ is determined from the above optimality system, a control $\mathbf{g}$ such that $\mathbf{g}=\mathbf{w}|_{\partial\Omega}+\mathbf{v}_\infty$ for $t\in(0,T)$. □

REMARK. The complexity of the optimality systems makes it nontrivial to study the regularity of solutions for these systems. The regularity of solutions will be studied elsewhere. □

REFERENCES

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

[2] F. ABERGEL AND R. TEMAM, *On some control problems in fluid mechanics*, Theoret. Comput. Fluid Dynamics, 1 (1990) pp. 303–325.

[3] V. ALEKSEEV, V. TIKHOMIROV, AND S. FOMIN, *Optimal Control*, Consultants Bureau, New York, 1987.

[4] O. BESOV, V. IL'IN, AND S. NIKOL'SKII, *Integral Representations of Functions and Imbedding Theorems*, Winston, Washington, D.C., 1979.

[5] P. CONSTANTIN AND C. FOIAS, *Navier–Stokes Equations*, Univ. of Chicago Press, Chicago, IL, 1988.

[6] C. CUVERLIER, *Optimal control of a system governed by the Navier-Stokes equations coupled with the heat equation*, in New Developments in Differential Equations, W. Eckhaus, ed., North-Holland, Amsterdam, 1976, pp. 81–98.

[7] H. FATTORINI AND S. SRITHARAN, *Existence of optimal controls for viscous flow problems*, Proc. Roy. Soc. London Ser. A, 439 (1992), pp. 81–102.

[8] H. FATTORINI AND S. SRITHARAN, *Necessary and sufficient conditions for optimal controls in viscous flow problems*, Proc. Roy. Soc. London Ser. A, 124 (1994), pp. 211–251.

[9] H. FATTORINI AND S. SRITHARAN, *Optimal chattering control for viscous flows*, Nonlinear Anal., 25 (1995), pp. 763–797.

[10] H. FATTORINI AND S. SRITHARAN, *Optimal control theory for viscous flow problems*, to appear.

[11] A. FURSIKOV, *On some control problems and results concerning the unique solvability of a mixed boundary value problem for the three-dimensional Navier-Stokes and Euler systems*, Soviet Math. Dokl., 21 (1980), pp. 889–893.

[12] A. FURSIKOV, *Control problems and theorems concerning the unique solvability of a mixed boundary value problem for the three-dimensional Navier-Stokes and Euler equations*, Math USSR Sb., 43 (1982), pp. 281–307.

[13] A. Fursikov, *Properties of solutions of some extremal problems connected with the Navier-Stokes system*, Math USSR Sb., 46 (1983), pp. 323–351.
[14] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier-Stokes Equations*, Springer-Verlag, Berlin, 1986.
[15] M. Gunzburger, L. Hou, and T. Svobodny, *Numerical approximation of an optimal control problem associated with the Navier-Stokes equations*, Appl. Math. Lett., 2 (1989), pp. 29–31.
[16] M. Gunzburger, L. Hou, and T. Svobodny, *Analysis and finite element approximation of optimal control problems for the stationary Navier-Stokes equations with Dirichlet controls*, Modél. Math. Anal. Numér., 25 (1991), pp. 711–748.
[17] M. Gunzburger, L. Hou, and T. Svobodny, *Analysis and finite element approximation of optimal control problems for the stationary Navier-Stokes equations with distributed and Neumann controls*, Math. Comp., 57 (1991), pp. 123–151.
[18] M. Gunzburger, L. Hou, and T. Svobodny, *Boundary velocity control of incompressible flow with an application to viscous drag reduction*, SIAM J. Control Optim., 30 (1992), pp. 167–181.
[19] O. Ladyzhenskaya, *The Mathematical Theory of Viscous Incompressible Flow*, Gordon and Breach, New York, 1963.
[20] J.-L. Lions, *Control of Distributed Singular systems*, Bordas, Paris, 1985.
[21] J.-L. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problems and Applications* I, Springer-Verlag, Berlin, 1972.
[22] P. Panagiotopoulos, *Inequality Problems in Mechanics and Applications*, Birkhäuser, Boston, 1985.
[23] J. Serrin, *Mathematical principles of classical fluid mechanics*, in Handbuch der Physik, VIII/1, S. Flügge and C. Truesdell, eds., 1959, pp. 1–125.
[24] S. Sritharan, *An optimal control problem in exterior hydrodynamics*, in Distributed Parameter Control Systems, New Trends and Applications, G. Chen, B. Lee, W. Littman, and L. Markus, eds., Marcel Dekker, New York, 1991, pp. 385–417.
[25] S. Sritharan, *Dynamic programming of the Navier-Stokes equations*, System Control Lett., 16 (1991), pp. 299–307.
[26] S. Sritharan, *An optimal control problem in exterior hydrodynamics*, Roy. Soc. Edinburgh Proc. A, 121 (1992), pp. 5–32.
[27] S. Sritharan, *On Hamilton-Jacobi equation in infinite dimensions*, in Nonlinear Problems in Aviation and Aerospace, S. Sivasundaram, ed., Embry-Riddle University Press, Daytona Beach, FL, 1994, pp. 631–638.
[28] R. Temam, *Navier-Stokes Equations*, North-Holland, Amsterdam, 1979.
[29] M. Triebel, *Interpolation Theory, Function Spaces, Differential Operators*, Veb Deutscher Verlag der Wissenschaften, Berlin, 1978; also, North-Holland, Amsterdam, New York, 1978.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 3, pp. 895–909, May 1998

005

EXISTENCE OF AN OPTIMAL SOLUTION OF A SHAPE CONTROL PROBLEM FOR THE STATIONARY NAVIER–STOKES EQUATIONS*

MAX D. GUNZBURGER† AND HONGCHUL KIM‡

Abstract. This paper is concerned with an optimal shape control problem for the stationary Navier–Stokes system. A two-dimensional channel flow of an incompressible, viscous fluid is examined to determine the shape of a bump on a part of the boundary that minimizes the viscous drag. After giving a precise formulation of the extremal problem in a function analytic setting, it is shown that optimal solutions exist.

Key words. shape control, optimal design, Navier–Stokes equations, drag minimization

AMS subject classifications. 49J20, 76D05

PII. S0363012994276123

1. Introduction.

The control of fluid flow has long been a subject of interest to engineers and scientists and, recently, due to interest in more complex technological applications, it has also become a subject of intense study in the mathematical community. Broadly speaking, optimal control problems can be characterized by some specified physical objective, e.g., drag reduction, and by a means of achieving that objective, i.e., by specifying the control mechanisms. The latter can be divided into two classes, namely, value controls and shape controls. Value controls include data adjustments such as external body forces, boundary stresses, boundary velocities, heat sources, and heat fluxes and temperatures on the boundary. Meeting the objective through the use of shape controls requires the identification of a shape of the domain, i.e., of the boundary, among a specified class of domains. In this sense, a shape control problem can be viewed as a "free boundary problem."

It is believed that the Navier–Stokes equations describe general flows of fluids ranging from certain gas motions to the lubrication of ball bearings. Thus, optimal shape control problems associated with the Navier–Stokes equations, if settled successfully, have wide and valuable application to aerodynamic and hydrodynamic problems such as the design of car hoods, airplane wings, inlet shapes for jet engines, etc. In this paper, we are concerned with some mathematical issues in a shape control problem associated with flows governed by the stationary, two-dimensional, incompressible Navier–Stokes equations. Such studies are in their infancy and at present, only a scant mathematical literature is available. One of the first studies devoted to an optimal shape design problem for the Navier–Stokes equations is found in [16]; there, attempts were made at determining a minimum drag profile submerged in a homogeneous, steady, viscous fluid by utilizing optimal control theories for distributed parameter systems. In [9], a finite difference method is used in some computational

*Received by the editors October 24, 1994; accepted for publication (in revised form) February 24, 1997.

http://www.siam.org/journals/sicon/36-3/27612.html

†Department of Mathematics, Iowa State University, Ames, IA 50011-2064 (gunzburg@iastate.edu). The research of this author was supported in part by Office of Naval Research grant N00014-91-J-1493 and by Air Force Office of Scientific Research grants AFOSR-93-1-0061, AFOSR-93-1-0280.

‡Department of Mathematics, Seoul National University, Seoul 151-742, Korea. Current address: Department of Mathematics, Kangnŭng National University, Kangnŭng 210-702, Korea. The research of this author was supported in part by the Basic Science Research Institute Program of the Korean Ministry of Education and by GARC.

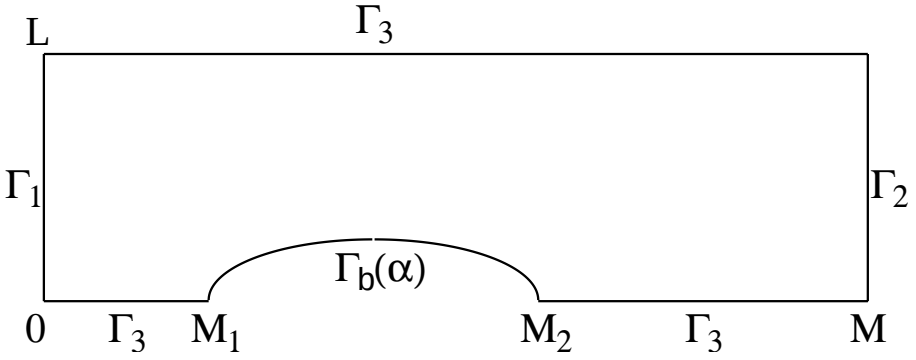


FIG. 1. *The domain* $\Omega(\alpha)$ *for flow through a channel with a bump.*

experiments for determining minimal drag profiles. A successful application of optimal shape theory in fluid mechanics can be found in the design of riblets as a drag reduction device by considering a simplified boundary layer approximation of the Navier–Stokes equations; see [1]. For drag reduction in linear Stokes flow, some rigorous mathematical results for the sensitivity analysis are given in [19]. While the general objective in these studies was drag reduction, one may also consider optimal shapes for other purposes, e.g., lift enhancement, the location of transitional points from laminar to turbulent flow, etc.

In this paper, we deal with a specific drag minimization problem in two dimensions. However, the approach used is discussed in general terms and is applicable to many other optimal control problems involving different objectives and classes of shape controls. Our aim here is to provide a systematic formulation of a problem in which the viscous drag is minimized through the use of shape modifications and to show the existence of optimal solutions. In subsequent papers we will discuss sensitivity and numerical analyses for the shape control problem; these analyses will include the derivation of a formula for the shape gradient that may be used within an optimization algorithm to determine optimal shapes.

2. The model problem. We consider the two-dimensional incompressible flow of a viscous fluid passing through a channel having a finite depth; see Figure 1. Let $\mathbf{g}_1$ and $\mathbf{g}_2$ be the preset velocities at the inflow Γ_1 and outflow Γ_2 of the channel, respectively. Along the bottom and top sides of the channel the velocity vanishes. (If some other type of boundary conditions, e.g., specifying some components of the stress vector, is specified along the left or right or top boundary, the results given below are still valid, although the analyses may be more complicated.) The arc $\Gamma_b(\alpha)$, which is part of the bottom boundary, represents the bump, which is to be determined.

Let the boundary shape corresponding to the bump be represented by the graph of the curve $\alpha : [M_1, M_2] \to [0, L)$. (We thus avoid the separation of the in- and out-flows.) The domain $\Omega(\alpha)$ is composed of two fixed rectangles and a domain with an unknown boundary. Thus, the domain $\Omega(\alpha)$ is determined by the shape of the unknown boundary $\Gamma_b(\alpha)$, which we assume is given by

$$\Gamma_b(\alpha) = \{(x_1, x_2) \in [M_1, M_2] \times [0, L) \mid x_2 = \alpha(x_1)\},$$

where $\alpha(x_1)$ is a function to be determined by the optimization process. Let $\Gamma(\alpha) = \partial\Omega(\alpha) = \cup_{i=1}^{3}\Gamma_i \cup \Gamma_b(\alpha)$, where $\Gamma_3 = \Gamma \backslash (\Gamma_1 \cup \Gamma_2 \cup \Gamma_b(\alpha))$. Assume that both end points

of $\Gamma_b(\alpha)$ are fixed (at $x_1 = M_1, x_2 = 0$ and at $x_1 = M_2, x_2 = 0$) for all admissible domains. Since the domain $\Omega(\alpha)$ is determined by the shape of $\Gamma_b(\alpha)$, one may define the admissible family of curves defining $\Gamma_b(\alpha)$ as follows:

$$\mathcal{U}_{ad} = \{\, \alpha \in C^{0,1}([M_1, M_2]) \mid 0 \le \alpha(x_1) \le L_0 < L,\ |\alpha(x_1) - \alpha(\overline{x}_1)| \le \beta |x_1 - \overline{x}_1|, \\ \forall x_1,\ \overline{x}_1 \in [M_1, M_2],\ \text{and}\ \alpha(M_1) = \alpha(M_2) = 0 \,\},$$

where β is a given *fixed* positive constant. We have denoted the set of Lipschitz continuous functions in $[M_1, M_2]$ by the symbol $C^{0,1}([M_1, M_2])$.

The condition $|\alpha(x_1) - \alpha(\overline{x}_1)| \le \beta |x_1 - \overline{x}_1|$ is invoked to prevent the "blow-up" of the boundary, i.e., to suppress *excessive* oscillations of $\Gamma_b(\alpha)$ and to support the uniform Lipschitz continuity of the moving boundary $\Gamma(\alpha)$. In [17], an example is provided illustrating the observation that when the boundary is allowed to oscillate, the limit of a sequence that minimizes the objective functional may have nothing to do with the given optimization problem. Furthermore, equipped with this condition, it is possible to show that all allowable domains $\Omega(\alpha)$ have the uniform extension property (see [5] and section 4).

We consider, for each $\alpha \in \mathcal{U}_{ad}$, the viscous, incompressible, stationary Navier–Stokes equations in nondimensional form in $\Omega(\alpha)$. Let $\mathbf{u} = (u_1, u_2)^T$ denote the velocity and p the pressure. Then, we have

$$-\nu \Delta \mathbf{u} + (\mathbf{u} \cdot \nabla)\mathbf{u} + \nabla p = \mathbf{f} \qquad \text{in } \Omega(\alpha) \tag{2.1}$$

and

$$\nabla \cdot \mathbf{u} = 0 \qquad \text{in } \Omega(\alpha) \tag{2.2}$$

along with the Dirichlet boundary conditions

$$\mathbf{u} = \mathbf{g} = \begin{cases} \mathbf{g}_1 \text{ on } \Gamma_1, \\ \mathbf{g}_2 \text{ on } \Gamma_2, \\ \mathbf{0} \text{ on } \Gamma_3 \cup \Gamma_b(\alpha), \end{cases} \tag{2.3}$$

where $\mathbf{f}$ and $\mathbf{g}_i$, $i = 1, 2$, are given, fixed functions. Here, Δ and ∇ denote the Laplacian and gradient operators in $\mathbb{R}$, respectively, $\mathbf{f}$ denotes the given external force, and, in the nondimensional form of the Navier–Stokes equations, ν denotes the reciprocal of the Reynolds number *Re*. Note that the constant density has been absorbed into the pressure and the body force. For the compatibility and regularity of solutions, we assume

$$\text{support of } \mathbf{g}_i \subset \Gamma_i \quad \text{and} \quad \int_{\Gamma_1} \mathbf{g}_1 \cdot \mathbf{n}\, d\Gamma + \int_{\Gamma_2} \mathbf{g}_2 \cdot \mathbf{n}\, d\Gamma = 0\,. \tag{2.4}$$

One can examine several objectives for determining the shape of the bump, e.g., the reduction of the drag due to viscosity or the identification of the velocity at a fixed vertical slit downstream of the bump. To fix ideas, we focus on the minimization of the cost functional (or, in the terminology of shape optimization, the design performance functional)

$$\begin{aligned} \mathcal{J}(\alpha) = \mathcal{J}\Big(\Omega(\alpha), \mathbf{u}(\alpha)\Big) &= 2\nu \int_{\Omega(\alpha)} D(\mathbf{u}) : D(\mathbf{u})\, d\Omega \\ &= \frac{\nu}{2} \sum_{i,j=1}^{2} \int_{\Omega(\alpha)} \left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right)^2 d\Omega\,, \end{aligned} \tag{2.5}$$

where $\mathbf{u}(\alpha)$ is a solution of (2.1)–(2.3) in $\Omega(\alpha)$ and $D(\mathbf{u}) = \frac{1}{2}(\nabla\mathbf{u} + (\nabla\mathbf{u})^T)$ is the deformation tensor for the flow $\mathbf{u}$. This functional represents the rate of energy dissipation due to deformation. Physically, except for an unimportant additive constant whose value depends on the boundary data $\mathbf{g}_1$ and $\mathbf{g}_2$, this functional represents the viscous drag of the flow. In (2.5), the colon denotes the scalar product operator between two tensors. (Again, our results remain valid if we consider other functionals such as the identification of the velocity at a location downstream of the bump.)

The extremal problem we consider is then given as follows:

$$\begin{aligned} &\min_{\alpha\in\mathcal{U}_{ad}} \mathcal{J}(\Omega(\alpha), \mathbf{u}(\alpha)) \quad \text{such that, for some } p(\alpha), \\ &(\mathbf{u}(\alpha), p(\alpha)) \text{ is a solution of (2.1)–(2.3) in } \Omega(\alpha)\,. \end{aligned} \tag{2.6}$$

In preparation for showing the existence of optimal solutions satisfying (2.6), we recast, in the next section, this problem into a precise function space setting.

3. Function space setting of extremal problem.

3.1. Notation. Throughout, depending on the context, $\mathcal{I}$ will denote the identity mapping or the identity matrix; C denotes a generic constant whose value also depends on context. We denote by $H^s(\mathcal{D})$, $s \in \mathbb{R}$, the standard Sobolev space of order s with respect to the set $\mathcal{D}$, which is either the flow domain Ω, or its boundary Γ, or part of its boundary. Whenever m is a nonnegative integer, the inner product over $H^m(\mathcal{D})$ is given by

$$(f,g)_{m,\mathcal{D}} = (f,g)_{0,\mathcal{D}} + \sum_{0<|\sigma|\le m} (D^\sigma f, D^\sigma g)_{0,\mathcal{D}}\,,$$

where $(f,g)_0 = \int_{\mathcal{D}} fg\,d\mathcal{D}$ denotes the inner product over $H^0(\mathcal{D}) = L^2(\mathcal{D})$ and σ denotes a multi-index. Hence, we naturally associate the norm on $H^m(\mathcal{D})$ with $\|f\|_{m,\mathcal{D}} = \sqrt{(f,f)_{m,\mathcal{D}}}$. Whenever there is no chance for confusion, we will, for the flow domain $\Omega(\alpha)$, let $(\cdot,\cdot)_{m,\Omega(\alpha)} = (\cdot,\cdot)_m$ and $\|\cdot\|_{m,\Omega(\alpha)} = \|\cdot\|_m$.

For vector-valued functions and spaces, we use boldface notation. For example, $\mathbf{H}^s(\Omega) = [H^s(\Omega)]^n$ denotes the space of $\mathbb{R}$-valued functions such that each component belongs to $H^s(\Omega)$. Of special interest to us is the space

$$\mathbf{H}^1(\Omega) = \left\{ v_j \in L^2(\Omega) \;\middle|\; \frac{\partial v_j}{\partial x_k} \in L^2(\Omega) \quad \text{for } j,k = 1,2 \right\}$$

equipped with the norm $\|\mathbf{v}\|_1 = (\sum_{i=1}^2 \|v_i\|_1^2)^{1/2}$. For $\Gamma_s \subset \Gamma = \partial\Omega$ with nonzero measure, we also consider the subspace

$$\mathbf{H}^1_{\Gamma_s}(\Omega) = \{\, \mathbf{v} \in \mathbf{H}^1(\Omega) \,|\, \mathbf{v} = \mathbf{0} \text{ on } \Gamma_s \,\};$$

we let $\mathbf{H}^1_0(\Omega) = \mathbf{H}^1_\Gamma(\Omega)$. For any $\mathbf{v} \in \mathbf{H}^1(\Omega)$, we let

$$|\!|\!|\mathbf{v}|\!|\!| = 2\left(\int_\Omega D(\mathbf{v}) : D(\mathbf{v})\,d\Omega\right)^{1/2} = \frac{1}{2}\Big(\sum_{i,j=1}^2 \|\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i}\|_0^2\Big)^{1/2}.$$

By applying Korn's inequality and a compactness argument, we obtain the following result.

LEMMA 3.1. *Let Ω be a Lipschitz continuous bounded domain and let Γ_s be a subset of Γ, the boundary of Ω, with a positive measure. Then, there exists a positive constant C such that*

$$|||\mathbf{v}||| \geq C\|\mathbf{v}\|_1 \quad \textit{for all } \mathbf{v} \in \mathbf{H}^1_{\Gamma_s}(\Omega)\,. \qquad \square \tag{3.1}$$

Note that the constant C in (3.1) is independent of the choice of $\mathbf{v}$. Thus, we have that $|||\cdot|||$ is a norm which is equivalent to the norm $\|\cdot\|_{1,\Omega}$ on $\mathbf{H}^1_{\Gamma_s}(\Omega)$. Hence, if we take the inner product on $\mathbf{H}^1_{\Gamma_s}(\Omega)$ as $((\mathbf{u},\mathbf{v}))_1 = 2(D(\mathbf{u}), D(\mathbf{v}))_0$, then $|||\mathbf{u}||| = ((\mathbf{u},\mathbf{u}))_1^{1/2}$.

For each $\alpha \in \mathcal{C}^{0,1}([M_1, M_2])$, let $\Gamma_0(\alpha) = \Gamma_3 \cup \Gamma_b(\alpha)$ and $\Gamma_g = \Gamma_1 \cup \Gamma_2$ so that $\overline{\Gamma} = \overline{\Gamma_0(\alpha)} \cup \overline{\Gamma_g}$. Since $\mathbf{u} = \mathbf{0}$ on $\Gamma_0(\alpha)$, we may take a generalized velocity space to be

$$\mathbf{V}_\alpha = \mathbf{H}^1_{\Gamma_0(\alpha)}(\Omega(\alpha)) = \{\, \mathbf{u} \in \mathbf{H}^1(\Omega(\alpha)) \;\mid\; \mathbf{u} = \mathbf{0} \text{ on } \Gamma_0(\alpha) \,\};$$

$\mathbf{V}_\alpha$ is the space of $\mathbf{H}^1(\Omega(\alpha))$-functions that vanish on $\Gamma_0(\alpha)$; i.e., $\mathbf{V}_\alpha$ is the space on which the *homogeneous* essential boundary conditions are imposed. Let $\mathbf{V}^*_\alpha$ be the dual space of $\mathbf{V}_\alpha$. Note that $\mathbf{V}^*_\alpha$ is a subspace of $\mathbf{H}^{-1}(\Omega(\alpha))$, where the latter is the dual space of $\mathbf{H}^1_0(\Omega(\alpha))$. The duality between $\mathbf{V}^*_\alpha$ and $\mathbf{V}_\alpha$ is denoted by $\langle\cdot,\cdot\rangle_{-1}$.

Let

$$\mathbf{L}^2_g(\Gamma) = \{\, \mathbf{s} \in \mathbf{L}^2(\Gamma) \mid \mathbf{s} = \mathbf{0} \text{ on } \Gamma_0(\alpha) \,\}$$

and let $\gamma_g : \mathbf{V}_\alpha \to \mathbf{L}^2_g(\Gamma)$ be the trace mapping. Let us define

$$\mathbf{W}_\alpha = \gamma_g(\mathbf{V}_\alpha)\,.$$

Let $\mathbf{W}^*_\alpha$ denote its dual space and let $\langle\cdot,\cdot\rangle_{-1/2,\Gamma_g}$ denote the duality pairing between $\mathbf{W}^*_\alpha$ and $\mathbf{W}_\alpha$. For each $l \geq 0$, we denote by $\mathbf{H}^l(\Gamma_g)$ the space of the restrictions to Γ_g of the functions of $\mathbf{H}^l(\Gamma)$, and by $\mathbf{H}^{-l}(\Gamma_g)$, its dual space. It is clear that the restrictions to Γ_g of the functions belonging to $\mathbf{W}_\alpha$ form a closed subspace of $\mathbf{H}^{1/2}(\Gamma_g)$.

Now, let $\mathbf{s}$ be an element of $\mathbf{W}_\alpha$. It is well known that $\mathbf{W}_\alpha$ is a Hilbert space with the norm

$$\|\mathbf{s}\|_{1/2,\Gamma_g} = \inf_{\mathbf{v}\in\mathbf{V}_\alpha,\, \gamma_g\mathbf{v}=\mathbf{s}} \|\mathbf{v}\|_{1,\Omega(\alpha)} \qquad \forall\, \mathbf{s} \in \mathbf{W}_\alpha\,.$$

Let $\mathbf{s}^*$ belong to $\mathbf{W}^*_\alpha$. By the definition of the dual norm, we note that

$$\|\mathbf{s}^*\|_{-1/2,\Gamma_g} = \sup_{\mathbf{s}\in\mathbf{W}_\alpha,\, \mathbf{s}\neq\mathbf{0}} \frac{\langle \mathbf{s}^*, \mathbf{s}\rangle_{-1/2,\Gamma_g}}{\|\mathbf{s}\|_{1/2,\Gamma_g}} \qquad \forall\, \mathbf{s}^* \in \mathbf{W}^*_\alpha\,.$$

For later use, we can derive an alternate definition for the norm $\|\cdot\|_{-1/2,\Gamma_g}$.

LEMMA 3.2. *It holds that*

$$\|\mathbf{s}^*\|_{-1/2,\Gamma_g} = \sup_{\mathbf{v}\in\mathbf{V}_\alpha,\, \mathbf{v}\neq\mathbf{0}} \frac{\langle \mathbf{s}^*, \gamma_g\mathbf{v}\rangle_{-1/2,\Gamma_g}}{\|\mathbf{v}\|_{1,\Omega(\alpha)}} \qquad \forall\, \mathbf{s}^* \in \mathbf{W}^*_\alpha\,. \tag{3.2}$$

Proof. Using the continuity of the trace mapping, it follows that for any $\mathbf{s}^* \in \mathbf{W}^*_\alpha$,

$$\begin{aligned}\langle \mathbf{s}^*, \gamma_g\mathbf{v}\rangle_{-1/2,\Gamma_g} &\leq \|\mathbf{s}^*\|_{-1/2,\Gamma_g}\|\gamma_g\mathbf{v}\|_{1/2,\Gamma_g} \\ &\leq \|\mathbf{s}^*\|_{-1/2,\Gamma_g}\|\mathbf{v}\|_{1,\Omega(\alpha)} \quad \forall\, \mathbf{v} \in \mathbf{V}_\alpha\,.\end{aligned}$$

Hence, we obtain

$$\sup_{\mathbf{v}\in\mathbf{V}_\alpha,\,\mathbf{v}\neq\mathbf{0}} \frac{\langle \mathbf{s}^*, \gamma_g\mathbf{v}\rangle_{-1/2,\Gamma_g}}{\|\mathbf{v}\|_{1,\Omega(\alpha)}} \le \|\mathbf{s}^*\|_{-1/2,\Gamma_g} \qquad \forall\, \mathbf{s}^* \in \mathbf{W}_\alpha^*\,. \tag{3.3}$$

Now, given $\mathbf{s}\in\mathbf{W}_\alpha$, let $\boldsymbol{\xi}$ be an element in $\mathbf{V}_\alpha$ such that $\gamma_g\boldsymbol{\xi}=\mathbf{s}$ and

$$\int_{\Omega(\alpha)}\Big(\nabla\boldsymbol{\xi} : \nabla\boldsymbol{\eta} + \boldsymbol{\xi}\cdot\boldsymbol{\eta}\Big)\,d\Omega = 0 \qquad \forall\,\boldsymbol{\eta}\in\mathbf{H}_0^1(\Omega(\alpha))\,.$$

Clearly, such a $\boldsymbol{\xi}$ is uniquely determined and $\|\boldsymbol{\xi}\|_{1,\Omega(\alpha)} \le \|\boldsymbol{\psi}\|_{1,\Omega(\alpha)}$ for all $\boldsymbol{\psi}\in\mathbf{V}_\alpha$ such that $\gamma_g\boldsymbol{\psi}=\mathbf{s}=\gamma_g\boldsymbol{\xi}$; then, by the definition of $\|\cdot\|_{1/2,\Gamma_g}$, it follows that $\|\mathbf{s}\|_{1/2,\Gamma_g} = \|\boldsymbol{\xi}\|_{1,\Omega(\alpha)}$. As a result, we have that for all $\mathbf{s}\in\mathbf{W}_\alpha$ and $\mathbf{s}^*\in\mathbf{W}_\alpha^*$,

$$\frac{\langle \mathbf{s}^*, \mathbf{s}\rangle_{-1/2,\Gamma_g}}{\|\mathbf{s}\|_{1/2,\Gamma_g}} = \frac{\langle \mathbf{s}^*, \gamma_g\boldsymbol{\xi}\rangle_{-1/2,\Gamma_g}}{\|\boldsymbol{\xi}\|_{1,\Omega(\alpha)}} \le \sup_{\mathbf{v}\in\mathbf{V}_\alpha,\,\mathbf{v}\neq\mathbf{0}} \frac{\langle \mathbf{s}^*, \gamma_g\mathbf{v}\rangle_{-1/2,\Gamma_g}}{\|\mathbf{v}\|_{1,\Omega(\alpha)}}\,.$$

Then, from the definition of $\|\cdot\|_{-1/2,\Gamma_g}$, we have that

$$\|\mathbf{s}^*\|_{-1/2,\Gamma_g} \le \sup_{\mathbf{v}\in\mathbf{V}_\alpha,\,\mathbf{v}\neq\mathbf{0}} \frac{\langle \mathbf{s}^*, \gamma_g\mathbf{v}\rangle_{-1/2,\Gamma_g}}{\|\mathbf{v}\|_{1,\Omega(\alpha)}} \qquad \forall\,\mathbf{s}^*\in\mathbf{W}_\alpha^*\,. \tag{3.4}$$

The combination of (3.3) and (3.4) yields the desired result. ∎

Since the pressure is determined only up to a constant in the mathematical formulation of the Navier–Stokes equations with velocity boundary conditions, we define the space of generalized pressures to be

$$S = \Big\{\, p\in L^2(\Omega(\alpha)) \;\Big|\; \int_{\Omega(\alpha)} p\,d\Omega = 0 \,\Big\}\,.$$

Thus, S consists of square integrable functions having zero mean over $\Omega(\alpha)$.

3.2. Weak variational formulation of the state equations. For the weak variational formulation, we will use the forms

$$\begin{aligned} a(\mathbf{u},\mathbf{v}) &= ((\mathbf{u},\mathbf{v}))_1 = 2\int_{\Omega(\alpha)} D(\mathbf{u}) : D(\mathbf{v})\,d\Omega \\ &= \frac{1}{2}\sum_{i,j=1}^{2}\int_{\Omega(\alpha)}\left(\frac{\partial u_i}{\partial x_j}+\frac{\partial u_j}{\partial x_i}\right)\left(\frac{\partial v_i}{\partial x_j}+\frac{\partial v_j}{\partial x_i}\right)d\Omega \quad \forall\,\mathbf{u},\,\mathbf{v}\in\mathbf{H}^1(\Omega(\alpha)), \end{aligned}$$

$$b(\mathbf{v},q) = -\int_{\Omega(\alpha)} q\,\nabla\cdot\mathbf{v}\,d\Omega = -\sum_{i=1}^{2}\int_{\Omega(\alpha)} q\,\frac{\partial v_i}{\partial x_i}\,d\Omega \quad \forall\,\mathbf{u}\in\mathbf{H}^1(\Omega(\alpha))\,, q\in L^2(\Omega(\alpha))\,,$$

and

$$c(\mathbf{w},\mathbf{u},\mathbf{v}) = \int_{\Omega(\alpha)} (\mathbf{w}\cdot\nabla)\mathbf{u}\cdot\mathbf{v}\,d\Omega = \sum_{i,j=1}^{2}\int_{\Omega(\alpha)} w_j\frac{\partial u_i}{\partial x_j}v_i\,d\Omega \quad \forall\,\mathbf{u},\,\mathbf{v},\,\mathbf{w}\in\mathbf{H}^1(\Omega(\alpha))\,.$$

Obviously, $a(\cdot,\cdot)$ is a continuous bilinear form on $\mathbf{H}^1(\Omega(\alpha))\times\mathbf{H}^1(\Omega(\alpha))$ and $b(\cdot,\cdot)$ is a continuous bilinear form on $\mathbf{H}^1(\Omega(\alpha))\times L^2(\Omega(\alpha))$; also, $c(\cdot,\cdot,\cdot)$ is a continuous trilinear form on $\mathbf{H}^1(\Omega(\alpha))\times\mathbf{H}^1(\Omega(\alpha))\times\mathbf{H}^1(\Omega(\alpha))$ which can be verified by the

Sobolev embedding of $\mathbf{H}^1(\Omega(\alpha)) \subset \mathbf{L}^4(\Omega(\alpha))$ and Hölder's inequality. Moreover, as a consequence of (3.1), we have the coercivity property

$$a(\mathbf{v},\mathbf{v}) \geq C\|\mathbf{v}\|_1^2 \quad \forall\, \mathbf{v} \in \mathbf{V}_\alpha$$

and the inf-sup condition (or LBB condition)

$$\inf_{q\in S} \sup_{\mathbf{v}\in \mathbf{H}_0^1(\Omega)} \frac{b(\mathbf{v},q)}{\|\mathbf{v}\|_1 \|q\|_0} \geq C\,. \tag{3.5}$$

For details concerning these forms and their properties, one may consult [7], [8], [10], [13], [21], or [22].

One can show that (2.1)–(2.3) have the following weak formulation: for each $\alpha \in \mathcal{U}_{ad}$, find $\mathbf{u} \in \mathbf{V}_\alpha$, $p \in S$, and $\mathbf{t} \in \mathbf{W}_\alpha^*$ satisfying

$$\nu a(\mathbf{u},\mathbf{v}) + c(\mathbf{u},\mathbf{u},\mathbf{v}) + b(\mathbf{v},p) - \langle \mathbf{t}, \gamma_g \mathbf{v}\rangle_{-1/2,\Gamma_g} = \langle \mathbf{f},\mathbf{v}\rangle_{-1} \quad \forall \mathbf{v} \in \mathbf{V}_\alpha\,, \tag{3.6}$$

$$b(\mathbf{u},q) = 0 \quad \forall q \in S\,, \tag{3.7}$$

and

$$\langle \mathbf{s},\mathbf{u}\rangle_{-1/2,\Gamma_g} = \langle \mathbf{s},\mathbf{g}\rangle_{-1/2,\Gamma_g} \quad \forall \mathbf{s} \in \mathbf{W}_\alpha^*\,. \tag{3.8}$$

In showing that (3.6) is a weak formulation of (2.1), it is convenient to replace the viscous term in the latter with $2\nu\nabla\cdot(D(\mathbf{u}))$; the equivalence of the two forms of the viscous terms follows from the incompressibility constraint (2.2). Note that the boundary condition on the velocity is enforced weakly through the use of Lagrange multipliers; see [2], [11], and [12].

It can be shown that, in the sense of distributions, $\mathbf{t}$ is the stress vector on Γ_g; i.e.,

$$\mathbf{t} = -p\mathbf{n} + 2\nu D(\mathbf{u})\cdot\mathbf{n} = -p\mathbf{n} + \nu(\nabla\mathbf{u} + (\nabla\mathbf{u})^T)\cdot\mathbf{n} \quad \text{on } \Gamma_g\,.$$

Existence and uniqueness results for solutions of the system (3.6)–(3.8) are contained in the following theorem; for a proof, one may consult [8], [11], [13], [18], [21], or [22].

THEOREM 3.3. *Let $\alpha \in \mathcal{U}_{ad}$ be fixed and let the data satisfy $\mathbf{f} \in \mathbf{V}_\alpha^*$, $\mathbf{g} \in \mathbf{W}_\alpha$, and the compatibility condition* (2.4). *Then,*

I. *there exists at least one solution $(\mathbf{u},p,\mathbf{t}) \in \mathbf{V}_\alpha \times S \times \mathbf{W}_\alpha^*$ of* (3.6)–(3.8)*;*

II. *if $\mathcal{S}$ denotes a set of velocity fields that are solutions of* (3.6)–(3.8)*, then $\mathcal{S}$ is closed in $\mathbf{H}^1(\Omega(\alpha))$ and is compact in $\mathbf{L}^2(\Omega(\alpha))$; and*

III. *if $\nu > \nu_0(\Omega(\alpha);\mathbf{f},\mathbf{g})$ for some positive constant ν_0 whose value is determined by the given data, then $\mathcal{S}$ is composed of a single element.* □

Note that the solutions of (3.6)–(3.8) exist for any Reynolds number; however, III implies that uniqueness can be guaranteed only for "large enough" values of ν or for "small enough" values of the data $(\mathbf{f},\mathbf{g}) \in \mathbf{V}_\alpha^* \times \mathbf{W}_\alpha$.

3.3. The extremal problem. In the notation introduced in section 3.2, the cost functional $\mathcal{J}$ defined in (2.5) can be expressed in the form

$$\mathcal{J}(\alpha) = \mathcal{J}\Big(\Omega(\alpha),\mathbf{u}(\alpha)\Big) = 2\nu \int_{\Omega(\alpha)} D(\mathbf{u}) : D(\mathbf{u})\, d\Omega = \nu((\mathbf{u},\mathbf{u}))_1 = \nu|\!|\!|\mathbf{u}|\!|\!|^2\,. \tag{3.9}$$

We introduce the admissibility set of controls and velocities

$$\mathcal{V}_{ad} = \Big\{ (\alpha, \mathbf{u}(\alpha)) \in \mathcal{U}_{ad} \times \mathbf{V}_\alpha \;\Big|\; \mathcal{J}(\alpha, \mathbf{u}(\alpha)) < \infty, \text{ and there exist } p(\alpha) \in S \text{ and } \mathbf{t}(\alpha) \in \mathbf{W}_\alpha^* \text{ such that } (\mathbf{u}(\alpha), p(\alpha), \mathbf{t}(\alpha)) \text{ is a solution of (3.6)–(3.8)} \Big\}.$$

Then, the extremal problem (2.6) can be restated in the following precise form:

$$\min_{(\alpha, \mathbf{u}(\alpha)) \in \mathcal{V}_{ad}} \mathcal{J}(\alpha, \mathbf{u}(\alpha)) \,. \tag{3.10}$$

4. Convergence notions in sequences of domains. We introduce some concepts dealing with convergence in function spaces and domains.

Let X be a normed vector space, X^* its dual space, and let $\langle \cdot, \cdot \rangle_{X^*}$ denote the duality pairing between functions belonging to X^* and X. We use the notation "$x_n \rightharpoonup x$" to denote the *weak convergence* of a sequence $\{x_n\}$ in X to x, i.e.,

$$x_n \rightharpoonup x \iff \langle f, x_n \rangle_{X^*} \xrightarrow{n \to \infty} \langle f, x \rangle_{X^*} \quad \forall\, f \in X^* \,.$$

Let Y be a subspace of X. Y is called a *weakly closed subspace* of X if, for every sequence $\{x_n\}$ in Y, whenever $x_n \rightharpoonup x^*$ in X, we have $x^* \in Y$. In connection with optimal controls, the following result (see [20]) is very useful for verifying the weak convergence of sequences.

LEMMA 4.1. *Let X be a normed vector space. A sequence $\{x_n\}$ in X converges weakly to $x \in X$ if and only if $\sup_n \|x_n\|_X < \infty$ and $\langle f, x_n \rangle \to \langle f, x \rangle$ for each $f \in F$, where F is a linear span of a set which is dense in X^*. Moreover, if X is a reflexive Banach space, each bounded sequence in X contains a weakly convergent subsequence.* □

In general, the most crucial concept in optimization is semicontinuity, especially when the cost functional contains the gradient of the function. Let S be a subset of X and $\mathcal{K}$ be a real functional on S. We say that $\mathcal{K}$ is (*weakly*) *lower semicontinuous* if, for every sequence $\{x_n\}$ in S, whenever

$$x_n \longrightarrow x \;\; (x_n \rightharpoonup x) \quad \text{in } X \,,$$

we have

$$\liminf_{n \to \infty} \mathcal{K}(x_n) \;\geq\; \mathcal{K}(x) \,.$$

Note that the notion of (weak) lower semicontinuity is a local property.

To deal with domain optimization, we need to define an appropriate convergence criterion with respect to domains. Since domains and corresponding function spaces are changing, we need a fixed domain $\widehat{\Omega}$ such that $\cup_{\alpha \in \mathcal{U}_{ad}} \Omega(\alpha) \subset \widehat{\Omega}$ in order to discuss the convergence of domains and corresponding functions. Thus, in the current setting, we let $\widehat{\Omega}$ and Ω_0 denote the interiors of the rectangular shaded regions depicted in the top and bottom pictures in Figure 2, respectively, so that

$$\bigcup_{\alpha \in \mathcal{U}_{ad}} \Omega(\alpha) \subset \overline{\widehat{\Omega}} \qquad \text{and} \qquad \bigcup_{\alpha \in \mathcal{U}_{ad}} \Gamma_b(\alpha) \subset \overline{\Omega}_0 \,.$$

Note that the definition of $\mathcal{U}_{ad}$ implies that the graph of α for each $\alpha \in \mathcal{U}_{ad}$ lies in the rectangular region Ω_0.

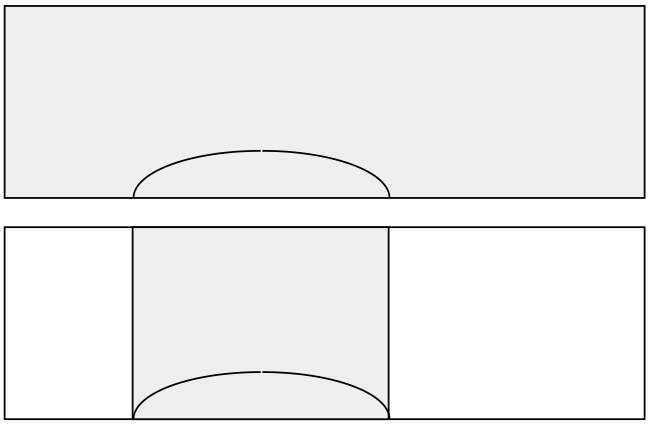


FIG. 2. *The domains* $\widehat{\Omega}$ (*top*) *and* Ω_0 (*bottom*).

A domain class for which optimal shape problems usually have an optimal solution has been studied in [5] and [6]. It was shown in [5] that the set of domains with the cone property is compact for the strong $L^2(\widehat{\Omega})$-topology of the characteristic functions of its elements. Let χ_Ω denote the characteristic function of the domain Ω which is included in $\widehat{\Omega}$. The convergence of the sequence $\{\Omega_m\}$ of domains having the cone property may be defined by

$$\Omega_m \to \Omega \iff \int_{\widehat{\Omega}} |\chi_{\Omega_m} - \chi_\Omega|^2 \, d\Omega \to 0 \quad \text{as } m \to \infty \,.$$

This method of convergence using characteristic functions is often used to solve some specific shape optimization problems such as the transmission problem governed by a pair of different elliptic equations over adjacent regions; see, e.g., [4] and [16]. However, since the convergence of characteristic functions does not preserve the regularity of domains, it is not appropriate for dealing with general shape optimization problems in which the regularity of domains is a concern.

In our case, the domains $\{\Omega(\alpha)\}_{\alpha \in \mathcal{U}_{ad}}$ are determined by the variable part $\Gamma_b(\alpha)$ of the boundary $\Gamma(\alpha)$. Thus, it is more natural to define the convergence of domains in terms of functions α belonging to $\mathcal{U}_{ad}$. Let $\{\alpha_n\}$ be a sequence in $\mathcal{U}_{ad}$. For each $\alpha_n \in \mathcal{U}_{ad}$, let $\Omega_n = \Omega(\alpha_n)$. We define the convergence of Ω_n to $\Omega(\alpha)$ by

$$\Omega_n \to \Omega(\alpha) \iff \|\alpha_n - \alpha\|_\infty = \max_{M_1 \le x_1 \le M_2} |\alpha_n(x_1) - \alpha(x_1)| \to 0 \,.$$

Remark. In general shape optimization problems, more stringent topologies are often introduced to enforce the convergence of geometrical elements; see, e.g., [14] and [16]. When the inclusive relation between subdomains of $\mathbb{R}$ is the main issue, as in problems of domain identification, the topology induced by the following *Hausdorff metric* is widely used: let A and B be two closed subsets of $\mathbb{R}$ and define the Hausdorff metric δ by

$$\delta(A,B) = \max\{\rho(A,B), \rho(B,A)\} \,, \qquad \text{where} \qquad \rho(A,B) = \sup_{x \in A} \inf_{y \in B} |x-y|_{\mathbb{R}} \,.$$

Then, the topology on the closed subsets of $\mathbb{R}$ is defined by

$$A_m \longrightarrow A \iff \delta(A_m, A) \to 0 \,.$$

The most important property of this topology is that it preserves the relation of domain inclusions.

Remark. Another important topology can be used in conjunction with mapping techniques. Let A be a fixed domain in $\mathbb{R}$. Suppose domain perturbations are described by a family of bijective mappings having some regularities, for example, $\mathcal{F}_k = \{T(A) \mid T \in C^k \text{ and } T \text{ is bijective}\}$. Then, the convergence of domains can be defined using the minimal norm $\|T - \mathcal{I}\| + \|T^{-1} - \mathcal{I}\|$ among the mappings T such that $T(A) \in \mathcal{F}_k$.

Recall that $\alpha \in \mathcal{C}^{0,1}([M_1, M_2])$ so that $\Omega(\alpha)$ is Lipschitz continuous. In fact, $\Omega(\alpha)$ is a uniformly Lipschitz continuous domain for each $\alpha \in \mathcal{U}_{ad}$ so that the possibility of domains $\{\Omega(\alpha)\}_{\alpha \in \mathcal{U}_{ad}}$ having a *cusp* is excluded. Hence, for each $\alpha \in \mathcal{U}_{ad}$, the domain $\Omega(\alpha)$ has the uniform extension property which we now discuss.

Let Γ denote the boundary of the domain Ω. Ω is said to have a *cusp* at $x \in \Gamma$ if no affine image in $\overline{\Omega}$ of a finite cone has a vertex at x. We observe that if a certain domain has the *cone property*, then using the homogeneity along a line segment emanating from the vertex of the cone, a function with small support in a neighborhood of the vertex may be perturbed into the whole cone. Based on this fact, the following result is given in [5].

LEMMA 4.2. *The open sets satisfying the cone property are the uniform Lipschitz sets.* □

For domain perturbations, the following extension property given in, e.g., [15], plays a central role.

PROPOSITION 4.3 (Calderón's extension theorem). *For every uniform Lipschitz domain Ω in $\mathbb{R}$ and every positive integer m there exists a linear continuous extension operator*

$$P : H^m(\Omega) \longrightarrow H^m(\mathbb{R}) \tag{4.1}$$

such that for each $f \in H^m(\Omega)$

$$\|Pf\|_{m,\mathbb{R}} \leq C \|f\|_{m,\Omega}\,, \tag{4.2}$$

where the positive constant C depends only on the cone embedded in Ω. □

Let γ_Ω denote the restriction operator from $\mathbb{R}$ to Ω. Then, note that for each m, $\gamma_\Omega \circ P = \mathcal{I}_{H^m(\Omega)}$, the identity map over $H^m(\Omega)$. For this reason, P is often called a *lifting*. For a domain $\Omega \subset \mathbb{R}$, if there exists an extension operator P satisfying (4.1) and (4.2) for m, then the domain Ω is said to have an *m-extension property*. Proposition 4.3 states that Lipschitz continuous domains have the m-extension property for each m. Hence, we have the following compact embedding property for bounded Lipschitz continuous domains which can be proved using Rellich's theorem for the compact embedding of $H_0^{m+1}(\Omega)$ into $H_0^m(\Omega)$ and an approriately defined extension operator P.

LEMMA 4.4. *For a bounded Lipschitz continuous domain Ω, the natural injection of $H^{m+1}(\Omega)$ into $H^m(\Omega)$ is compact.* □

Now, let us return to the setting of the extremal problem (3.10). For any $\mathbf{v} \in \mathbf{V}_\alpha$, let $\widehat{\mathbf{v}}$ be its Calderón extension to $\mathbf{H}^1(\widehat{\Omega})$, i.e.,

$$\widehat{\mathbf{v}} = P_{\widehat{\Omega}} \mathbf{v}\,,$$

where $P_{\widehat{\Omega}}$ is the Calderón extension operator (defined on $\Omega(\alpha)$) to $\widehat{\Omega}$. Then, by Proposition 4.3, there exists a positive constant C such that $\|\widehat{\mathbf{v}}\|_{1,\widehat{\Omega}} \leq C \|\mathbf{v}\|_{1,\Omega(\alpha)}$.

Let $\{\alpha_n\}$ be a sequence in $\mathcal{U}_{ad}$ and let

$$\mathbf{V}_{\alpha_n} = \mathbf{H}^1_{\Gamma_0(\alpha_n)}(\Omega(\alpha_n)) = \{\, \mathbf{u} \in \mathbf{H}^1(\Omega(\alpha_n)) \;\mid\; \mathbf{u} = \mathbf{0} \text{ on } \Gamma_0(\alpha_n) \,\}\,.$$

If $\mathbf{v}_n \in \mathbf{V}_{\alpha_n}$ and $\mathbf{v} \in \mathbf{V}_{\alpha}$, the convergence "$\mathbf{v}_n \to \mathbf{v}$" is defined by

$$\mathbf{v}_n \longrightarrow \mathbf{v} \iff \widehat{\mathbf{v}}_n = P_{\widehat{\Omega}}\mathbf{v}_n \rightharpoonup P_{\widehat{\Omega}}\mathbf{v} \equiv \widehat{\mathbf{v}} \quad \text{in } \mathbf{H}^1(\widehat{\Omega})\,.$$

A certain class of functionals with the lower semicontinuity property in domain optimization problems was studied in [6]. We state one useful result.

LEMMA 4.5. *Let Ω and $\{\Omega_m\}$ be bounded domains having the cone property. Let u and u_m be elements in $H^1(\Omega)$ and $H^1(\Omega_m)$ such that $u_m \to u$. Assume that $f(s)$ is continuous, nonnegative, and convex for $s \in \mathbb{R}$. Then, the inequality*

$$\int_{\Omega} f(\nabla u(x))\, d\Omega \le \liminf_{m\to\infty} \int_{\Omega_m} f(\nabla u_m(x))\, d\Omega$$

holds. ∎

5. Existence of optimal solutions. We now turn to the question of existence of optimal solutions for the problem (3.10). We will use what is called *a direct method* in the calculus of variations; i.e., we will try to minimize the cost functional directly rather than to solve the Euler–Lagrange equations.

THEOREM 5.1. *There exists at least one optimal solution $(\alpha^*, \mathbf{u}(\alpha^*)) \in \mathcal{V}_{ad}$ for the problem* (3.10).

Proof. The nonemptiness of $\mathcal{V}_{ad}$ follows from Theorem 3.3 for the existence of solutions of the weak variational formulation (3.6)–(3.8) of the Navier–Stokes system.

We define $\mathbf{u}_n = \mathbf{u}(\alpha_n)$, where $\{\alpha_n, \mathbf{u}(\alpha_n)\}$ is a sequence in $\mathcal{V}_{ad}$. Let $\Omega_n = \Omega(\alpha_n)$ and let $p_n = p(\alpha_n)$ and $\mathbf{t}_n = \mathbf{t}(\alpha_n)$, where $(\mathbf{u}(\alpha_n), p(\alpha_n), \mathbf{t}(\alpha_n))$ is a solution of (3.6)–(3.7) for $\Omega(\alpha_n)$. Since $\mathcal{J}(\alpha, \mathbf{u}(\alpha))$ is obviously bounded from below for every $\alpha \in \mathcal{U}_{ad}$, there exists a minimizing subsequence, which is denoted by the same notation $\{(\alpha_n, \mathbf{u}_n)\}$; i.e, there exists a sequence $\{(\alpha_n, \mathbf{u}_n)\} \in \mathcal{V}_{ad}$ such that

$$\lim_{n\to\infty} \mathcal{J}(\alpha_n, \mathbf{u}_n) = \inf_{(\alpha, \mathbf{u}(\alpha)) \in \mathcal{V}_{ad}} \mathcal{J}(\alpha, \mathbf{u}(\alpha))\,.$$

Since $\Omega_0(\alpha_n)$ is contained in $\overline{\widehat{\Omega}}$ for $\{\alpha_n\} \subset \mathcal{U}_{ad}$, the latter is a family of uniformly bounded equicontinuous functions. Hence, by the definition of $\mathcal{U}_{ad}$ and the Ascoli–Arzelà theorem, there exists a subsequence of $\{\alpha_n\}$, which we again denote by the same notation $\{\alpha_n\}$, and an $\alpha^* \in \mathcal{U}_{ad}$ such that $\alpha_n \to \alpha^*$ uniformly in $[M_1, M_2]$.

Note that for any $(\alpha, \mathbf{u}(\alpha)) \in \mathcal{V}_{ad}$, we have that

$$\nu C_1 \|\mathbf{u}(\alpha)\|^2_{1,\Omega(\alpha)} \le \mathcal{J}(\alpha, \mathbf{u}(\alpha)) = \nu \,|\!|\!|\mathbf{u}(\alpha)|\!|\!|^2 \le \nu C_2 \|\mathbf{u}(\alpha)\|^2_{1,\Omega(\alpha)}\,, \tag{5.1}$$

where C_1 and C_2 are positive constants. The first inequality follows from (3.1), and the last, from the Cauchy–Schwarz inequality. Note that the constant C_2 is independent of $\mathbf{u}(\alpha)$ and $\Omega(\alpha)$; i.e., its value is independent of α. According to (5.1) and the definition of $\mathcal{V}_{ad}$, there exists a positive constant K such that $\|\mathbf{u}_n\|_{1,\Omega_n} < K < \infty$ for all n. Furthermore, due to the uniform extension property we can choose an extension $\widehat{\mathbf{u}}_n$ of $\mathbf{u}_n$ to $\widehat{\Omega}$ and a positive constant C that is independent of n such that

$$\|\widehat{\mathbf{u}}_n\|_{1,\widehat{\Omega}} \le C \|\mathbf{u}_n\|_{1,\Omega_n}\,.$$

Thus, $\|\widehat{\mathbf{u}}_n\|_{1,\widehat{\Omega}}$ is uniformly bounded in $\mathbf{H}^1(\widehat{\Omega})$. From (3.5), (3.6), and the fact that $\|\mathbf{u}_n\|_{1,\Omega_n}$ is uniformly bounded, we have that $\|p_n\|_{0,\Omega_n}$ is uniformly bounded. Let $\widehat{p}_n$ be an extension by zero of p_n to $\widehat{\Omega}$, i.e., $\widehat{p}_n|_{\Omega_n} = p_n$ and $\widehat{p}_n|_{\widehat{\Omega}/\Omega_n} = 0$; clearly, $\|\widehat{p}_n\|_{0,\widehat{\Omega}} = \|p_n\|_{0,\Omega_n}$ so that $\|\widehat{p}_n\|_{0,\widehat{\Omega}}$ is uniformly bounded. From (3.6), Lemma 3.2, and the facts that $\|\mathbf{u}_n\|_{1,\Omega_n}$ and $\|p_n\|_{0,\Omega_n}$ are uniformly bounded, we have that $\|\mathbf{t}_n\|_{-1/2,\Gamma_g}$ is uniformly bounded. Consequently, using the compactness of the continuous embeddings $\mathbf{H}^1(\widehat{\Omega}) \subset \mathbf{L}^2(\widehat{\Omega})$ and $\mathbf{H}^{1/2}(\Gamma_g) \subset \mathbf{L}^2(\Gamma_g)$, one may extract from the sequence $\{\widehat{\mathbf{u}}_n, \widehat{p}_n, \mathbf{t}_n\}$ a subsequence (denoted again by $\{\widehat{\mathbf{u}}_n, \widehat{p}_n, \mathbf{t}_n\}$) in $\mathbf{H}^1(\widehat{\Omega}) \times L^2(\widehat{\Omega}) \times \mathbf{H}^{-1/2}(\Gamma_g)$ such that

$$
\begin{aligned}
\widehat{\mathbf{u}}_n &\rightharpoonup \widehat{\mathbf{u}} && \text{in } \mathbf{H}^1(\widehat{\Omega}),\\
\widehat{\mathbf{u}}_n &\to \widehat{\mathbf{u}} && \text{in } \mathbf{L}^2(\widehat{\Omega}),\\
\widehat{p}_n &\rightharpoonup \widehat{p} && \text{in } L^2(\widehat{\Omega}),\\
\mathbf{t}_n &\rightharpoonup \mathbf{t} && \text{in } \mathbf{H}^{-1/2}(\Gamma_g),\\
\gamma_g\widehat{\mathbf{u}}_n &\rightharpoonup \gamma_g(\widehat{\mathbf{u}}) && \text{in } \mathbf{H}^{1/2}(\Gamma_g),\\
\gamma_g\widehat{\mathbf{u}}_n &\to \gamma_g(\widehat{\mathbf{u}}) && \text{in } \mathbf{L}^2(\Gamma_g)
\end{aligned}
\tag{5.2}
$$

for some $(\widehat{\mathbf{u}}, \widehat{p}, \mathbf{t}) \in \mathbf{H}^1(\widehat{\Omega}) \times \mathbf{L}^2(\widehat{\Omega}) \times \mathbf{H}^{-1/2}(\Gamma_g)$.

Now, define $\mathbf{u}(\alpha^*) = \widehat{\mathbf{u}}\Big|_{\Omega(\alpha^*)}$, $p(\alpha^*) = \widehat{p}\Big|_{\Omega(\alpha^*)}$, and $\mathbf{t}(\alpha^*) = \mathbf{t}$. We wish to show that $(\mathbf{u}(\alpha^*), p(\alpha^*), \mathbf{t}(\alpha^*))$ is a solution of (3.6)–(3.8) over $\Omega(\alpha^*)$. To this end, let us define the function spaces

$$\mathcal{W}_n = \{\, \boldsymbol{\phi} \in [C^\infty(\overline{\Omega}_n)]^2 \;\mid\; \boldsymbol{\phi} = \mathbf{0} \text{ in a neighborhood of } \Gamma_0(\alpha_n) = \Gamma_3 \cup \Gamma(\alpha_n) \,\}$$

and

$$\mathcal{W} = \{\, \boldsymbol{\phi} \in [C^\infty(\overline{\Omega}(\alpha^*))]^2 \;\mid\; \boldsymbol{\phi} = \mathbf{0} \text{ in a neighborhood of } \Gamma_0(\alpha^*) = \Gamma_3 \cup \Gamma(\alpha^*) \,\}\,.$$

Then, it is clear that

$$
\begin{aligned}
\mathbf{V}_{\alpha_n} &= \mathbf{H}^1_{\Gamma_0(\alpha_n)}(\Omega_n) = \text{the closure of } \mathcal{W}_n \text{ in } \mathbf{H}^1(\Omega_n) \quad \text{and}\\
\mathbf{V}^*_{\alpha_n} &= \mathbf{H}^1_{\Gamma_0(\alpha^*)}(\Omega(\alpha^*)) = \text{the closure of } \mathcal{W} \text{ in } \mathbf{H}^1(\Omega(\alpha^*))\,.
\end{aligned}
$$

We may consider $\mathbf{H}^1_{\Gamma_0(\alpha^*)}(\Omega(\alpha^*))$ as a closed subspace of

$$\mathbf{H}^1_L(\widehat{\Omega}) = \{\, \mathbf{u} \in \mathbf{H}^1(\widehat{\Omega}) \;\mid\; \mathbf{u}(x_1, 0) = \mathbf{u}(x_1, L) = \mathbf{0} \,\}$$

by extending all the elements of $\mathbf{H}^1_{\Gamma_0(\alpha^*)}(\Omega(\alpha^*))$ by $\mathbf{0}$ in $\overline{\widehat{\Omega}} - \Omega(\alpha^*)$. Let us take $\boldsymbol{\phi} = (\phi_1, \phi_2)^T \in \mathcal{W}$. Since $\alpha_n \to \alpha^*$ uniformly, $\boldsymbol{\phi} \in \mathcal{W}_m$ for sufficiently large m, say, for $m \geq m_0$. We first consider the equation (3.6) over Ω_m for $m \geq m_0$. If we substitute $\boldsymbol{\phi}$ for $\mathbf{v}$, we obtain that

$$
\begin{aligned}
2\nu \int_{\Omega_m} D(\mathbf{u}_m) : D(\boldsymbol{\phi})\, d\Omega + \int_{\Omega_m} ((\mathbf{u}_m \cdot \nabla)\, \mathbf{u}_m) \cdot \boldsymbol{\phi}\, d\Omega\\
- \int_{\Omega_m} p_m\, \nabla \cdot \boldsymbol{\phi}\, d\Omega - \langle \mathbf{t}_m, \boldsymbol{\phi} \rangle_{-1/2,\Gamma_g} = \langle \mathbf{f}, \boldsymbol{\phi} \rangle_{-1,\Omega_m}\,.
\end{aligned}
\tag{5.3}
$$

We examine each term separately. We first note that

$$\begin{aligned}\int_{\Omega_m} D(\mathbf{u}_m) : D(\boldsymbol{\phi})\, d\Omega &= \int_{\widehat{\Omega}} D(\widehat{\mathbf{u}}_m) : D(\boldsymbol{\phi})\, d\Omega \qquad \text{(by the extension of } \mathbf{u}_m \text{ to } \widehat{\Omega}) \\ &\xrightarrow{m\to\infty} \int_{\widehat{\Omega}} D(\widehat{\mathbf{u}}) : D(\boldsymbol{\phi})\, d\Omega \qquad \text{(by the first equation of (5.2))} \\ &= \int_{\Omega(\alpha^*)} D(\mathbf{u}(\alpha^*)) : D(\boldsymbol{\phi})\, d\Omega \qquad \text{(since we chose } \boldsymbol{\phi} \in \mathcal{W}) .\end{aligned}$$

In a similar fashion, using the third equation of (5.2),

$$\int_{\Omega_m} p_m \nabla\cdot\boldsymbol{\phi}\, d\Omega \xrightarrow{m\to\infty} \int_{\Omega(\alpha^*)} p(\alpha^*) \nabla\cdot\boldsymbol{\phi}\, d\Omega$$

and, using the fourth equation of (5.2),

$$\langle \mathbf{t}_m, \boldsymbol{\phi}\rangle_{-1/2,\Gamma_g} \xrightarrow{m\to\infty} \langle \mathbf{t}(\alpha^*), \boldsymbol{\phi}\rangle_{-1/2,\Gamma_g} .$$

Next, we estimate the nonlinear convective term. Since $\mathbf{u}_m = \mathbf{0}$ on $\Gamma_0(\alpha_m)$ for every m, using integration by parts, we have that

$$\begin{aligned}\int_{\Omega_m} ((\mathbf{u}_m\cdot\nabla)\,\mathbf{u}_m)\cdot\boldsymbol{\phi}\, d\Omega &= \int_{\Gamma_g} (\mathbf{u}_m\cdot\mathbf{n})\,(\mathbf{u}_m\cdot\boldsymbol{\phi})\, d\Gamma \\ &\quad - \int_{\Omega_m} (\nabla\cdot\mathbf{u}_m)\,(\mathbf{u}_m\cdot\boldsymbol{\phi})\, d\Omega - \int_{\Omega_m} ((\mathbf{u}_m\cdot\nabla)\,\boldsymbol{\phi})\cdot\mathbf{u}_m\, d\Omega .\end{aligned} \tag{5.4}$$

Note that the outward unit normal vector $\mathbf{n}$ along Γ_g is fixed throughout the domain perturbations. It follows from the last two equations of (5.2) that

$$\begin{aligned}\int_{\Gamma_g} (\mathbf{u}_m\cdot\mathbf{n})\,(\mathbf{u}_m\cdot\boldsymbol{\phi})\, d\Gamma &= \int_{\Gamma_g} (\mathbf{g}\cdot\mathbf{n})\,(\mathbf{g}\cdot\boldsymbol{\phi})\, d\Gamma \\ &= \int_{\Gamma_g} (\widehat{\mathbf{u}}_m\cdot\mathbf{n})\,(\widehat{\mathbf{u}}_m\cdot\boldsymbol{\phi})\, d\Gamma \\ &\xrightarrow{m\to\infty} \int_{\Gamma_g} (\mathbf{u}(\alpha^*)\cdot\mathbf{n})\,(\mathbf{u}(\alpha^*)\cdot\boldsymbol{\phi})\, d\Gamma .\end{aligned}$$

For the second and third terms of (5.4), we use the fact that every one of the components ϕ_i and $(\nabla\boldsymbol{\phi})_{ij}$ belongs to $L^\infty(\Omega(\alpha^*))$. Since $\|\nabla\cdot\widehat{\mathbf{u}}_m\|_{L^2(\widehat{\Omega})} \le \|\widehat{\mathbf{u}}_m\|_{1,\widehat{\Omega}} < \infty$ for all m, we may extract a subsequence, which is again denoted by $\nabla\cdot\widehat{\mathbf{u}}_m$, such that

$$\nabla\cdot\widehat{\mathbf{u}}_m \rightharpoonup \nabla\cdot\widehat{\mathbf{u}} \qquad \text{in } L^2(\widehat{\Omega}) . \tag{5.5}$$

We note that

$$\begin{aligned}&\int_{\widehat{\Omega}} (\nabla\cdot\widehat{\mathbf{u}}_m)\,(\widehat{\mathbf{u}}_m\cdot\boldsymbol{\phi})\, d\Omega - \int_{\widehat{\Omega}} (\nabla\cdot\widehat{\mathbf{u}})\,(\widehat{\mathbf{u}}\cdot\boldsymbol{\phi})\, d\Omega \\ &= \int_{\widehat{\Omega}} (\nabla\cdot\widehat{\mathbf{u}}_m)(\widehat{\mathbf{u}}_m - \widehat{\mathbf{u}})\cdot\boldsymbol{\phi}\, d\Omega + \int_{\widehat{\Omega}} (\nabla\cdot\widehat{\mathbf{u}}_m - \nabla\cdot\widehat{\mathbf{u}})(\widehat{\mathbf{u}}\cdot\boldsymbol{\phi})\, d\Omega \xrightarrow{m\to\infty} 0 ,\end{aligned}$$

using that $\nabla \cdot \widehat{\mathbf{u}}_m$ is uniformly bounded in $L^2(\widehat{\Omega})$. Consequently, using (5.5) and the uniform extension property, it holds that

$$\begin{aligned}\int_{\Omega_m} (\nabla \cdot \mathbf{u}_m)\,(\mathbf{u}_m \cdot \boldsymbol{\phi})\,d\Omega &= \int_{\widehat{\Omega}} (\nabla \cdot \widehat{\mathbf{u}}_m)\,(\widehat{\mathbf{u}}_m \cdot \boldsymbol{\phi})\,d\Omega \\ &\xrightarrow{m\to\infty} \int_{\widehat{\Omega}} (\nabla \cdot \widehat{\mathbf{u}})\,(\widehat{\mathbf{u}} \cdot \boldsymbol{\phi})\,d\Omega \\ &= \int_{\Omega(\alpha^*)} (\nabla \cdot \mathbf{u}(\alpha^*))\,(\mathbf{u}(\alpha^*) \cdot \boldsymbol{\phi})\,d\Omega\,.\end{aligned}$$

In a similar fashion, we have that

$$\int_{\Omega_m} ((\mathbf{u}_m \cdot \nabla)\,\boldsymbol{\phi}) \cdot \mathbf{u}_m\,d\Omega \xrightarrow{m\to\infty} \int_{\Omega(\alpha^*)} ((\mathbf{u}(\alpha^*) \cdot \nabla)\,\boldsymbol{\phi}) \cdot \mathbf{u}(\alpha^*)\,d\Omega\,.$$

Therefore, combining all the results for the three terms in (5.4), we have that

$$\begin{aligned}&\int_{\Omega_m} ((\mathbf{u}_m \cdot \nabla)\,\mathbf{u}_m) \cdot \boldsymbol{\phi}\,d\Omega \\ &\quad= \int_{\Gamma_g} (\mathbf{u}_m \cdot \mathbf{n})\,(\mathbf{u}_m \cdot \boldsymbol{\phi})\,d\Gamma - \int_{\Omega_m} (\nabla \cdot \mathbf{u}_m)\,(\mathbf{u}_m \cdot \boldsymbol{\phi})\,d\Omega - \int_{\Omega_m} ((\mathbf{u}_m \cdot \nabla)\,\boldsymbol{\phi}) \cdot \mathbf{u}_m\,d\Omega \\ &\quad\xrightarrow{m\to\infty} \int_{\Gamma_g} (\mathbf{u}(\alpha^*) \cdot \mathbf{n})\,(\mathbf{u}(\alpha^*) \cdot \boldsymbol{\phi})\,d\Gamma - \int_{\Omega(\alpha^*)} (\nabla \cdot \mathbf{u}(\alpha^*))\,(\mathbf{u}(\alpha^*) \cdot \boldsymbol{\phi})\,d\Omega \\ &\qquad - \int_{\Omega(\alpha^*)} ((\mathbf{u}(\alpha^*) \cdot \nabla)\,\boldsymbol{\phi}) \cdot \mathbf{u}(\alpha^*)\,d\Omega = \int_{\Omega(\alpha^*)} ((\mathbf{u}(\alpha^*) \cdot \nabla)\,\mathbf{u}(\alpha^*)) \cdot \boldsymbol{\phi}\,d\Omega\,.\end{aligned}$$

Collecting various results, we have, up to the present, shown that

$$\begin{aligned}\nu \int_{\Omega(\alpha^*)} D(\mathbf{u}(\alpha^*)) : D(\boldsymbol{\phi})\,d\Omega + \int_{\Omega(\alpha^*)} ((\mathbf{u}(\alpha^*) \cdot \nabla)\,\mathbf{u}(\alpha^*)) \cdot \boldsymbol{\phi}\,d\Omega \\ + \int_{\Omega(\alpha^*)} p(\alpha^*)\,\nabla \cdot \boldsymbol{\phi}\,d\Omega - \langle \mathbf{t}, \boldsymbol{\phi} \rangle_{-1/2,\Gamma_g} = \langle \mathbf{f}, \boldsymbol{\phi} \rangle_{-1,\Omega(\alpha^*)}\end{aligned}$$

for any $\boldsymbol{\phi} \in \mathcal{W}$. Since $\mathcal{W}$ is dense in $\mathbf{H}^1_{\Gamma_0(\alpha^*)}(\Omega(\alpha^*))$, we can then conclude that $(\mathbf{u}(\alpha^*), p(\alpha^*), \mathbf{t}(\alpha^*))$ satisfies (3.6) over $\Omega(\alpha^*)$.

In a similar manner it can be shown that $(\mathbf{u}(\alpha^*), p(\alpha^*), \mathbf{t}(\alpha^*))$ also satisfies (3.7) and (3.8) over $\Omega(\alpha^*)$. Therefore, $(\alpha^*, \mathbf{u}(\alpha^*)) \in \mathcal{V}_{ad}$ and this implies that $\mathcal{V}_{ad}$ is weakly closed. Also, $\mathcal{J}(\alpha, \mathbf{u}(\alpha))$ is coercive and strongly continuous over $\mathbf{V}_\alpha$ for each $\alpha \in \mathcal{U}_{ad}$. Moreover, it readily follows that $\mathcal{J}(\alpha, \mathbf{u}(\alpha))$ is convex with respect to $\mathbf{u}$. Hence $\mathcal{J}(\alpha, \mathbf{u}(\alpha))$ is weakly lower semicontinuous by Lemma 4.5. Since $(\alpha_n, \mathbf{u}_n)$ is a minimizing sequence such that $(\alpha_n, \mathbf{u}_n) \to (\alpha^*, \mathbf{u}(\alpha^*))$ in $\mathcal{V}_{ad}$,

$$\begin{aligned}\inf_{(\alpha,\mathbf{u}(\alpha))\in\mathcal{V}_{ad}} \mathcal{J}(\alpha, \mathbf{u}(\alpha)) &= \liminf_{n\to\infty} \mathcal{J}(\alpha_n, \mathbf{u}_n) \\ &\geq \mathcal{J}(\alpha^*, \mathbf{u}(\alpha^*)) \geq \inf_{(\alpha,\mathbf{u}(\alpha))\in\mathcal{V}_{ad}} \mathcal{J}(\alpha, \mathbf{u}(\alpha))\,.\end{aligned}$$

Consequently we have $\mathcal{J}(\alpha^*, \mathbf{u}(\alpha^*)) = \inf_{(\alpha,\mathbf{u}(\alpha))\in\mathcal{V}_{ad}} \mathcal{J}(\alpha, \mathbf{u}(\alpha))$ and $(\alpha^*, \mathbf{u}(\alpha^*))$ is an optimal solution for the problem (3.10). ∎

Remark. Since the steady-state Navier–Stokes equations have multiple solutions for large Reynolds numbers, we cannot expect a unique optimal solution. Even when the state equation has a unique solution, the optimal shape need not be unique. This was indicated in [3] for optimal shape control problems for elliptic state equations. The same argument can also be applied to our case.

Acknowledgements. The authors wish to acknowledge the many helpful comments and suggestions of one of the referees that resulted in an improved paper.

REFERENCES

[1] G. Armugan and O. Pironneau, *On the problem of riblets as a drag reduction device*, Optim. Control Appl. Meth., 10 (1989), pp. 93–112.
[2] I. Babuška, *The finite element method with Lagrange multipliers*, Numer. Math., 20 (1973), pp. 179–192.
[3] D. Begis and R. Glowinski, *Applications de la méthode des élements finis a l'approximation d'un problème de domaine optimal. Méthodes de résolution des problèmes approches*, Appl. Math. Optim., 2 (1975), pp. 130–169.
[4] J. Cea, *Problems of shape optimal design*, in Optimization of Distributed Parameter Structures, J. Cea and E.J. Haug, eds., Sijthoff and Noordhoff, Alphen aan den Rijn, the Netherlands, 1981, pp. 1049–1088.
[5] D. Chenais, *On the existence of a solution in a domain identification problem*, J. Math. Anal. Appl., 52 (1975), pp. 189–219.
[6] N. Fujii, *Lower semicontinuity in domain optimization problems*, J. Optim. Theory Appl., 57 (1988), pp. 407–422.
[7] A. Fursikov, *Properties of solutions of some extremal problems connected with the Navier–Stokes system*, Math. USSR Sb., 46 (1983), pp. 323–351.
[8] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier–Stokes Equations*, Springer-Verlag, Berlin, 1986.
[9] R. Glowinski and O. Pironneau, *Toward the computation of minimum drag profiles in viscous laminar flow*, Appl. Math. Modelling, 1 (1976), pp. 58–66.
[10] M. Gunzburger, *Finite Element Methods for Incompressible Viscous Flows: A Guide to Theory, Practice and Algorithms*, Academic Press, Boston, 1989.
[11] M. Gunzburger and L. Hou, *Treating inhomogeneous essential boundary conditions in finite element methods and the calculation of boundary stresses*, SIAM J. Numer. Anal., 29 (1992), pp. 390–424
[12] M. Gunzburger, L. Hou, and T. Svobodny, *Analysis and finite element approximation of optimal control problems for the stationary Navier–Stokes equations with Dirichlet controls*, RAIRO Modél. Math. Anal. Numér., 25 (1991), pp. 711–748.
[13] O. Ladyzhenskaya, *The Mathematical Theory of Incompressible Viscous Flows*, Gordon and Breach, New York, 1969.
[14] W. Liu and J. Rubio, *Local convergences and optimal shape design*, SIAM J. Control Optim., 30 (1992), pp. 49–62.
[15] J. Marti, *Introduction to Sobolev Spaces and Finite Element Solution of Elliptic Boundary Value Problems*, Academic Press, London, 1986.
[16] O. Pironneau, *On optimal design in fluid mechanics*, J. Fluid Mech., 64 (1974), pp. 97–110.
[17] O. Pironneau, *Optimal Shape Design for Elliptic Systems*, Springer-Verlag, Berlin, 1984.
[18] J. Saut and R. Temam, *Generic properties of Navier–Stokes equations: Genericity with respect to the boundary values*, Indiana Univ. Math. J., 29 (1980), pp. 427–446.
[19] J. Simon, *Domain variation for drag in Stokes flow*, in Control Theory of Distributed Parameter Systems and Applications, Lecture Notes in Control and Inform. Sci. 159, X. Li and J. Yang, eds., Springer-Verlag, Berlin, 1990, pp. 28–42.
[20] A. Taylof and D. Lay, *Introduction to Functional Analysis*, Krieger, Malabar, 1986.
[21] R. Temam, *Navier–Stokes Equations*, North-Holland, Amsterdam, 1979.
[22] R. Temam, *Navier–Stokes Equations and Nonlinear Functional Analysis*, SIAM, Philadelphia, PA, 1983.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 3, pp. 910–928, May 1998

006

CENTRAL LIMIT THEOREM AND LAW OF ITERATED LOGARITHM FOR LEAST SQUARES ALGORITHMS IN ADAPTIVE TRACKING*

BERNARD BERCU†

Abstract. In autoregressive adaptive tracking, we prove that the least squares and the weighted least squares algorithms possess the same asymptotic properties, sharing the same central limit theorem and the same law of iterated logarithm. We also obtain the same asymptotic behavior and show the limitations of these results in the autoregressive with moving average framework.

Key words. linear regression, least squares, central limit theorem, law of iterated logarithm

AMS subject classifications. 62J05, 93E24, 60F05, 60F15

PII. S0363012995294183

Notations. For any square matrix A, $\mathrm{tr}(A)$ is the trace of A and $\det(A)$ denotes the determinant of A. In addition, $\lambda_{min}A$ and $\lambda_{max}A$ are the minimum and the maximum eigenvalues of A, respectively. Finally, for any vectorial sequence $X=(X_n)$ and any integer $p\geq 1$, $X_n^p=(X_n^t,\ldots,X_{n-p+1}^t)$.

1. Introduction. Let $(\Omega,\mathcal{A},P)$ be a probability space endowed with a filtration $\mathbf{F}=(\mathcal{F}_n)_{n\geq 0}$, where $\mathcal{F}_n$ is the σ-algebra of the events occurring up to time n. Consider the controlled autoregressive with moving average (ARMA) model of order (p,r) given, for all $n\geq 0$, by

$$X_{n+1}=\theta^t\Psi_n+U_n+\varepsilon_{n+1}, \tag{1}$$

where X_n, U_n, and ε_n are, respectively, the d-dimensional system output, input, and driven noise and $\Psi_n=(X_n^p,\varepsilon_n^r)^t$. In order to estimate the unknown $\delta\times d$ matrix θ with $\delta=d\,(p+r)$, we use the weighted least squares (WLS) algorithm that satisfies, for all $n\geq 0$,

$$\hat{\theta}_{n+1}=\hat{\theta}_n+a_nS_n^{-1}(a)\Phi_n\left(X_{n+1}-U_n-\hat{\theta}_n^t\Phi_n\right)^t, \tag{2}$$

$$S_n(a)=\sum_{k=0}^{n}a_k\Phi_k\Phi_k^t+S, \tag{3}$$

$$\hat{\varepsilon}_{n+1}=X_{n+1}-U_n-\hat{\theta}_{n+1}^t\Phi_n,\qquad \Phi_n=(X_n^p,\hat{\varepsilon}_n^r)^t, \tag{4}$$

where the initial value $\hat{\theta}_0$ is arbitrarily chosen and S is a deterministic, symmetric, and positive definite matrix. We set

$$S_n=\sum_{k=0}^{n}\Phi_k\Phi_k^t+S,\qquad s_n=\mathrm{tr}(S_n). \tag{5}$$

The choice of the weighted sequence $a=(a_n)$ is crucial. If

$$a_n=1 \tag{6}$$

*Received by the editors November 3, 1995; accepted for publication (in revised form) March 5, 1997.

http://www.siam.org/journals/sicon/36-3/29418.html

†Laboratoire de Statistiques, Batiment 425 Mathématiques, Université de Paris-Sud, 91 405 Orsay Cedex, France (bernard.bercu@math.u-psud.fr).

we find again the extended least squares (ELS) algorithm. Otherwise, if

$$a_n = \left(\frac{1}{\log s_n}\right)^{1+\gamma} \tag{7}$$

with $\gamma > 0$, we obtain the WLS algorithm proposed by Duflo and Bercu [4], [5]. For these two algorithms, a wide literature concerning the strong consistency and the optimality in adaptive tracking is available (see, e.g., [4], [5], [6], [8], [9], [10], [11], [13], [15], [20], [26]). In these papers, it is always necessary to establish an excitation property for the regressive sequence $\Phi = (\Phi_n)$. To be more precise, for the strong consistency, one has to prove that

$$\lambda_{min} S_n \longrightarrow +\infty, \qquad \log \lambda_{max} S_n = o(\lambda_{min} S_n) \qquad \text{almost surely (a.s.)} \tag{8}$$

and, for the optimality, that $s_n = O(n)$ a.s. In fact, one always has to show that $n = O(\lambda_{min} S_n)$ and $\lambda_{max} S_n = O(n)$ a.s. In autoregressive (AR) adaptive tracking with $r = 0$, we improve the previous results showing the almost sure convergence

$$\frac{S_n}{n} \longrightarrow L_p, \tag{9}$$

$L_p = \text{diag}\,(\Gamma, \ldots, \Gamma)$, where Γ is the conditional covariance matrix of the driven noise. This convergence allows us to obtain a central limit theorem (CLT) and a law of iterated logarithm (LIL) for both LS and WLS algorithms. Since the WLS introduces less weight to the more recent information than the LS, one may expect that WLS may be inferior to LS in asymptotic properties. However, we prove that in the AR framework, the LS and WLS algorithms possess the same asymptotic properties, sharing the same CLT,

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, L_p^{-1} \otimes \Gamma), \tag{10}$$

and the same LIL. In addition, we also obtain that the ELS and WLS algorithms have the same asymptotic behavior in the ARMA framework. Finally, there is no loss in asymptotic efficiency by using WLS, which has many other advantages [4], [5], [17], [22] over LS or ELS in adaptive control theory.

The paper is organized as follows. In section 2, we establish in the AR framework the same CLT and LIL for LS and WLS algorithms. In AR adaptive tracking, the limit matrix given in (9) is positive definite, while this is no longer true in ARMA adaptive tracking. In the ARMA framework, in order to obtain strong consistency results, it is necessary to introduce an excitation on the adaptive tracking control. In section 3, we prove that the effect of this excitation is to make the limit matrix in (9) positive definite. Therefore, for the ARMA models of orders one, we establish the same CLT and LIL for ELS and WLS algorithms. In section 4, we show by simulations the limitation of these last results if the ARMA orders are greater than one. A short conclusion is given in section 5. All technical proofs are collected in the Appendices.

2. AR adaptive tracking. We first consider the AR framework with $r = 0$. Let $x = (x_n)$ be a predictable reference trajectory, to track, step by step, by the observation $X = (X_n)$. To this end, we use the adaptive tracking control proposed by Aström and Wittenmark [1] given, for all $n \geq 0$, by

$$U_n = x_{n+1} - \hat{\theta}_n^t \Phi_n. \tag{11}$$

Relation (1) can then be rewritten as

$$X_{n+1} - x_{n+1} = \pi_n + \varepsilon_{n+1}, \tag{12}$$

where $\pi_n = (\theta - \hat{\theta}_n)^t \Phi_n$. Throughout the following, we assume that the reference trajectory x satisfies

$$\sum_{k=1}^{n} \| x_k \|^2 = o(n) \qquad \text{a.s.} \tag{13}$$

We also assume that the driven noise $\varepsilon = (\varepsilon_n)$ is a martingale difference sequence with

$$E\left[\varepsilon_{n+1} \varepsilon_{n+1}^t \mid \mathcal{F}_n \right] = \Gamma, \tag{14}$$

where Γ is a positive definite deterministic covariance matrix. Finally, we assume that ε satisfies the strong law of large numbers; i.e., if

$$\Gamma_n = \frac{1}{n} \sum_{k=1}^{n} \varepsilon_k \varepsilon_k^t, \tag{15}$$

Γ_n converges a.s. to Γ. This is the case if, for example, ε has finite conditional moment of order > 2 or ε is a white noise, i.e., if ε is independant and identically distributed with mean 0 and covariance matrix Γ. Let (C_n) be the average cost matrix sequence defined by

$$C_n = \frac{1}{n} \sum_{k=1}^{n} (X_k - x_k)(X_k - x_k)^t. \tag{16}$$

The adaptive tracking is said to be optimal if C_n converges a.s. to Γ. Let L_p be the block diagonal square matrix of order $\delta_p = dp$,

$$L_p = \operatorname{diag}(\Gamma, \ldots, \Gamma). \tag{17}$$

THEOREM 2.1. *Consider the AR framework with $r = 0$. Assume that ε has finite conditional moment of order > 2. Then, for the LS algorithm, we have*

$$\frac{S_n}{n} \longrightarrow L_p \qquad \textit{a.s.} \tag{18}$$

In addition, the tracking is optimal:

$$\| C_n - \Gamma_n \| = O\left(\frac{\log n}{n} \right) \qquad \textit{a.s.} \tag{19}$$

We can be more precise in (19) *as follows*

$$\frac{1}{\log n} \sum_{k=1}^{n} (X_k - x_k - \varepsilon_k)(X_k - x_k - \varepsilon_k)^t \longrightarrow \delta_p \Gamma \qquad \textit{a.s.} \tag{20}$$

Finally, $\hat{\theta}_n$ is a strongly consistent estimator of θ:

$$\| \hat{\theta}_n - \theta \|^2 = O\left(\frac{\log n}{n} \right) \qquad \textit{a.s.} \tag{21}$$

Proof. The proof is given in Appendix A. □

THEOREM 2.2. *Consider the AR framework with $r=0$. Assume that either ε is a white noise or ε has finite conditional moment of order >2. Then, for the WLS algorithm with $a_n^{-1}=(\log s_n)^{1+\gamma}$, where $\gamma>0$, we have*

$$(\log n)^{1+\gamma}\frac{S_n(a)}{n} \longrightarrow L_p \qquad a.s. \tag{22}$$

In addition, the tracking is optimal:

$$\| C_n - \Gamma_n \| = o\left(\frac{(\log n)^{1+\gamma}}{n}\right) \qquad a.s. \tag{23}$$

Finally, $\hat{\theta}_n$ is a strongly consistent estimator of θ:

$$\| \hat{\theta}_n - \theta \|^2 = O\left(\frac{(\log n)^{1+\gamma}}{n}\right) \qquad a.s. \tag{24}$$

Proof. The proof is given in Appendix B. □

Remark. Theorem 2.2 is similar to Theorem 2.1. On the one hand, it is not necessary to require a conditional moment of order >2 for the noise ε. On the other hand, we note a loss in $(\log n)^{\gamma}$ in the rates of convergence.

THEOREM 2.3. *Consider the AR framework with $r=0$. Assume that ε has finite conditional moment of order $\alpha>2$ and that x has the same regularity in norm as ε; i.e., for all $2<\beta<\alpha$,*

$$\sum_{k=1}^{n} \| x_k \|^{\beta} = O(n) \qquad a.s. \tag{25}$$

Then, the LS and the WLS algorithms share the same CLT,

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, L_p^{-1} \otimes \Gamma), \tag{26}$$

with $L_p^{-1} \otimes \Gamma = \mathrm{diag}(\Gamma^{-1} \otimes \Gamma, \ldots, \Gamma^{-1} \otimes \Gamma)$. In addition, for any vectors $u \in \mathbb{R}^d$ and $v \in \mathbb{R}^{dp}$, they also share the same LIL,

$$\begin{aligned} \limsup_{n\to\infty} \left(\frac{n}{2\log\log n}\right)^{1/2} v^t(\hat{\theta}_n - \theta)u &= -\liminf_{n\to\infty} \left(\frac{n}{2\log\log n}\right)^{1/2} v^t(\hat{\theta}_n - \theta)u \\ &= (v^t L_p^{-1} v)^{1/2}(u^t \Gamma u)^{1/2} \qquad a.s. \end{aligned} \tag{27}$$

In particular,

$$\left(\frac{\lambda_{min}\Gamma}{\lambda_{max}\Gamma}\right) \le \limsup_{n\to\infty} \left(\frac{n}{2\log\log n}\right) \| \hat{\theta}_n - \theta \|^2 \le \left(\frac{\lambda_{max}\Gamma}{\lambda_{min}\Gamma}\right) \qquad a.s. \tag{28}$$

Proof. The proof is given in Appendix C. □

Remark. First, one can realize that (28) improves Theorem 3.1 of Guo [16] for the LS algorithm. Next, we can also prove that Theorem 2.3 holds for the cost matrix sequence (C_n). To be more precise, assume that ε satisfies the following CLT:

$$\sqrt{n}(\Gamma_n - \Gamma) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Lambda), \tag{29}$$

where Λ is an appropriate deterministic covariance matrix. Then, by (19) or (23), it immediately follows that

$$\sqrt{n}(C_n - \Gamma) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Lambda) \tag{30}$$

for both LS and WLS algorithms. Moreover, via (19) or (23), we can also obtain an LIL for the sequence (C_n). Finally, in AR adaptive tracking, we can avoid the restrictive assumption (13) on the reference trajectory x. Using the same approach developed in Appendix A, we only need to assume that x satisfies the strong law of large numbers

$$\frac{1}{n}\sum_{k=1}^{n} x_k x_k^t \longrightarrow \Delta \qquad \text{a.s.}, \tag{31}$$

where Δ is a deterministic covariance matrix. Then, we just have to replace Γ by $\Gamma+\Delta$ in relation (17).

3. ARMA adaptive tracking. We now consider the ARMA framework. We always use the adaptive tracking control given, for all $n \geq 0$, by

$$U_n = x_{n+1} - \hat{\theta}_n^t \Phi_n, \tag{32}$$

where the reference trajectory x satisfies (13). Relation (1) can be rewritten as

$$X_{n+1} - x_{n+1} = \pi_n + \varepsilon_{n+1}, \tag{33}$$

where $\pi_n = \theta^t \Psi_n - \hat{\theta}_n^t \Phi_n$. Let L_r be the block diagonal square matrix of order $\delta_r = dr$,

$$L_r = \mathrm{diag}(\Gamma, \ldots, \Gamma). \tag{34}$$

For $s = \inf\{p, r\}$, let K be the rectangular matrix of dimension $\delta_p \times \delta_r$ with all coefficients equal to 0 except its left superior block, which is the block diagonal square matrix of order ds, L_s. Finally, let L be the square matrix of order $\delta = \delta_p + \delta_r$:

$$L = \begin{pmatrix} L_p & K \\ K^t & L_r \end{pmatrix}. \tag{35}$$

Throughout the following, we make use of the traditional assumption of passivity: if C is the matrix polynomial associated with the moving average (MA) part of (1) and I_d is the identity matrix of order d,

$$C^{-1} - \frac{1}{2} I_d \tag{P}$$

is strictly positive real. In the ARMA framework, many results concerning the tracking optimality are available (see, e.g., [2], [5], [11], [15], [16]). It is also well known that we can't directly obtain strong consistency for both ELS or WLS algorithms (see, e.g., [5], [7], [11], [13], [17]). Nevertheless, we prove in the following lemma that convergences (18) and (22) still hold here, replacing L_p by L. This can lead to interesting asymptotic properties.

LEMMA 3.1. *For the ARMA model, assume that* (P) *is satisfied. Then, for the ELS algorithm, if ε has finite conditional moment of order > 2, we have*

$$\frac{S_n}{n} \longrightarrow L \quad \text{a.s.} \tag{36}$$

In addition, for the WLS algorithm with $a_n^{-1} = (\log s_n)^{1+\gamma}$, *where* $\gamma > 0$, *if* ε *is a white noise or if* ε *has finite conditional moment of order* > 2, *we have*

$$(\log n)^{1+\gamma} \frac{S_n(a)}{n} \longrightarrow L \quad a.s. \tag{37}$$

Proof. The proof is given in Appendix D. □

THEOREM 3.2. *For the ARMA model, assume that* (P) *is satisfied and consider the regulation problem with* $x = 0$. *Assume that* ε *has finite conditional moment of order* > 2. *For a positive, nonincreasing, and deterministic sequence* (α_n) *such that* $\alpha_n = O(n)$, *assume that* $\| \varepsilon_n \|^2 = O(\alpha_n)$. *Let* (λ_n) *be a positive, nonincreasing, and deterministic sequence such that* $n^c \alpha_n = O(\lambda_n)$, $n^{1+c}\alpha_n = O(\lambda_n^2)$ *with* $0 < c < 1$ *for the ELS algorithm, and* $c = 0$ *for the WLS algorithm. Finally, assume that*

$$\| \Gamma_n - \Gamma \| = o\left(\frac{\lambda_n}{n}\right) \quad a.s. \tag{38}$$

Then, for both ELS and WLS algorithms, the tracking is optimal:

$$\| C_n - \Gamma \| = o\left(\frac{\lambda_n}{n}\right) \quad a.s. \tag{39}$$

Moreover, we also have

$$\left\| \frac{S_n}{n} - L \right\| = o\left(\frac{\lambda_n}{n}\right) \quad a.s. \tag{40}$$

Finally, on the one hand, it results for the ELS estimator that

$$\| L^{1/2}(\hat{\theta}_n - \theta) \|^2 = o\left(\lambda_n \frac{\log n}{n}\right) \quad a.s. \tag{41}$$

On the other hand, we have for the WLS estimator that

$$\| L^{1/2}(\hat{\theta}_n - \theta) \|^2 = o\left(\frac{\lambda_n}{n}\right) \quad a.s. \tag{42}$$

Remark. If ε has finite conditional moment of order $\alpha > 2$, we can take by Chow's lemma (see, e.g., Corollary 2.8.5 of Stout [23] or [12]) the sequence (λ_n) such that

$$\sum_{k=1}^{\infty} \left(\frac{1}{\lambda_n}\right)^{\alpha/2} < +\infty. \tag{43}$$

We can choose, for example, $\lambda_n = n^\beta$ with $2\alpha^{-1} < \beta < 1$. One can realize that (41) improves Theorem 3.2 (i) of Guo [16].

Proof. By Theorem 1 of Guo and Chen [15] and Theorem 5 of Bercu [5] on the prediction errors sequence (π_n), respectively, we have

$$\sum_{k=0}^{n} \| \pi_k \|^2 = o(n^c \alpha_n) \quad \text{a.s.} \tag{44}$$

with $c > 0$ for the ELS algorithm and

$$\sum_{k=0}^{n} \| \pi_k \|^2 = o(a_n^{-1} + \alpha_n) \quad \text{a.s.} \tag{45}$$

for the WLS algorithm. Then, for these two algorithms, we find that

$$\sum_{k=0}^{n} \| \pi_k \|^2 = o(\lambda_n) \qquad \text{a.s.} \tag{46}$$

By (33), we also have

$$\| C_n - \Gamma_n \| = O\left(\frac{1}{n}\sum_{k=1}^{n} \| \pi_{k-1} \|^2\right) \qquad \text{a.s.,} \tag{47}$$

and we immediately obtain relation (39). Therefore (33), (44), and (45), together with the second assumption on the sequence $\lambda = (\lambda_n)$, imply (40). Finally, for the ELS estimator, by Theorem 1 of Lai and Wei [19], we have $\| \hat{\theta}_{n+1} - \theta \|^2 = O(\log n)$ a.s. Moreover, by Theorem 1 of Bercu [5], the WLS estimator is always a.s. bounded, $\| \hat{\theta}_{n+1} - \theta \|^2 = O(1)$. Therefore, (40) clearly implies (41) and (42), completing the proof of Theorem 3.2. □

In order to obtain strong consistency for ELS and WLS algorithms, we are brought to introduce an excitation on the adaptive tracking control. As one can see below, the effect of this excitation is to make the limit matrix in Lemma 3.1 positive definite. First, we use the continually disturbed control given, for all $n \geq 0$, by

$$U_n = x_{n+1} - \hat{\theta}_n^t \Phi_n + \xi_{n+1}, \tag{48}$$

where the reference trajectory x satisfies (13) and ξ is an exogenous noise of dimension d, adapted to $\mathbf{F}$, with mean 0 and positive definite covariance matrix Λ. In addition, we assume that ξ is independent of ε, of x, and of the initial state of the system. Let

$$\Delta_n = \frac{1}{n}\sum_{k=1}^{n} (\varepsilon_k + \xi_k)(\varepsilon_k + \xi_k)^t. \tag{49}$$

Assume that ξ satisfies the strong law of large numbers, so Δ_n converges a.s. to $\Gamma + \Lambda$. Relation (1) can be rewritten as

$$X_{n+1} - x_{n+1} = \pi_n + \varepsilon_{n+1} + \xi_{n+1}. \tag{50}$$

The adaptive tracking is said to be residually optimal if C_n converges a.s. to $\Gamma + \Lambda$. Let H be the square matrix of order $\delta = \delta_p + \delta_r$,

$$H = \begin{pmatrix} H_p & K \\ K^t & L_r \end{pmatrix}, \tag{51}$$

where H_p is the block diagonal square matrix of order $\delta_p = dp$:

$$H_p = \operatorname{diag}(\Gamma + \Lambda, \ldots, \Gamma + \Lambda). \tag{52}$$

THEOREM 3.3. *For the ARMA model, assume that* (P) *is satisfied. Assume that* ε *has finite conditional moment of order* > 2. *Then, for the ELS algorithm, we have*

$$\frac{S_n}{n} \longrightarrow H \qquad \textit{a.s.} \tag{53}$$

In addition, the tracking is residually optimal:

$$\| C_n - \Delta_n \| = O\left(\frac{\log n}{n}\right) \qquad \textit{a.s.} \tag{54}$$

Finally, $\hat{\theta}_n$ is a strongly consistent estimator of θ:

$$\| \hat{\theta}_n - \theta \|^2 = O\left(\frac{\log n}{n}\right) \qquad a.s. \tag{55}$$

Proof. The proof is given in Appendix D. □

THEOREM 3.4. *For the ARMA model, assume that* (P) *is satisfied. Assume that either ε is a white noise or ε has finite conditional moment of order > 2. Then, for the WLS algorithm with $a_n^{-1} = (\log s_n)^{1+\gamma}$, where $\gamma > 0$, we have*

$$(\log n)^{1+\gamma} \frac{S_n(a)}{n} \longrightarrow H \qquad a.s. \tag{56}$$

In addition, the tracking is residually optimal:

$$\| C_n - \Delta_n \| = o\left(\frac{(\log n)^{1+\gamma}}{n}\right) \qquad a.s. \tag{57}$$

Finally, $\hat{\theta}_n$ is a strongly consistent estimator of θ:

$$\| \hat{\theta}_n - \theta \|^2 = O\left(\frac{(\log n)^{1+\gamma}}{n}\right) \qquad a.s. \tag{58}$$

Proof. The proof is given in Appendix D. □

Remark. We note that Theorems 3.3 and 3.4 are similar to Theorems 2.1 and 2.2. In addition, it is easy to see that the matrix H is positive definite. In fact, if $p \leq r$, then $\det H = (\det \Gamma)^r (\det \Lambda)^p$, and if $p > r$, then $\det H = (\det \Gamma)^r (\det \Lambda)^r (\det(\Gamma + \Lambda))^{p-r}$.

THEOREM 3.5. *For the ARMA model, assume that* (P) *is satisfied, with p and r equal to* 1. *Assume that ε and ξ have finite conditional moments of order $\alpha > 2$. On the one hand, assume that x satisfies* (25) *and*

$$\sum_{k=1}^{n} \| x_k \|^2 = o\left(\frac{n}{\log n}\right) \qquad a.s. \tag{59}$$

for the ELS algorithm. On the other hand, assume that x satisfies (25) *and*

$$\sum_{k=1}^{n} \| x_k \|^2 = o\left(\frac{n}{(\log n)^{2+2\gamma}}\right) \qquad a.s. \tag{60}$$

for the WLS algorithm with $a_n^{-1} = (\log s_n)^{1+\gamma}$, where $\gamma > 0$. Then, the ELS and the WLS algorithms share the same CLT,

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, H^{-1} \otimes \Gamma). \tag{61}$$

For any vectors $u \in \mathbb{R}^d$ and $v \in \mathbb{R}^\delta$, they also share the same LIL,

$$\begin{aligned} \limsup_{n\to\infty} \left(\frac{n}{2\log\log n}\right)^{1/2} v^t(\hat{\theta}_n - \theta)u &= -\liminf_{n\to\infty} \left(\frac{n}{2\log\log n}\right)^{1/2} v^t(\hat{\theta}_n - \theta)u \\ &= (v^t H^{-1} v)^{1/2} (u^t \Gamma u)^{1/2} \qquad a.s. \end{aligned} \tag{62}$$

In particular,

$$\left(\frac{\lambda_{min}\Gamma}{\lambda_{max}H}\right) \leq \limsup_{n\to\infty} \left(\frac{n}{2\log\log n}\right) \| \hat{\theta}_n - \theta \|^2 \leq \left(\frac{\lambda_{max}\Gamma}{\lambda_{min}H}\right) \quad a.s. \tag{63}$$

Proof. The proof is given in Appendix E. □

4. Simulations. The goal of this section is to show that Theorem 3.5 is no longer true if the orders p or r are greater than 1. From relations (1) and (2), we have

$$S_{n-1}(a)(\hat{\theta}_n - \theta) = M_n(a) - R_{n-1}(a)\theta, \tag{64}$$

$$M_n(a) = M_0 + \sum_{k=1}^{n} a_{k-1}\Phi_{k-1}\varepsilon_k^t, \qquad R_n(a) = \sum_{k=0}^{n} a_k\Phi_k(\Phi_k - \Psi_k)^t \tag{65}$$

with $M_0 = S(\hat{\theta}_0 - \theta)$. By Lemmas C.1 or C.2 in Appendix C, we know how to deal with $M_n(a)$. The remainder $R_n(a)$ is much more complicated to study. This remainder vanishes in the AR framework. Consequently, we can easily establish CLT and LIL as in Theorem 2.3. In order to obtain similar results in the ARMA framework, we have to prove that the remainder $R_n(a)$ can be neglected. This was done with p and r equal to 1 in Theorem 3.5. Unfortunately, if p or r is greater than 1, $R_n(a)$ plays a prominent part and is really very complicated to study. We shall now show it by simulations for the ELS algorithm. Consider the following two models:

$$X_{n+1} = \frac{5}{4}X_n + \frac{1}{2}X_{n-1} + U_n + \frac{3}{4}\varepsilon_n + \varepsilon_{n+1}, \tag{I}$$

$$X_{n+1} = \frac{5}{4}X_n + U_n + \frac{3}{4}\varepsilon_n + \frac{1}{4}\varepsilon_{n-1} + \varepsilon_{n+1}, \tag{II}$$

where ε is a Gaussian white noise $N(0,1)$. For simplicity, we study the regulation problem taking the reference trajectory $x = 0$. Therefore, we use the continually disturbed control

$$U_n = -\hat{\theta}_n^t\Phi_n + \xi_{n+1}, \tag{66}$$

where ξ is an exogenous Gaussian white noise $N(0,1)$. We base our simulations on $M = 500$ realizations of sample size $N = 10000$. In order to keep this section brief, we focus our attention on the behavior of the statistic

$$Z_N = \sqrt{N}H^{1/2}(\hat{\theta}_N - \theta), \tag{67}$$

where the matrix H is for models (I) and (II), respectively:

$$\begin{pmatrix} 2 & 0 & 1 \\ 0 & 2 & 0 \\ 1 & 0 & 1 \end{pmatrix}, \qquad \begin{pmatrix} 2 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{68}$$

We expect at least that each component of Z_N has $N(0,1)$ distribution. Figure 1 represents the three coordinates of Z_N in model (I). One can realize that the second coordinate is not $N(0,1)$. Figure 2 represents the three coordinates of Z_N in model (II). One can realize that the third coordinate is not $N(0,1)$. Next, if we consider an ARMA model of orders $p = 2$ and $r = 2$, we can also see that the second and the fourth coordinates of Z_N are not $N(0,1)$. We can conclude that if p or r is greater than 1, $R_n(a)$ plays a prominent part which can't be neglected. It would be very nice to clarify the behavior of $R_n(a)$ in ARMA adaptive tracking.

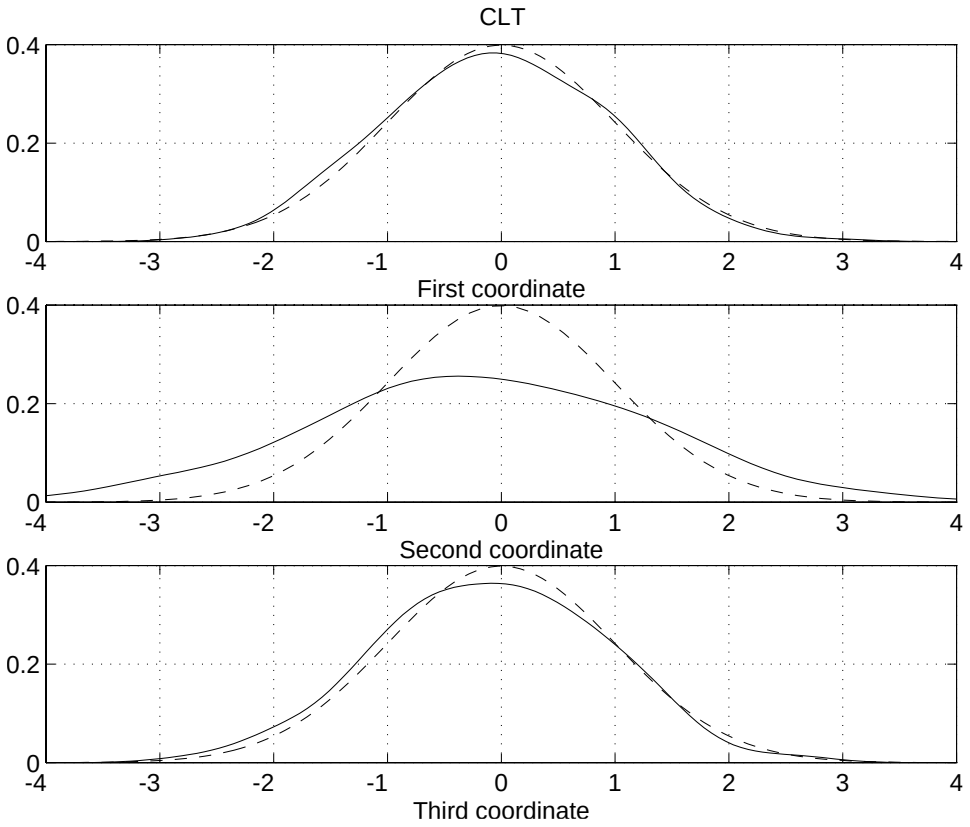


FIG. 1.

5. Conclusion. In AR adaptive tracking, we have proved that the LS and the WLS algorithms share the same CLT and LIL. We have also extended and shown the limitations of these results in the ARMA framework. One can ask the natural question: Why make use of the WLS algorithm?

• First, we have seen in this paper that WLS performs as well as ELS for parameter estimation when the system is persistently excited. There is no loss in asymptotic efficiency by using the WLS algorithm.

• Next, as it was shown in [5], the WLS algorithm is more convenient than the ELS in the analysis of autoregressive with moving average and exogenous control (ARMAX) adaptive tracking thanks to the behavior of the prediction errors sequence. The convergence rates proved for the tracking optimality are in general better for the WLS [5] than for the ELS [15].

• In the ARMAX framework, the leading matrix associated with the control is usually called the high frequency gain. For ARX models with known or unknown high frequency gain, strong consistency and tracking optimality results have been established in [16]. It is reasonable to conjecture that CLT and LIL could also be proved for ARX models with known high frequency gain. However, it would be extremely difficult in the general case.

• Finally, Guo [17] has recently proved the almost sure self-convergence of the WLS algorithm. This property can lead to various applications in adaptive control theory such as adaptive pole-placement and LQG problems [17], [22] for ARMAX models.

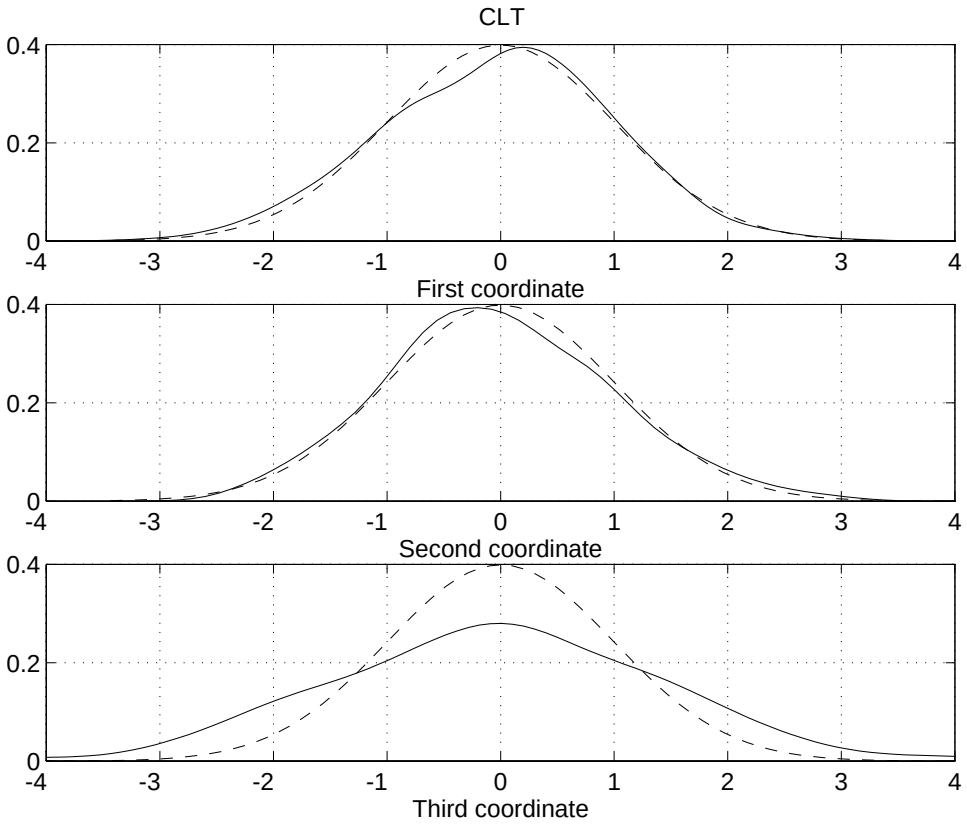


FIG. 2.

Appendix A.

Proof of Theorem 2.1. By the strong law of large numbers and relation (12), we easily prove that $n = O(s_n)$. By Lemma 1 of Guo and Chen [15] or Theorem 1 of Bercu [3] on the prediction errors sequence (π_n), we have

$$\sum_{k=1}^{n}(1 - f_k) \parallel \pi_k \parallel^2 = O(\log s_n) \qquad \text{a.s.}, \tag{A.1}$$

where $f_n = \Phi_n^t S_n^{-1} \Phi_n$. If ε has finite conditional moment of order $\alpha > 2$, using the same approach as Chen and Guo [11], [15], we can show by (A.1) that $\parallel \Phi_n \parallel^2 = O(s_n^\beta)$ with $2\alpha^{-1} < \beta < 1$. We also find by (A.1) and (12) that

$$\sum_{k=1}^{n} \parallel \pi_k \parallel^2 = o(s_n^\beta \log s_n) \qquad \text{a.s.}, \tag{A.2}$$

$$\sum_{k=1}^{n} \parallel X_{k+1} \parallel^2 = o(s_n^\beta \log s_n) + O(n) \qquad \text{a.s.} \tag{A.3}$$

Finally, we obtain that $s_n = o(s_n) + O(n)$, so $s_n = O(n)$. Consequently, we prove the

tracking optimality, as by (12) and (A.2),

$$\| C_n - \Gamma_n \| = O(\frac{1}{n} \sum_{k=1}^{n} \| \pi_{k-1} \|^2) \qquad \text{a.s.}, \tag{A.4}$$

$$\sum_{k=1}^{n} \| \pi_k \|^2 = o(n) \qquad \text{a.s.} \tag{A.5}$$

We still have to establish the convergence rate given in (19). As the reference trajectory x satisfies (13), we have already proven the almost sure convergence

$$\frac{1}{n} \sum_{k=0}^{n} X_k X_k^t \longrightarrow \Gamma. \tag{A.6}$$

Recalling (12), we have for $1 \leq i \leq p-1$ that

$$\sum_{k=1}^{n} X_k X_{k-i}^t = \sum_{k=1}^{n} (\pi_{k-1} + x_k) X_{k-i}^t + \sum_{k=1}^{n} \varepsilon_k X_{k-i}^t. \tag{A.7}$$

The right-hand side of (A.7) is a regressive sequence. Therefore, we have a.s.

$$\left\| \sum_{k=1}^{n} X_k X_{k-i}^t \right\| \leq \sum_{k=1}^{n} \| \pi_{k-1} + x_k \| \| X_{k-i} \| + o\left(\sum_{k=1}^{n} \| X_{k-i} \|^2 \right). \tag{A.8}$$

We prove, by (13) and (A.5), together with the Cauchy–Schwarz inequality, that

$$\sum_{k=1}^{n} X_k X_{k-i}^t = o(n) \qquad \text{a.s.}, \tag{A.9}$$

which implies the convergence (18). As the matrix L_p is positive definite, it clearly follows that $n = O(\lambda_{min} S_n)$, $\| \Phi_n \|^2 = o(n)$, and f_n tends a.s. towards 0. Then, by (A.1), we find that

$$\sum_{k=1}^{n} \| \pi_k \|^2 = O(\log n) \qquad \text{a.s.}, \tag{A.10}$$

and consequently, we obtain the relation (19). By a well-known result established in Theorem 1 of Lai and Wei [19], [20] for the LS estimator, we also have

$$\| \hat{\theta}_{n+1} - \theta \|^2 = O\left(\frac{\log s_n}{\lambda_{min} S_n} \right) \qquad \text{a.s.}, \tag{A.11}$$

which implies (21). Moreover, if $\check{\theta}_n = \hat{\theta}_n - \theta$, we immediately deduce from (28) that

$$\| S_{n-1}^{1/2} \check{\theta}_n \|^2 = o(\log n) \qquad \text{a.s.} \tag{A.12}$$

By Duflo, Senoussi, and Touati [14, p. 560], we also have the almost sure convergence

$$\frac{1}{\log n} \left(\check{\theta}_n^t S_{n-1} \check{\theta}_n + \sum_{k=0}^{n-1} (1 - f_k) \pi_k \pi_k^t \right) \longrightarrow \delta_p \Gamma. \tag{A.13}$$

Finally, (A.12) and (A.13) imply (20), completing the proof of Theorem 2.1. □

Appendix B.

Proof of Theorem 2.2. By Theorem 1 of Bercu and Duflo [4], [5] on the prediction errors sequence (π_n), we have

$$\sum_{n=1}^{\infty} a_n(1-f_n(a)) \parallel \pi_n \parallel^2 < +\infty \qquad \text{a.s.}, \tag{B.1}$$

where $f_n(a) = a_n \Phi_n^t S_n^{-1}(a) \Phi_n$. Then, as $a_n^{-1} = O(s_n)$, we find by (B.1) together with Kronecker's lemma that

$$\sum_{k=1}^{n} \parallel \pi_k \parallel^2 = o(s_n) \qquad \text{a.s.} \tag{B.2}$$

Contrary to the LS algorithm, we can easily prove that $s_n = O(n)$. In fact, (B.2) and (12) immediately imply

$$\sum_{k=1}^{n} \parallel X_{k+1} \parallel^2 = o(s_n) + O(n) \qquad \text{a.s.} \tag{B.3}$$

Therefore, $s_n = o(s_n) + O(n)$, so $s_n = O(n)$. Finally, we have established the tracking optimality. In Appendix A, we have also shown that $n^{-1}S_n$ converges a.s. to L_p. Consequently, as the weighting sequence $a = (a_n)$ is nonincreasing, it results that $a_n S_n \leq S_n(a)$ so $na_n = O(\lambda_{min} S_n(a))$ and $f_n(a)$ tends a.s. towards 0. We can conclude by (B.1) that

$$\sum_{k=1}^{n} \parallel \pi_k \parallel^2 = o(a_n^{-1}) \qquad \text{a.s.}, \tag{B.4}$$

which implies relation (23) as s_n has the same order as n, so a_n^{-1} is a.s. equivalent to $(\log n)^{1+\gamma}$. We can also deduce (24), as by Theorem 1 of Bercu and Duflo [4], [5],

$$\parallel \hat{\theta}_{n+1} - \theta \parallel^2 = O\left(\frac{1}{\lambda_{min} S_n(a)}\right) \qquad \text{a.s.} \tag{B.5}$$

Now, we have

$$S_n(a) = a_{n+1} S_n + \sum_{k=1}^{n} b_k \frac{S_k}{k} + R \tag{B.6}$$

with $b_n = n(a_n - a_{n+1})$ and $R = S_0(a) - a_1 S_0$. In addition,

$$\sum_{k=1}^{n} b_k = \sum_{k=1}^{n} a_k - n a_{n+1}. \tag{B.7}$$

Next, as a_n^{-1} is a.s. equivalent to $(\log n)^{1+\gamma}$,

$$\sum_{k=1}^{n} b_k \sim (1+\gamma) \frac{n a_n}{\log n}, \qquad \sum_{k=1}^{n} b_k = o(n a_n) \qquad \text{a.s.} \tag{B.8}$$

Finally, (B.6), together with Toeplitz's lemma, imply the convergence (22), completing the proof of Theorem 2.2. □

Appendix C.

Proof of Theorem 2.3. In order to prove Theorem 2.3, we need the two following lemmas on regressive sequences. They result from the CLT on triangular arrays [18], [21], [25] and from the LIL on martingales [14], [23], [24]. Let $\varepsilon = (\varepsilon_n)$ be a d-dimensional noise, adapted to $\mathbf{F}$, which satisfies (14) where Γ is a deterministic covariance matrix. Let $\varphi = (\varphi_n)$ be a δ-dimensional sequence of random vectors, adapted to $\mathbf{F}$. Set, for $n \geq 0$,

$$M_n = M_0 + \sum_{k=1}^{n} \varphi_{k-1}\varepsilon_k^t, \qquad S_n = \sum_{k=0}^{n} \varphi_k \varphi_k^t + S.$$

LEMMA C.1. *Let (c_n) be a deterministic real sequence increasing to infinity. Assume that, for all $\epsilon > 0$,*

$$(H_1) \qquad c_n^{-1} S_{n-1} \xrightarrow{\mathcal{P}} L,$$

$$(H_2) \qquad c_n^{-1} \sum_{k=1}^{n} E\left[\| \Delta M_k \|^2 \mathbf{1}_{\left\{\|\Delta M_k\| \geq \epsilon\sqrt{c_n}\right\}} \mid \mathcal{F}_{k-1} \right] \xrightarrow{\mathcal{P}} 0,$$

where $\Delta M_n = M_n - M_{n-1}$. Then, $c_n^{-1} M_n$ tends a.s. towards 0 and

$$\frac{1}{\sqrt{c_n}} M_n \xrightarrow{\mathcal{L}} \mathcal{N}(0, L \otimes \Gamma).$$

In addition, if L is positive definite, we have the CLT

$$\sqrt{c_n} S_{n-1}^{-1} M_n \xrightarrow{\mathcal{L}} \mathcal{N}(0, L^{-1} \otimes \Gamma).$$

Remark. Assume that ε has finite conditional moment of order > 2. Then, Lindeberg's condition (H_2) is satisfied if $\| \varphi_n \|^2 = o(c_n)$ a.s.

LEMMA C.2. *Let (c_n) be a deterministic real sequence increasing to infinity. Assume that the noise ε has finite conditional moment of order $\alpha > 2$. Also assume that*

$$(H_3) \qquad c_n^{-1} S_{n-1} \longrightarrow L \quad \text{a.s.},$$

$$(H_4) \qquad \sum_{n=1}^{\infty} \left(\frac{\| \varphi_n \|}{\sqrt{c_n}} \right)^{\beta} < +\infty \quad \text{a.s.},$$

with $2 < \beta \leq \alpha$. Then, for any vector $u \in \mathbb{R}^d$ and $v \in \mathbb{R}^\delta$ such that $v^t L v > 0$, we have

$$\limsup_{n\to\infty} \left(\frac{1}{2c_n \log\log c_n} \right)^{1/2} v^t M_n u = -\liminf_{n\to\infty} \left(\frac{1}{2c_n \log\log c_n} \right)^{1/2} v^t M_n u$$
$$= (v^t L v)^{1/2} (u^t \Gamma u)^{1/2} \quad \text{a.s.}$$

In addition, if L is positive definite, we have the LIL

$$\limsup_{n\to\infty} \left(\frac{c_n}{2\log\log c_n} \right)^{1/2} v^t S_{n-1}^{-1} M_n u = -\liminf_{n\to\infty} \left(\frac{c_n}{2\log\log c_n} \right)^{1/2} v^t S_{n-1}^{-1} M_n u$$
$$= (v^t L^{-1} v)^{1/2} (u^t \Gamma u)^{1/2} \quad \text{a.s.}$$

Theorem 2.3 is a direct application of Lemmas C.1 and C.2. By the relations (1) and (2), we have

$$\hat{\theta}_n - \theta = S_{n-1}^{-1}(a) M_n(a), \tag{C.1}$$

$$M_n(a) = M_0 + \sum_{k=1}^{n} a_{k-1} \Phi_{k-1} \varepsilon_k^t \tag{C.2}$$

with $M_0 = S(\hat{\theta}_0 - \theta)$. On the one hand, for the LS algorithm, we choose $\varphi_n = \Phi_n$ and $c_n = n$. On the other hand, for the WLS algorithm, we take $\varphi_n = a_n \Phi_n$ and $c_n = n/(\log n)^{2+2\gamma}$. First, for the LS algorithm, (26) can be clearly deduced via Lemma C.1 together with (18) and equation (C.1). In addition, if ε has finite conditional moment of order $\alpha > 2$, for all $2 < \beta < \alpha$, we have by Chow's lemma (see, e.g., Corollary 2.8.5 of Stout [23]) that

$$\sum_{k=1}^{n} \| \varepsilon_k \|^\beta = O(n) \qquad \text{a.s.} \tag{C.3}$$

Since the reference trajectory x satisfies (25), we show by (A.10) that

$$\sum_{k=1}^{n} \| X_k \|^\beta = O(n), \qquad \sum_{k=1}^{n} \| \Phi_k \|^\beta = O(n) \quad \text{a.s.} \tag{C.4}$$

Therefore, as $\beta > 2$, (C.4) implies

$$\sum_{n=1}^{\infty} \left(\frac{\| \Phi_n \|}{\sqrt{n}} \right)^\beta < +\infty \qquad \text{a.s.} \tag{C.5}$$

Finally, we find (27) via Lemma C.2 together with (18) and equation (C.1). Next, for the WLS algorithm, set

$$Q_n(a) = \sum_{k=0}^{n} a_k^2 \Phi_k \Phi_k^t + S. \tag{C.6}$$

As in (22), we prove that $c_n^{-1} Q_n(a)$ converges a.s. to L_p. Hence, (22) and equation (C.1) clearly imply (26). In addition, if ε has finite conditional moment of order $\alpha > 2$, for all $2 < \beta < \alpha$, we have by Chow's lemma [23], together with (B.4),

$$\sum_{k=1}^{n} (a_k \| \Phi_k \|)^\beta = O(n) \qquad \text{a.s.} \tag{C.7}$$

Then, it follows from (C.7) that

$$\sum_{n=1}^{\infty} \left(\frac{a_n \| \Phi_n \|}{\sqrt{c_n}} \right)^\beta < +\infty \qquad \text{a.s.} \tag{C.8}$$

Finally, we prove (27) via Lemma C.2 together with (22) and equation (C.1), completing the proof of Theorem 2.3. □

Appendix D.

Proof of Theorems 3.3 and 3.4. First, we prove Lemma 3.1 for both ELS and WLS algorithms. On the one hand, relation (A.1) holds for the ELS algorithm in the ARMAX framework [3], [11], [15]. In addition, by Theorem 1 of Lai and Wei [19], [20], we also have

$$\sum_{k=1}^{n} \| \Phi_k - \Psi_k \|^2 = O(\log s_n) \qquad \text{a.s.} \tag{D.1}$$

If ε has finite conditional moment of order $\alpha > 2$, we can show as in Appendix A that $\| \Phi_n \|^2 = O(s_n^\beta)$ with $2\alpha^{-1} < \beta < 1$. Hence, we find by (A.1) and (33) that

$$\sum_{k=1}^{n} \| \pi_k \|^2 = o(s_n^\beta \log s_n) \qquad \text{a.s.,} \tag{D.2}$$

$$\sum_{k=1}^{n} \| X_{k+1} \|^2 = o(s_n^\beta \log s_n) + O(n) \qquad \text{a.s.} \tag{D.3}$$

Finally, (D.1) together with (D.3) imply that $s_n = o(s_n) + O(n)$, so $s_n = O(n)$. Consequently, we find by (D.2) that

$$\sum_{k=1}^{n} \| \pi_k \|^2 = o(n) \qquad \text{a.s.} \tag{D.4}$$

We now recall that the reference trajectory x satisfies (13). Therefore, exactly as in Appendix A, (33), (D.1), and (D.4) imply the convergence (36) for the ELS algorithm. On the other hand, concerning the WLS algorithm, relation (B.1) holds in the ARMAX framework [4], [5]. In addition, we also have, by Theorem 1 of Bercu [5],

$$\sum_{n=1}^{\infty} a_n \| \Phi_n - \Psi_n \|^2 < +\infty \qquad \text{a.s.} \tag{D.5}$$

As $a_n^{-1} = (\log s_n)^{1+\gamma}$ with $\gamma > 0$, we find by (33), (B.2), and (D.5) together with Kronecker's lemma that $s_n = o(s_n) + O(n)$, so $s_n = O(n)$. Consequently, we immediately obtain by (B.2) that

$$\sum_{k=1}^{n} \| \pi_k \|^2 = o(n) \qquad \text{a.s.} \tag{D.6}$$

Therefore, (33), (D.5), and (D.6) imply the convergence (36) for the WLS algorithm. Finally, via (B.6)–(B.8), we also find the convergence (37) for the WLS algorithm, completing the proof of Lemma 3.1. We now prove Theorems 3.3 and 3.4. We can easily switch to the continually disturbed tracking situation. Indeed, as ξ is an exogenous noise that satisfies the strong law of large numbers, we prove by (50) the convergences (53) and (56) exactly as in Lemma 3.1. Furthermore, since the matrix H is positive definite, we find for the ELS algorithm that $n = O(\lambda_{min} S_n)$, and for the WLS algorithm, that $n a_n = O(\lambda_{min} S_n(a))$. Finally, as the relations (A.1), (A.11) and (B.1), (B.5) hold in the ARMAX framework, Theorems 3.3 and 3.4 are established. □

Appendix E.

Proof of Theorem 3.5. We finally prove Theorem 3.5 for both ELS and WLS algorithms. On the one hand, for the ELS algorithm, by (1) and (2), we have

$$S_{n-1}(\hat{\theta}_n - \theta) = M_n - R_{n-1}\theta, \tag{E.1}$$

$$M_n = M_0 + \sum_{k=1}^{n} \Phi_{k-1}\varepsilon_k^t, \qquad R_n = \sum_{k=0}^{n} \Phi_k(\Phi_k - \Psi_k)^t \tag{E.2}$$

with $M_0 = S(\hat{\theta}_0 - \theta)$. In order to study the remainder R_n, it is enough by (D.1) to work on

$$P_n = \sum_{k=0}^{n} X_k\check{\varepsilon}_k^t, \qquad Q_n = \sum_{k=0}^{n} \varepsilon_k\check{\varepsilon}_k^t, \tag{E.3}$$

where $\check{\varepsilon}_n = \hat{\varepsilon}_n - \varepsilon_n$. The first equality of (4) can be rewritten as

$$\check{\varepsilon}_{n+1} = (1 - f_n)\pi_n - f_n\varepsilon_{n+1} \tag{E.4}$$

with $f_n = \Phi_n^t S_n^{-1}\Phi_n$. By (A.1) and (E.4) together with Chow's lemma [23], we have the almost sure convergence

$$\frac{1}{\log n}\sum_{k=0}^{n} \varepsilon_k\check{\varepsilon}_k^t \longrightarrow -\delta\Gamma. \tag{E.5}$$

Therefore, we immediately obtain $Q_n = o(\sqrt{n})$ a.s. In addition, by (A.1), we also have

$$\left\|\sum_{k=1}^{n} \pi_{k-1}\check{\varepsilon}_k^t\right\| = O(\log n) \qquad \text{a.s.} \tag{E.6}$$

Finally, as the trajectory x satisfies relation (59), we can conclude by (50), (A.1), and the Cauchy–Schwarz inequality that $P_n = o(\sqrt{n})$, so $R_n = o(\sqrt{n})$ a.s. Lemmas C.1 and C.2 together with (53) lead to (61) and (62) for the ELS algorithm. On the other hand, for the WLS algorithm, by (1) and (2), we have

$$S_{n-1}(a)(\hat{\theta}_n - \theta) = M_n(a) - R_{n-1}(a)\theta, \tag{E.7}$$

$$M_n(a) = M_0 + \sum_{k=1}^{n} a_{k-1}\Phi_{k-1}\varepsilon_k^t, \qquad R_n(a) = \sum_{k=0}^{n} a_k\Phi_k(\Phi_k - \Psi_k)^t \tag{E.8}$$

with $M_0 = S(\hat{\theta}_0 - \theta)$. Set

$$P_n(a) = \sum_{k=0}^{n} a_k X_k\check{\varepsilon}_k^t, \qquad Q_n(a) = \sum_{k=0}^{n} a_k\varepsilon_k\check{\varepsilon}_k^t. \tag{E.9}$$

The first equality of (4) can be rewritten as

$$\check{\varepsilon}_{n+1} = (1 - f_n(a))\pi_n - f_n(a)\varepsilon_{n+1} \tag{E.10}$$

with $f_n(a) = a_n\Phi_n^t S_n^{-1}(a)\Phi_n$. The main property of the weighted sequence $a = (a_n)$ is that

$$\sum_{n=1}^{\infty} a_n f_n(a) < +\infty \qquad \text{a.s.} \tag{E.11}$$

Therefore, as $a=(a_n)$ is nonincreasing, we find by (B.1), (E.10), and (E.11) that

$$\left\|\sum_{k=0}^{n} a_k \varepsilon_k \check{\varepsilon}_k^t\right\| = o((\log n)^{1+\gamma}) \qquad \text{a.s.}, \tag{E.12}$$

so $Q_n(a)=o(\sqrt{c_n})$ a.s. with $c_n=n/(\log n)^{2+2\gamma}$. In addition, by (B.1) and (D.5), as $f_n(a)$ tends a.s. towards 0,

$$\left\|\sum_{k=1}^{n} \pi_{k-1} \check{\varepsilon}_k^t\right\| = O(1) \qquad \text{a.s.} \tag{E.13}$$

Finally, as the trajectory x satisfies relation (60), we can conclude by (50), (B.1), and the Cauchy–Schwarz inequality that $P_n(a)=o(\sqrt{c_n})$, so $R_n(a)=o(\sqrt{c_n})$ a.s. Lemmas C.1 and C.2, together with (56), lead to (61) and (62) for the WLS algorithm, completing the proof of Theorem 3.5. □

REFERENCES

[1] K. J. Aström and B. Wittenmark, *On self tuning regulators*, Automatica, 9 (1973), pp. 185–199.

[2] A. H. Becker, P. R. Kumar, and C. Z. Wei, *Adaptive control with the stochastic approximation algorithm: Geometry and convergence*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 330–338.

[3] B. Bercu, *Sur l'estimateur des moindres carrés généralisé d'un modèle ARMAX. Application à l'identification des modèles ARMA*, Ann. Inst. H. Poincaré, 27 (1991), pp. 425–443.

[4] B. Bercu and M. Duflo, *Moindres carrés pondérés et poursuite*, Ann. Inst. Henri Poincaré, 28 (1992), pp. 403–430.

[5] B. Bercu, *Weighted estimation and tracking for ARMAX models*, SIAM J. Control Optim., 33 (1995), pp. 89–106.

[6] B. Bercu, *Théorème de limite centrale et loi du logarithme itéré pour les algorithmes des moindres carrés en poursuite adaptative*, Note au C. R. Acad. Sci. Paris Sér. I, 320 (1995), pp. 493–496.

[7] P. E. Caines, *Linear Stochastic Systems*, John Wiley, New York, 1988.

[8] H. F. Chen, *Recursive Estimation and Control for Stochastic Systems*, John Wiley, New York, 1985.

[9] H. F. Chen and L. Guo, *Convergence rate of least squares identification and adaptive control for stochastic systems*, Internat. J. Control, 44 (1986), pp. 1459–1476.

[10] H. F. Chen and J. F. Zhang, *Convergence rates in stochastic adaptive tracking*, Internat. J. Control, 49 (1989), pp. 1915–1935.

[11] H. F. Chen and L. Guo, *Identification and Stochastic Adaptive Control*, Birkhäuser, Boston, 1991.

[12] Y. S. Chow and H. Teicher, *Probability Theory: Independence, Interchangeability and Martingales*, Springer-Verlag, Berlin, 1978.

[13] M. Duflo, *Random Iterative Methods*, Springer-Verlag, Berlin, 1996.

[14] M. Duflo, R. Senoussi, and A. Touati, *Sur la loi des grands nombres pour les martingales vectorielles et l'estimateur des moindres carrés d'un modéle de régression*, Ann. Inst. H. Poincaré, 26 (1990), pp. 549–566.

[15] L. Guo and H. F. Chen, *The Aström-Wittenmark self-tuning regulator revisited and ELS-based adaptive trackers*, IEEE Trans. Automat. Control, 36 (1991), pp. 802–812.

[16] L. Guo, *Further results on least squares based adaptive minimum variance control*, SIAM J. Control Optim., 32 (1994). pp. 187–212.

[17] L. Guo, *Self convergence of weighted least squares with applications to stochastic adaptive control*, IEEE Trans. Automat. Control, 41 (1996), pp. 79–89.

[18] D. Hall and C. Heyde, *Martingale Limit Theory and Its Applications*, Academic Press, New York, 1980.

[19] T. L. Lai and C. Z. Wei, *Extended least squares and their applications to adaptive control and prediction in linear systems*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 898–906.

[20] T. L. Lai and C. Z. Wei, *On the concept of excitation in least squares identification and adaptive control*, Stochastics, 16 (1986), pp. 227–254.
[21] R. S. Liptser and A. N. Shiryaev, *A functional central limit theorem for semimartingales*, Theory Probab. Appl., 25 (1980), pp. 667–688.
[22] K. Nassiri-Toussi and W. Ren, *Indirect adaptive pole-placement control of MIMO stochastic systems: Self-tuning results*, IEEE Trans. Automat. Control, 42 (1997), pp. 38–52.
[23] W. F. Stout, *Almost Sure Convergence*, Academic Press, New York, 1974.
[24] W. F. Stout, *A martingale analogue of Kolmogorov's law of the iterated logarithm*, Z. Wahrscheinlichkeitstheorie, 15 (1970), pp. 279–290.
[25] A. Touati, *On the functional convergence in distribution of sequences of semimartingales to a mixture of brownian motions*, Theory Probab. Appl., 36 (1991), pp. 752–771.
[26] A. Touati, *Vitesse de convergence en loi de l'estimateur des moindres carrés d'un modèle autoregressif (cas mixte)*, Ann. Inst. H. Poincaré, 32 (1996), pp. 211–230.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 3, pp. 929–947, May 1998

007

STOCHASTIC NEAR-OPTIMAL CONTROLS: NECESSARY AND SUFFICIENT CONDITIONS FOR NEAR-OPTIMALITY*

XUN YU ZHOU†

Abstract. Near-optimization is as sensible and important as optimization for both theory and applications. This paper concerns dynamic near-optimization, or near-optimal controls, for systems governed by the Ito stochastic differential equations (SDEs), where both the drift and diffusion terms are allowed to depend on controls and the systems are allowed to be degenerate. Necessary and sufficient conditions for a control to be near-optimal are studied. It is shown that any near-optimal control nearly maximizes the "$\mathcal{H}$-function" (which is a generalization of the usual Hamiltonian and is quadratic with respect to the diffusion coefficients) in some integral sense, and vice versa if certain additional concavity conditions are imposed. Error estimates for both the near-optimality of the controls and the near-maximum of the $\mathcal{H}$-function are obtained, based on some delicate estimates of the adjoint processes. Examples are presented to demonstrate the results.

Key words. stochastic near-optimal control, necessary and sufficient condition, adjoint equation, $\mathcal{H}$-function, Hamiltonian, Ekeland's principle

AMS subject classifications. 93E, 49K

PII. S0363012996302664

1. Introduction.

This paper is one in a series of papers studying near-optimal controls. In view of both theory and applications, near-optimality makes as good sense as the (exact) optimality. First, many more near-optimal controls are available than optimal ones. Indeed, optimal controls may not even exist in many situations, while near-optimal controls *always* exist. Second, it is usually much easier to obtain near-optimal controls than optimal ones, both analytically and numerically. For example, optimal production controls for a stochastic two-machine flowshop may involve very complicated switching curves. However, Sethi and Zhou [11] showed that a near-optimal control can be found in the class of the so-called threshold-type policies which involve only two real parameters. Therefore, the original problem can be greatly simplified in analysis, computation, and implementation by considering near-optimal controls. Third, since there are many more candidates for near-optimal controls, it is possible to select among them appropriate ones that are easier for analysis and implementation. For example, optimal feedback controls for linear systems are usually of "bang-bang" type, as is well known. These controls are not continuous in state, making it very difficult to handle analytically. Indeed, even the existence of system states under such controls are not clear in general. However, one can always modify the bang-bang controls into Lipschitz continuous controls with only a small loss in the objective value. Fourth, an optimal control is usually very sensitive for the external perturbation because it is too "greedy" (by the nature of "optimality"!) and does not usually leave allowance to accommodate changing situations. This pitfall becomes significant for stochastic systems, where uncertainties and perturbations are inherent. A good example is optimal production planning for manufacturing systems. The so-called zero-inventory policy is in general optimal in deterministic cases [6], but is no

*Received by the editors April 29, 1996; accepted for publication (in revised form) March 5, 1997. This work was supported by the RGC Earmarked Grant CUHK 249/94E.

http://www.siam.org/journals/sicon/36-3/30266.html

†Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, Hong Kong (xyzhou@se.cuhk.edu.hk).

longer optimal in stochastic cases, even for the simplest one-machine, one-product systems [1]. The reason is that the zero-inventory policy, while resulting in zero surplus cost, is prone to cause shortage (due to various uncertainties involved) which is usually even more costly. Last but not least, for many real systems, insisting on optimal solutions is not only unrealistic but also *unnecessary*, because a near-optimal solution can satisfactorily serve the ultimate purpose of the decision makers in most practical situations.

In the first two papers [13, 14] in this series, near-optimal controls for deterministic dynamic systems (governed by ordinary differential equations) are investigated. Necessary and sufficient conditions of near-optimality are derived and dynamic programming approach is employed to study the behavior of those "good-enough" controls. Starting from this paper, we proceed to treat the stochastic cases. As mentioned earlier, near-optimization makes even better sense in stochastic systems than in deterministic ones due to the presence of uncertainties.

The stochastic systems under consideration in this paper are of the diffusion type, namely, they are governed by the Ito SDEs. As is well known, this kind of models occur in many real-world systems, including those of finance, economy, and manufacturing. The specific purpose of this paper is to derive necessary and sufficient conditions of near-optimality for the controlled diffusion processes. Elliott and Kohlmann [4] studied the necessity part, but as in almost all the previous papers studying the similar problem for deterministic cases based on the Ekeland variational principle, their necessary conditions were derived only for *some* near-optimal controls (see [4, Theorem 4.4]). In this paper, we shall derive necessary conditions for *all* near-optimal controls. More specifically, we will show that *any* ε-optimal control nearly maximizes the so-called $\mathcal{H}$-function in an integral form with an error order of "almost" $\varepsilon^{\frac{1}{3}}$ (see Theorem 4.1 for the precise meaning). Moreover, we prove that under certain concavity conditions, an ε-maximum condition in terms of the $\mathcal{H}$-function in the integral form is sufficient for the near-optimality of order $\varepsilon^{\frac{1}{2}}$. Our results are based on some delicate estimates for the solutions of adjoint equations, which are linear backward SDEs, along with the Ekeland principle.

The plan of the rest of the paper is as follows. In section 2, we formulate the problem and define near-optimality. In section 3, we study some continuity of the adjoint processes with respect to a prescribed metric in the set of admissible controls. Sections 4 and 5 are devoted to the necessary and sufficient conditions for near-optimal controls, respectively. Section 6 discusses the results obtained and presents an example to demonstrate them. Finally, section 7 concludes the paper.

2. Problem formulation and preliminaries. We consider in this paper stochastic optimal control problems of the following kind. For a given $s \in [0, T]$, by the set of admissible controls $U_{ad}[s, T]$ we mean the collection of (i) standard probability spaces $(\Omega, \mathcal{F}, P)$ along with l-dimensional Brownian motions $B = \{B(t) : s \leq t \leq T\}$ with $B(s) = 0$, and (ii) Γ-valued $\mathcal{F}_t^s$-adapted measurable processes $u(\cdot) = \{u(t) : s \leq t \leq T\}$, where $\mathcal{F}_t^s$ is the natural filtration generated by $B(t)$ augmented by all the P-null sets in $\mathcal{F}$, and Γ is a given closed set in some Euclidean space R^m. We denote $(\Omega, \mathcal{F}, P, B; u(\cdot)) \in U_{ad}[s, T]$, but occasionally we will write only $u(\cdot) \in U_{ad}[s, T]$ if no ambiguity arises.

Let $(s, y) \in [0, T) \times R^n$ be given, representing the initial time and initial state, respectively, of the system. For each $(\Omega, \mathcal{F}, P, B; u(\cdot)) \in U_{ad}[s, T]$, the corresponding cost is

$$J(s,y;u(\cdot)) = E\left\{\int_s^T L(t,x(t),u(t))dt + h(x(T))\right\}, \tag{2.1}$$

where $x(\cdot) = \{x(t) : s \le t \le T\}$ is the solution of the following Ito SDE on the filtered space $(\Omega, \mathcal{F}, P; \mathcal{F}_t^s)$:

$$\begin{cases} dx(t) = f(t,x(t),u(t))dt + \sigma(t,x(t),u(t))dB(t), \\ x(s) = y. \end{cases} \tag{2.2}$$

The solution $x(\cdot)$ of the above SDE is called the response of the control $u(\cdot) \in U_{ad}[s,T]$, and $(x(\cdot),u(\cdot))$ is called an *admissible* pair. The objective of the optimal control problem is to minimize the cost function $J(s,y;u(\cdot))$, for a given $(s,y) \in [0,T) \times R^n$, over all $u(\cdot) \in U_{ad}[s,T]$. We denote the above problem by $C_{s,y}$ to recall the dependence on the initial time s and the initial state y. The value function is defined as

$$V(s,y) = \inf_{u(\cdot)\in U_{ad}[s,T]} J(s,y;u(\cdot)). \tag{2.3}$$

An admissible pair $(x^*(\cdot),u^*(\cdot))$ is called *optimal* for $C_{s,y}$ if $u^*(\cdot)$ achieves the infimum of $J(s,y;u(\cdot))$ over $U_{ad}[s,T]$.

Since the objective of this paper is to study near-optimal rather than optimal controls of the system, we give here the precise definition of the near-optimality, following [13].

DEFINITION 2.1. *For a given $\varepsilon > 0$, an admissible pair $(x^\varepsilon(\cdot),u^\varepsilon(\cdot))$, or simply $u^\varepsilon(\cdot)$, is called* ε-optimal *with respect to (s,y) if*

$$|J(s,y;u^\varepsilon(\cdot)) - V(s,y)| \le \varepsilon.$$

DEFINITION 2.2. *Both a family of admissible pairs $\{(x^\varepsilon(\cdot),u^\varepsilon(\cdot))\}$ parameterized by $\varepsilon > 0$ and any element $(x^\varepsilon(\cdot),u^\varepsilon(\cdot))$, or simply $u^\varepsilon(\cdot)$, in the family are called* near-optimal *with respect to (s,y) if*

$$|J(s,y;u^\varepsilon(\cdot)) - V(s,y)| \le r(\varepsilon)$$

holds for sufficiently small ε, where r is a function of ε satisfying $r(\varepsilon) \to 0$ as $\varepsilon \to 0$. The estimate $r(\varepsilon)$ is called an error bound. *If $r(\varepsilon) = c\varepsilon^\delta$ for some $\delta > 0$ independent of the constant c, then $u^\varepsilon(\cdot)$ is called near-optimal with order ε^δ.*

In the above definitions, the terms "admissible," "optimal," "ε-optimal," and "near-optimal" are dependent on the initial time s and initial state y. In the following discussion, however, the phrase "with respect to (s,y)" may be omitted if no confusion would occur.

Notation. We make use of the following notation in this paper:

$a \cdot b$	:	the inner product of any two vectors a and b;
$\lvert a\rvert$	:	$= \lvert a^1\rvert + \cdots + \lvert a^n\rvert$ for any vector $a = (a^1, \ldots, a^n)$;
M^*	:	the transpose of any vector or matrix M;
ρ_x	:	the gradient or Jacobian of a function ρ with respect to the variable x;

ρ_{xx}	:	the Hessian of a scalar function ρ with respect to the variable x;
χ_A	:	the indicator function of a set A;
$X+Y$	:	$=\{x+y : x\in X, y\in Y\}$ for any set X and Y;
$C, C_i, i=1,2,\dots$	:	multiplicative constants required in the analysis.

Given a probability space $(\Omega,\mathcal{F},P)$ with a filtration $\{\mathcal{F}_t : a\le t\le b\}$ $(-\infty\le a<b\le+\infty)$, a Hilbert space X with the norm $\|\cdot\|_X$, and p $(1\le p\le+\infty)$, define the Banach space $L^p_{\mathcal{F}}(a,b;X)=\{\phi(\cdot)=\{\phi(t,\omega) : a\le t\le b\}|\ \phi(\cdot)$ is an $\mathcal{F}_t$ - adapted, X-valued measurable process on $[a,b]$, and $E\int_a^b \|\phi(t,\omega)\|_X^p\ dt<+\infty\}$, with the norm

$$\|\phi(\cdot)\|_{\mathcal{F},p}=\left(E\int_a^b \|\phi(t,\omega)\|_X^p\ dt\right)^{\frac{1}{p}}.$$

In the rest of this paper, we shall employ the usual convention of suppressing the ω-dependence of all random functions whenever no confusion arises.

Assumptions. The following basic assumptions will be in force throughout this paper:

(A1) $f:[0,T]\times R^n\times\Gamma\to R^n$, $\sigma:[0,T]\times R^n\times\Gamma\to R^{n\times l}$, and $L:[0,T]\times R^n\times\Gamma\to R^1$ are measurable in (t,x,u), twice continuously differentiable in x for each (t,u), and there exists a constant $C>0$ such that for $\rho=f,\sigma,L$,

$$|\rho(t,x,u)|\le C(1+|x|), \tag{2.4}$$

$$|\rho(t,x,u)-\rho(t,x',u)|+|\rho_x(t,x,u)-\rho_x(t,x',u)|\le C|x-x'|. \tag{2.5}$$

(A2) $h:R^n\to R^1$ is twice continuously differentiable, and

$$|h(x)|\le C(1+|x|), \tag{2.6}$$

$$|h(x)-h(x')|+|h_x(x)-h_x(x')|\le C|x-x'|. \tag{2.7}$$

Remark 2.1. Under (A1) and (A2), the existence and uniqueness of (strong) solutions of (2.2) hold for any given $(\Omega,\mathcal{F},P,B;u(\cdot))\in U_{ad}[s,T]$. Moreover, it is well known that for every solution $x(\cdot)$ of (2.2) and any $p\ge 0$, it holds that

$$E\sup_{s\le t\le T}|x(t)|^p\le C(p), \tag{2.8}$$

where $C(p)$ is a constant depending only on p.

Before concluding this section, let us recall the definition of the Clarke generalized gradient as well as the Ekeland variational principle.

DEFINITION 2.3 (Clarke [2]). *Let X be a convex set in R^d and let $\eta(\cdot):X\to R^1$ be a locally Lipschitz function. The* generalized gradient *of v at $\hat{x}\in X$, denoted by $\partial_x\eta(\hat{x})$, is a set defined by*

$$\partial_x\eta(\hat{x})=\{p\in R^d | p\cdot\xi\le\eta^0(\hat{x};\xi),\ \textit{for any } \xi\in R^d\},$$

where $\eta^0(\hat{x};\xi) = \limsup_{x\in X, x+h\xi\in X, x\to\hat{x}, h\to 0+} \frac{\eta(x+h\xi)-\eta(x)}{h}$.

LEMMA 2.4 (Ekeland's principle [3]). *Let* (S,d) *be a complete metric space and* $\rho(\cdot): S\to R^1$ *be lower-semicontinuous and bounded from below. For* $\varepsilon\geq 0$, *suppose* $u^\varepsilon\in S$ *satisfies*

$$\rho(u^\varepsilon)\leq \inf_{u\in S}\rho(u)+\varepsilon.$$

Then for any $\lambda>0$, *there exists* $u^\lambda\in S$ *such that*

$$\begin{aligned}&\rho(u^\lambda)\leq\rho(u^\varepsilon),\\ &d(u^\lambda,u^\varepsilon)\leq\lambda,\ \textit{and}\\ &\rho(u^\lambda)\leq\rho(u)+\tfrac{\varepsilon}{\lambda}d(u,u^\lambda)\ \textit{for all}\ u\in S.\end{aligned}$$

3. Adjoint equations: Backward SDEs. From now on, let us assume that the initial time s and initial state y of the system are fixed, and the Brownian motion B is one dimensional for notational simplicity. We also set $\mathcal{F}_t=\mathcal{F}_t^s$. Define a metric on $U_{ad}[s,T]$:

$$d(u(\cdot),u'(\cdot))=\tilde{P}\{(t,\omega)\in[s,T]\times\Omega: u(t,\omega)\neq u'(t,\omega)\},\tag{3.1}$$

where $\tilde{P}$ is the product measure of the Lebesgue measure and P. Since Γ is closed, it can be shown similarly to [4, Lemma 3.2] that $U_{ad}[s,T]$ is a complete metric space under d.

As is well known, the study of adjoint equations plays a key role in deriving the necessary and sufficient conditions of optimality. In the stochastic (diffusion) case, there are first-order and second-order adjoint processes corresponding to each given admissible control. The equations which they satisfy are (linear) backward SDEs with the terminal conditions given. This section is mainly devoted to investigating certain continuity of the adjoint processes with respect to the metric d.

Our first lemma below, however, is concerned with the continuity of the state processes under d.

LEMMA 3.1. *For any* $0<\alpha<1$ *and* $p\geq 0$ *satisfying* $\alpha p<1$, *there is a constant* $C_1=C_1(\alpha,p)>0$ *such that for any* $u(\cdot),u'(\cdot)\in U_{ad}[s,T]$ *along with the corresponding trajectories* $x(\cdot),x'(\cdot)$, *it holds that*

$$E\sup_{s\leq t\leq T}|x(t)-x'(t)|^{2p}\leq C_1 d(u(\cdot),u'(\cdot))^{\alpha p}.\tag{3.2}$$

Proof. In the proof below, all the constants C_i are understood to be dependent only on α and p. First we assume $p\geq 1$. We can compute, for any $r\geq s$,

$$\begin{aligned}&E\sup_{s\leq t\leq r}|x(t)-x'(t)|^{2p}\\ \leq\ & C_2E\textstyle\int_s^r\{|f(t,x(t),u(t))-f(t,x'(t),u'(t))|^{2p}\\ &+|\sigma(t,x(t),u(t))-\sigma(t,x'(t),u'(t))|^{2p}\}dt\\ \leq\ & C_3E\textstyle\int_s^r\{|f(t,x(t),u(t))-f(t,x(t),u'(t))|^{2p}\\ &+|\sigma(t,x(t),u(t))-\sigma(t,x(t),u'(t))|^{2p}\}\chi_{u(t)\neq u'(t)}(t)dt\\ &+C_3E\textstyle\int_s^r\{|f(t,x(t),u'(t))-f(t,x'(t),u'(t))|^{2p}\\ &+|\sigma(t,x(t),u'(t))-\sigma(t,x'(t),u'(t))|^{2p}\}dt.\end{aligned}\tag{3.3}$$

Taking $q' = \frac{1}{\alpha p} > 1$ and $p' > 1$ such that $\frac{1}{p'} + \frac{1}{q'} = 1$, and applying the Cauchy–Schwarz inequality, we obtain

$$
\begin{aligned}
&E\textstyle\int_s^r \{|f(t,x(t),u(t)) - f(t,x(t),u'(t))|^{2p}\}\chi_{u(t)\neq u'(t)}(t)dt \\
&\le \big\{E\textstyle\int_s^r |f(t,x(t),u(t)) - f(t,x(t),u'(t))|^{2pp'}\big\}^{\frac{1}{p'}} \big\{E\textstyle\int_s^r \chi_{u(t)\neq u'(t)}(t)dt\big\}^{\frac{1}{q'}} \\
&\le C_4\big\{E\textstyle\int_s^r (1+|x(t)|^{2pp'})dt\big\}^{\frac{1}{p'}} d(u(\cdot),u'(\cdot))^{\alpha p} \\
&\le C_5 d(u(\cdot),u'(\cdot))^{\alpha p}.
\end{aligned} \tag{3.4}
$$

The same inequality holds if f above is replaced by σ. Therefore, by noting (A1), we conclude from (3.3) that

$$
E\sup_{s\le t\le r}|x(t)-x'(t)|^{2p} \le C_6\left\{\int_s^r E\sup_{s\le t\le\theta}|x(t)-x'(t)|^{2p}d\theta + d(u(\cdot),u'(\cdot))^{\alpha p}\right\}.
$$

Hence (3.2) follows from the Gronwall inequality.

Now assume $0 \le p < 1$. Then the Cauchy–Schwarz inequality yields

$$
\begin{aligned}
E\sup_{s\le t\le T}|x(t)-x'(t)|^{2p} &\le \{E\sup_{s\le t\le T}|x(t)-x'(t)|^2\}^p \\
&\le \{C_1 d(u(\cdot),u'(\cdot))^{\alpha}\}^p \\
&= C_1^p d(u(\cdot),u'(\cdot))^{\alpha p}.
\end{aligned}
$$

This completes the proof. □

For any $u(\cdot) \in U_{ad}[s,T]$ and the corresponding state trajectory $x(\cdot)$, we define the first-order adjoint process $\psi(\cdot)$ and the second-order adjoint process $Q(\cdot)$ as the ones satisfying the following two backward SDEs, respectively:

$$
\begin{cases}
d\psi(t) = & -\{f_x(t,x(t),u(t))^*\psi(t) + \sigma_x(t,x(t),u(t))^*K(t) \\
& \quad + L_x(t,x(t),u(t))\}dt + K(t)dB(t), \\
\psi(T) = & h_x(x(T)),
\end{cases} \tag{3.5}
$$

(3.6)

$$
\begin{cases}
dQ(t) = & -\{f_x(t,x(t),u(t))^*Q(t) + Q(t)f_x(t,x(t),u(t)) \\
& \quad + \sigma_x(t,x(t),u(t))^*Q(t)\sigma_x(t,x(t),u(t)) + \sigma_x(t,x(t),u(t))^*R(t) \\
& \quad + R(t)\sigma_x(t,x(t),u(t)) + \Lambda(t)\}dt + R(t)dB(t), \\
Q(T) = & h_{xx}(x(T)),
\end{cases}
$$

where $\Lambda(t) = L_{xx}(t,x(t),u(t)) + \sum_{i=1}^n \{\psi^i(t)f^i_{xx}(t,x(t),u(t)) + K^i(t)\sigma^i_{xx}(t,x(t),u(t))\}$.

Note that under assumptions (A1) and (A2), the first-order adjoint equation (3.5) admits one and only one $\mathcal{F}_t$-*adapted* solution *pair* $(\psi,K) \in L^2_{\mathcal{F}}(0,1;R^n)\times L^2_{\mathcal{F}}(0,1;R^n)$, and the second-order adjoint equation (3.6) admits one and only one $\mathcal{F}_t$-*adapted* solution *pair* $(Q,R) \in L^2_{\mathcal{F}}(0,1;R^{n\times n}) \times L^2_{\mathcal{F}}(0,1;R^{n\times n})$. Moreover, since f_x, σ_x, L_x, and h_x are bounded by C by assumptions (A1) and (A2), there exists a constant $C_1 > 0$, independent of $(x(\cdot),u(\cdot))$, such that the solutions of (3.5) and (3.6) have the following estimates:

$$
E\left\{\sup_{s\le t\le T}|\psi(t)|^2 + \int_s^T |K(t)|^2dt\right\} \le C_1, \tag{3.7}
$$

$$E\left\{\sup_{s\le t\le T}|Q(t)|^2+\int_s^T|R(t)|^2dt\right\}\le C_1; \tag{3.8}$$

(see [12, Corollaries 2.2, 2.4]). It should also be noted that *no* supremum estimates for $K(t)$ and $R(t)$ are available.

The following lemma gives the pth moment continuity of the solutions to the adjoint equations with respect to the metric d. It plays a key role in proving the necessary condition in the next section.

LEMMA 3.2. *For any $0<\alpha<1$ and $1<p<2$ satisfying $(1+\alpha)p<2$, there is a constant $C_1=C_1(\alpha,p)>0$ such that for any $u(\cdot),u'(\cdot)\in U_{ad}[s,T]$ along with the corresponding trajectories $x(\cdot),x'(\cdot)$ and the solutions $(\psi(\cdot),K(\cdot),Q(\cdot),R(\cdot)),(\psi'(\cdot),K'(\cdot),Q'(\cdot),R'(\cdot))$ of the corresponding adjoint equations, it holds that*

$$E\int_s^T\{|\psi(t)-\psi'(t)|^p+|K(t)-K'(t)|^p\}dt\le C_1d(u(\cdot),u'(\cdot))^{\frac{\alpha p}{2}} \tag{3.9}$$

and

$$E\int_s^T\{|Q(t)-Q'(t)|^p+|R(t)-R'(t)|^p\}dt\le C_1d(u(\cdot),u'(\cdot))^{\frac{\alpha p}{2}}. \tag{3.10}$$

Proof. Once again, all the constants C_i below depend only on α and p. Note that $(\bar\psi(t),\bar K(t))\equiv(\psi(t)-\psi'(t),K(t)-K'(t))$ satisfies the following backward SDE:

$$\begin{cases} d\bar\psi(t)= & -\{f_x(t,x(t),u(t))^*\bar\psi(t)+\sigma_x(t,x(t),u(t))^*\bar K(t)\\ & \quad+\bar f(t)\}dt+\bar K(t)dB(t),\\ \bar\psi(T)= & h_x(x(T))-h_x(x'(T)),\end{cases} \tag{3.11}$$

where

$$\begin{aligned}\bar f(t)= & \{f_x(t,x(t),u(t))^*-f_x(t,x'(t),u'(t))^*\}\psi'(t)\\ & +\{\sigma_x(t,x(t),u(t))^*-\sigma_x(t,x'(t),u'(t))^*\}K'(t)\\ & +\{L_x(t,x(t),u(t))-L_x(t,x'(t),u'(t))\}.\end{aligned} \tag{3.12}$$

Now let ρ be the solution of the following linear (forward) SDE:

$$\begin{cases} d\rho(t)= & \{f_x(t,x(t),u(t))\rho(t)+|\bar\psi(t)|^{p-1}\mathrm{sgn}(\bar\psi(t))\}dt\\ & +\{\sigma_x(t,x(t),u(t))\rho(t)+|\bar K(t)|^{p-1}\mathrm{sgn}(\bar K(t))\}dB(t),\\ \rho(s)=0,\end{cases}$$

where $\mathrm{sgn}(a)\equiv(\mathrm{sgn}(a^1),\dots,\mathrm{sgn}(a^n))^*$ for a vector $a=(a^1,\dots,a^n)^*$. Note that the existence and uniqueness of solutions to the above equation are verified by (A1) and the fact that

$$E\int_s^T\left\{\left||\bar\psi(t)|^{p-1}\mathrm{sgn}(\bar\psi(t))\right|^2+\left||\bar K(t)|^{p-1}\mathrm{sgn}(\bar K(t))\right|^2\right\}dt<+\infty.$$

Since $1<p<2$, there is $q>2$ such that $\frac1p+\frac1q=1$. So

$$\begin{aligned}E\sup_{s\le t\le T}|\rho(t)|^q\le & \ C_2E\textstyle\int_s^T\{|\bar\psi(t)|^{pq-q}+|\bar K(t)|^{pq-q}\}dt\\ = & \ C_2E\textstyle\int_s^T\{|\bar\psi(t)|^p+|\bar K(t)|^p\}dt.\end{aligned} \tag{3.13}$$

Note that the right side of (3.13) is bounded due to (3.7). On the other hand, by applying the Ito formula to $\bar{\psi}(t)\cdot\rho(t)$ and taking expectations, we obtain

$$\begin{aligned}
& E\int_s^T\Big\{\bar{\psi}(t)\cdot[|\bar{\psi}(t)|^{p-1}\mathrm{sgn}(\bar{\psi}(t))]+\bar{K}(t)\cdot[|\bar{K}(t)|^{p-1}\mathrm{sgn}(\bar{K}(t))]\Big\}dt\\
= \;& E\Big\{\int_s^T\bar{f}(t)\cdot\rho(t)dt+[h_x(x(T))-h_x(x'(T))]\cdot\rho(T)\Big\}\\
\le \;& \{E\int_s^T|\bar{f}(t)|^pdt\}^{\frac{1}{p}}\{E\int_s^T|\rho(t)|^qdt\}^{\frac{1}{q}}+\{E|h_x(x(T))-h_x(x'(T))|^p\}^{\frac{1}{p}}\{E|\rho(T)|^q\}^{\frac{1}{q}}\\
\le \;& C_3\Big\{E\int_s^T[|\bar{\psi}(t)|^p+|\bar{K}(t)|^p]dt\Big\}^{\frac{1}{q}}\Big\{[E\int_s^T|\bar{f}(t)|^pdt]^{\frac{1}{p}}+[E|h_x(x(T))-h_x(x'(T))|^p]^{\frac{1}{p}}\Big\}.
\end{aligned}$$

Noting that the left side of the above inequality is equal to $E\int_s^T[|\bar{\psi}(t)|^p+|\bar{K}(t)|^p]dt$, we deduce

$$E\int_s^T\{|\bar{\psi}(t)|^p+|\bar{K}(t)|^p\}dt\le C_4E\left\{\int_s^T|\bar{f}(t)|^pdt+|h_x(x(T))-h_x(x'(T))|^p\right\}. \tag{3.14}$$

We proceed to estimate the right side of (3.14). First, noting that $\frac{\alpha p}{2}<1-\frac{p}{2}<1$, we can apply Lemma 3.1 to get

$$E|h_x(x(T))-h_x(x'(T))|^p\le C^pE|x(T)-x'(T)|^p\le C_5d(u(\cdot),u'(\cdot))^{\frac{\alpha p}{2}}. \tag{3.15}$$

Next, by repeatedly applying the Cauchy–Schwarz inequality, we can estimate

$$\begin{aligned}
& E\int_s^T|f_x(t,x(t),u(t))^*-f_x(t,x'(t),u'(t))^*|^p|\psi'(t)|^pdt\\
\le \;& C_6E\int_s^T\{|f_x(t,x(t),u(t))^*-f_x(t,x(t),u'(t))^*|^p|\psi'(t)|^p\\
& \qquad +|f_x(t,x(t),u'(t))^*-f_x(t,x'(t),u'(t))^*|^p|\psi'(t)|^p\}dt\\
\le \;& C_7E\int_s^T\{\chi_{u(t)\neq u'(t)}(t)|\psi'(t)|^p+|x(t)-x'(t)|^p|\psi'(t)|^p\}dt\\
\le \;& C_7\Big\{E\int_s^T|\psi'(t)|^2dt\Big\}^{\frac{p}{2}}d(u(\cdot),u'(\cdot))^{1-\frac{p}{2}}\\
& +C_7\Big\{E\int_s^T|\psi'(t)|^2dt\Big\}^{\frac{p}{2}}\Big\{E\int_s^T|x(t)-x'(t)|^{\frac{2p}{2-p}}dt\Big\}^{1-\frac{p}{2}}.
\end{aligned} \tag{3.16}$$

Noting that $1-\frac{p}{2}>\frac{\alpha p}{2}$ and $d(u(\cdot),u'(\cdot))\le1$, we know that the first term of the right side of (3.16) is bounded by $C_8d(u(\cdot),u'(\cdot))^{\frac{\alpha p}{2}}$. Further, since $\frac{\alpha p}{2-p}<1$, we conclude from Lemma 3.1 that

$$E\int_s^T|x(t)-x'(t)|^{\frac{2p}{2-p}}dt\le C_9d(u(\cdot),u'(\cdot))^{\frac{\alpha p}{2-p}}.$$

Hence the second term in the right side of (3.16) is also dominated by $C_{10}d(u(\cdot),u'(\cdot))^{\frac{\alpha p}{2}}$. So, (3.16) yields

$$E\int_s^T|f_x(t,x(t),u(t))^*-f_x(t,x'(t),u'(t))^*|^p|\psi'(t)|^pdt\le C_{11}d(u(\cdot),u'(\cdot))^{\frac{\alpha p}{2}}. \tag{3.17}$$

Notice that the estimate in (3.16) involved only $E\int_s^T|\psi'(t)|^2dt$ and therefore is adaptable to terms involving $K'(t)$. Hence we can similarly prove that

$$E\int_s^T|\sigma_x(t,x(t),u(t))^*-\sigma_x(t,x'(t),u'(t))^*|^p|K'(t)|^pdt\le C_{12}d(u(\cdot),u'(\cdot))^{\frac{\alpha p}{2}}. \tag{3.18}$$

Similarly (in fact, more easily) one can prove that

$$E\int_s^T |L_x(t,x(t),u(t)) - L_x(t,x'(t),u'(t))|^p dt \le C_{13} d(u(\cdot),u'(\cdot))^{\frac{\alpha p}{2}}. \tag{3.19}$$

It follows from (3.12), (3.17), (3.18), and (3.19) that

$$E\int_s^T |\bar f(t)|^p dt \le C_{14} d(u(\cdot),u'(\cdot))^{\frac{\alpha p}{2}}.$$

The desired result (3.9) then follows immediately from (3.14) and (3.15). Similarly one can prove (3.10). □

4. Necessary conditions of near-optimality. Define the (usual) Hamiltonian

$$H(t,x,u,p,q) = -L(t,x,u) - p\cdot f(t,x,u) - q\cdot\sigma(t,x,u) \tag{4.1}$$

for $(t,x,u,p,q)\in[s,T]\times R^n\times\Gamma\times R^n\times R^n$. Furthermore, we define the $\mathcal{H}$-function *corresponding to* a given admissible pair $(x(\cdot),u(\cdot))$ as follows:

$$\begin{aligned}\mathcal{H}^{(x(\cdot),u(\cdot))}(t,x,u) &= H(t,x,u,\psi(t),K(t)-Q(t)\sigma(t,x(t),u(t)))\\ &\quad -\frac{1}{2}\sigma(t,x,u)^*Q(t)\sigma(t,x,u)\end{aligned} \tag{4.2}$$

for $(t,x,u)\in[s,T]\times R^n\times\Gamma$, where $\psi(t)$, $K(t)$, and $Q(t)$ are determined by adjoint equations (3.5) and (3.6) corresponding to $(x(\cdot),u(\cdot))$.

THEOREM 4.1. *For any $\gamma\in[0,\frac{1}{3})$, there exists a constant $C_1=C_1(\gamma)>0$ such that for any $\varepsilon>0$ and any ε-optimal pair $(x^\varepsilon(\cdot),u^\varepsilon(\cdot))$ of the problem $C_{s,y}$, it holds that*

$$\begin{aligned}E\int_s^T \Big\{&\tfrac{1}{2}[\sigma(t,x^\varepsilon(t),u)-\sigma(t,x^\varepsilon(t),u^\varepsilon(t))]^*Q^\varepsilon(t)[\sigma(t,x^\varepsilon(t),u)-\sigma(t,x^\varepsilon(t),u^\varepsilon(t))]\\ &+\psi^\varepsilon(t)\cdot[f(t,x^\varepsilon(t),u)-f(t,x^\varepsilon(t),u^\varepsilon(t))]\\ &+K^\varepsilon(t)\cdot[\sigma(t,x^\varepsilon(t),u)-\sigma(t,x^\varepsilon(t),u^\varepsilon(t))]\\ &+L(t,x^\varepsilon(t),u)-L(t,x^\varepsilon(t),u^\varepsilon(t))\Big\}dt \ge -C_1\varepsilon^\gamma,\end{aligned} \tag{4.3}$$

where $(\psi^\varepsilon(\cdot),K^\varepsilon(\cdot))$ and $(Q^\varepsilon(\cdot),R^\varepsilon(\cdot))$ are the solutions to (3.5) *and* (3.6)*, respectively, corresponding to $(x^\varepsilon(\cdot),u^\varepsilon(\cdot))$.*

Proof. First note that the multiplicative constants C_i required in the analysis below do not depend on ε. By assumptions (A1) and (A2), it is easy to see that $J(s,y;u(\cdot))$ is continuous on $U_{ad}[s,T]$ endowed with the metric d defined by (3.1). By the Ekeland principle (Lemma 2.1) with $\lambda=\varepsilon^{\frac{2}{3}}$, there is an admissible pair $(\tilde x^\varepsilon(\cdot),\tilde u^\varepsilon(\cdot))$ such that

$$d(u^\varepsilon(\cdot),\tilde u^\varepsilon(\cdot)) \le \varepsilon^{\frac{2}{3}} \tag{4.4}$$

and

$$\tilde J(s,y;\tilde u^\varepsilon(\cdot)) \le \tilde J(s,y;u(\cdot)) \text{ for any } u(\cdot)\in U_{ad}[s,T], \tag{4.5}$$

where

$$\tilde J(s,y;u(\cdot)) = J(s,y;u(\cdot)) + \varepsilon^{\frac{1}{3}} d(u(\cdot),\tilde u^\varepsilon(\cdot)). \tag{4.6}$$

This means that $(\tilde{x}^\varepsilon(\cdot), \tilde{u}^\varepsilon(\cdot))$ is an optimal pair for the system (2.2) with a new cost function $\tilde{J}$. Next we use the spike variation technique to derive a "maximum principle" for $(\tilde{x}^\varepsilon(\cdot), \tilde{u}^\varepsilon(\cdot))$. To this end, let $\bar{t} \in [s, T)$ and $u \in \Gamma$ be fixed. For any $\delta > 0$, define $u_\delta \in U_{ad}[s, T]$ by

$$u_\delta(t) = \begin{cases} u, & t \in [\bar{t}, \bar{t} + \delta], \\ \tilde{u}^\varepsilon(t), & t \in [s, T] \setminus [\bar{t}, \bar{t} + \delta]. \end{cases} \tag{4.7}$$

The fact that

$$\tilde{J}(s, y; \tilde{u}^\varepsilon(\cdot)) \le \tilde{J}(s, y; u_\delta(\cdot))$$

and

$$d(u_\delta(\cdot), \tilde{u}^\varepsilon(\cdot)) \le \delta$$

imply that

$$-\varepsilon^{\frac{1}{3}}\delta \le J(s, y; u_\delta(\cdot)) - J(s, y; \tilde{u}^\varepsilon(\cdot)). \tag{4.8}$$

However, by (5.14) in [12] (with ε there replaced by δ), the right-hand side of the above inequality is equal to

$$\begin{aligned} E\int_{\bar{t}}^{\bar{t}+\delta} \Big\{ & \tfrac{1}{2}[\sigma(t, \tilde{x}^\varepsilon(t), u) - \sigma(t, \tilde{x}^\varepsilon(t), \tilde{u}^\varepsilon(t))]^* \tilde{Q}^\varepsilon(t)[\sigma(t, \tilde{x}^\varepsilon(t), u) - \sigma(t, \tilde{x}^\varepsilon(t), \tilde{u}^\varepsilon(t))] \\ & + \tilde{\psi}^\varepsilon(t) \cdot [f(t, \tilde{x}^\varepsilon(t), u) - f(t, \tilde{x}^\varepsilon(t), \tilde{u}^\varepsilon(t))] \\ & + \tilde{K}^\varepsilon(t) \cdot [\sigma(t, \tilde{x}^\varepsilon(t), u) - \sigma(t, \tilde{x}^\varepsilon(t), \tilde{u}^\varepsilon(t))] \\ & + [L(t, \tilde{x}^\varepsilon(t), u) - L(t, \tilde{x}^\varepsilon(t), \tilde{u}^\varepsilon(t))] \Big\} dt + o(\delta). \end{aligned}$$

Dividing (4.8) by δ and sending δ to 0, we conclude that

$$\begin{aligned} E\Big\{ & \tfrac{1}{2}[\sigma(t, \tilde{x}^\varepsilon(t), u) - \sigma(t, \tilde{x}^\varepsilon(t), \tilde{u}^\varepsilon(t))]^* \tilde{Q}^\varepsilon(t)[\sigma(t, \tilde{x}^\varepsilon(t), u) - \sigma(t, \tilde{x}^\varepsilon(t), \tilde{u}^\varepsilon(t))] \\ & + \tilde{\psi}^\varepsilon(t) \cdot [f(t, \tilde{x}^\varepsilon(t), u) - f(t, \tilde{x}^\varepsilon(t), \tilde{u}^\varepsilon(t))] \\ & + \tilde{K}^\varepsilon(t) \cdot [\sigma(t, \tilde{x}^\varepsilon(t), u) - \sigma(t, \tilde{x}^\varepsilon(t), \tilde{u}^\varepsilon(t))] \\ & + [L(t, \tilde{x}^\varepsilon(t), u) - L(t, \tilde{x}^\varepsilon(t), \tilde{u}^\varepsilon(t))] \Big\} \ge -\varepsilon^{\frac{1}{3}}. \end{aligned} \tag{4.9}$$

Now we are to derive an estimate for the term similar to the left side of (4.9) with all the $\tilde{x}^\varepsilon(t), \tilde{u}^\varepsilon(t)$, etc. replaced by $x^\varepsilon(t), u^\varepsilon(t)$, etc. To this end, we first estimate the following difference:

$$\begin{aligned} & E\int_s^T \Big\{ \tilde{K}^\varepsilon(t) \cdot [\sigma(t, \tilde{x}^\varepsilon(t), u) - \sigma(t, \tilde{x}^\varepsilon(t), \tilde{u}^\varepsilon(t))] \\ & \qquad - K^\varepsilon(t) \cdot [\sigma(t, x^\varepsilon(t), u) - \sigma(t, x^\varepsilon(t), u^\varepsilon(t))] \Big\} dt \\ & = E\int_s^T \{\tilde{K}^\varepsilon(t) - K^\varepsilon(t)\} \cdot \{\sigma(t, \tilde{x}^\varepsilon(t), u) - \sigma(t, \tilde{x}^\varepsilon(t), \tilde{u}^\varepsilon(t))\} dt \\ & \quad + E\int_s^T K^\varepsilon(t) \cdot \{\sigma(t, \tilde{x}^\varepsilon(t), u) - \sigma(t, x^\varepsilon(t), u)\} dt \\ & \quad - E\int_s^T K^\varepsilon(t) \cdot \{\sigma(t, \tilde{x}^\varepsilon(t), \tilde{u}^\varepsilon(t)) - \sigma(t, x^\varepsilon(t), u^\varepsilon(t))\} dt \\ & = I_1 + I_2 + I_3, \text{ say.} \end{aligned}$$

For any $\gamma \in [0, \frac{1}{3})$, let $\alpha = 3\gamma \in [0, 1)$. Fix a $p \in (1, 2)$ so that $(1 + \alpha)p < 2$. Take $q \in (2, +\infty)$ with $\frac{1}{p} + \frac{1}{q} = 1$. Then, appealing to Lemma 3.2,

$$\begin{aligned} I_1 \leq\ & \big\{E\textstyle\int_s^T |\tilde{K}^\varepsilon(t) - K^\varepsilon(t)|^p dt\big\}^{\frac{1}{p}} \big\{E\textstyle\int_s^T |\sigma(t, \tilde{x}^\varepsilon(t), u) - \sigma(t, \tilde{x}^\varepsilon(t), \tilde{u}^\varepsilon(t))|^q dt\big\}^{\frac{1}{q}} \\ & \leq \{C_2 d(u^\varepsilon(\cdot), \tilde{u}^\varepsilon(\cdot))^{\frac{\alpha p}{2}}\}^{\frac{1}{p}} \big\{C_2 E\textstyle\int_s^T (1 + |\tilde{x}^\varepsilon(t)|^q) dt\big\}^{\frac{1}{q}} \\ & \leq C_3 \varepsilon^{\frac{\alpha}{3}} \text{ (note (4.4))} \\ & = C_3 \varepsilon^\gamma. \end{aligned}$$

Next,

$$\begin{aligned} I_2 \leq\ & \big\{E\textstyle\int_s^T |K^\varepsilon(t)|^2 dt\big\}^{\frac{1}{2}} \big\{E\textstyle\int_s^T |\sigma(t, \tilde{x}^\varepsilon(t), u) - \sigma(t, x^\varepsilon(t), u))|^2 dt\big\}^{\frac{1}{2}} \\ & \leq C_4 \big\{E\textstyle\int_s^T |\tilde{x}^\varepsilon(t) - x^\varepsilon(t)|^2 dt\big\}^{\frac{1}{2}} \\ & \leq C_5 \{d(u^\varepsilon(\cdot), \tilde{u}^\varepsilon(\cdot))^\alpha\}^{\frac{1}{2}} \\ & \leq C_5 \{\varepsilon^{\frac{2\alpha}{3}}\}^{\frac{1}{2}} \\ & = C_5 \varepsilon^\gamma. \end{aligned}$$

Further,

$$\begin{aligned} I_3 =\ & - E\textstyle\int_s^T K^\varepsilon(t) \cdot \{\sigma(t, \tilde{x}^\varepsilon(t), \tilde{u}^\varepsilon(t)) - \sigma(t, \tilde{x}^\varepsilon(t), u^\varepsilon(t))\} dt \\ & - E\textstyle\int_s^T K^\varepsilon(t) \cdot \{\sigma(t, \tilde{x}^\varepsilon(t), u^\varepsilon(t)) - \sigma(t, x^\varepsilon(t), u^\varepsilon(t))\} dt \\ \leq\ & \big\{E\textstyle\int_s^T |K^\varepsilon(t)|^2 dt\big\}^{\frac{1}{2}} \big\{E\textstyle\int_s^T |\sigma(t, \tilde{x}^\varepsilon(t), \tilde{u}^\varepsilon(t)) \\ & \qquad - \sigma(t, \tilde{x}^\varepsilon(t), u^\varepsilon(t))|^2 \chi_{\tilde{u}^\varepsilon(t) \neq u^\varepsilon(t)}(t) dt\big\}^{\frac{1}{2}} + C_5 \varepsilon^\gamma. \end{aligned}$$

By the Cauchy–Schwarz inequality, one has

$$\begin{aligned} & E\textstyle\int_s^T |\sigma(t, \tilde{x}^\varepsilon(t), \tilde{u}^\varepsilon(t)) - \sigma(t, \tilde{x}^\varepsilon(t), u^\varepsilon(t))|^2 \chi_{\tilde{u}^\varepsilon(t) \neq u^\varepsilon(t)}(t) dt \\ \leq\ & \big\{E\textstyle\int_s^T |\sigma(t, \tilde{x}^\varepsilon(t), \tilde{u}^\varepsilon(t)) - \sigma(t, \tilde{x}^\varepsilon(t), u^\varepsilon(t))|^4 dt\big\}^{\frac{1}{2}} \big\{E\textstyle\int_s^T \chi_{\tilde{u}^\varepsilon(t) \neq u^\varepsilon(t)}(t) dt\big\}^{\frac{1}{2}} \\ \leq\ & C_6 d(\tilde{u}^\varepsilon(\cdot), u^\varepsilon(\cdot))^{\frac{1}{2}} \\ \leq\ & C_6 \varepsilon^{\frac{1}{3}} \\ \leq\ & C_6 \varepsilon^\gamma. \end{aligned}$$

Thus, we have proved that

$$\begin{aligned} & E\textstyle\int_s^T \Big\{\tilde{K}^\varepsilon(t) \cdot [\sigma(t, \tilde{x}^\varepsilon(t), u) - \sigma(t, \tilde{x}^\varepsilon(t), \tilde{u}^\varepsilon(t))] \\ & \qquad - K^\varepsilon(t) \cdot [\sigma(t, x^\varepsilon(t), u) - \sigma(t, x^\varepsilon(t), u^\varepsilon(t))]\Big\} dt \\ \leq\ & C_7 \varepsilon^\gamma. \end{aligned} \tag{4.10}$$

Similarly,

$$\begin{aligned} & E\textstyle\int_s^T \Big\{\big\{\frac{1}{2}[\sigma(t, \tilde{x}^\varepsilon(t), u) - \sigma(t, \tilde{x}^\varepsilon(t), \tilde{u}^\varepsilon(t))]^* \tilde{Q}^\varepsilon(t) [\sigma(t, \tilde{x}^\varepsilon(t), u) - \sigma(t, \tilde{x}^\varepsilon(t), \tilde{u}^\varepsilon(t))] \\ & - \frac{1}{2}[\sigma(t, x^\varepsilon(t), u) - \sigma(t, x^\varepsilon(t), u^\varepsilon(t))]^* Q^\varepsilon(t) [\sigma(t, x^\varepsilon(t), u) - \sigma(t, x^\varepsilon(t), u^\varepsilon(t))]\big\} \\ & + \big\{\tilde{\psi}^\varepsilon(t) \cdot [f(t, \tilde{x}^\varepsilon(t), u) - f(t, \tilde{x}^\varepsilon(t), \tilde{u}^\varepsilon(t))] \\ & \quad - \psi^\varepsilon(t) \cdot [f(t, x^\varepsilon(t), u) - f(t, x^\varepsilon(t), u^\varepsilon(t))]\big\} \\ & + \big\{[L(t, \tilde{x}^\varepsilon(t), u) - L(t, \tilde{x}^\varepsilon(t), \tilde{u}^\varepsilon(t))] \\ & \quad - [L(t, x^\varepsilon(t), u) - L(t, x^\varepsilon(t), u^\varepsilon(t))]\big\}\Big\} dt \\ \leq\ & C_8 \varepsilon^\gamma. \end{aligned} \tag{4.11}$$

The inequality (4.3) therefore follows from combining (4.9), (4.10) and (4.11). □

COROLLARY 4.2. *Under the conditions of Theorem* 4.1, *it holds that*

$$E\int_s^T \mathcal{H}^{(x^\varepsilon(\cdot),u^\varepsilon(\cdot))}(t,x^\varepsilon(t),u^\varepsilon(t))dt \geq \sup_{u(\cdot)\in U_{ad}[s,T]} E\int_s^T \mathcal{H}^{(x^\varepsilon(\cdot),u^\varepsilon(\cdot))}(t,x^\varepsilon(t),u(t))dt - C_1\varepsilon^\gamma. \tag{4.12}$$

Proof. In the definition of the perturbed control $u_\delta(\cdot)$ in (4.7), the point $u \in \Gamma$ can be replaced by any admissible control $u(\cdot) \in U_{ad}[s,T]$, and the subsequent argument still goes through. So (4.3) holds with u replaced by $u(t)$ for any $u(\cdot) \in U_{ad}[s,T]$, which is an easy variant of (4.12). □

Let us now look at an example.

Example 4.1. Let $n = l = 1, s = y = 0, T = 1, f = 0, \sigma = u, L = -u, h = \frac{1}{2}x^2, \Gamma = [0,1]$. For a given admissible pair $(x^\varepsilon(\cdot), u^\varepsilon(\cdot))$, the corresponding second-order adjoint equation is

$$\begin{cases} dQ^\varepsilon(t) = R^\varepsilon(t)dB(t), \\ Q^\varepsilon(1) = 1. \end{cases}$$

By the uniqueness of its solutions, $(Q^\varepsilon(t), R^\varepsilon(t)) = (1,0)$. Then for any admissible control $u(\cdot)$ we have

$$\begin{aligned} \mathcal{H}^{(x^\varepsilon(\cdot),u^\varepsilon(\cdot))}(t,x^\varepsilon(t),u(t)) &= u(t) - (K^\varepsilon(t) - Q^\varepsilon(t)u^\varepsilon(t))u(t) - \tfrac{1}{2}Q^\varepsilon(t)u^2(t) \\ &= -\tfrac{1}{2}(u(t) - u^\varepsilon(t) + K^\varepsilon(t) - 1)^2 + \tfrac{1}{2}(u^\varepsilon(t) - K^\varepsilon(t) + 1)^2 \end{aligned}$$

and

$$\mathcal{H}^{(x^\varepsilon(\cdot),u^\varepsilon(\cdot))}(t,x^\varepsilon(t),u^\varepsilon(t)) = \frac{1}{2}[u^\varepsilon(t)]^2 - (K^\varepsilon(t) - 1)u^\varepsilon(t).$$

Hence a simple calculation shows that if

$$u^\varepsilon(t) - K^\varepsilon(t) + 1 \in \Gamma, \tag{4.13}$$

then (4.12) gives

$$E\int_0^1 (K^\varepsilon(t) - 1)^2 dt \leq C_1\varepsilon^\gamma. \tag{4.14}$$

The above condition reveals the "minimum" qualification for the pair $(x^\varepsilon(\cdot), u^\varepsilon(\cdot))$ to be ε-optimal. For example, the controls $u^\varepsilon(t) \equiv 1 - \varepsilon^{\frac{1}{2}}$ are candidates for ε-optimality. To see this, note that the first-order adjoint equation is

$$\begin{cases} d\psi^\varepsilon(t) = K^\varepsilon(t)dB(t), \\ \psi^\varepsilon(1) = x^\varepsilon(1). \end{cases}$$

It is clear that if we choose $u^\varepsilon(t) \equiv 1 - \varepsilon^{\frac{1}{2}}$ with the corresponding $x^\varepsilon(t) = (1 - \varepsilon^{\frac{1}{2}})B(t)$, then the unique solution pair of the first-order adjoint equation will be $(\psi^\varepsilon(t), K^\varepsilon(t)) = ((1 - \varepsilon^{\frac{1}{2}})B(t), 1 - \varepsilon^{\frac{1}{2}})$. Hence (4.13) and (4.14) will be satisfied.

Remark 4.1. Corollary 4.1 says that *any* ε-optimal control nearly maximizes the $\mathcal{H}$-function (in the integral sense) with an error bound of order of "almost" $\varepsilon^{\frac{1}{3}}$. We believe, although we are not able to prove at this moment, that the error bound can be improved.

Remark 4.2. The necessary conditions of near-optimal controls are derived in terms of the near-maximum condition of the $\mathcal{H}$-function in an integral form. It is well known that, for exact optimality, the integral form and the pointwise form of the maximum condition are equivalent (cf. [7]), but it is certainly not the case for near-optimality. On the other hand, when $\varepsilon = 0$, Theorem 4.1 reduces to the stochastic maximum principle [9, 12].

5. Sufficient conditions of near-optimality. In this section, we will show that, under certain concavity conditions, the near-maximum condition of the $\mathcal{H}$-function in the integral form is sufficient for near-optimality, with the following additional assumption:

(A3) ρ is differentiable in u for $\rho = f, \sigma, L$, and there is a constant $C > 0$ such that

$$|\rho(t,x,u) - \rho(t,x,u')| + |\rho_u(t,x,u) - \rho_u(t,x,u')| \le C|u - u'|. \tag{5.1}$$

THEOREM 5.1. *Let $(x^\varepsilon(\cdot), u^\varepsilon(\cdot))$ be an admissible pair, and $(\psi^\varepsilon(\cdot), K^\varepsilon(\cdot))$ be the solution to* (3.5) *corresponding to $(x^\varepsilon(\cdot), u^\varepsilon(\cdot))$. Assume that $H(t,\cdot,\cdot,\psi^\varepsilon(t), K^\varepsilon(t))$ is concave for a.e. $t \in [s,T]$, $P-a.s.$, and $h(\cdot)$ is convex. If, for some $\varepsilon > 0$,*

$$E\int_s^T \mathcal{H}^{(x^\varepsilon(\cdot),u^\varepsilon(\cdot))}(t, x^\varepsilon(t), u^\varepsilon(t))dt \ge \sup_{u(\cdot)\in U_{ad}[s,T]} E\int_s^T \mathcal{H}^{(x^\varepsilon(\cdot),u^\varepsilon(\cdot))}(t, x^\varepsilon(t), u(t)) - \varepsilon, \tag{5.2}$$

then

$$J(s,y;u^\varepsilon(\cdot)) \le \inf_{u(\cdot)\in U_{ad}[s,T]} J(s,y;u(\cdot)) + C_1\varepsilon^{\frac{1}{2}}, \tag{5.3}$$

where $C_1 > 0$ is a constant independent of ε.

Proof. All the constants C_i appearing in this proof are independent of ε. The key step in the proof is to show that $H_u(t, x^\varepsilon(t), u^\varepsilon(t), \psi^\varepsilon(t), K^\varepsilon(t))$ is very small and to estimate it in terms of ε. To this end, let us fix an $\varepsilon > 0$. Define a new metric $\tilde{d}$ on $U_{ad}[s,T]$ as follows:

$$\tilde{d}(u(\cdot), u'(\cdot)) = E\int_s^T \nu^\varepsilon(t)|u(t) - u'(t)|dt, \tag{5.4}$$

where

$$\nu^\varepsilon(t) = 1 + |\psi^\varepsilon(t)| + |K^\varepsilon(t)| + |Q^\varepsilon(t)| + |Q^\varepsilon(t)||x^\varepsilon(t)| \ge 1. \tag{5.5}$$

Obviously $\tilde{d}$ is a metric, and it is a complete metric as a weighted L^1 norm.

Define a functional W on $U_{ad}[s,T]$ as follows:

$$W(u(\cdot)) = E\int_s^T \mathcal{H}^{(x^\varepsilon(\cdot),u^\varepsilon(\cdot))}(t, x^\varepsilon(t), u(t))dt.$$

A simple calculation shows that

$$|W(u(\cdot)) - W(u'(\cdot))| \le C_2 E\int_s^T \nu^\varepsilon(t)|u(t) - u'(t)|dt.$$

Therefore, W is continuous on $U_{ad}[s,T]$ with respect to $\tilde{d}$. By (5.2) and the Ekeland principle, there exists a $\tilde{u}^\varepsilon(\cdot)\in U_{ad}[s,T]$ such that

$$\tilde{d}(u^\varepsilon(\cdot),\tilde{u}^\varepsilon(\cdot))\le \varepsilon^{\frac{1}{2}} \tag{5.6}$$

and

$$E\int_s^T \tilde{\mathcal{H}}(t,x^\varepsilon(t),\tilde{u}^\varepsilon(t))dt=\max_{u(\cdot)\in U_{ad}[s,T]} E\int_s^T \tilde{\mathcal{H}}(t,x^\varepsilon(t),u(t))dt, \tag{5.7}$$

where

$$\tilde{\mathcal{H}}(t,x,u)=\mathcal{H}^{(x^\varepsilon(\cdot),u^\varepsilon(\cdot))}(t,x,u)-\varepsilon^{\frac{1}{2}}\nu^\varepsilon(t)|u-\tilde{u}^\varepsilon(t)|. \tag{5.8}$$

The integral-form maximum condition (5.7) implies a pointwise maximum condition, namely, for a.e. $t\in[s,T]$ and $P-$ a.s.,

$$\tilde{\mathcal{H}}(t,x^\varepsilon(t),\tilde{u}^\varepsilon(t))=\max_{u\in\Gamma}\tilde{\mathcal{H}}(t,x^\varepsilon(t),u). \tag{5.9}$$

This, in turn, yields [2, Proposition 2.3.2]

$$0\in\partial_u\tilde{\mathcal{H}}(t,x^\varepsilon(t),\tilde{u}^\varepsilon(t)). \tag{5.10}$$

By (5.8) and the fact that the generalized gradient of the sum of two functions is contained in the sum of the generalized gradients of the two functions [2, Proposition 2.3.3], we deduce

$$\begin{aligned}
&\partial_u\tilde{\mathcal{H}}(t,x^\varepsilon(t),\tilde{u}^\varepsilon(t))\\
\subset\ &\partial_u\tilde{\mathcal{H}}^{(x^\varepsilon(\cdot),u^\varepsilon(\cdot))}(t,x^\varepsilon(t),\tilde{u}^\varepsilon(t))+[-\varepsilon^{\frac{1}{2}}\nu^\varepsilon(t),\varepsilon^{\frac{1}{2}}\nu^\varepsilon(t)]\\
=\ &\partial_u H(t,x^\varepsilon(t),\tilde{u}^\varepsilon(t),\psi^\varepsilon(t),K^\varepsilon(t))+[-\varepsilon^{\frac{1}{2}}\nu^\varepsilon(t),\varepsilon^{\frac{1}{2}}\nu^\varepsilon(t)]\\
&+\{\sigma_u(t,x^\varepsilon(t),\tilde{u}^\varepsilon(t))^*Q^\varepsilon(t)(\sigma(t,x^\varepsilon(t),u^\varepsilon(t))-\sigma(t,x^\varepsilon(t),\tilde{u}^\varepsilon(t)))\}.
\end{aligned}$$

Since H is differentiable in u by assumption (A3), the inclusion (5.10) implies that there is $\beta^\varepsilon(t)\in[-\varepsilon^{\frac{1}{2}}\nu^\varepsilon(t),\varepsilon^{\frac{1}{2}}\nu^\varepsilon(t)]$ such that

$$\begin{aligned}
&H_u(t,x^\varepsilon(t),\tilde{u}^\varepsilon(t),\psi^\varepsilon(t),K^\varepsilon(t))\\
=\ &-\beta^\varepsilon(t)-\sigma_u(t,x^\varepsilon(t),\tilde{u}^\varepsilon(t))^*Q^\varepsilon(t)(\sigma(t,x^\varepsilon(t),u^\varepsilon(t))-\sigma(t,x^\varepsilon(t),\tilde{u}^\varepsilon(t))).
\end{aligned} \tag{5.11}$$

Consequently, by noting (A3),

$$\begin{aligned}
&|H_u(t,x^\varepsilon(t),u^\varepsilon(t),\psi^\varepsilon(t),K^\varepsilon(t)|\\
\le\ &|H_u(t,x^\varepsilon(t),u^\varepsilon(t),\psi^\varepsilon(t),K^\varepsilon(t))-H_u(t,x^\varepsilon(t),\tilde{u}^\varepsilon(t),\psi^\varepsilon(t),K^\varepsilon(t))|\\
&+|H_u(t,x^\varepsilon(t),\tilde{u}^\varepsilon(t),\psi^\varepsilon(t),K^\varepsilon(t))|\\
\le\ &C_3\nu^\varepsilon(t)|u^\varepsilon(t)-\tilde{u}^\varepsilon(t)|+|\beta^\varepsilon(t)|\\
&+|\sigma_u(t,x^\varepsilon(t),\tilde{u}^\varepsilon(t))^*Q^\varepsilon(t)(\sigma(t,x^\varepsilon(t),u^\varepsilon(t))-\sigma(t,x^\varepsilon(t),\tilde{u}^\varepsilon(t)))|\\
\le\ &C_4\nu^\varepsilon(t)|u^\varepsilon(t)-\tilde{u}^\varepsilon(t)|+\varepsilon^{\frac{1}{2}}\nu^\varepsilon(t).
\end{aligned} \tag{5.12}$$

By the concavity of $H(t,\cdot,\cdot,\psi^\varepsilon(t),K^\varepsilon(t))$, we have

$$\begin{aligned}
&H(t,x(t),u(t),\psi^\varepsilon(t),K^\varepsilon(t))-H(t,x^\varepsilon(t),u^\varepsilon(t),\psi^\varepsilon(t),K^\varepsilon(t))\\
\le\ &H_x(t,x^\varepsilon(t),u^\varepsilon(t),\psi^\varepsilon(t),K^\varepsilon(t))(x(t)-x^\varepsilon(t))\\
&+H_u(t,x^\varepsilon(t),u^\varepsilon(t),\psi^\varepsilon(t),K^\varepsilon(t))(u(t)-u^\varepsilon(t))
\end{aligned} \tag{5.13}$$

for any admissible pair $(x(\cdot), u(\cdot))$. Upon taking integrations on both sides and noting (5.1) and (5.12), we obtain

(5.14)
$$\begin{aligned} & E\int_s^T \{H(t,x(t),u(t),\psi^\varepsilon(t),K^\varepsilon(t)) - H(t,x^\varepsilon(t),u^\varepsilon(t),\psi^\varepsilon(t),K^\varepsilon(t))\}dt \\ \leq\ & E\int_s^T H_x(t,x^\varepsilon(t),u^\varepsilon(t),\psi^\varepsilon(t),K^\varepsilon(t))(x(t)-x^\varepsilon(t))dt + C_5\{\varepsilon^{\frac{1}{2}} + \tilde{d}(u^\varepsilon(\cdot),\tilde{u}^\varepsilon(\cdot))\} \\ \leq\ & E\int_s^T H_x(t,x^\varepsilon(t),u^\varepsilon(t),\psi^\varepsilon(t),K^\varepsilon(t))(x(t)-x^\varepsilon(t))dt + C_6\varepsilon^{\frac{1}{2}}. \end{aligned}$$

On the other hand, the state equation (2.2) can be rewritten as

$$\begin{aligned} & d(x(t)-x^\varepsilon(t)) \\ =\ & f_x^\varepsilon(t)(x(t)-x^\varepsilon(t))dt + \sigma_x^\varepsilon(t)(x(t)-x^\varepsilon(t))dB(t) \\ & + [-f_x^\varepsilon(t)(x(t)-x^\varepsilon(t)) + f(t,x(t),u(t)) - f(t,x^\varepsilon(t),u^\varepsilon(t))]dt \\ & + [-\sigma_x^\varepsilon(t)(x(t)-x^\varepsilon(t)) + \sigma(t,x(t),u(t)) - \sigma(t,x^\varepsilon(t),u^\varepsilon(t))]dB(t), \end{aligned}$$

where $f_x^\varepsilon(t) = f_x(t,x^\varepsilon(t),u^\varepsilon(t))$, etc. By applying the Ito formula to $\psi^\varepsilon(t)\cdot(x(t)-x^\varepsilon(t))$ we have

$$\begin{aligned} & E\{\int_s^T L_x^\varepsilon(t)\cdot(x(t)-x^\varepsilon(t))dt + h_x(x^\varepsilon(T))\cdot(x(1)-x^\varepsilon(1))\} \\ =\ & E\int_s^T \Big\{\psi^\varepsilon(t)\cdot[-f_x^\varepsilon(t)(x(t)-x^\varepsilon(t)) + f(t,x(t),u(t)) - f(t,x^\varepsilon(t),u^\varepsilon(t))] \\ & \quad + K^\varepsilon(t)\cdot[-\sigma_x^\varepsilon(t)(x(t)-x^\varepsilon(t)) + \sigma(t,x(t),u(t)) - \sigma(t,x^\varepsilon(t),u^\varepsilon(t))]\Big\}dt. \end{aligned}$$

Hence,

$$\begin{aligned} & E\int_s^T \{H_x(t,x^\varepsilon(t),u^\varepsilon(t),\psi^\varepsilon(t),K^\varepsilon(t))(x(t)-x^\varepsilon(t)) \\ & \qquad + \psi^\varepsilon(t)\cdot[f(t,x(t),u(t)) - f(t,x^\varepsilon(t),u^\varepsilon(t))] \\ & \qquad + K^\varepsilon(t)\cdot[\sigma(t,x(t),u(t)) - \sigma(t,x^\varepsilon(t),u^\varepsilon(t))]\}dt \\ =\ & E\{h_x(x^\varepsilon(T))(x(T)-x^\varepsilon(T))\} \\ \leq\ & E\{h(x(T)) - h(x^\varepsilon(T))\}. \end{aligned} \tag{5.15}$$

Combining (5.14) and (5.15), we arrive at

$$J(u^\varepsilon(\cdot)) \leq J(u(\cdot)) + C_1\varepsilon^{\frac{1}{2}}.$$

Since $u(\cdot)$ is arbitrary, the desired result follows. □

COROLLARY 5.2. *Under the assumptions of Theorem* 5.1, *a sufficient condition for an admissible pair* $(x^\varepsilon(\cdot), u^\varepsilon(\cdot))$ *to be* ε*-optimal is*

$$\begin{aligned} & E\int_s^T \mathcal{H}^{(x^\varepsilon(\cdot),u^\varepsilon(\cdot))}(t,x^\varepsilon(t),u^\varepsilon(t))dt \\ & \geq \sup_{u(\cdot)\in U_{ad}[s,T]} E\int_s^T \mathcal{H}^{(x^\varepsilon(\cdot),u^\varepsilon(\cdot))}(t,x^\varepsilon(t),u(t)) - \left(\frac{\varepsilon}{C_1}\right)^2. \end{aligned} \tag{5.16}$$

Example 5.1. Consider the problem presented in Example 4.1. Since the Hamiltonian

$$H(t,x,u,p,q) = u - qu$$

is concave in (x, u), we can apply Corollary 5.1 to conclude that $u^\varepsilon(t) \equiv 1 - (\frac{\varepsilon}{C_1})^2$ (which in view of Theorem 4.1 is a candidate for ε-optimality for sufficiently small ε) is indeed an ε-optimal control.

Remark 5.1. Theorem 5.1 asserts that the ε-maximum condition of the $\mathcal{H}$-function implies near-optimality with order $\varepsilon^{\frac{1}{2}}$. In Example 4.1, the control $u^\varepsilon(t) = 1 - \varepsilon^{\frac{1}{2}}$ (with the corresponding $K^\varepsilon(t) = 1 - \varepsilon^{\frac{1}{2}}$) satisfies the ε-maximum condition of the $\mathcal{H}$-function; see the left-hand side of (4.14). However, the underlying cost function

$$J = -E\int_0^1 u(t)dt + \frac{1}{2}E\int_0^1 u(t)^2 dt = \frac{1}{2}E\int_0^1 [u(t)-1]^2 dt - \frac{1}{2}.$$

Therefore, the minimum value of the cost function is $-\frac{1}{2}$. Moreover, we see that $u^\varepsilon(t) = 1 - \varepsilon^{\frac{1}{2}}$ is actually a near-optimal control with order ε. This suggests that we may lose some sharpness in estimating the error bound when applying Theorem 5.1. just as with Theorem 4.1. The setback basically comes from the Ekeland principle (Lemma 2.1), where a smaller value of λ results in a better knowledge about the position of u^λ at the cost of a less sharp estimate on the value of $\rho(u^\lambda)$.

Remark 5.2. If (5.2) holds with $\varepsilon = 0$, then Theorem 5.1 becomes a sufficiency theorem for (exact) optimality. In this case, the sufficient condition (5.2) is equivalent to Zhou's sufficient condition [15]:

$$\mathcal{H}^{(x^\varepsilon(\cdot),u^\varepsilon(\cdot))}(t, x^\varepsilon(t), u^\varepsilon(t)) = \max_{u\in\Gamma} \mathcal{H}^{(x^\varepsilon(\cdot),u^\varepsilon(\cdot))}(t, x^\varepsilon(t), u),\ P-\text{ a.s., a.e. } t \in [s, T].$$

6. Some discussions. In sections 4 and 5, we established necessary and sufficient conditions, respectively, for stochastic near-optimal controls in terms of a small parameter ε. Here ε may appear in two different situations. First, it may reflect the loss in the objective value allowed by the decision maker, who may have set this "tolerance level" before he started to seek a near-optimal policy. Second, ε may be a parameter representing the complexity of the original decision problem that can be approximated by simpler models as ε is sufficiently small. For the second situation, two good examples are hierarchical production models (see, e.g., [10, 11, 16]) and discrete approximation (see, e.g., [5, 8]).

Many practical systems are so complicated that it is simply impossible to obtain their optimal controls. A commonly used approach is to approximate the original optimal control problems by simpler ones and then construct controls based on the optimal controls of the approximating problems. Theorems 4.1 and 5.1 provide a possible way to show analytically that the constructed controls are near-optimal for the original complicated stochastic control problems. The idea behind it, in some abstract form, is as follows. Suppose we have a family of stochastic control problems $\mathcal{P}^\varepsilon$ parameterized by $\varepsilon > 0$ and a stochastic control problem $\mathcal{P}$ which is much easier to solve than each $\mathcal{P}^\varepsilon$. Let $U_{ad,\varepsilon}[s, T]$ and $U_{ad}[s, T]$ be the sets of admissible controls for $\mathcal{P}^\varepsilon$ and $\mathcal{P}$, respectively. Now we solve $\mathcal{P}$ and obtain an optimal or near-optimal (in terms of ε) control $u^*(\cdot)$. Then the stochastic maximum principle [9, 12] or Corollary 4.1 of this paper yields

$$\mathcal{H}^{(x^*(\cdot),u^*(\cdot))}(t, x^*(t), u^*(t)) = \max_{u\in\Gamma} \mathcal{H}^{(x^*(\cdot),u^*(\cdot))}(t, x^*(t), u),\quad P-\text{ a.s., a.e. } t. \tag{6.1}$$

or

$$
\begin{aligned}
&E\int_s^T \mathcal{H}^{(x^*(\cdot),u^*(\cdot))}(t,x^*(t),u^*(t))dt \\
&\qquad \geq \sup_{u(\cdot)\in U_{ad}[s,T]} E\int_s^T \mathcal{H}^{(x^*(\cdot),u^*(\cdot))}(t,x^*(t),u(t))dt - C_1\varepsilon^{\gamma},
\end{aligned} \tag{6.2}
$$

depending on whether $u^*(\cdot)$ is optimal or near-optimal. Here $\mathcal{H}^{(x^*(\cdot),u^*(\cdot))}$ is the $\mathcal{H}$-function associated with problem $\mathcal{P}$ and $(x^*(\cdot),u^*(\cdot))$. We then construct $u^\varepsilon(\cdot)$ for $\mathcal{P}^\varepsilon$ from $u^*(\cdot)$ (the way of constructions depends on each particular situation). Let $\mathcal{H}_\varepsilon^{(x^\varepsilon(\cdot),u^\varepsilon(\cdot))}$ be the $\mathcal{H}$-function associated with problem $\mathcal{P}^\varepsilon$ and $(x^\varepsilon(\cdot),u^\varepsilon(\cdot))$. If we can prove that

$$
\begin{aligned}
& E\textstyle\int_s^T \mathcal{H}_\varepsilon^{(x^\varepsilon(\cdot),u^\varepsilon(\cdot))}(t,x^\varepsilon(t),u^\varepsilon(t))dt \\
\geq\ & E\textstyle\int_s^T \mathcal{H}^{(x^*(\cdot),u^*(\cdot))}(t,x^*(t),u^*(t))dt - C'\varepsilon^{\alpha} \\
\geq\ & \sup_{u(\cdot)\in U_{ad}[s,T]} E\textstyle\int_s^T \mathcal{H}^{(x^*(\cdot),u^*(\cdot))}(t,x^*(t),u(t))dt - C'\varepsilon^{\beta} \\
\geq\ & \sup_{u(\cdot)\in U_{ad,\varepsilon}[s,T]} E\textstyle\int_s^T \mathcal{H}_\varepsilon^{(x^\varepsilon(\cdot),u^\varepsilon(\cdot))}(t,x^\varepsilon(t),u(t))dt - C''\varepsilon^{\delta},
\end{aligned} \tag{6.3}
$$

where the second inequality above is due to (6.1) or (6.2) and the first and third inequalities may be obtained by some estimates on the difference between the trajectories $x^\varepsilon(\cdot)$ and $x^*(\cdot)$, then $u^\varepsilon(\cdot)$ is near-optimal for $\mathcal{P}^\varepsilon$ with an error bound of order $\varepsilon^{\delta/2}$ if all the other assumptions of Theorem 5.1 are satisfied.

The above general idea has been applied to the hierarchical controls of stochastic manufacturing systems [16] (although the controlled process there is piecewise deterministic rather than of diffusion type). To be more specific, in [16], $\mathcal{P}^\varepsilon$ is an optimal production planning problem with stochastic machine capacity and ε representing the reciprocal of the fluctuation rate of the machine capacity process, and $\mathcal{P}$ is a deterministic problem where the random machine capacity has been averaged out. Certainly, $\mathcal{P}$ is much easier to solve analytically than $\mathcal{P}^\varepsilon$. A near-optimal control for $\mathcal{P}^\varepsilon$ is constructed based on an optimal control for $\mathcal{P}$ and its near-optimality is proved via a sequence of inequalities similar to (6.3) (see [16, section 5]).

Here we present another example which realizes the above idea.

Example 6.1. Consider the following problem:

$$
\begin{aligned}
&\text{minimize} \quad J^\varepsilon(u(\cdot)) = E\{\textstyle\int_0^1 \varepsilon g(u(t))dt + \frac{1}{2}x(1)^2\} \\
&\text{subject to} \quad \begin{cases} dx(t) = u(t)dt + u(t)dB(t), \\ x(0) = x_0, \end{cases}
\end{aligned} \tag{6.4}
$$

where $\varepsilon > 0$ is a small parameter and g (independent of ε) is a nonlinear, convex function satisfying assumption (A3) (see (5.1)). Let the control region $\Gamma = R^1$ and denote the problem by $\mathcal{P}^\varepsilon$. The running cost is nonlinear, which becomes small if ε becomes small. If one insists on solving the problem optimally, then he has to take the nonlinearity into full account, and it may turn out to be very hard to obtain an analytical solution. Let us do it differently. What we are going to show is that we can easily get a near-optimal (feedback) control analytically based on the optimal control of a simpler problem, called $\mathcal{P}$, which is obtained by setting $\varepsilon = 0$ in (6.4), namely,

by neglecting the nonlinearity. Indeed, $\mathcal{P}$ can be solved directly. To see this, applying Ito's formula yields

$$d(e^{t-1}x(t)^2) = e^{t-1}(u(t)+x(t))^2dt + e^{t-1}u(t)dB(t).$$

Hence

$$Ex(1)^2 = Ee^{-1}x_0^2 + E\int_0^1 e^{t-1}(u(t)+x(t))^2dt.$$

It follows that the optimal control for problem $\mathcal{P}$ is a feedback $u(t) = -x(t)$. Now we are to prove that the *same* feedback control is near-optimal for the original problem $\mathcal{P}^\varepsilon$ when ε becomes sufficiently small. First, notice that the system dynamics are the same with $\mathcal{P}^\varepsilon$ and $\mathcal{P}$ (the only difference lies in the cost functions), thus the state trajectories and all adjoint processes coincide for both problems under the same controls. Denote optimal state and control under the feedback $u(t) = -x(t)$ with initial x_0 by $(x^*(\cdot), u^*(\cdot))$ and the corresponding solutions to the first- and second-adjoint equations by $(\psi(\cdot), K(\cdot))$ and $(Q(\cdot), R(\cdot))$, respectively. It is easy to show that $Q(t) \equiv 1$. Then the $\mathcal{H}$-function for $\mathcal{P}$ is

$$\mathcal{H}^{(x^*(\cdot),u^*(\cdot))}(t,x,u) = -\frac{1}{2}u^2 - (\psi(t)+K(t)-u^*(t))u.$$

Since $u^*(\cdot)$ is optimal, by stochastic maximum principle, it is necessary that $u^*(t)$ maximizes the $\mathcal{H}$-function a.s., namely,

$$-u^*(t) - (\psi(t)+K(t)-u^*(t)) = 0, \quad P- \text{ a.s., a.e. } t.$$

So

$$\psi(t)+K(t) = 0, \quad P- \text{ a.s., a.e. } t.$$

However, the $\mathcal{H}$-function for $\mathcal{P}^\varepsilon$ is

$$\begin{aligned}\mathcal{H}_\varepsilon^{(x^*(\cdot),u^*(\cdot))}(t,x,u) &= -\tfrac{1}{2}u^2 - (\psi(t)+K(t)-u^*(t))u - \varepsilon g(u)\\ &= -\tfrac{1}{2}u^2 + u^*(t)u - \varepsilon g(u).\end{aligned}$$

The above function is maximized at u^ε, which satisfies

$$u^\varepsilon = u^*(t) - \varepsilon\dot{g}(u^\varepsilon). \tag{6.5}$$

Hence

$$\begin{aligned}&\mathcal{H}_\varepsilon^{(x^*(\cdot),u^*(\cdot))}(t,x,u^*(t)) - \max_{u\in\Gamma}\mathcal{H}_\varepsilon^{(x^*(\cdot),u^*(\cdot))}(t,x,u)\\ =\ &\mathcal{H}_\varepsilon^{(x^*(\cdot),u^*(\cdot))}(t,x,u^*(t)) - \mathcal{H}_\varepsilon^{(x^*(\cdot),u^*(\cdot))}(t,x,u^\varepsilon)\\ =\ &\tfrac{1}{2}(u^\varepsilon - u^*(t))^2 + \varepsilon[g(u^\varepsilon) - g(u^*(t))]\\ \le\ &\tfrac{1}{2}\varepsilon^2|\dot{g}(u^\varepsilon)|^2 + C\varepsilon^2|\dot{g}(u^\varepsilon)|\\ \le\ &C'\varepsilon^2.\end{aligned}$$

Moreover, the Hamiltonian for problem $\mathcal{P}^\varepsilon$ is

$$H(t,x,u,\psi(t),K(t)) = -\varepsilon g(u) - \psi(t)u - K(t)u = -\varepsilon g(u),$$

which is concave. It follows then by Theorem 5.1 that $u^*(\cdot)$ is near-optimal for $\mathcal{P}^\varepsilon$ with an error order of ε when ε is sufficiently small.

It is worth mentioning that the idea in the above example may also apply to some cases where there are nonlinearities in system dynamics. If those nonlinearities are "small", then one considers a (usually) simpler linear system (by ignoring the nonlinearities) and solves the original nonlinear problem near-optimally based on the solutions to the linear system. In a more general setting, many optimal control problems are difficult because some small components of the systems are difficult to handle. The whole problems might be insurmountable if we insist on taking those components into consideration. However, the problems may become easily solvable if we ignore those small but difficult parts and consider near-optimality instead. It also should be noted that by saying some components are "small" we meant that they are small compared with other components of the systems; so it is in a relative sense. In hierarchical production planning, for instance, ε being small means that the machine capacity process changes much faster than other processes involved, such as discounting process. See [10, 11, 16] for details.

7. Concluding remarks. In this paper, we established necessary and sufficient conditions for near-optimal stochastic controls in terms of a small parameter ε. The theory developed in this paper is parallel to the classical Pontryagin maximum principle in exact optimization. It would be interesting to investigate the dynamic programming approach applied to near-optimality. Another challenging problem is to improve the error bounds obtained in this paper. We hope to study these problems in forthcoming papers.

REFERENCES

[1] R. Akella and P. R. Kumar, *Optimal control of production rate in a failure prone manufacturing system*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 116–126.

[2] F. H. Clarke, *Optimization and Nonsmooth Analysis*, SIAM, Philadelphia, 1990.

[3] I. Ekeland, *On the variational principle*, J. Math. Anal. Appl., 47 (1974), pp. 324–353.

[4] R. J. Elliott and M. Kohlmann, *The variational principle and stochastic optimal control*, Stochastics, 3 (1980), pp. 229–241.

[5] R. Gabasov, F. M. Kirillova, and B. SH. Mordukhovich, *The ε-maximum principle for suboptimal controls*, Sov. Math. Dokl., 27 (1983), pp. 95–99.

[6] R. W. Hall, *Zero Inventories*, Dow Jones-Irwin Press, Homewood, IL, 1983.

[7] H. J. Kushner, *Necessary conditions for continuous parameter stochastic optimization problems*, SIAM J. Control, 10 (1972), pp. 550–565.

[8] B. SH. Mordukhovich, *Approximation Methods in Problems of Optimization and Control*, Nauka, Moscow, 1988.

[9] S. Peng, *A general stochastic maximum principle for optimal control problems*, SIAM J. Control Optim., 28 (1990), pp. 966–979.

[10] S. Sethi, Q. Zhang, and X. Y. Zhou, *Hierarchical controls in stochastic manufacturing systems with machines in tandem*, Stochastics Stochastics Rep., 41 (1992), pp. 89–118.

[11] S. Sethi and X. Y. Zhou, *Asymptotic optimal feedback controls in stochastic dynamic two-machine flowshops*, Lecture Notes in Control and Inform. Sci. 214, G. Yin and Q. Zhang, eds., Springer-Verlag, New York, 1996, pp. 147–180.

[12] X. Y. Zhou, *A unified treatment of maximum principle and dynamic programming in stochastic controls*, Stochastics Stochastics Rep., 36 (1991), pp. 137–161.

[13] X. Y. Zhou, *Deterministic near-optimal controls, part* I: *Necessary and sufficient conditions for near-optimality*, J. Optim. Theory Appl., 85 (1995), pp. 473–488.

[14] X. Y. Zhou, *Deterministic near-optimal controls, part* II: *Dynamic programming and viscosity solution approach*, Math. Oper. Res., 21 (1996), pp. 655–674.

[15] X. Y. Zhou, *Sufficient conditions of optimality for stochastic systems with controllable diffusions*, IEEE Trans. Automat. Control, AC-41 (1996), pp. 1176–1179.

[16] X. Y. Zhou and S. Sethi, *A sufficient condition for near optimal stochastic controls and its applications to manufacturing systems*, Appl. Math. Optim., 29 (1994), pp. 67–92.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 3, pp. 948–959, May 1998

008

THE SYMMETRIC RENDEZVOUS-EVASION GAME*

STEVE ALPERN† AND WEI SHI LIM‡

Abstract. E. J. Anderson and R. R. Weber, *J. Appl. Probab.*, 28 (1990), pp. 839–851, considered the problem of two rendezvousers, R_1, R_2, randomly placed among n indistinguishable locations, who seek to meet in least expected time, using the same mixed strategy. We retain their dynamics but modify the rendezvousers' aim to meeting each other before either encounters an enemy searcher S. We solve this zero-sum game in minimal space (3 locations) and time (2 steps after placement), and find that optimal play requires that the rendezvous team use a mixture over behavioral strategies. While such complicated strategies are known to be necessary in principal for team games (the theory of Isbell and Alpern), we believe this is the first naturally occuring game where such a solution is derived. (An earlier paper by Lim solved a similar game in which R_1 and R_2 were allowed to use different strategies and joint randomization.)

Key words. rendezvous search, zero-sum game

AMS subject classifications. 90D05, 90D10

PII. S0363012996309770

1. Introduction.

The problem dealt with in this paper can best be introduced in terms of the "telephone problem" posed by the first author about 20 years ago: In two similar rooms in New York and San Francisco, there are n telephones strewn about, which are randomly connected in pairs. The original problem asked how two friends, placed in rooms in New York and San Francisco, should choose which phones to pick up at a sequence of discrete time points. Their aim is to minimize the expected time when they first pick up paired phones, subject to the restriction that they must use the same mixed strategy (otherwise one would stay at a fixed phone, the other would pick a permutation of the phones). This problem was reformulated by Anderson and Weber [3] in terms of two players moving among n distinct locations, who meet when they first occupy the same location. They found a family of strategies that give an expected meeting time of $0.8288n$ for large n. This is markedly better than the random strategy, which has expected meeting time n.

The rendezvous-evasion game, posed a few years ago by the first author, can be put into the "telephone" setting by adding an enemy "phone-tapper" S (for Searcher) who is placed in a Chicago room through which all n cables pass. At the times when the rendezvousers R_1, R_2 pick up phones and say something, S picks a cable to "tap." The rendezvous team wins the game if they pick corresponding phones before either talks on a line overheard by S. In terms of the distinct location version of Anderson and Weber, the rendezvous team wins if they meet before either meets S (and before the time runs out, if it is limited to m steps). Thus this game is something of a hybrid between a search game in the sense of Gal [4] and a rendezvous problem in the sense of Alpern [1]. (It may be viewed more generally as a geometric game, in the sense of Ruckle [10].) The *asymmetric* version of the rendezvous-evasion game, in which R_1 and R_2 are allowed to use *different* strategies, has been studied and solved

*Received by the editors September 25, 1996; accepted for publication (in revised form) March 11, 1997.

http://www.siam.org/journals/sicon/36-3/30977.html

†Mathematics Department, London School of Economics, Houghton Street, London WC2A 2AE, UK (alpern@vax.lse.ac.uk).

‡Faculty of Business Administration, National University of Singapore, 10 Kent Ridge Road, Singapore.

for some cases by Lim [6]. (Unlike the pure rendezvous problem, where asymmetric players could simply stay put and search exhaustively, respectively, Lim found that complex joint randomization was required for optimal play by the rendezvous team.) This problem has some similarities to that studied by Nakai [8], where two searchers compete to find a stationary target first. In that problem, the target has a given distribution and there is the added complexity of an overlook probability. Another rendezvous problem involving more than two agents is studied in [7], but in that version all agents have the common objective of minimizing the time required for a group meeting.

The present paper deals with the *symmetric* version of the rendezvous-evasion game on discrete locations. We consider the minimal problem in terms of space and time, by assuming three locations, and a time limit of two steps after placement of the players at distinct locations. (If no meeting has occured after two steps, we say that S has won.) We consider that the three locations are placed in the plane, so that all three agents have a common (clockwise) ordering of them. This common ordering allows each individual strategy to simply be expressed as a sequence of positions relative to the initial position.

The symmetry requirement for the rendezvous team can be viewed in terms of a "team coordinator," who broadcasts a strategy which is received by R_1, R_2 (but not S) in each play of the game. The broadcast strategy must be behavioral, i.e., require R_1, R_2 to (independently) randomize before each move; if they followed the same *pure* strategy they would always stay in the same relative position and hence never meet. Since the rendezvous team coordinator can also randomize before deciding which behavioral strategy to broadcast, the rendezvous team's most general strategy is a mixture over behavioral strategies. In fact we shall see that such a complex strategy is indeed required for optimal play.

Since the time-limited rendezvous-evasion game is apparently a finite, two-person, zero-sum game, the necessity of mixtures over behavioral strategies requires some explanation. Actually there are two explanations, depending on whether it is viewed as an infinite game or a finite team game (game with repeated decisions, as in [2]). If one takes the infinite game viewpoint, one says that the rendezvous coordinator's pure strategies (those to be actually used in an individual play of the game) are in fact the behavioral strategies that he broadcasts. So the general theory of infinite (compact) games says that some mixture over these strategies is optimal. The more sophisticated viewpoint is to say that the game is finite, but it is a team game which is equivalent to one with "repeated decisions." This means that it is equivalent to one with an extensive form where some paths pass through an information set more than once (in fact, once for R_1 and once for R_2). The theory of such games [5], [2] says that finite mixtures over behavioral strategies may be required, and [2] gives a bound on the maximum number of behavioral strategies that must be used. In the game considered here we find that the rendezvous team must mix over three behavioral strategies. While the possible necessity of such mixed-behavioral strategies has been noted in the theory of team games (and illustrated by examples), we believe this is their first occurrence in a naturally arising game. We should observe that, for games of the type considered here, no general solution algorithms have been developed. So a pair of proposed solutions must be individually shown to ensure the same value. In this sense the game studied here is like an infinite, compact game.

The paper is organized as follows. In section 2 we formally define and solve the two-step, symmetric rendezvous-evasion game on three locations. In section 3 we consider certain qualitative aspects of the solution obtained in section 2. We explain

the need for the rendezvous team to use a mixture over behavioral strategies in terms of the information structure of the game, and we show that no single behavioral strategy is adequate. We also show that the optimal rendezvous strategy given in section 2 is stable with respect to unilateral deviations by an individual rendezvous agent, and thus not subject to the Piccione–Rubinstein's "absent-minded driver paradox." Finally, we show that, for asymmetric or symmetric rendezvous-evasion games on n locations with $m \leq \infty$ steps, the optimal solution is the same as for a pure rendezvous game (i.e., with no enemy searcher) where the agents receive a fixed prize when they meet, which they time discount with a common factor. This formulation shows that the rendezvous team does better against the searcher when they are allowed to jointly randomize over pure strategy pairs, winning with probability 36/81 instead of 24/81 for the symmetric problem. In section 4 we discuss what happens in the rendezvous-evasion game if we drop the assumption of a common knowledge ordering of the three locations. We show that the resulting game is better for the searcher in that the value (optimal probability that the searcher wins) increases.

We conclude this introduction with a small technical note for those readers familiar with the Anderson–Weber formulation [3]. They begin their problem with a random initial placement of the two agents, including the possibility that they are placed on the same location. We consider a different but equivalent start in which the three agents are placed according to a random permutation. This is what the Anderson–Weber start gives in the nontrivial case that no two are at the same location. Thus step k in our model corresponds to step $k+1$ in theirs.

2. The rendezvous-evasion game. In this section we define and solve a rendezvous-evasion game Γ played between a searcher S and a rendezvous team R consisting of two agents called R_1 and R_2. The three are initially placed randomly at three distinct locations, which we view as the vertices of a triangle. We assume they all have a common notion of a clockwise direction around these vertices but otherwise have no common labeling. (In the formal notation of rendezvous search described in [1], this corresponds to the case where the given group G of symmetries is the group of the three rotations of the triangle, the group generated by any cycle of the three locations.) At each of two time steps, they may each move to any vertex (including the one they currently occupy). This is a zero-sum game which is won by the rendezvous team R (with payoff $\pi = 0$) if R_1 and R_2 meet each other before either meets the enemy searcher S and before the game ends at the end of step two; otherwise the searcher S wins (with $\pi = 1$). Thus the value of the game, if it exists, is the probability that S wins, given optimal play on both sides. We assume that the agents R_1 and R_2 have not met prior to play to agree which role each should play (e.g., one stationary and one searching), and so must play the same mixed strategy.

To formally describe strategies, we assume that each agent labels his initial position as 0 and labels the remaining two in a clockwise direction as 1 and 2. The possibility of such a labeling is justified by the symmetry assumptions described above, formalized in the definition of G. Thus there are nine pure strategies,

$$\{(0,0),(1,2),(2,1),(0,1),(1,0),(2,2),(0,2),(1,1),(2,0)\},$$

which we number in this order as $w_i, i = 1, \ldots, 9$. Note that in this numbering,

$$\text{if } w_i = (x_1, x_2), i = 1, \ldots, 3, \quad \text{then } w_{i+3j} = (x_1, x_2 + j \bmod 3). \tag{2.1}$$

A probability distribution b over these nine strategies, with b_i interpreted as the probability of playing the pair w_i, will be called a behavioral strategy. (The way

we have described it would usually be called a mixed strategy, but it is certainly equivalent to a behavioral strategy giving a distribution for x_1 and a conditional distribution for x_2; we have reasons for preferring our chosen terminology.)

The times of meeting between pairs of agents depend on the random initial placement and the strategies used. Specifically, suppose that the initial location of $R_1$and R_2 in S-labels is denoted by (r_1, r_2), which is either $(1, 2)$ or $(2, 1)$. If the pure strategies played by R_1, R_2, and S are denoted x, y, z, then the rendezvous meeting time and the searcher meeting time are given, respectively, by

$$T_R = \min\{t : x_t - y_t = r_2 - r_1\}, \tag{2.2}$$
$$T_S = \min\{t : z_t = r_1 + x_t \text{ or } r_2 + y_t\}.$$

The payoff π to the maximizing searcher S is given by

$$\pi = \begin{cases} 0 & \text{if} \quad T_R \le \min(T_S - 1, 2), \\ 1 & \text{otherwise.} \end{cases}$$

We will be concerned with the four following behavioral strategies, the first three for R and the fourth random strategy for S.

$$\begin{aligned} \hat{b}^1 &= \tfrac{1}{3}(1,1,1,0,0,0,0,0,0), \\ \hat{b}^2 &= \tfrac{1}{3}(0,0,0,1,1,1,0,0,0), \\ \hat{b}^3 &= \tfrac{1}{3}(0,0,0,0,0,0,1,1,1), \\ \hat{s} &= \tfrac{1}{9}(1,1,1,1,1,1,1,1,1). \end{aligned}$$

We will show that $\hat{s}$ is optimal for the searcher S and that the mixture $\beta = \frac{1}{3}\hat{b}^1 + \frac{1}{3}\hat{b}^2 + \frac{1}{3}\hat{b}^3$ of three behavioral strategies is optimal for the rendezvous team R. (An optimal strategy for the controller of the rendezvous agents is to choose equiprobably among the $\hat{b}^k$ and to broadcast the chosen one for the two rendezvousers to use, with each of them randomizing independently.)

We begin our analysis of the strategy β by showing that it has two important properties; it has early rendezvous (T_R small), and unpredictable paths for its agents. These properties are established in the following two lemmas.

LEMMA 2.1. *If the two rendezvous agents both use the same behavioral strategy $\hat{b}^k$ for some $k = 1, 2, 3$, then*

$$\Pr(T_R = 1) = \Pr(T_R = 2) = \frac{1}{3}.$$

Proof. The simplest proof is a straightforward checking of the nine equiprobable cases that arise from the common use of $\hat{b}^k$. We consider $\hat{b}^1$ in detail and then show how the remaining two behavioral strategies give the same result. For $\hat{b}^1$ the nine equiprobable cases are listed below, and for each pair the probability of rendezvous in one or two steps is given.

	$(0,0)$	$(1,2)$	$(2,1)$
$(0,0)$	$0, 0$	$1/2, 1/2$	$1/2, 1/2$
$(1,2)$	$1/2, 1/2$	$0, 0$	$1/2, 1/2$
$(2,1)$	$1/2, 1/2$	$1/2, 1/2$	$0, 0$

To see how the entries were calculated, consider for example the situation of R_1 follows the path $(0, 0)$ and stays still, while R_2 follows path $(1, 2)$, a clockwise circuit.

If R_2 was initially placed one unit clockwise of R_2, then $T_R = 2$; otherwise $T_R = 1$. Since six of the nine pairs have entry $1/2, 1/2$, we have that

$$\Pr(T_R = 1) = \Pr(T_R = 2) = \frac{6}{9} * \frac{1}{2} = \frac{1}{3}.$$

The corresponding result for the behavioral strategies $\hat{b}^i, i = 2, 3$ follows from that of $\hat{b}^1$ and the equation noted above, $w_{i+3j} = (x_1, x_2 + j \bmod 3)$. The behavior on step 1 is identical, while the behavior on step 2 is rotated by the same amount for both players. (An equal rotation of two paths preserves their distance and hence their rendezvous time.) □

LEMMA 2.2. *If an agent is following the mixed behavioral strategy* $\beta = \frac{1}{3}\hat{b}^1 + \frac{1}{3}\hat{b}^2 + \frac{1}{3}\hat{b}^3$, *then*

1. *his position at step* 1 *is random, and*
2. *his position at step* 2 *is random and independent of his position at step* 1.

Proof. Property 1 is obviously true of each constituent behavioral strategy $\hat{b}^i$ and hence also true for any mixture. Property 2 is not true of the constituent strategies but is true of the mixture. Both properties can also be seen to hold from observing that β choose each of the nine pure strategy pairs equiprobably. □

THEOREM 2.3. *If the rendezvous team R follows the mixed behavioral strategy* $\beta = \frac{1}{3}\hat{b}^1 + \frac{1}{3}\hat{b}^2 + \frac{1}{3}\hat{b}^3$, *then it wins against any pure strategy x of the searcher with probability* $\frac{24}{81}$. *That is,*

$$\pi(\beta, x) = \frac{57}{81}.$$

Proof. The searcher S can win the game in three mutually exclusive ways: by meeting a rendezvouser on step 1, by meeting a rendezvouser on step 2, or if no meetings at all have occurred by the end of step 2. Hence we have

$$\Pr(S \text{ wins}) = \Pr(T_S = 1) + \Pr(T_S = 2 \wedge T_R > 1) + \Pr(T_S > 2 \wedge T_R > 2).$$

If we condition the above equation according to the three events $T_R = 1, 2, > 2$, which all have probability $1/3$ by Lemma 2.1, we obtain

$$\begin{aligned} \Pr(S \text{ wins}) &= \tfrac{1}{3}\left(\tfrac{1}{3} + 0 + 0\right) + \tfrac{1}{3}\left(\tfrac{2}{3} + \tfrac{1}{9} + 0\right) + \tfrac{1}{3}(1) \\ &= \tfrac{57}{81}. \end{aligned}$$

The above equation is explained easily as follows. If $T_R = 1$, then by Lemma 2.2, part 1, the search will be at their common meeting location on step 1 with probability $1/3$, regardless of the search strategy. If $T_R = 2$, then similarly at step 1 the searcher will be at one of the two locations occupied by the rendezvousers with probability $2/3$. Furthermore, if the searcher is not at either of these locations (which occurs with probability $1/3$), then at step 2 he will be at their common meeting place with probability $1/3$, by the second part of Lemma 2.2. (The independence asserted in that part of the lemma is needed to show that the searcher cannot use any knowledge gained in step 1 to improve his chances at step 2.) Finally, the third term is simply the statement that S is defined to be the winner if no meeting has occurred in the first two steps. □

We now turn to a consideration of the searcher's optimal strategy, which is simply the random strategy denoted by $\hat{s}$, which picks the location on each step randomly and independently of previous choices. According to the theory of games of this type, we

must evaluate $\hat{s}$ against any behavioral strategy of the opponent, since pure strategies in general will do worse.

LEMMA 2.4. *Suppose the searcher is employing the random strategy $\hat{s}$. Then the rendezvous team R will not do worse by replacing any behavioral strategy b by the strategy b', which is the average of b on each of the three groups $\{w_{3k+1}, w_{3k+2}, w_{3k+3}\}, k = 0, 1, 2$. That is, for any behavioral strategy b, we have*

$$\pi(b', \hat{s}) \leq \pi(b, \hat{s}), \text{ where } b'_{3k+m} = \sum_{j=1}^{3} b_{3k+j}/3,\ k = 0, 1, 2,\ m = 1, 2, 3.$$

Furthermore, equality holds only for the three strategies $\hat{b}^i, i = 1, 2, 3$.

Proof. For $i, j = 1, \ldots, 9$, let a_{ij} denote the probability that the random searcher strategy $\hat{s}$ wins when R_1 and R_2 follow the paths w_i and w_j, respectively. If R_1 and R_2 follow a pair of paths for which the probability that $T_R = t$ is denoted $p_t, t = 1, 2$, then they win on step 1 if S misses their meeting place (probability $2/3$) and they win on step 2 if S missed both their positions at step 1 (probability $1/3$) and also missed their common position on step 2 (probability $2/3$). Hence the team R wins with probability $1 - \pi = p_1(2/3) + p_2(2/9)$. It turns out that there are only four possible values for the pair (p_1, p_2), and the following table gives the value of π for these pairs.

p_1	p_2	π
0	0	9/9
0	1/2	8/9
1/2	0	6/9
1/2	1/2	5/9

These pairs occur in the following situations. When both agents of R use the same path, they cannot meet, and this gives the 9s on the diagonal in the matrix $9A$ below. If they do the same thing on step 1 and different things on step 2, then we are in the second row of the table and $\pi = 8/9$. Pairs that differ in step 1 then meet with probability $1/2$ and either never or always meet on step 2, giving the respective values of $6/9$ and $5/9$ for π (rows three and four of the table). The complete matrix $A = \{a_{ij}\}$ is given below, with the fractions cleared.

$$9A = \begin{array}{c} \\ (0,0) \\ (1,2) \\ (2,1) \\ (0,1) \\ (1,0) \\ (2,2) \\ (0,2) \\ (1,1) \\ (2,0) \end{array} \begin{array}{c} \begin{array}{ccccccccc} (0,0) & (1,2) & (2,1) & (0,1) & (1,0) & (2,2) & (0,2) & (1,1) & (2,0) \end{array} \\ \begin{pmatrix} 9 & 5 & 5 & 8 & 6 & 6 & 8 & 6 & 6 \\ 5 & 9 & 5 & 6 & 8 & 6 & 6 & 8 & 6 \\ 5 & 5 & 9 & 6 & 6 & 8 & 6 & 6 & 8 \\ 8 & 6 & 6 & 9 & 5 & 5 & 8 & 6 & 6 \\ 6 & 8 & 6 & 5 & 9 & 5 & 6 & 8 & 6 \\ 6 & 6 & 8 & 5 & 5 & 9 & 6 & 6 & 8 \\ 8 & 6 & 6 & 8 & 6 & 6 & 9 & 5 & 5 \\ 6 & 8 & 6 & 6 & 8 & 6 & 5 & 9 & 5 \\ 6 & 6 & 8 & 6 & 6 & 8 & 5 & 5 & 9 \end{pmatrix} \end{array}$$

Thus for any behavioral strategy (probability 9-vector) b, we have

$$\pi(b, \hat{s}) = \sum_{i,j} b_i b_j a_{ij} = b^\top A b.$$

Observe that we may write the averaged strategy b' in terms of b according to the equation

$$b' = Cb, \text{where } C = \begin{pmatrix} M & 0 & 0 \\ 0 & M & 0 \\ 0 & 0 & M \end{pmatrix}, \text{with } M = (1/3)\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix},$$

and the 0s in the matrix C denote the 3×3 matrix of zeros. Thus we have that

$$\pi(b, \hat{s}) - \pi(b', \hat{s}) = b^\top Ab - (Cb)^\top A(Cb) = b^\top (A - C^\top AC)b.$$

The matrix $A - C^\top AC$ has 9 real eigenvalues 2,2,2,2,0,0,0,8,8. Since all are nonnegative, it is positive semi-definite. It follows that the right-hand side of the above equation is nonnegative for any behavioral strategy b. Since there are only three zero eigenvalues, there can be at most three linearly independent solutions to the equation $\pi(b, \hat{s}) - \pi(b', \hat{s})$, and these must be the strategies $\hat{b}^i, i = 1, 2, 3$. □

THEOREM 2.5. *For any behavioral strategy b employed by the rendezvous team R, the random searcher strategy wins with probability at least 57/81. Specifically, $\pi(b, \hat{s}) \geq 57/81$, with equality only for the three behavioral strategies $\hat{b}^k, k = 1, 2, 3$.*

Proof. By the previous lemma, we may assume that the strategy b is already averaged over each of the three groups, that is, $b = (x, x, x, y, y, y, z, z, z)$. It is easily calculated that

$$\begin{aligned} \pi(b, \hat{s}) &= [x(19x + 20y + 20z) + y(20x + 19y + 20z) + z(20x + 20y + 19z)]/3 \\ &= (1/3)(x, y, z)\begin{pmatrix} 19 & 20 & 20 \\ 20 & 19 & 20 \\ 20 & 20 & 19 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix}, \end{aligned}$$

which has a minimum, over the set $0 \leq x, y, z \leq 1/3$, $x+y+z = 1/3$, of $19/27 = 57/81$. This minimum is obtained exactly when one of the three variables is equal to $1/3$, which corresponds to one of the three strategies $\hat{b}^i, i = 1, 2, 3$. □

Combining the last two theorems, we obtain the following corollary.

COROLLARY 2.6. *The value of the game Γ is 57/81. An optimal strategy for the rendezvous team is the equiprobable mixture of the three behavioral strategies $\hat{b}^k, k = 1, 2, 3$, which optimize against the random strategy for the searcher. An optimal strategy for the searcher is the random strategy.*

3. Properties of the solution. In this section we consider some qualitative and interpretative questions related to our proposed solution $(\beta, \hat{s})$ to the rendezvous-evasion game Γ on three locations with two time steps. We consider whether and why the rendezvous strategy β needs to be so complicated (mixed behavioral), and whether one of the rendezvous agents would wish to deviate from it if allowed to do so. We also show that the game Γ is in some sense equivalent to a pure rendezvous problem, with no enemy searcher, in which the agents get a fixed prize when they meet, which is discounted with respect to the meeting time. Thus this paper has something new to say about the pure rendezvous problem, as well as extending it to a game theoretic setting.

When the rendezvous players are viewed as agents of a team, this team player has imperfect recall because neither agent is aware of the other's previous choices. Actually, the situation is even worse, as an agent who is asked by a referee to make a choice does not know if he is the first or second agent to be asked such a question. That is, the paths of the extensive form game will enter an information set more than once.

Games of this type have been studied by Isbell [5] and Alpern [2], who have shown that mixtures over behavioral strategies may be required. The latter paper also gives a bound on the number of behavioral strategies needed, which for the three-location, two-step game Γ solved in section 2 turns out to be 27. (We found that in fact only three behavioral strategies were needed.) Recent interest in this problem has been generated by the provocative paper of Piccione and Rubinstein [9], who consider models of single-person decision making with this type of forgetting. Later in this section we will consider whether our problem is subject to an intertemporal paradox they find for certain games of this type. The first qualitative question we consider in this section is whether a single behavioral strategy will suffice for the rendezvous team in the sense of ensuring the value of the game. The following result answers this question in the negative.

THEOREM 3.1. *No behavioral strategy b for the rendezvous team ensures that the searcher cannot win with probability greater than the value, i.e., that $\pi(b,s) \leq 57/81$, for all s.*

Proof. Suppose that b is a behavioral strategy with $\pi(b,s) \leq 57/81$ for every behavioral strategy s of the searcher. Taking s equal to the random strategy $\hat{s}$, we see that the only possibilities for b are the three strategies $\hat{b}^k, k = 1,2,3$, by the last part of Theorem 2.5. However, the searcher has an effective answer to each of these strategies. Against $\hat{b}^1$, the pure strategy $w_2 = (1,2)$ is quite effective. The following matrix gives the probability that w_2 wins against each of the nine cases arising from the use of $\hat{b}^1$ by the rendezvousers.

	$(0,0)$	$(1,2)$	$(2,1)$
$(0,0)$	1	3/4	3/4
$(1,2)$	3/4	1	1/2
$(2,1)$	3/4	1/2	1

It follows that

$$\pi\left(\hat{b}^1, w_2\right) = \left(3 * 1 + 4 * \frac{3}{4} + 2 * \frac{1}{2}\right)/9 = 7/9 = 63/81 > 57/81.$$

A similar argument works for $\hat{b}^2$ and $\hat{b}^3$. □

We now consider whether our proposed solution $(\beta, \hat{s})$ is subject to a variation of the absent-minded driver paradox of [9]. More precisely, we consider whether a rendezvous agent who has heard the announcement of one of the component behavioral strategies $\hat{b}^i, 1,2,3$, and believes that his partner is following this behavioral strategy, can improve the team's payoff by deviating from the announced $\hat{b}^i$. A solution subject to this problem would still be considered a valid solution to the symmetric form of the rendezvous-evasion game as defined in section 2 but would restrict the interpretations of this game to scenarios in which the agents were constrained to follow orders. However, we find that our solution is in fact immune to this problem, as stated below.

THEOREM 3.2. *The solution $(\beta, \hat{s})$ is stable for the rendezvous agents in the following sense. Once a particular behavioral strategy $\hat{b}^i, i = 1,2,3$, is randomly selected for common use, neither of the rendezvous agents can benefit from a unilateral deviation.*

Proof. It can be seen from the matrix $9A$ given in the proof of Lemma 2.4 that if the searcher plays the random strategy $\hat{s}$ and the row rendezvous agent chooses the first three rows with probability 1/3 each ($\hat{b}^1$), then the values of $9 \cdot 3 \cdot \pi$ corresponding

to the nine columns are given by

$$(19, 19, 19, 20, 20, 20, 20, 20, 20)\,.$$

Consequently the column rendezvous agent minimizes π by choosing any behavioral strategy concentrated on the first three columns; in particular, by choosing $\hat{b}^1$. A similar argument applies to $\hat{b}^i, i = 2, 3$. □

In the rendezvous-evasion game Γ of section 2, the enemy searcher S is introduced to give an incentive to the rendezvous team to meet quickly. However, it may be thought by some that the searcher unnecessarily complicates the basic rendezvous problem. We now show that the game Γ is in fact strategically equivalent to a pure rendezvous problem (with no enemy searcher). This result can be stated in somewhat greater generality than our main result; we need no restriction on the number of locations or time steps, except that the former still needs to be finite.

THEOREM 3.3. *The symmetric rendezvous-evasion game $\Gamma(n, m)$ on n locations with rendezvous required by step $m \leq \infty$ is equivalent to the following pure rendezvous problem $R(n, m)$, in the sense that an optimal rendezvous strategy for one determines an optimal rendezvous strategy for the other.*

> *$R(n, m)$: Two indistinguishable agents rendezvous on n locations, for m steps. They each get a fixed prize when they meet, which they discount in time by the same factor.*

Proof. Suppose the rendezvous team R in the game $\Gamma(n, m)$ follows a strategy b (either behavioral or mixed behavioral) where the probability that they meet for the first time at time $i = 1, 2, \ldots$ is denoted by p_i,

$$\Pr(T_R = i) = p_i.$$

Suppose that the rendezvous team modifies this strategy by rotating each agent's position on each step i by a common amount e_i, where the vector e is chosen (by the controller) randomly according to the product measure on $\prod_{i=1,m} \{0, 1, 2\}$. It can be seen that the modified strategy b' has the same probabilities p_i, by referring to the formula (2.2) for the meeting time T_R, and observing that $x_t - y_t = (x_t + e_t) - (y_t + e_t)$. The modified strategy satisfies the same properties given in Lemma 2.2 for the mixed behavioral strategy β (which in this language is the modification of $\hat{b}^1$). That is, the position of each agent at each time is random (equiprobable) and independent of previous positions. Consequently it follows exactly as in the proof of Theorem 2.3 that every pure searcher strategy does equally well against this modified rendezvous strategy. Hence each pure searcher strategy is an optimal response, and therefore so is the random searcher strategy (an average of optimal strategies is optimal).

In fact there is a more intuitive way of seeing that the random searcher strategy is optimal. Note that if a random sequence of rotations e_t is applied to any pure searcher strategy z_t, the resulting strategy $z_t + e_t$ is what we have called the random searcher strategy. It follows that if the rendezvous team applies a common random sequence of rotations $-e_t$ to their strategies, it is strategically equivalent to the situation where they use their original strategies and the searcher is using his original strategy modified by $+e_t$, i.e., a random strategy. Hence it is as if the rendezvous team can force the searcher to use a random strategy, without any restrictions on their own strategy. Hence if the searcher had a better strategy than the random strategy, the rendezvous team could change that to a random strategy, so no other strategy is better.

Let q_i denote the probability that team R wins at the i'th step,

$$q_i = \Pr(T_R = i \wedge T_S > i)\,.$$

Then the probability that R wins, using b when S plays randomly, is

$$V = 1 - \pi(b, \hat{s}) = \sum_{i=1,m} q_i,$$

$$\begin{aligned}\Pr(T_R = i \wedge T_S > i-1) &= \Pr(T_R = i) \Pr(T_S > i-1 \backslash T_R = i) \\ &= p_i \left(\frac{n-2}{n}\right)^{i-1},\end{aligned}$$

$$\begin{aligned}q_i &= \Pr(T_R = i \wedge T_S > i-1) * \Pr(T_S > i \backslash T_R = i \wedge T_S > i-1) \\ &= p_i \left(\frac{n-2}{n}\right)^{i-1} * \left(\frac{n-1}{n}\right) = \left(\frac{n-1}{n}\right) p_i \left(\frac{n-2}{n}\right)^{i-1}.\end{aligned}$$

Therefore the payoff function $V(b)$ is the same as if the team R is given $(n-1)/n$ dollars when they meet, assuming their discount factor is $(n-2)/n$, and step 1 takes place at time zero and so is not discounted.

$$V(b) = \frac{n-1}{n} \sum_{i=1,m} p_i \left(\frac{n-2}{n}\right)^{i-1}. \tag{3.1}$$

This is exactly the payoff function for the pure rendezvous problem $R(n, m)$, where the agents get a fixed prize when they meet which is discounted by both at $\left(\frac{n-2}{n}\right)$ per step, and so the rendezvousers' aim for the problem or the game is to attain the $p_i's$ which maximize (3.1). □

The above reasoning applies equally well for an asymmetric rendezvous, in which case the unmodified strategy b can be a pair of pure strategies (w_k, w_j). Consider the same setting as that of section 2, namely $n = 3$ locations and $m = 2$ time steps. In this case it is clear that the maximum values of p_1 and p_2 are $1/2$ each, obtained by the strategy pair $((0,0),(1,2))$, where one rendezvous agent waits and the other searches. (An optimal strategy is then obtained by applying a common random rotation e_1, e_2, giving the strategy pair $(e_1, e_2), (1+e_1, 2+e_2)$, which is an average over 9 pairs.) Substituting into (3.1) gives the value $V = \frac{2}{3}(\frac{1}{2} * 1 + \frac{1}{2}\frac{1}{3}) = \frac{4}{9}$, which corresponds to an optimal payoff (probability of S winning) $\pi = 1 - V = 5/9 = 45/81$, which is much less than the $57/81$ obtainable optimally when the rendezvous team acts symmetrically. By the way, one optimal strategy for this game is an equiprobable joint randomization over the pairs giving a 5 in the matrix $9A$.

4. Rendezvous with full isometry group. In formalizing the rendezvous-evasion game Γ discussed in section 2, we took the assumption that the players had a common notion of a clockwise direction around the three locations. This would be reasonable if the three locations were on a surface where the players had a common notion of up. The associated group of isometries for the game Γ was the rotation group generated by the unit's clockwise rotation (with the identity and the anticlockwise rotation as the other two elements). We now consider how the problem may change if the full isometry group (six permutations of the three locations) is given. This would be the relevant group to describe the problem if the three locations were three arbitrary general position points in space. We denote the rendezvous-evasion game on three locations, with two time steps and the full permutation group as Γ', and its value by V'. We show that the searcher benefits from enlarging the set of symmetries, in that V' is greater than V. (In both cases the value is the optimal probability that the searcher wins.)

When the rendezvous evasion game on three locations is played with the full isometry group of the 3!=6 permutations of the three locations, the pure strategies of the players can be condensed into the following form. For each player (actually agent, in the case of the rendezvous team) let the initial position be labeled a, and the subsequent positions b and c, if reached. (This is the labeling procedure used by Anderson and Weber [3].) There are five possible strategies, $\{aa, bc, ab, bb, ba\}$, which we label as $z_i, i = 1, \ldots, 5$. For example, the strategy $z_3 = ab$ says to stay still on step 1 and then go equiprobably to one of the two remaining locations at step 2. Strategy $z_2 = bc$ says to move equiprobably to one of the other two locations at step 1 and then to the remaining location at step 2. Since each of the z_i involves at most one binary choice, there are at most eight equiprobable scenarios associated with any triple of strategies. Thus the probabilities $A_k(i, j)$ that the searcher S wins when R_1 follows z_i, R_2 follows z_j, and S follows z_k are integers divided by eight. The rationalized matrices $8A_k$ are given below.

$$8A_1 = \begin{pmatrix} 8 & 4 & 4 & 4 & 4 \\ 4 & 8 & 4 & 8 & 8 \\ 4 & 4 & 8 & 4 & 4 \\ 4 & 8 & 4 & 8 & 6 \\ 4 & 8 & 4 & 6 & 8 \end{pmatrix}, 8A_2 = \begin{pmatrix} 8 & 4 & 6 & 4 & 4 \\ 4 & 6 & 4 & 8 & 8 \\ 6 & 4 & 6 & 4 & 4 \\ 4 & 8 & 4 & 8 & 7 \\ 4 & 8 & 4 & 7 & 8 \end{pmatrix}, 8A_3 = \begin{pmatrix} 8 & 4 & 8 & 6 & 6 \\ 4 & 6 & 4 & 8 & 8 \\ 6 & 4 & 6 & 4 & 4 \\ 4 & 8 & 4 & 8 & 7 \\ 4 & 8 & 4 & 7 & 8 \end{pmatrix},$$

$$8A_4 = \begin{pmatrix} 8 & 4 & 8 & 6 & 6 \\ 4 & 6 & 6 & 5 & 6 \\ 8 & 6 & 8 & 5 & 6 \\ 6 & 5 & 5 & 6 & 5 \\ 6 & 6 & 6 & 5 & 6 \end{pmatrix}, 8A_5 = \begin{pmatrix} 8 & 4 & 8 & 6 & 6 \\ 4 & 6 & 6 & 6 & 5 \\ 8 & 6 & 8 & 6 & 5 \\ 6 & 6 & 6 & 6 & 5 \\ 6 & 5 & 5 & 5 & 6 \end{pmatrix}.$$

For any behavioral strategies u (for the rendezvous team) and v (for the searcher), i.e., distributions over the z_i, the payoff (probability that S wins) is given by

$$\pi(u, v) = uA_v u^\top, \text{ where } A_v = \sum_i v_i A_i.$$

The theory of "games with forgetting" [5], [2] says that since no path passes through an information set of S more than once, some single behavioral strategy is optimal for S. This means that

$$V' = \max_v \min_u \pi(u, v) \leq \min_u \max_v \pi(u, v),$$

where V' is the value of the game Γ' with the full isometry group. (In order to ensure equality, we must minimize over mixed behavioral strategies in the rightmost expression.)

THEOREM 4.1. *With optimal play, the searcher S does better in Γ' (with the full permutation group of isometries) than in the game Γ (with the rotation subgroup). In particular, we have*

$$\frac{57}{81} = V < \frac{58}{81} < V' \leq \frac{59}{81},$$

where the values are the optimal probabilities of S winning.

Proof. To obtain the upper bound for V', assume that the rendezvous team plays randomly. In this case (as in section 2) the best that S can do is also play randomly.

Thus the rendezvous team wins at step 1 if they meet (1/3) and they don't meet S (2/3, conditionally). The rendezvous team wins at step 2 if the game proceeds to step 2 ($\frac{2}{3} * \frac{1}{3}$) and they meet each other (1/3) but not S (2/3, conditionally). Thus the rendezvous team wins with probability

$$\frac{1}{3}\frac{2}{3} + \left(\frac{2}{3}\ \frac{1}{3}\right)\left(\frac{1}{3}\frac{2}{3}\right) = \frac{22}{81}$$

under random play, or S wins with probability $\frac{59}{81}$, as claimed.

Observe that in the current notation, random play by the searcher S is denoted by the probability distribution $(1/9, 2/9, 2/9, 2/9, 2/9)$. If we perturb this a bit in the z_2 direction, say (-e,4e,-e,-e,-e) with e=1/18, the resulting distribution is $\bar{v} = (1, 8, 3, 3, 3)/18$. The minimum over u of the payoff function $uA_{\bar{v}}u^{\top}$ is easily calculated to be at $u = (73, 184, 0, 4, 4)/265$, with a minimum of 191/265, which is more than 58/81. A good approximation to the value V' and an optimal strategy for the searcher could be obtained by analyzing the quadratic program $V' = \max_v \min_u \pi(u, v)$, but we know of no algorithm for finding the optimal mixed behavioral strategy for the rendezvous team. □

REFERENCES

[1] S. Alpern, *The rendezvous search problem*, SIAM J. Control Optim., 33 (1995), pp. 673–683.
[2] S. Alpern, *Games with repeated decisions*, J. Control Optim., 26 (1988), pp. 468–477.
[3] E. J. Anderson and R. R. Weber, *The rendezvous problem on discrete locations*, J. Appl. Probab., 28 (1990), pp. 839–851.
[4] S. Gal, *Search Games*, Academic Press, New York, 1980.
[5] J. Isbell, *Finitary games*, in Contributions to the Theory of Games III, Princeton University Press, Princeton, NJ, 1957.
[6] W. S. Lim, *Rendezvous–evasion games on discrete locations, with joint randomization*, Adv. in Appl. Probab., 29 (1997), pp. 1004–1017.
[7] W. S. Lim, S. Alpern, and A. Beck, *Rendezvous search on the line with more than two players*, Oper. Res., 45 (1997), pp. 357–364.
[8] T. Nakai, *A search game with one object and two searchers*, J. Appl. Probab., 23 (1986), pp. 696–707.
[9] M. Piccione and A. Rubinstein, *On the interpretation of decision problems with imperfect recall*, Working paper 9-94, Sackler Institute, Tel Aviv University, 1994.
[10] W. H. Ruckle, *Geometric Games and Their Applications*, Pitman, Boston, MA, 1983.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 3, pp. 960–980, May 1998

009

AN INTEGRAL INVARIANCE PRINCIPLE FOR DIFFERENTIAL INCLUSIONS WITH APPLICATIONS IN ADAPTIVE CONTROL*

E. P. RYAN†

Abstract. The Byrnes–Martin integral invariance principle for ordinary differential equations is extended to differential inclusions on $\mathbb{R}^N$. The extended result is applied in demonstrating the existence of adaptive stabilizers and servomechanisms for a variety of nonlinear system classes.

Key words. adaptive control, differential inclusions, invariance principles, nonlinear systems, universal servomechanisms

AMS subject classifications. 93D05, 93D09, 93D15, 93D21, 34D05

PII. S0363012996301701

1. Introduction.

Suppose that $\dot{x} = f(x)$ generates a semidynamical system on $\mathbb{R}^N$ with semiflow φ, and so, for each $x^0 \in \mathbb{R}^N$, $x(\cdot) = \varphi(\cdot, x^0)$ is the unique maximal forward-time solution of the initial-value problem $\dot{x} = f(x)$, $x(0) = x^0$. In [2], Byrnes and Martin prove the following integral invariance principle: if $\varphi(\cdot, x^0)$ is bounded and $\int_0^\infty l(\varphi(t, x^0))dt < \infty$ for some continuous function $l : \mathbb{R}^N \to \mathbb{R}_+ := [0, \infty)$, then $\varphi(t, x^0)$ tends, as $t \to \infty$, to the largest invariant (with respect to the differential equation) set in $l^{-1}(0)$, the zero level set of l. This result has ramifications in adaptive control, some of which are highlighted in the present paper. However, we wish to consider the (adaptive) control problem in a fairly general setting that allows time variation in the underlying differential equations, possible nonuniqueness of solutions, and discontinuous feedback strategies; each of these features places the problem outside the scope of [2]. For this reason we develop, in Theorem 2.10, an integral invariance principle for initial-value problems of the form $\dot{x} \in X(x)$, $x(0) = x^0$, where the set-valued map X is defined on some open domain $G \subset \mathbb{R}^N$ and is assumed to be upper semicontinuous with nonempty, convex, and compact values. In the case $G = \mathbb{R}^N$, Theorem 2.10 contains the following generalization of the Byrnes–Martin result: if $x(\cdot) : \mathbb{R}_+ \to \mathbb{R}^N$ is a bounded solution and $\int_0^\infty l(x(s))ds < \infty$ for some lower semicontinuous $l : \mathbb{R}^N \to \mathbb{R}_+$, then $x(t)$ tends, as $t \to \infty$, to the largest weakly invariant (with respect to the differential inclusion) set in $l^{-1}(0)$. One particular consequence of Theorem 2.10 is to facilitate the derivation of a nonsmooth extension, to differential inclusions, of LaSalle's invariance principle for differential equations; this extension may be of independent interest and is presented in Theorem 2.11. The remainder of the paper is devoted to the application (in a collection of five lemmas) of the generalized integral invariance principle to demonstrate, by construction and for a variety of nonlinear system classes, the existence of a single adaptive controller that achieves (without system identification, parameter estimation, or injection of probing signals) some prescribed objective for every system in the underlying class.

*Received by the editors April 10, 1996; accepted for publication (in revised form) March 12, 1997.

http://www.siam.org/journals/sicon/36-3/30170.html

†Department of Mathematical Sciences, University of Bath, Bath BA2 7AY, UK (epr@maths.bath.ac.uk).

2. Differential inclusions. Some known facts (tailored[1] to our immediate purpose) pertaining to differential inclusions are first assembled.

2.1. Maximal solutions. Consider the nonautonomous initial-value problem

$$\dot{x}(t) \in X(t, x(t)), \quad x(t) \in G, \quad x(t_0) = x^0, \tag{2.1}$$

where $G \neq \emptyset$ is an open subset of $\mathbb{R}^N$. The set-valued map $(t, x) \mapsto X(t, x) \subset \mathbb{R}^N$ in (2.1) is assumed to be upper semicontinuous[2] on $\mathbb{R} \times G$, with nonempty, convex, and compact values. This is sufficient (see, for example, [1, Chapter 2, Theorem 3]) to ensure that, for each $(t_0, x^0) \in \mathbb{R} \times G$, (2.1) admits a solution: an X-arc[3] $x \in AC([t_0, \omega); G)$, with $x(t_0) = x^0$.

DEFINITION 2.1. *A solution x of* (2.1) *is said to be* maximal *if it does not have a proper right extension which is also a solution of* (2.1).

PROPOSITION 2.2. *Every solution of* (2.1) *can be extended to a maximal solution.*

DEFINITION 2.3. *A solution $x \in AC([t_0, \omega); G)$ of* (2.1) *is* precompact *if it is maximal and the closure* $\mathrm{cl}(x([t_0, \omega)))$ *of its trajectory is a compact subset of G.*

PROPOSITION 2.4. *If $x \in AC([t_0, \omega); G)$ is a precompact solution of* (2.1)*, then $\omega = \infty$.*

2.2. Limit sets. Here, we specialize to the autonomous case of (2.1), rewritten as

$$\dot{x}(t) \in X(x(t)), \quad x(t) \in G, \quad x(0) = x^0, \tag{2.2}$$

where, without loss of generality, $t_0 = 0$ is assumed. The map $x \mapsto X(x) \subset \mathbb{R}^N$ (with domain G) is upper semicontinuous with nonempty, convex, and compact values.

DEFINITION 2.5. *Let $x \in AC([0, \omega); G)$ be a maximal solution of* (2.2)*. A point $\bar{x} \in \mathbb{R}^N$ is an* ω-limit point of x *if there exists an increasing sequence $(t_n) \subset [0, \omega)$ such that $t_n \to \omega$ and $x(t_n) \to \bar{x}$ as $n \to \infty$. The set $\Omega(x)$ of all ω-limit points of x is the* ω-limit set of x.

DEFINITION 2.6. *Let $C \subset \mathbb{R}^N$ be nonempty. A function $x \in AC([0, \omega); G)$ is said to* approach *C if $d_C(x(t)) \to 0$ as $t \to \omega$, where d_C is the (Euclidean) distance function for C defined (on $\mathbb{R}^N$) by $d_C(v) := \inf\{\|v - c\| \mid c \in C\}$.*

DEFINITION 2.7. *Relative to* (2.2)*, $S \subset \mathbb{R}^N$ is said to be a* weakly invariant set *if, for each $x^0 \in S \cap G$, there exists at least one maximal solution $x \in AC([0, \omega); G)$ of* (2.2) *with $\omega = \infty$ and with trajectory $x([0, \omega))$ in S.*

PROPOSITION 2.8. *If x is a precompact solution of* (2.2)*, then $\Omega(x)$ is a nonempty, compact, connected subset of G. Moreover, $\Omega(x)$ is the smallest closed set approached by x and is weakly invariant.*

2.3. Invariance principles. For later use, the following fact (a specialization of a more general result [4, Theorem 3.1.7]) is first recorded.

---

[1] Variants of Propositions 2.2, 2.4, and 2.8 can be found in, for example, [8], [19]; for general treatments of differential inclusions and related topics in set-valued analysis, nonsmooth control, and optimization, see [1], [4], [6], [7], [8], and [12].

[2] The set-valued map X is upper semicontinuous if it is upper semicontinuous at every point $\bar{\xi}$ of its domain in the sense that, for each $\varepsilon > 0$ there exists $\delta > 0$ such that $X(\xi) \subset X(\bar{\xi}) + \varepsilon\mathbb{B}$ for all ξ with $\|\xi - \bar{\xi}\| < \delta$, where $\mathbb{B}$ (with closure $\bar{\mathbb{B}}$) denotes the open unit ball centred at 0 in $\mathbb{R}^N$.

[3] For an interval $I \subset \mathbb{R}$ and $S \subset \mathbb{R}^N$, $AC(I; S)$ denotes the space of functions $I \to S$ that are absolutely continuous on compact subintervals of I. For simplicity, we write $AC(I)$ in place of $AC(I; I)$; the same notational convention applies to other function spaces. A function $x \in AC(I; G)$ is said to be an X-arc if it satisfies the differential inclusion in (2.1) almost everywhere.

PROPOSITION 2.9. *Let $I = [a, b]$, let nonempty $K \subset G$ be compact. If $(x_n) \subset AC(I; K)$ is a sequence of X-arcs and there exists a scalar Φ such that, for all n, $\|\dot{x}_n(t)\| \leq \Phi$ for almost all $t \in I$, then (x_n) has a subsequence converging uniformly to an X-arc $x \in AC(I; K)$.*

We now arrive at the main result, which generalizes [2, Theorem 1.2].

THEOREM 2.10. *Let $l : G \to \mathbb{R}$ be lower semicontinuous. Suppose that $U \subset G$ is nonempty and that $l(z) \geq 0$ for all $z \in U$. If x is a precompact solution of (2.2) with trajectory in U and $l \circ x \in L^1(\mathbb{R}_+)$, then x approaches the largest weakly-invariant set in $\Sigma := \{z \in \mathrm{cl}(U) \cap G |\ l(z) \leq 0\}$.*

Proof. By Proposition 2.4, x has maximal interval of existence $\mathbb{R}_+ = [0, \infty)$ and, by Proposition 2.8, has nonempty ω-limit set $\Omega(x)$. Clearly $\Omega(x) \subset \mathrm{cl}(U) \cap G$. Let $\bar{x} \in \Omega(x)$, and so there exists $(t_n) \subset \mathbb{R}_+$ with $t_n \to \infty$ and $x(t_n) \to \bar{x}$ as $n \to \infty$. Define $K := \mathrm{cl}(x(\mathbb{R}_+))$. By upper semicontinuity of X together with compactness of its values, $X(K)$ is compact, and so $\dot{x} \in L^\infty(\mathbb{R}_+; \mathbb{R}^N)$, with norm $\|\dot{x}\|_\infty$. Write $I = [0, 1]$ and define a sequence (x_n) of X-arcs $x_n \in AC(I; K)$ by $x_n(s) := x(s + t_n)$. Evidently, $x_n(0) \to \bar{x}$ as $n \to \infty$. By Proposition 2.9 (with $\Phi = \|\dot{x}\|_\infty$), (x_n) has a subsequence, which we do not relabel, converging uniformly to an X-arc $x^* \in AC(I; K)$, with $x^*(0) = \bar{x}$.

Define $\lambda^* \in L^1(I; \mathbb{R})$ by $\lambda^*(s) := l(x^*(s))$ and the sequence $(\lambda_n) \subset L^1(I; \mathbb{R})$ by $\lambda_n(s) := l(x_n(s))$. By lower semicontinuity of l, together with continuity of x, x^* and (uniform) convergence of (x_n) to x^*, it follows that λ^* and λ_n, $n \in \mathbb{N}$, are lower semicontinuous with $\liminf_{n\to\infty} \lambda_n(s) \geq \lambda^*(s)$ for all $s \in I$. By Fatou's lemma,

$$\int_0^t \lambda^*(s)ds \leq \int_0^t \liminf_{n\to\infty} \lambda_n(s)\, ds \leq \liminf_{n\to\infty} \int_0^t \lambda_n(s)\, ds \quad \forall\ t \in I.$$

By the hypotheses, $l(x(t)) \geq 0$ for all t and $\int_0^\infty l(x(t))dt < \infty$. Therefore,

$$W(t) := \int_t^\infty l(x(s))ds \quad \forall\ t \geq 0$$

defines a monotone function $W \in AC(\mathbb{R}_+)$ with $W(t) \downarrow 0$ as $t \to \infty$. Hence,

$$\begin{aligned} 0 = \lim_{n\to\infty} W(t_n) - \lim_{n\to\infty} W(t + t_n) &= \lim_{n\to\infty} \int_{t_n}^{t+t_n} l(x(s))ds \\ &= \lim_{n\to\infty} \int_0^t l(x(s + t_n))ds = \lim_{n\to\infty} \int_0^t \lambda_n(s)ds \geq \int_0^t \lambda^*(s)ds \quad \forall\ t \in I. \end{aligned}$$

Seeking a contradiction, suppose $\epsilon := \lambda^*(0) > 0$. Then, by lower semicontinuity of λ^*, there exists $t \in (0, 1]$ such that $\lambda^*(s) > \lambda^*(0) - \frac{1}{2}\epsilon = \frac{1}{2}\epsilon$ for all $s \in (0, t)$, whence the contradiction

$$0 \geq \int_0^t \lambda^*(s)ds > \frac{1}{2}\epsilon t > 0.$$

Therefore, $0 \geq \lambda^*(0) = l(x^*(0)) = l(\bar{x})$, and so $\bar{x} \in \Sigma$. Since $\bar{x} \in \Omega(x)$ is arbitrary, we have $\Omega(x) \subset \Sigma$. By Proposition 2.8, x approaches $\Omega(x)$ and the latter is a weakly invariant set. Therefore, x approaches the largest weakly invariant set in Σ. ∎

The next result is a nonsmooth extension, to differential inclusions, of LaSalle's theorem [11, Chapter 2, Theorem 6.4]; a smooth version (that is, restricted to smooth functions V) is given in [19, Theorem 1] and a nonsmooth version is proved in [23,

Theorem 3]; the alternative proof given below is considerably simpler, by virtue of its use of the integral invariance principle. First, we give Clarke's [4] definition of a generalized directional derivative $V^o(z;\phi)$ of a locally Lipschitz function $V : G \to \mathbb{R}$ at z in direction ϕ:

$$V^o(z;\phi) := \limsup_{\substack{y \to z \\ h \downarrow 0}} \frac{V(y+h\phi) - V(y)}{h} .$$

The map $(z,\phi) \mapsto V^o(z;\phi)$ is upper semicontinuous (in the sense of real-valued functions) and, for each z, the map $\phi \mapsto V^o(z;\phi)$ is Lipschitz continuous.

THEOREM 2.11. *Let $V : G \to \mathbb{R}$ be locally Lipschitz. Define*

$$u : G \to \mathbb{R}, \quad z \mapsto u(z) := \max\{V^o(z;\phi) |\ \phi \in X(z)\}.$$

Suppose that $U \subset G$ is non-empty and that $u(z) \le 0$ for all $z \in U$. If x is a precompact solution of (2.2) with trajectory in U, then, for some constant $c \in V(\mathrm{cl}(U) \cap G)$, x approaches the largest weakly invariant set in $\Sigma \cap V^{-1}(c)$, where

$$\Sigma = \{z \in \mathrm{cl}(U) \cap G |\ u(z) \ge 0\}.$$

Proof. This result is essentially a corollary to Theorem 2.10 insofar as the essence of the proof is to show that the hypotheses of Theorem 2.10 hold with $l := -u$.

We first show that u is upper semicontinuous (and so, $l \equiv -u$ is lower semicontinuous). Let $z \in G$ be arbitrary and let $(z_n) \subset G$ be such that $z_n \to z$ as $n \to \infty$. From $(u(z_n))$ we extract a subsequence $(u(z_{n_k}))$ with $u(z_{n_k}) \to \limsup_{n\to\infty} u(z_n)$ as $k \to \infty$. For each k, let ϕ_k be a maximizer of continuous $V^o(z_{n_k};\cdot)$ over compact $X(z_{n_k})$, and so $u(z_{n_k}) = V^o(z_{n_k};\phi_k)$. Let $\varepsilon > 0$. By upper semicontinuity of X, we have $X(z_{n_k}) \subset X(z) + \varepsilon\mathbb{B}$ for all k sufficiently large. Since $\phi_k \in X(z_{n_k})$ and $X(z)$ is compact, (ϕ_k) contains a subsequence converging to $\phi^* \in \mathrm{cl}(X(z) + \varepsilon\mathbb{B})$. Since $\varepsilon > 0$ is arbitrary and $X(z)$ is compact, $\phi^* \in X(z)$. Thus, invoking upper semicontinuity of $V^o(\cdot;\cdot)$, we may conclude that $\limsup_{n\to\infty} u(z_n) \le V^o(z;\phi^*) \le u(z)$, whence upper semicontinuity of u.

Observe that, for all $z \in U$,

$$V_+(z;\phi) := \liminf_{h\downarrow 0} \frac{V(z+h\phi) - V(z)}{h} \le V^o(z;\phi) \le u(z) \le 0 \quad \forall\ \phi \in X(z). \tag{2.3}$$

By Proposition 2.4, x has interval of existence $\mathbb{R}_+$. Let $\mathcal{O} \subset \mathbb{R}_+$ denote the set of measure zero on which the derivative $\dot{x}(t)$ fails to exist. Since V is locally Lipschitz, for each $t \in \mathbb{R}_+\backslash\mathcal{O}$ there exists a constant L_t such that, for all $h > 0$ sufficiently small,

$$\begin{aligned} V(x(t+h)) - V(x(t)) &\le V(x(t) + h\dot{x}(t)) - V(x(t)) + L_t|x(t+h) - x(t) - h\dot{x}(t)| \\ &\le V(x(t+h)) - V(x(t)) + 2L_t|x(t+h) - x(t) - h\dot{x}(t)|. \end{aligned}$$

Therefore,

$$\liminf_{h\downarrow 0} \frac{V(x(t+h)) - V(x(t))}{h} = V_+(x(t);\dot{x}(t)) \quad \forall\ t \in \mathbb{R}_+\backslash\mathcal{O}. \tag{2.4}$$

Next, we prove that $V \circ x :\ t \mapsto V(x(t))$ is nonincreasing on $\mathbb{R}_+$. This we do by showing that $V \circ x$ is nonincreasing on every compact subinterval. Let $[\alpha,\beta] \subset \mathbb{R}_+$,

and let $K \subset G$ be compact and such that $x([\alpha,\beta]) \subset K$. Since V is locally Lipschitz on G, it is Lipschitz on K. Thus, the restriction of $V \circ x$ to $[\alpha,\beta]$ is a composition of a Lipschitz function and an absolutely continuous function and so is itself absolutely continuous. It now follows from (2.3) and (2.4) that

$$V(x(t)) - V(x(\tau)) \;=\; \int_\tau^t (V \circ x)' \;\leq\; \int_\tau^t u(x(s))ds \leq 0 \quad \forall\, t,\tau \in [\alpha,\beta],\ t \geq \tau.$$

Therefore, $t \mapsto V(x(t))$ is nonincreasing on $[\alpha,\beta]$. Since $[\alpha,\beta] \subset \mathbb{R}_+$ is arbitrary, we conclude that $V \circ x$ is nonincreasing on $\mathbb{R}_+$ with

$$V(x(t)) - V(x(\tau)) \leq \int_\tau^t u(x(s))ds \leq 0 \quad \forall\, t,\tau \in \mathbb{R}_+,\ t \geq \tau. \tag{2.5}$$

By continuity of V and precompactness of x, we conclude that $V(x(\cdot))$ is bounded. Therefore, $V(x(t)) \downarrow c := \inf_{t\in\mathbb{R}_+} V(x(t)) \in \mathbb{R}$ as $t \to \infty$. It follows that (i) $\Omega(x) \subset V^{-1}(c)$ and (ii) for all $t \geq 0$,

$$\int_0^t l(x(s))ds \equiv -\int_0^t u(x(s))ds \leq V(x(0)) - V(x(t)) \leq V(x(0)) - c < \infty.$$

An application of Theorem 2.10 completes the proof. □

3. Adaptive control. Approaches to adaptive control may be classified into methods that—either implicitly or explicitly—exhibit some aspect of identification of the process to be controlled and methods that seek only to control. The latter approach, to be adopted here and sometimes referred to as universal control, has its origins in the work of Byrnes and Willems [3], [27], Mårtensson [13], [14], [15], Morse [16], Nussbaum [17] and others (see [9] for a survey and comprehensive bibliography). In common with its above-cited precursors, this section of the present paper is concerned with demonstrating the existence—under relatively weak assumptions—of a single controller that achieves some prescribed objective for every system in the underlying class. In contrast with its above-cited precursors (which deal mainly with classes of linear systems, possibly subject to "mild" nonlinear perturbations, in a context of adaptive linear feedback), the present paper considers strongly nonlinear systems and nonsmooth feedback (for an overview of adaptive control of nonlinear systems in the context of *smooth* feedback, see [18]). In essence, the ensuing two subsections provide a unified analysis—unified through its use of the integral invariance principle—of various problems in nonlinear adaptive control (some closely related problems have been individually investigated, via alternative analyses, in [5], [10], [20], and [21]).

3.1. Scalar systems. First, consider scalar systems of the form

$$\dot{y}(t) = f(p(t),y(t)) + bu(t), \quad y(t),u(t) \in \mathbb{R},\ p(t) \in \mathbb{R}^P, \quad y(t_0) = y^0, \tag{3.1}$$

where parameters $b \in \mathbb{R}$, $P \in \mathbb{N}$, and functions f, p are unknown. The state $y(t)$ is available for feedback. We will identify (3.1) with the quadruple (b,f,p,P).

For any function $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ that is both continuous and positive definite ($\phi(s) > 0$ for all $s \neq 0$), we denote, by $\mathcal{N}_\phi$, the set of system quadruples (b,f,p,P) satisfying the following three assumptions.

Assumption A. $b \neq 0$.

Assumption B. $(p,y) \mapsto f(p,y)$, $\mathbb{R}^P \times \mathbb{R} \to \mathbb{R}$ is a continuous function and is ϕ-bounded uniformly with respect to p in compact sets; precisely, for every compact $K \subset \mathbb{R}^P$, there exists scalar μ_K such that $|f(p,y)| \leq \mu_K \phi(|y|)$ for all $(p,y) \in K \times \mathbb{R}$.

Assumption C. $p(\cdot) \in L^\infty(\mathbb{R}; \mathbb{R}^P)$.

By virtue of Assumption C, without loss of generality, $t_0 = 0$ may be assumed in (3.1); this we will do, without further comment, throughout this subsection.

Examples. (a) Let $\phi : |y| \mapsto \exp(|y|)$. Then all polynomial systems, of arbitrary degree, of the form $\dot{y}(t) = p_1(t) + p_2(t)y(t) + \cdots + p_P(t)y^{P-1}(t) + bu(t)$, with $b \neq 0$ and coefficient functions $p_i(\cdot) \in L^\infty(\mathbb{R})$, are of class $\mathcal{N}_\phi$.

(b) Suppose that Assumptions A and C hold and that the only a priori information on continuous f is its behavior "at infinity," captured in the following manner: for some known continuous $\hat{\phi} : \mathbb{R}_+ \to \mathbb{R}_+$, $(p, y) \mapsto f(p, y)$ is $O(\hat{\phi}(|y|))$ as $|y| \to \infty$, uniformly with respect to p in compact sets in the sense that, for every compact $K \subset \mathbb{R}^P$, there exist scalars c_K and C_K such that, for all $p \in K$, $|f(p, y)| \leq c_K \hat{\phi}(|y|)$ for all $|y| > C_K$. Then Assumption B holds with $\phi := 1 + \hat{\phi}$, and so $(b, f, p, P) \in \mathcal{N}_\phi$.

3.1.1. Adaptive stabilizer. Let $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ be a continuous, positive-definite function. Assuming only that the function ϕ and the instantaneous state $y(t)$ are available for control purposes, we will show that the following adaptive feedback strategy (appropriately interpreted) is a $\mathcal{N}_\phi$-universal stabilizer in the sense that it assures that the state of (3.1) approaches $\{0\}$ for all quadruples $(b, f, p, P) \in \mathcal{N}_\phi$ while maintaining boundedness of the controller function $\lambda(\cdot)$:

$$u(t) = \nu(\lambda(t))\phi(|y(t)|)\mathrm{sgn}(y(t)), \quad \dot{\lambda}(t) = \phi(|y(t)|)|y(t)|, \quad \lambda(0) = \lambda^0, \tag{3.2}$$

where ν is any continuous function $\mathbb{R} \to \mathbb{R}$ with the properties

$$\text{(a)} \quad \limsup_{\eta \to \infty} \frac{1}{\eta} \int_0^\eta \nu = +\infty, \qquad \text{(b)} \quad \liminf_{\eta \to \infty} \frac{1}{\eta} \int_0^\eta \nu = -\infty. \tag{3.3}$$

For example, $\nu : \theta \mapsto \theta^2 \cos\theta$ suffices.

In view of the discontinuous nature of the feedback (however, note that, if $\phi(0) = 0$, then the feedback is continuous), we interpret the strategy (3.2) in the set-valued sense

$$u(t) \in \nu(\lambda(t))\phi(|y(t)|)\sigma(y(t)), \quad \dot{\lambda}(t) = \phi(|y(t)|)|y(t)|, \ \lambda(0) = \lambda^0 \tag{3.4}$$

with $y \mapsto \sigma(y) \subset \mathbb{R}$ given by

$$\sigma(y) := \begin{cases} \{\mathrm{sgn}(y)\}, & y \neq 0, \\ [-1, +1], & y = 0\ . \end{cases} \tag{3.5}$$

Let $(b, f, p, P) \in \mathcal{N}_\phi$. By properties of $f(\cdot, \cdot)$ and boundedness of $p(\cdot)$, there exists a scalar μ such that $|f(p(t), y)| \leq \mu\phi(|y|)$ for all (t, y).

We embed the feedback-controlled system in a differential inclusion on $\mathbb{R}^2$:

$$\dot{x}(t) \in X(x(t)), \quad x(t) = (y(t), \lambda(t)) \in G := \mathbb{R}^2, \quad x(0) = x^0 = (y^0, \lambda^0), \tag{3.6}$$

where $x \mapsto X(x) \subset \mathbb{R}^2$ is given by

$$X(x) \equiv X(y, \lambda) := \{v + bu| \ |v| \leq \mu\phi(|y|), \ u \in \nu(\lambda)\phi(|y|)\sigma(y)\} \times \{\phi(|y|)|y|\}.$$

X is upper semicontinuous on $\mathbb{R}^2$ with nonempty, convex, and compact values. Therefore for each $x^0 \in \mathbb{R}^2$, the initial-value problem (3.6) has a solution and, by Proposition 2.2, every solution can be extended to a maximal solution.

LEMMA 3.1. *Let $x^0 \in \mathbb{R}^2$ be arbitrary and let $x(\cdot) = (y(\cdot), \lambda(\cdot))$ be a maximal solution of* (3.6)*, defined on its maximal interval of existence $[0, \omega)$. Then* (i) *$\omega = \infty$;* (ii) *$\lim_{t\to\infty} \lambda(t)$ exists and is finite;* (iii) *$y(t) \to 0$ as $t \to \infty$.*

Proof. The essence of the proof is to establish boundedness of $x(\cdot)$, whence, by Proposition 2.4, assertion (i) and, by monotonicity, assertion (ii): state convergence to zero (assertion (iii)) is then an immediate consequence of Theorem 2.10.

For almost all $t \in [0, \omega)$, we have

$$y(t)\dot{y}(t) \leq [\mu + b\nu(\lambda(t))]\,\phi(|y(t)|)|y(t)| = [\mu + b\nu(\lambda(t))]\dot{\lambda}(t), \tag{3.7}$$

which, on integration, yields

$$0 \leq y^2(t) \leq y^2(\tau) + 2\mu[\lambda(t) - \lambda(\tau)] + 2b\int_{\lambda(\tau)}^{\lambda(t)} \nu \quad \forall\ t, \tau \in [0, \omega),\ t \geq \tau. \tag{3.8}$$

Seeking a contradiction, suppose that solution component $\lambda(\cdot)$ (monotone increasing) is unbounded. Fix τ such that $\lambda(\tau) \geq 1$. Dividing by $\lambda(t) \geq \lambda(\tau) \geq 1$ in (3.8) gives

$$0 \leq \text{constant} + \frac{b}{\lambda(t)}\int_{\lambda(\tau)}^{\lambda(t)} \nu \quad \forall\ t \in [\tau, \omega).$$

Recalling that $b \neq 0$ and taking limit inferior as $t \uparrow \omega$ ($\lambda(t) \uparrow \infty$) lead to a contradiction of one or the other of properties (3.3). Hence, $\lambda(\cdot)$ is bounded, and so, by (3.8), $y(\cdot)$ is also bounded. Therefore, $x(\cdot) = (y(\cdot), \lambda(\cdot)) \in AC([0, \omega); \mathbb{R}^2)$ is a precompact solution of (3.6), and so, by Proposition 2.4, $\omega = \infty$. Defining $l : \mathbb{R}^2 \to \mathbb{R}_+$, $x \equiv (y, \lambda) \mapsto \phi(|y|)|y|$, for which $l^{-1}(0) = \{0\} \times \mathbb{R}$ and $\dot{\lambda} = l \circ x$, we may conclude, by boundedness of $\lambda(\cdot)$, that $l \circ x \in L^1(\mathbb{R}_+)$, and so, by Theorem 2.10 (with $U = \mathbb{R}^2$), $x(\cdot) = (y(\cdot), \lambda(\cdot))$ approaches the set $\{0\} \times \mathbb{R}$. In particular, $y(t) \to 0$ as $t \to \infty$ and, by boundedness and monotonicity of $\lambda(\cdot)$, $\lim_{t\to\infty} \lambda(t)$ exists and is finite. ∎

3.1.2. Adaptive servomechanism. We now turn attention to the servomechanism problem for scalar systems (3.1), that is, the construction of controls that cause the state to track, asymptotically, reference signals $r(\cdot)$ of some given class in the sense that $|y(t) - r(t)| \to 0$ as $t \to \infty$. For the class of reference signals (previously adopted in [20], [10], [21]) we take the (Sobolev) space $\mathcal{R} = W^{1,\infty}(\mathbb{R})$ of functions $r \in (AC \cap L^\infty)(\mathbb{R})$ with essentially bounded derivative $\dot{r} \in L^\infty(\mathbb{R})$, equipped with the norm $\|r\|_{1,\infty} = \|r\|_\infty + \|\dot{r}\|_\infty$, where $\|\ \|_\infty$ denotes the norm on $L^\infty(\mathbb{R})$.

We impose a stronger assumption on the function f by requiring that Assumption B should hold for some known, continuous, positive-definite, nondecreasing function ϕ with the additional property that, for each $R \geq 0$, there exists a scalar ρ_R such that

$$\phi(|e + r|) \leq \rho_R \phi(|e|) \quad \forall\ (e, r) \in \mathbb{R} \times [-R, R]. \tag{3.9}$$

Note that, by positive definiteness of ϕ together with property (3.9), $\phi(0) > 0$.

Example. Let $\phi : |y| \mapsto \exp(|y|)$, which has the property (3.9), and so all polynomial systems (of arbitrary degree and with coefficients in $L^\infty(\mathbb{R})$), as cited in a previous Example, remain admissible.

Let ϕ be a continuous, positive-definite, nondecreasing function with property (3.9). We claim that, in order to assure that the tracking error $e(\cdot) = y(\cdot) - r(\cdot)$ approaches $\{0\}$ for all reference signals $r \in \mathcal{R}$ and all quadruples $(b, f, p, P) \in \mathcal{N}_\phi$, it suffices to replace every occurrence of $y(t)$ in (3.4) by $e(t)$. Proof of this claim follows.

Let $(b, f, p, P) \in \mathcal{N}_\phi$. Write $\tilde{P} = P + 2$ and define the continuous function

$$\tilde{f} : \mathbb{R}^{\tilde{P}} \times \mathbb{R} \to \mathbb{R}, \quad (\tilde{p}, e) \equiv (p, r, s, e) \mapsto f(p, e + r) - s.$$

Let $\tilde{K} \subset \mathbb{R}^{\tilde{P}}$ be compact, and so there exist compact $K \subset \mathbb{R}^P$ and $R > 0$ such that $\tilde{K} \subset K \times [-R, R]^2$. By properties of f and ϕ, there exist constants μ_K and ρ_R such that, for all $(p, r, s) \equiv \tilde{p} \in \tilde{K} \subset K \times [-R, R]^2$,

$$|\tilde{f}(\tilde{p}, e)| \leq |f(p, e + r)| + |s| \leq \mu_K \phi(|e + r|) + R \leq \mu_K \rho_R \phi(|e|) + R \leq \tilde{\mu}_{\tilde{K}} \phi(|e|),$$

with $\tilde{\mu}_{\tilde{K}} := \mu_K \rho_R + R/\phi(0)$. Thus, $\tilde{f}$ is ϕ-bounded uniformly with respect to $\tilde{p}$ in compact sets.

Let $r \in \mathcal{R}$ be arbitrary. Then $\tilde{p}(\cdot) := (p(\cdot), r(\cdot), \dot{r}(\cdot)) \in L^\infty(\mathbb{R}; \mathbb{R}^{\tilde{P}})$, and so $(b, \tilde{f}, \tilde{p}, \tilde{P}) \in \mathcal{N}_\phi$. Expressed in terms of the tracking error $e(t) = y(t) - r(t)$, the system dynamics have the form

$$\dot{e}(t) = \tilde{f}(\tilde{p}(t), e(t)) + bu(t), \quad \tilde{p}(t) = (p(t), r(t), \dot{r}(t)) \in \mathbb{R}^{\tilde{P}}, \quad e(0) = e^0.$$

We are now in precisely the same context, modulo notation, as in the case of an adaptive stabilizer, and so, replacing all occurrences of $y(t)$ in (3.4) with $e(t)$ to yield

$$u(t) \in \nu(\lambda(t))\phi(|e(t)|)\sigma(e(t)), \quad \dot{\lambda}(t) = \phi(|e(t)|)|e(t)|, \ \lambda(0) = \lambda^0, \tag{3.10}$$

then the same argument (as used to establish Lemma 3.1) applies *mutatis mutandis* to conclude that (3.10) is an $(\mathcal{R}, \mathcal{N}_\phi)$-universal servomechanism: for each $r(\cdot) \in \mathcal{R}$ and $(b, f, p, P) \in \mathcal{N}_\phi$, every solution $(e(\cdot), \lambda(\cdot))$ of the controlled system has maximal interval of existence $\mathbb{R}_+$ with $e(t) \to 0$ as $t \to \infty$, and $\lim_{t\to\infty} \lambda(t)$ exists and is finite.

3.1.3. Practical stabilization and tracking by continuous feedback. The adaptive strategies outlined above are (generically) of a discontinuous feedback nature. From a viewpoint of practical utility, this feature might be regarded as unpleasant. Here, we investigate the possibility of adopting smooth approximations to the discontinuous feedbacks. Of course, in so doing, one would expect to pay a price. It will be shown that, if the objective of attractivity of the zero state (in the stabilization case) or asymptotic tracking (in the case of a servomechanism) is weakened to requiring global attractivity of any (arbitrarily small) prescribed neighborhood of zero or, for the servomechanism problem, tracking to within any prescribed (arbitrarily small but nonzero) error margin, then such approximations are feasible. We will present this result only in the context of the stabilization problem (imposing the additional property (3.9), the corresponding result for the servomechanism problem is readily inferred by analogy with section 3.1.2). The naïve idea, as developed in [10] and [21], is to inhibit the adaption whenever the state lies within the prescribed neighborhood of zero.

Let $\epsilon > 0$ be arbitrary and let d_ϵ denote the distance function for the set $[-\epsilon, \epsilon]$; thus, $d_\epsilon(x) := |x| - \epsilon$ if $|x| \geq \epsilon$; $d_\epsilon(x) = 0$ if $|x| < \epsilon$.

Let $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ be continuous and positive-definite. Let $\mathrm{sat}_\epsilon : \mathbb{R} \to \mathbb{R}$ be any continuous function (arbitrarily smooth) such that (i) $|\mathrm{sat}_\epsilon(x)| \leq 1$ for all x and (ii) $\mathrm{sat}_\epsilon(x) = \mathrm{sgn}(x)$ whenever $|x| \geq \epsilon$. We will show that the following strategy assures that the state of (3.1) approaches the interval $[-\epsilon, \epsilon]$ for all quadruples $(b, f, p, P) \in \mathcal{N}_\phi$:

$$u(t) = \nu(\lambda(t))\phi(|y(t)|)\mathrm{sat}_\epsilon(y(t)), \quad \dot{\lambda}(t) = \phi(|y(t)|)d_\epsilon(y(t)), \quad \lambda(0) = \lambda^0,$$

where, as before, ν is any continuous function with properties (3.3).

Let $(b, f, p, P) \in \mathcal{N}_\phi$, and so there exists constant μ such that $|f(p(t), y)| \le \mu\phi(|y|)$ for all (t, y). Define the set-valued map $y \mapsto \sigma_\epsilon(y)$ by

$$\sigma_\epsilon(y) := \begin{cases} \{\mathrm{sgn}(y)\}, & |y| \ge \epsilon, \\ [-1, +1], & |y| < \epsilon. \end{cases}$$

Evidently, sat_ϵ is a continuous selection from σ_ϵ. We now embed the smooth-feedback–controlled system in the following differential inclusion on $\mathbb{R}^2$:

$$\dot{x}(t) \in X_\epsilon(x(t)), \quad x(t) = (y(t), \lambda(t)) \in \mathbb{R}^2, \quad x(0) = x^0 = (y^0, \lambda^0), \tag{3.11}$$

where $x \mapsto X_\epsilon(x) \subset \mathbb{R}^2$ is given by

$$X_\epsilon(x) \equiv X_\epsilon(y, \lambda) := \{v + bu|\ |v| \le \mu\phi(|y|),\ u \in \nu(\lambda)\phi(|y|)\sigma_\epsilon(y)\} \times \{\phi(|y|)d_\epsilon(y)\}.$$

X_ϵ is upper semicontinuous on $\mathbb{R}^2$ with nonempty, convex, and compact values. Therefore for each $x^0 \in \mathbb{R}^2$, the initial-value problem (3.11) has a solution and every solution has a maximal extension.

LEMMA 3.2. *Let $x^0 \in \mathbb{R}^2$ be arbitrary and let $x(\cdot) = (y(\cdot), \lambda(\cdot))$ be a maximal solution of* (3.11), *defined on its maximal interval of existence* $[0, \omega)$. *Then* (i) $\omega = \infty$; (ii) $\lim_{t\to\infty} \lambda(t)$ *exists and is finite;* (iii) $d_\epsilon(y(t)) \to 0$ *as* $t \to \infty$.

Proof. For almost all $t \in [0, \omega)$,

$$\begin{aligned}(d/dt)(\frac{1}{2}d_\epsilon^2(y(t))) &= d_\epsilon(y(t))\mathrm{sat}_\epsilon(y(t))\dot{y}(t) \\ &\le [\mu + b\nu(\lambda(t))]\,\phi(|y(t)|)d_\epsilon(y(t)) = [\mu + b\nu(\lambda(t))]\dot{\lambda}(t),\end{aligned}$$

which, on integration, yields

$$0 \le d_\epsilon^2(y(t)) \le d_\epsilon^2(y(\tau)) + 2\mu[\lambda(t) - \lambda(\tau)] + 2b\int_{\lambda(\tau)}^{\lambda(t)} \nu \tag{3.12}$$

valid for all $t, \tau \in [0, \omega)$, with $t \ge \tau$. By precisely the same contradiction argument as employed previously in the case of the discontinuous stabilizer, we may deduce that $x(\cdot) = (y(\cdot), \lambda(\cdot)) \in AC([0, \omega); \mathbb{R}^2)$ is a precompact solution of (3.11), and so, $\omega = \infty$. Defining $l : \mathbb{R}^2 \to \mathbb{R}$, $x \equiv (y, \lambda) \mapsto \phi(|y|)d_\epsilon(y)$ (and so, $\dot{\lambda} = l \circ x$), we conclude, by boundedness of $\lambda(\cdot)$, that $l \circ x \in L^1(\mathbb{R}_+)$). Therefore, by Theorem 2.10 (with $U = \mathbb{R}^2$), $x(\cdot) = (y(\cdot), \lambda(\cdot))$ approaches the set $l^{-1}(0) \equiv \{(y, \lambda)|\ d_\epsilon(y) = 0\}$. In particular, $d_\epsilon(y(t)) \to 0$ as $t \to \infty$ and, by monotonicity of bounded $\lambda(\cdot)$, $\lim_{t\to\infty} \lambda(t)$ exists and is finite. ∎

3.1.4. Dynamically perturbed scalar systems. Let $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ be continuous and positive definite. Let $\Sigma_1 = (b, f, p, P) \in \mathcal{N}_\phi$. We wish to consider the situation wherein Σ_1 is subject to perturbations generated through its interconnection with a dynamical system Σ_2 (Fig. 3.1).

System Σ_2 is assumed to correspond to a differential equation (driven by the state of the scalar system Σ_1) on $\mathbb{R}^N$ of the form

$$\Sigma_2 : \qquad \dot{\zeta}(t) = g(y(t), \zeta(t)), \quad w(t) = h(\zeta(t)), \quad \zeta(0) = \zeta^0 \in \mathbb{R}^N, \tag{3.13}$$

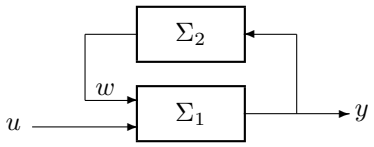


FIG. 3.1.

with input $y(t)$, and scalar output $w(t)$ perturbing Σ_1. Notationally, we identify the system Σ_2 with the triple (g, h, N). The overall system has representation (on $\mathbb{R}\times\mathbb{R}^N$)

$$\begin{cases} \dot{y}(t) = f(p(t), y(t)) + h(\zeta(t)) + bu(t), \\ \dot{\zeta}(t) = g(y(t), \zeta(t)), \quad (y(0), \zeta(0)) = (y^0, \zeta^0). \end{cases} \tag{3.14}$$

We will define, via Assumption D below, a class $\mathcal{P}_\psi$ of admissible systems $\Sigma_2 = (g, h, N)$, such that the $\mathcal{N}_\phi$-universal stabilizer of section 3.1.1 is readily modified to yield a $(\mathcal{N}_\phi, \mathcal{P}_\psi)$-universal stabilizer. Before stating Assumption D, we cite Sontag's concept of input-to-state stability [24], [25] (see also [26]) in the context of (3.13), with g assumed to be locally Lipschitz and with $y(\cdot)$ regarded as an independent input of class $L^\infty_{\rm loc}(\mathbb{R}_+;\mathbb{R})$. System (3.13) is *input-to-state stable* (ISS) if there exist a continuous, strictly increasing function $\gamma : \mathbb{R}_+ \to \mathbb{R}_+$, with $\gamma(0) = 0$, and a continuous function $\beta : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$, with $\beta(0, t) = 0$ for all t and having the properties that, for each $t \geq 0$, $\beta(\cdot, t)$ is strictly increasing and, for each $s \geq 0$, $\beta(s, t) \downarrow 0$ as $t \to \infty$, such that, for every $\zeta^0 \in \mathbb{R}^N$ and every $y(\cdot) \in L^\infty_{\rm loc}(\mathbb{R}_+;\mathbb{R})$, the (unique) maximal solution $\zeta(\cdot)$ of the initial-value problem (3.13) satisfies $\|\zeta(t)\| \leq \beta(\|\zeta^0\|, t)+\gamma(\|y_t\|_\infty)$ for all $t \geq 0$, where y_t denotes the truncation of y at t, that is, $y_t(s) = y(s)$ if $s \leq t$ and $y_t(s) = 0$ if $s > t$. If (3.13) is ISS, then it is forward complete and has the convergent-input, convergent-state property: for each $\zeta^0 \in \mathbb{R}^N$ and $y \in L^\infty_{\rm loc}(\mathbb{R}_+;\mathbb{R})$, the unique solution $\zeta(\cdot)$ of the initial-value problem has maximal interval of existence $\mathbb{R}_+$ and, if $y(t) \to 0$ as $t \to \infty$, then $\zeta(t) \to 0$ as $t \to \infty$.

For continuous $\psi : \mathbb{R}_+ \to \mathbb{R}_+$, we denote by $\mathcal{P}_\psi$ the set of system triples $\Sigma_2 = (g, h, N)$ satisfying the following assumption.

Assumption D. (i) $g : \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}^N$ is locally Lipschitz; (ii) system (3.13) is ISS; (iii) $h : \mathbb{R}^N \to \mathbb{R}$ is continuous; (iv) there exist a function $\alpha_0 : \mathbb{R}_+ \to \mathbb{R}_+$ and scalar $\alpha_1 > 0$, such that, for each $(\zeta^0, y(\cdot)) \in \mathbb{R}^N \times L^\infty_{\rm loc}(\mathbb{R}_+;\mathbb{R})$, the (unique) solution $\zeta(\cdot)$ of (3.13) satisfies

$$\int_0^t h(\zeta(s))y(s)ds \leq \alpha_0(\|\zeta^0\|) + \alpha_1 \int_0^t \psi(|y(s)|)|y(s)|ds \quad \forall\ t \geq 0.$$

While Assumption D is rather restrictive, it is not difficult to identify nontrivial classes of systems for which the assumption holds. Three such examples follow, the first of which is easily seen; the second and third can be verified by arguments invoking [22, Theorem 2].

Examples. (a) Let $\psi : |y| \mapsto |y|$ and suppose (g, h, N) defines a linear system with $g : (y, \zeta) \mapsto A\zeta + By$, $h : \zeta \mapsto C\zeta$. If A has spectrum, spec(A), in the open left half complex plane $\mathbb{C}_-$, then $(g, h, N) \in \mathcal{P}_\psi$.

(b) More generally, let $\psi : |y| \mapsto |y|^k$, $k \geq 1$. Assume $g : \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}^N$ is locally Lipschitz and $h : \mathbb{R}^N \to \mathbb{R}$ is continuous. In addition, assume g and h are positively homogeneous of degree k. If $\{0\}$ is an asymptotically stable equilibrium of

the unforced system $\dot{\zeta}(t) = g(0, \zeta(t))$, then $(g, h, N) \in \mathcal{P}_\psi$. For example, with $k = 3$ and $\psi : |y| \mapsto |y|^3$, systems (with $N = 1$ and with unknown real parameters a_i) of the form $\dot{\zeta} = a_1\zeta^3 + a_2\zeta^2 y + a_3\zeta y^2 + a_4 y^3$, with output $w = a_5\zeta^3$, are of class $\mathcal{P}_\psi$, provided that $a_1 < 0$.

(c) Let $k > 0$, $k_h \geq 1$ and $\psi : |y| \mapsto \frac{1}{2}\left(|y|^{(k+1)k_h - 1} + |y|^{1/k}\right)$. Assume $g : \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}^N$ is locally Lipschitz and that $h : \mathbb{R}^N \to \mathbb{R}$ is continuous and positively homogeneous of degree k_h. If, in addition, g is positively homogeneous of degree k_g, with $1 \leq k_g \leq (k+1)k_h$, and $\{0\}$ is an asymptotically stable equilibrium of the unforced system $\dot{\zeta}(t) = g(0, \zeta(t))$, then $(g, h, N) \in \mathcal{P}_\psi$. For example, systems (with $N = 1$ and unknown parameters ρ, $a_i \in \mathbb{R}$) of the form $\dot{\zeta} = |\zeta|^\rho[a_1\zeta + a_2 y]$, with linear output $y = a_3\zeta$ (in which case $k_h = 1$), are of class $\mathcal{P}_\psi$, provided that $a_1 < 0$ and $0 \leq \rho \leq k$.

Let $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ be continuous and positive definite. Let $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ be continuous. We will show that, for $(\mathcal{N}_\phi, \mathcal{P}_\psi)$-universal stabilization, it suffices to replace both occurrences of ϕ in (3.4) with $\phi + \psi$ to yield

$$u(t) \in \nu(\lambda(t))(\phi + \psi)(|y(t)|)\sigma(y(t)),\ \dot{\lambda}(t) = (\phi + \psi)(|y(t)|)|y(t)|,\ \lambda(0) = \lambda^0. \tag{3.15}$$

Let $(b, f, p, P) \in \mathcal{N}_\phi$ and $(g, h, N) \in \mathcal{P}_\psi$. Then there exists constant μ such that $|f(p(t), y)| \leq \mu\phi(|y|)$ for all (t, y). The feedback-controlled system (3.14)–(3.15) can be embedded in a differential inclusion on $\mathbb{R}^{N+2}$:

$$\begin{cases} \dot{x}(t) \in X(x(t)), \quad x(t) = (y(t), \zeta(t), \lambda(t)) \in G := \mathbb{R}^{N+2}, \\ x(0) = x^0 = (y^0, \zeta^0, \lambda^0), \end{cases} \tag{3.16}$$

where $x = (y, \zeta, \lambda) \mapsto X(x) \subset \mathbb{R}^{N+2}$ is given by

$$\begin{aligned} X(x) \equiv X(y, \zeta, \lambda) := \{h(\zeta) + v + bu|\ |v| \leq \mu\phi(|y|),\ u \in \nu(\lambda)(\phi + \psi)(|y|)\sigma(y)\} \\ \times \{g(y, \zeta)\} \times \{(\phi + \psi)(|y|)|y|\}. \end{aligned}$$

X is upper semicontinuous on $\mathbb{R}^{N+2}$ with nonempty, convex, and compact values. Therefore, for each $x^0 \in \mathbb{R}^{N+2}$, the initial-value problem (3.16) has a solution and every solution can be maximally extended.

LEMMA 3.3. *Let $x^0 \in \mathbb{R}^{N+2}$ be arbitrary and let $x(\cdot) = (y(\cdot), \zeta(\cdot), \lambda(\cdot))$ be a maximal solution of* (3.16)*, defined on its maximal interval of existence $[0, \omega)$. Then* (i) *$\omega = \infty$;* (ii) *$\lim_{t\to\infty} \lambda(t)$ exists and is finite;* (iii) *$(y(t), \zeta(t)) \to (0, 0)$ as $t \to \infty$.*

Proof. For almost all $t \in [0, \omega)$, we have

$$\begin{aligned} y(t)\dot{y}(t) &\leq h(\zeta(t))y(t) + \mu\phi(|y(t)|)|y(t)| + b\nu(\lambda(t))(\phi + \psi)(|y(t)|)|y(t)| \\ &\leq h(\zeta(t))y(t) + [\mu + b\nu(\lambda(t))]\dot{\lambda}(t) \end{aligned}$$

which, on integration, yields

$$0 \leq y^2(t) \leq y^2(\tau) + 2\int_\tau^t h(\zeta(s))y(s)ds + 2\mu[\lambda(t) - \lambda(\tau)] + 2b\int_{\lambda(\tau)}^{\lambda(t)} \nu$$

valid for all $t, \tau \in [0, \omega)$, with $t \geq \tau$. Invoking Assumption D, we have

$$0 \leq y^2(t) \leq y^2(\tau) + 2\alpha_0(\|\zeta(\tau)\|) + 2[\alpha_1 + \mu][\lambda(t) - \lambda(\tau)] + 2b\int_{\lambda(\tau)}^{\lambda(t)} \nu. \tag{3.17}$$

By the same contradiction argument as that employed in the proof of Lemma 3.1, we may deduce that $\lambda(\cdot)$ is bounded. Boundedness of $y(\cdot)$ then follows from (3.17). That $\zeta(\cdot)$ is bounded is a consequence of the ISS property of $\Sigma_2 = (g, h, N)$. Therefore, $x(\cdot) = (y(\cdot), \zeta(\cdot), \lambda(\cdot)) \in AC([0, \omega); \mathbb{R}^{N+2})$ is a precompact solution of (3.16), and so, $\omega = \infty$. Defining $l : \mathbb{R}^{N+2} \to \mathbb{R}, \quad x \equiv (y, \zeta, \lambda) \mapsto (\phi + \psi)(|y|)|y|$ (and so $\dot{\lambda} = l \circ x$), we may conclude, by boundedness of $\lambda(\cdot)$, that $l \circ x \in L^1(\mathbb{R}_+)$. Therefore by Theorem 2.10 (with $U = \mathbb{R}^{N+2}$), $x(\cdot)$ approaches the set $\{(y, \zeta, \lambda) |\ y = 0\}$. In particular, $y(t) \to 0$ as $t \to \infty$, and so, by the convergent-input, convergent-state property of the ISS system $\Sigma_2 = (g, h, N)$, we may also conclude that $\zeta(t) \to 0$ as $t \to \infty$. Finally, by boundedness and monotonicity of $\lambda(\cdot)$, $\lim_{t\to\infty} \lambda(t)$ exists and is finite. □

Example. Linear, minimum-phase systems of relative degree one. This class has played a central role in the development of universal adaptive control. In appropriate coordinates, such systems have state space representations of the form

$$\dot{y} = A_1 y + A_2 \zeta + CBu, \quad \dot{\zeta} = A_3 y + A_4 \zeta. \tag{3.18}$$

In the single-input, single-output case, we identify (3.18) and (3.14) with

$$f : (p, y) \mapsto A_1 y, \quad h : \zeta \mapsto A_2 \zeta, \quad b = CB, \quad g : (y, \zeta) \mapsto A_3 y + A_4 \zeta.$$

By the relative-degree-one assumption, $b = CB \neq 0$ and, by the minimum-phase assumption, $\mathrm{spec}(A_4) \subset \mathbb{C}_-$. Defining $\phi \equiv \psi : |y| \mapsto \frac{1}{2}|y|$, we see that every single-input, single-output, linear, minimum-phase system of relative degree one is of class $(\mathcal{N}_\phi, \mathcal{P}_\psi)$ and the control (3.15) reduces to the ubiquitous Byrnes–Willems strategy: $u(t) = \nu(\lambda(t))y(t)$, $\dot{\lambda}(t) = y^2(t)$, $\lambda(0) = \lambda^0$.

Remarks. In considering the case of dynamically perturbed scalar systems, we treated only the problem of adaptive stabilization. The adaptive servomechanism of section 3.1.2 can also be modified to incorporate dynamically perturbed systems when the dynamic perturbations are generated by linear systems $\dot{\zeta} = A\zeta + By$, $w = C\zeta$, $\mathrm{spec}(A) \subset \mathbb{C}_-$, as described in the previous Example. For such perturbations, the (modified) servomechanism ensures convergence to zero of the tracking error, convergence to a finite limit of the adapting parameter, and boundedness of the evolution $t \mapsto \zeta(t)$ of the perturbing system. We omit full details here.

3.2. Planar systems. In all applications of the integral invariance principle in section 3.1 above, the conclusion that $x(t)$ tends, as $t \to \infty$, to the zero level set $l^{-1}(0)$ proved sufficient for our purposes; the additional property that $x(\cdot)$ approaches the largest weakly invariant subset of $l^{-1}(0)$ was redundant. Here, we treat a class of systems for which the latter property can be fruitfully exploited. We consider planar systems (with scalar control u) described by a second-order differential equation:

$$\begin{cases} \ddot{y}(t) = d(t)\dot{y}(t) + f(p(t), y(t)) + bu(t), \quad y(t), u(t), d(t) \in \mathbb{R}, \quad p(t) \in \mathbb{R}^P, \\ (y(t_0), \dot{y}(t_0)) = (y^0, v^0), \end{cases} \tag{3.19}$$

where the parameters $b \in \mathbb{R}$, $P \in \mathbb{N}$, and functions d, f, p are unknown. The variable $y(t)$, but not its derivative $\dot{y}(t)$, is available for feedback. We identify (3.19) with the quintuple (b, d, f, p, P). For $\delta > 0$ and continuous, positive-definite, nondecreasing function $\phi : \mathbb{R}_+ \to \mathbb{R}_+$, we denote, by $\mathcal{N}_{\delta,\phi}$, the set of system quintuples (b, d, f, p, P) for which Assumption A (that is, $b \neq 0$) holds, together with the following three assumptions.

Assumption E. $d \in (C \cap L^\infty)(\mathbb{R})$ and, for some $\epsilon > 0$, $d(t) \leq -\delta - 2\epsilon$ for all t.

Assumption F. $(p, y) \mapsto f(p, y)$, $\mathbb{R}^P \times \mathbb{R} \to \mathbb{R}$ is continuous and is continuously differentiable in its first argument. Both f and $D_1 f$ ($\equiv \partial f/\partial p$) are ϕ-bounded uniformly with respect to p in compact sets; precisely, for every compact $K \subset \mathbb{R}^P$, there exists scalar μ_K such that $|f(p, y)| + \|D_1 f(p, y)\| \leq \mu_K \phi(|y|)$ for all $(p, y) \in K \times \mathbb{R}$.

Assumption G. $p(\cdot) \in W^{1,\infty}(\mathbb{R}; \mathbb{R}^P)$.

By virtue of Assumptions E and G, without loss of generality, $t_0 = 0$ may be assumed in (3.19); this we will do, without further comment, throughout.

In section 3.2.1 below, we will show that $(b, d, f, p, P) \in \mathcal{N}_{\delta,\phi}$, for some known $\delta > 0$ and known continuous positive-definite nondecreasing function ϕ, is sufficient *a priori* information for adaptive stabilizability of (3.19) by feedback of the variable $y(t)$ alone; in essence, Assumption E compensates for the inaccessibility of the velocity variable $\dot{y}(t)$ by requiring that the system should exhibit dissipitive dependence (loosely quantifiable by the known constant δ) on that variable.

Example. As motivation for (3.19), consider a single-degree-of-freedom mechanical system with position, but not velocity, available for feedback and with some constant (but unknown) natural damping d quantified by a known parameter δ in the sense that $d < -\delta < 0$. If we suppose that Assumption F holds with $\phi : |y| \mapsto 1 + |y|^3$, then, for example, the following particular realizations are admissible.

Nonlinear pendulum with disturbed support (disturbance $p(\cdot) \in W^{1,\infty}(\mathbb{R})$):

$$\ddot{y}(t) = d\dot{y}(t) + (a + p(t)) \sin(y(t)) + bu(t), \quad a, b \in \mathbb{R},\ b \neq 0.$$

Duffing equation with extraneous disturbance ($p(\cdot) \in W^{1,\infty}(\mathbb{R})$):

$$\ddot{y}(t) = d\dot{y}(t) + a_1 y(t) + a_2 y^3(t) + p(t) + bu(t), \quad a_1, a_2, b \in \mathbb{R},\ b \neq 0.$$

In the absence of control, such systems are potentially "chaotic."

3.2.1. Adaptive stabilizer. Let $\delta > 0$ and let $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ be a continuous, positive-definite, nondecreasing function. Assuming only that δ, ϕ and the instantaneous state $y(t)$ are available for control purposes, our goal is to demonstrate the existence of an adaptive feedback strategy that provides $\mathcal{N}_{\delta,\phi}$-universal stabilization in the sense that it assures that the state of (3.19) approaches $\{0\}$ for all quintuples $(b, d, f, p, P) \in \mathcal{N}_{\delta,\phi}$ while maintaining boundedness of the controller function $\lambda(\cdot)$.

Define the continuous function $\gamma : \mathbb{R}_+ \to \mathbb{R}_+$ by $\gamma(\xi) := \xi + \phi(\xi)$ and let Γ denote its indefinite integral: $\Gamma : \mathbb{R}_+ \to \mathbb{R}_+$, $\xi \mapsto \int_0^\xi \gamma$. We claim that the following (formal) strategy is a $\mathcal{N}_{\delta,\phi}$-universal stabilizer:

$$\begin{cases} u(t) = \nu(\eta(t))\gamma(|y(t)|)\mathrm{sgn}(y(t)), \quad \eta(t) := \delta\lambda(t) + \Gamma(|y(t)|), \\ \dot{\lambda}(t) = \gamma(|y(t)|)|y(t)|, \quad \lambda(0) = \lambda^0 \in \mathbb{R}, \end{cases} \tag{3.20}$$

where ν is any continuous function with properties (3.3).

Let $(b, d, f, p, P) \in \mathcal{N}_{\delta,\phi}$. Introducing the coordinate transformation $z(t) = \dot{y}(t) + \delta y(t)$, we may express (3.19) in the form

$$\begin{cases} \dot{y}(t) = -\delta y(t) + z(t), \\ \dot{z}(t) = (\delta + d(t))(z(t) - \delta y(t)) + f(p(t), y(t)) + bu(t), \\ (y(0), z(0)) = (y^0, z^0). \end{cases} \tag{3.21}$$

The feedback (3.20) is interpreted in the set-valued sense:

$$\begin{cases} u(t) \in \nu(\eta(t))\gamma(|y(t)|)\sigma(y(t)), \quad \eta(t) = \delta\lambda(t) + \Gamma(|y(t)|), \\ \dot{\lambda}(t) = \gamma(|y(t)|)|y(t)|, \quad \lambda(0) = \lambda^0, \end{cases} \tag{3.22}$$

with the map $y \mapsto \sigma(y)$ defined as before in (3.5).

Writing $x(t) = (y(t), z(t), \lambda(t))$, the overall adaptive feedback controlled system may be embedded in the following differential inclusion on $\mathbb{R}^3$:

$$\dot{x}(t) \in X(t, x(t)), \quad x(0) = x^0 = (y^0, z^0, \lambda^0), \tag{3.23}$$

where $X : (t, x) \equiv (t, y, z, \lambda) \mapsto \{-\delta y + z\} \times X_2(t, x) \times \{\gamma(|y|)|y|\} \subset \mathbb{R}^3$, with

$$X_2(t, x) := \{(\delta + d(t))(z - \delta y) + f(p(t), y) + bu |\ u \in \nu(\delta\lambda + \Gamma(|y|))\gamma(|y|)\sigma(y)\}.$$

X is upper semicontinuous on $\mathbb{R}\times\mathbb{R}^3$ and takes nonempty, convex, and compact values in $\mathbb{R}^3$. Therefore, for each $x^0 \in \mathbb{R}^3$, the initial-value problem (3.23) has a solution and every solution can be extended into a maximal solution.

LEMMA 3.4. *Let $x^0 \in \mathbb{R}^3$ be arbitrary and let $x(\cdot) = (y(\cdot), z(\cdot), \lambda(\cdot))$ be a maximal solution of* (3.23) *defined on its maximal interval of existence $[0, \omega)$. Then* (i) *$\omega = \infty$;* (ii) *$\lim_{t\to\infty} \lambda(t)$ exists and is finite;* (iii) *$(y(t), z(t)) \to (0, 0)$ as $t \to \infty$.*

Proof. On $\mathbb{R}$, define the locally Lipschitz function $\Phi : r \mapsto \Gamma(|r|)$, with directional derivative at r in direction s given by

$$\Phi^\dagger(r; s) = \lim_{h\downarrow 0} \frac{\Phi(r + hs) - \Phi(r)}{h} = \begin{cases} \gamma(|r|)\mathrm{sgn}(r)s, & r \neq 0, \\ \gamma(0)|s|, & r = 0. \end{cases} \tag{3.24}$$

Let $F_1 \in AC([0, \omega); \mathbb{R})$ denote the composition $\Phi \circ y$ and let $\mathcal{O}_1 \subset [0, \omega)$ be the set (of measure zero) of points t at which the derivative $\dot{F}_1(t)$ fails to exist. A straightforward argument (analogous to that yielding (2.4)) gives

$$\dot{F}_1(t) = \Phi^\dagger(y(t); \dot{y}(t)) \quad \forall\ t \in [0, \omega)\backslash\mathcal{O}_1. \tag{3.25}$$

By properties of f and p (Assumptions F and G), the function

$$F_2 : t \mapsto \int_0^{y(t)} f(p(t), \xi)\, d\xi$$

is of class $AC([0, \omega); \mathbb{R})$. Let $\mathcal{O}_2$ denote the set (of measure zero) of points t at which $\dot{p}(t)$ fails to exist. Again by properties of f and p, there exists $\mu > 0$ such that

$$\begin{aligned} \dot{F}_2(t) = \int_0^{y(t)} \langle D_1 f(p(t), \xi), \dot{p}(t)\rangle\, d\xi + f(p(t), y(t))\dot{y}(t) \\ \geq -\mu\phi(|y(t)|)|y(t)| + f(p(t), y(t))z(t) \quad \forall\ t \in [0, \omega)\backslash\mathcal{O}_2. \end{aligned} \tag{3.26}$$

Define $F_c \in AC([0, \omega); \mathbb{R})$, parameterized by $c > 0$, as $F_c : t \mapsto cF_1(t) - F_2(t)$. By Assumptions F and G and the definition of Γ, F_c is such that, for all c sufficiently large,

$$F_c(t) \geq \frac{1}{2} c y^2(t) \quad \forall\ t. \tag{3.27}$$

Moreover, by (3.25) and (3.26),

$$\begin{aligned}(3.28)\quad \dot{F}_c(t) &= c\dot{F}_1(t) - \dot{F}_2(t) = c[\dot{\eta}(t) - \delta\gamma(|y(t)|)|y(t)|] - \dot{F}_2(t)\\ &\le c\dot{\eta}(t) + (\mu - c\delta)\gamma(|y(t)|)|y(t)| - z(t)f(p(t), y(t)) \quad \forall\ t \in [0,\omega)\backslash(\mathcal{O}_1 \cup \mathcal{O}_2).\end{aligned}$$

Let $Y := \{t \in [0,\omega) |\ (y(t), \dot{y}(t)) \neq (0,0)\}$ be the set of points t at which $y(t)$ and $\dot{y}(t)$ are not both zero. Observe that (i) every point t of the subset $\mathcal{O}_0 := \{t \in Y |\ y(t) = 0\}$ is an isolated point implying that $\mathcal{O}_0$ is countable and so has measure zero, and (ii) $y(t) = \dot{y}(t) = z(t) = 0$ for all $t \in [0,\omega)\backslash Y$. From these observations, together with (3.24), (3.25), and writing $\mathcal{O} := \mathcal{O}_0 \cup \mathcal{O}_1 \cup \mathcal{O}_2$ (of measure zero), we deduce that

$$\begin{aligned}(3.29)\qquad \forall\ t \in [0,\omega)\backslash\mathcal{O}, \quad uz(t) &= \nu(\eta(t))[\delta\gamma(|y(t)|)|y(t)| + \Phi^{\dagger}(y(t); \dot{y}(t))]\\ &= \nu(\eta(t))\dot{\eta}(t) \quad \forall\ u \in \nu(\eta(t))\gamma(|y(t)|)\sigma(y(t)).\end{aligned}$$

Define $V_c \in AC([0,\omega);\mathbb{R})$ by $V_c(t) := F_c(t) + \frac{1}{2}z^2(t)$. Invoking (3.28), (3.29), and Assumption E,

$$\begin{aligned}(3.30)\qquad \dot{V}_c(t) &\le \dot{F}_c(t) - 2\epsilon z^2(t) - (d(t)+\delta)\delta y(t)z(t)\\ &\qquad + z(t)f(p(t), y(t)) + b\nu(\eta(t))\dot{\eta}(t)\\ &\le [\Delta^2/(4\epsilon) + \mu - c\delta]\gamma(|y(t)|)|y(t)| - \epsilon z^2(t) + [c + b\nu(\eta(t))]\dot{\eta}(t)\end{aligned}$$

for almost all $t \in [0,\omega)$, wherein we have used the fact that

$$-(d(t)+\delta)\delta y(t)z(t) \le \Delta|y(t)z(t)| \le \epsilon z^2(t) + \frac{\Delta^2}{4\epsilon}y^2(t) \le \epsilon z^2(t) + \frac{\Delta^2}{4\epsilon}\gamma(|y(t)|)|y(t)|,$$

with $\Delta := (\|d(\cdot)\|_\infty + \delta)\delta$. Now fix c sufficiently large so that $\Delta^2(4\epsilon)^{-1} + \mu - c\delta \le 0$ and (3.27) holds, in which case, $V_c(t) \ge \frac{1}{2}\left[cy^2(t) + z^2\big(t\big)\right]$ for all t and

$$(3.31)\qquad \dot{V}_c(t) \le [c + b\nu(\eta(t))]\dot{\eta}(t) \quad \text{for a.a. } t \in [0,\omega),$$

which, on integration, yields

$$(3.32)\qquad 0 \le \frac{1}{2}\left[cy^2(t) + z^2(t)\right] \le V_c(t) \le V_c(\tau) + c[\eta(t) - \eta(\tau)] + b\int_{\eta(\tau)}^{\eta(t)} \nu$$

for all $t, \tau \in [0,\omega)$, with $t \ge \tau$.

We first show that the function $\eta(\cdot)$ (and hence $\lambda(\cdot)$) is bounded. By properties (3.3) of ν, there exist increasing sequences $(\hat{\eta}_n)_{n\in\mathbb{N}}$ and $(\tilde{\eta}_n)_{n\in\mathbb{N}}$, with $\hat{\eta}_n \to \infty$ and $\tilde{\eta}_n \to \infty$ as $n \to \infty$, such that

$$(3.33)\qquad \text{(a)}\quad \frac{1}{\hat{\eta}_n}\int_{\hat{\eta}_1}^{\hat{\eta}_n} \nu(\theta)d\theta\ \to +\infty, \qquad \text{(b)}\quad \frac{1}{\tilde{\eta}_n}\int_{\tilde{\eta}_1}^{\tilde{\eta}_n} \nu(\theta)d\theta\ \to -\infty$$

as $n \to \infty$. Without loss of generality, we may assume $\hat{\eta}_1, \tilde{\eta}_1 \ge 1$. Seeking a contradiction, suppose that $\eta(\cdot)$ is unbounded. Now, $\eta(\cdot)$ is bounded from below (in particular, $\eta(t) \ge \delta\lambda^0$ for all $t \ge 0$), and so, by the supposition, $\eta(\cdot)$ is unbounded from above. Therefore, there exist increasing sequences $(\hat{t}_n), (\tilde{t}_n) \subset [0,\omega)$ such that, for all n,

$$\eta(\hat{t}_n) = \hat{\eta}_n \quad \text{and} \quad \eta(\tilde{t}_n) = \tilde{\eta}_n.$$

Now, either $b > 0$ or $b < 0$. If $b > 0$, then (3.32) and (3.33b) combine to yield the contradiction

$$0 \leq \text{constant} + \frac{b}{\tilde{\eta}_n} \int_{\tilde{\eta}_1}^{\tilde{\eta}_n} \nu \;\rightarrow\; -\infty \quad \text{as } n \rightarrow \infty.$$

If $b < 0$, then (3.32) and (3.33a) combine to yield the contradiction

$$0 \leq \text{constant} - \frac{|b|}{\hat{\eta}_n} \int_{\hat{\eta}_1}^{\hat{\eta}_n} \nu \;\rightarrow\; -\infty \quad \text{as } n \rightarrow \infty.$$

Therefore $\eta(\cdot)$ (and hence, $\lambda(\cdot)$) is bounded. Boundedness of $\eta(\cdot)$ and $\lambda(\cdot)$, together with (3.32), imply boundedness of $y(\cdot)$ and $z(\cdot)$. This establishes assertion (i) and assertion (ii) follows by monotonicity of $\lambda(\cdot)$. It remains to prove assertion (iii).

By boundedness of $d(\cdot)$, $p(\cdot)$, and $x(\cdot)$, there exists $\rho > 0$ such that $X_2(t, x(t)) \subset \rho\bar{\mathbb{B}}$ for all $t \in [0, \infty)$, and so, $x(\cdot) = (y(\cdot), z(\cdot), \lambda(\cdot))$ is a precompact solution of the autonomous initial-value problem

$$\dot{x}(t) \in \{-\delta y(t) + z(t)\} \times \rho\bar{\mathbb{B}} \times \{\gamma(|y(t)|)|y(t)|\}, \quad x(0) = x^0. \tag{3.34}$$

Moreover, by boundedness of $\lambda(\cdot)$,

$$\int_0^\infty y^2(s)ds \leq \int_0^\infty \dot{\lambda}(s)ds < \infty.$$

Therefore, by Theorem 2.10, $x(\cdot)$ approaches the largest weakly-invariant (relative to the autonomous differential inclusion (3.34)) set W in $\{(y, z, \lambda)|\ y = 0\}$. Let $\bar{w} = (0, \bar{z}, \bar{\lambda}) \in W$. By definition of weak invariance, the initial-value problem

$$\dot{w}(t) = (\dot{w}_1(t), \dot{w}_2(t), \dot{w}_3(t)) \in \{-\delta w_1(t) + w_2(t)\} \times \rho\bar{\mathbb{B}} \times \{\gamma(|w_1(t)|)|w_1(t)|\}, \quad w(0) = \bar{w}$$

has at least one solution $w(\cdot) = (w_1(\cdot), w_2(\cdot), w_3(\cdot))$ with maximal interval of existence $\mathbb{R}_+$ and with trajectory in $W \subset \{(y, z, \lambda)|\ y = 0\}$. Since $\dot{w}_1(t) = -\delta w_1(t) + w_2(t)$ and $w_1(\cdot) \equiv 0$ in W, it follows that $w_2(\cdot) \equiv 0$, and so, $\bar{z} = w_2(0) = 0$. Therefore, we conclude that the largest weakly-invariant set in W is contained in the set $\{(y, z, \lambda)|\ y = 0 = z\}$, and so the solution $x(\cdot)$ approaches the set $\{(0, 0)\} \times \mathbb{R}$. In particular, $(y(t), z(t)) \rightarrow (0, 0)$ as $t \rightarrow \infty$. □

3.2.2. Adaptive servomechanism. We now turn attention to the servomechanism problem for planar systems (3.21), that is, the construction of controls that cause the system to track, asymptotically, any reference signal $r(\cdot)$ of some given class, in the sense that both the tracking error $e(t) = y(t) - r(t)$ and its derivative $\dot{e}(t) = \dot{y}(t) - \dot{r}(t)$ tend to zero as $t \rightarrow \infty$. For the class of reference signals we take the (Sobolev) space $\mathcal{R} = W^{3,\infty}(\mathbb{R})$ of functions $r \in \left(C^2 \cap L^\infty\right)(\mathbb{R})$ with $\dot{r} \in \left(C^1 \cap L^\infty\right)(\mathbb{R})$ and $\ddot{r} \in W^{1,\infty}(\mathbb{R})$, equipped with the norm

$$\|r\|_{3,\infty} = \|r\|_\infty + \|\dot{r}\|_\infty + \|\ddot{r}\|_\infty + \|\dddot{r}\|_\infty.$$

For the servomechanism problem, we restrict the underlying class of systems by imposing a stronger assumption on the function f. Assumption F* below should hold for some known, continuous, positive-definite, nondecreasing function ϕ having the additional property (3.9). Specifically, for real $\delta > 0$ and continuous, positive-definite, nondecreasing function $\phi : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ with property (3.9), we denote, by $\mathcal{N}^*_{\delta,\phi}$, the set of quintuples (b, d, f, p, P) satisfying Assumptions A, E, and F*.

Assumption F*. $(p,y) \mapsto f(p,y)$, $\mathbb{R}^P \times \mathbb{R} \to \mathbb{R}$ is continuously differentiable. Both f and its gradient function $Df = (D_1 f, D_2 f)$ $(\equiv (\partial f/\partial p, \partial f/\partial y))$ are ϕ-bounded uniformly with respect to p in compact sets; precisely, for every compact $K \subset \mathbb{R}^P$, there exists scalar μ_K such that $|f(p,y)| + \|Df(p,y)\| \leq \mu_K \phi(|y|)$ for all $(p,y) \in K \times \mathbb{R}$.

Example. The function $\phi : |y| \mapsto 1 + |y|^3$ has property (3.9) and the mechanical systems described in the previous Example are admissible.

Let $\delta > 0$ and let ϕ be a continuous, positive-definite, nondecreasing function with property (3.9). We claim that, in order to assure convergence to zero of both the tracking error $e(t) = y(t) - r(t)$ and its derivative $\dot{e}(t)$ for all reference signals $r \in \mathcal{R}$ and all quintuples $(b,d,f,p,P) \in \mathcal{N}^*_{\delta,\phi}$, it suffices to replace every occurrence of $y(t)$ in (3.20) by $e(t)$. Proof of this claim follows.

Let $(b,d,f,p,P) \in \mathcal{N}^*_{\delta,\phi}$. Write $\tilde{P} = P + 3$ and define the continuous function

$$\tilde{f} : \mathbb{R}^{\tilde{P}} \times \mathbb{R} \to \mathbb{R}, \quad (\tilde{p}, e) \equiv (p, r, v, w, e) \mapsto f(p, e + r) - \delta v - w.$$

By properties of f, $\tilde{f}$ is continuously differentiable with respect to its first argument $(\tilde{p})$, with $D_1 \tilde{f}$ given by

$$D_1 \tilde{f}(\tilde{p}, e) \equiv (\partial \tilde{f}/\partial \tilde{p})(\tilde{p}, e) = (D_1 f(p, e + r), D_2 f(p, e + r), -\delta, -1).$$

Let $\tilde{K} \subset \mathbb{R}^{\tilde{P}}$ be compact, and so there exist compact $K \subset \mathbb{R}^P$ and $R > 0$ such that $\tilde{K} \subset K \times [-R, R]^3$. By properties of f and ϕ, there exist constants μ_K and ρ_R such that, for all $(p, r, v, w) \equiv \tilde{p} \in \tilde{K} \subset K \times [-R, R]^3$,

$$\begin{aligned} |\tilde{f}(\tilde{p}, e)| + \|D_1 \tilde{f}(\tilde{p}, e)\| &\leq |f(p, e + r)| + \delta|v| + |w| + \|Df(p, e + r)\| + \delta + 1 \\ &\leq \mu_K \phi(|e + r|) + (1 + \delta)(1 + R) \\ &\leq \mu_K \rho_R \phi(|e|) + (1 + \delta)(1 + R) \leq \tilde{\mu}_{\tilde{K}} \phi(|e|), \end{aligned}$$

with $\tilde{\mu}_{\tilde{K}} := \mu_K \rho_R + ((1 + \delta)(1 + R))/\phi(0)$. Therefore, $\tilde{f}$ satisfies Assumption F.

Let $r \in \mathcal{R} \equiv W^{3,\infty}(\mathbb{R})$. Then $t \mapsto \tilde{p}(t) := (p(t), r(t), \dot{r}(t), \ddot{r}(t)) \in \mathbb{R}^{\tilde{P}}$ is of class $W^{1,\infty}(\mathbb{R}; \mathbb{R}^{\tilde{P}})$, and so $(b, d, \tilde{f}, \tilde{p}, \tilde{P}) \in \mathcal{N}_{\delta,\phi}$. Expressed in terms of the tracking error $e(t) = y(t) - r(t)$ and adopting the coordinate transformation $z(t) = \dot{e}(t) + \delta e(t)$, the underlying dynamics have the form

$$\text{(3.35)} \qquad \begin{cases} \dot{e}(t) = -\delta e(t) + z(t), \\ \dot{z}(t) = (\delta + d(t))(z(t) - \delta e(t)) + \tilde{f}(\tilde{p}(t), e(t)) + bu(t), \\ (e(0), z(0)) = (e^0, z^0). \end{cases}$$

We are now in precisely the same context, modulo notation, as in the case of an adaptive stabilizer, and so, replacing all occurrences of $y(t)$ in (3.20) by $e(t)$, viz.

$$\text{(3.36)} \qquad \begin{cases} u(t) \in \nu(\eta(t))\gamma(|e(t)|)\sigma(e(t)), \quad \eta(t) = \delta\lambda(t) + \Gamma(|e(t)|), \\ \dot{\lambda}(t) = \gamma(|e(t)|)|e(t)|, \ \lambda(0) = \lambda^0, \end{cases}$$

then the same argument, as used to establish Lemma 3.4, applies *mutatis mutandis* to conclude that (3.36) is an $(\mathcal{R}, \mathcal{N}^*_{\delta,\phi})$-universal servomechanism; for each $r(\cdot) \in \mathcal{R}$ and $(b,d,f,p,P) \in \mathcal{N}^*_{\delta,\phi}$, every solution $(e(\cdot), z(\cdot), \lambda(\cdot))$ of the feedback controlled system has maximal interval of existence $\mathbb{R}_+$ with $(e(t), z(t)) \to (0,0)$ as $t \to \infty$ and, moreover, $\lim_{t\to\infty} \lambda(t)$ exists and is finite.

3.2.3. Dynamically perturbed planar systems. Let $\delta > 0$ and let the function $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ be continuous, positive definite, and nondecreasing. Let $\Sigma_1 = (b, d, f, p, P) \in \mathcal{N}_{\delta,\phi}$. Here we consider the case where Σ_1 is subject to perturbations generated through its interconnection with a dynamical system Σ_2 (as depicted in Figure 3.1).

The system Σ_2 is assumed to correspond to a differential equation (driven by the variable $y(t)$ of system Σ_1) on $\mathbb{R}^N$ of the form (3.13) with input $y(t)$, and scalar output $w(t)$ perturbing Σ_1. As before, we identify the system Σ_2 with the triple (g, h, N). Writing $z(t) = \dot{y}(t) + \delta y(t)$, the overall system has representation (on $\mathbb{R} \times \mathbb{R} \times \mathbb{R}^N$)

$$\text{(3.37)} \quad \begin{cases} \dot{y}(t) = -\delta y(t) + z(t), \\ \dot{z}(t) = (\delta + d(t))z(t) - (\delta + d(t))\delta y(t) + f(p(t), y(t)) + h(\zeta(t)) + bu(t), \\ \dot{\zeta}(t) = g(y(t), \zeta(t)), \quad (y(0), z(0), \zeta(0)) = (y^0, z^0, \zeta^0). \end{cases}$$

We will define, via Assumption H below, a class $\mathcal{P}_\psi$ of admissible systems $\Sigma_2 = (g, h, N)$, such that the $\mathcal{N}_{\delta,\phi}$-universal stabilizer of section 3.2.1 is readily modified to yield a $(\mathcal{N}_{\delta,\phi}, \mathcal{P}_\psi)$-universal stabilizer.

For continuous $\psi : \mathbb{R}_+ \to \mathbb{R}_+$, we denote, by $\mathcal{P}_\psi$ the set of system triples $\Sigma_2 = (g, h, N)$ satisfying the following.

Assumption H. (i) $g : \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}^N$ is locally Lipschitz; (ii) system (3.13) is ISS; (iii) $h : \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}$ is continuous; (iv) there exist a function $\alpha_0 : \mathbb{R}_+ \to \mathbb{R}_+$ and a scalar $\alpha_1 > 0$, such that, for each $(\zeta^0, y(\cdot)) \in \mathbb{R}^N \times L^\infty_{\text{loc}}(\mathbb{R}_+)$, the (unique) solution $\zeta(\cdot)$ of (3.13) satisfies

$$\int_0^t h^2(\zeta(s))ds \le \alpha_0(\|\zeta^0\|) + \alpha_1 \int_0^t \psi(|y(s)|)|y(s)|ds.$$

Examples. (a) Assumption H holds for the class of linear systems considered in the first Example of section 3.1.4.

(b) More generally, assume $g : \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}^N$ is locally Lipschitz and $h : \mathbb{R}^N \to \mathbb{R}$ is continuous. Assume further that h is positively homogeneous of degree $k \ge 1$ and that g is positively homogeneous of degree k_g, $1 \le k_g \le 2k$. Let $\psi : |y| \mapsto |y|^{2k-1}$. If $\{0\}$ is an asymptotically stable equilibrium of the unforced system $\dot{\zeta}(t) = g(0, \zeta(t))$, then it can be shown (by an argument invoking [22, Theorem 2]) that $(g, h, N) \in \mathcal{P}_\psi$. For example, if $\psi : |y| \mapsto |y|$, then systems (with $N = 1$ and with unknown real parameters a_i) of the form $\dot{\zeta} = a_1\zeta|\zeta| + a_2y^2$, $w = a_3\zeta$, are of class $\mathcal{P}_\psi$, provided that $a_1 < 0$.

Let $\delta > 0$ and let $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ be continuous, positive definite, and nondecreasing. As before, let $\gamma : \xi \mapsto \xi + \phi(\xi)$ and $\Gamma : \xi \mapsto \int_0^\xi \gamma$. Let $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ be continuous with indefinite integral $\Psi : \mathbb{R}_+ \to \mathbb{R}_+$, $\xi \mapsto \int_0^\xi \psi$.

We will show that, for $(\mathcal{N}_{\delta,\phi}, \mathcal{P}_\psi)$-universal stabilization, it suffices to replace both occurrences of γ in (3.20) by $\gamma + \psi$ and to replace the single occurrence of Γ with $\Gamma + \Psi$ to yield

$$\text{(3.38)} \quad \begin{cases} u(t) \in \nu(\eta(t))(\gamma + \psi)(|y(t)|)\sigma(y(t)), \quad \eta(t) = \delta\lambda(t) + (\Gamma + \Psi)(|y(t)|), \\ \dot{\lambda}(t) = (\gamma + \psi)(|y(t)|)|y(t)|, \ \lambda(0) = \lambda^0. \end{cases}$$

Let $(b, d, f, p, P) \in \mathcal{N}_{\delta,\phi}$ and $(g, h, N) \in \mathcal{P}_\psi$. The feedback-controlled system (3.37)–

(3.38) can be embedded in a differential inclusion on $\mathbb{R}^{N+3}$:

$$\text{(3.39)}\quad \begin{cases} \dot{x}(t) \in X(t, x(t)), \\ x(t) = (y(t), z(t), \zeta(t), \lambda(t)) \in G := \mathbb{R}^{N+3}, \quad x(0) = x^0 = (y^0, z^0, \zeta^0, \lambda^0), \end{cases}$$

where the set-valued map $(t, x) \equiv (t, y, z, \zeta, \lambda) \mapsto X(t, x) \subset \mathbb{R}^{N+3}$ is given by

$$X(t, x) = \{-\delta y + z\} \times X_2(t, x) \times \{g(y, \zeta)\} \times \{(\gamma + \psi)(|y|)|y|\},$$

$$X_2(t, x) := \{(\delta + d(t))z - (\delta + d(t))\delta y + f(p(t), y) + h(\zeta) + bu| \\ u \in \nu(\delta\lambda + (\Gamma + \Psi)(|y|)\sigma(y)\}.$$

X is upper semicontinuous on $\mathbb{R} \times \mathbb{R}^{N+3}$ and takes nonempty, convex, and compact values in $\mathbb{R}^{N+3}$. Therefore, for each $x^0 \in \mathbb{R}^{N+3}$, the initial-value problem (3.39) has a solution, and every solution can be maximally extended.

LEMMA 3.5. *Let $x^0 \in \mathbb{R}^{N+3}$ be arbitrary and let $x(\cdot) = (y(\cdot), z(\cdot), \zeta(\cdot), \lambda(\cdot))$ be a maximal solution of* (3.39) *defined on its maximal interval of existence $[0, \omega)$. Then* (i) *$\omega = \infty$;* (ii) *$\lim_{t\to\infty} \lambda(t)$ exists and is finite;* (iii) *$(y(t), z(t), \zeta(t)) \to (0, 0, 0)$ as $t \to \infty$.*

Proof. Let F_c and V_c, parameterized by $c > 0$, be defined as in the proof of Lemma 3.4. By an argument essentially the same as that adopted in the proof of Lemma 3.4 and choosing c sufficiently large, we arrive at a counterpart to (3.30):

$$\text{(3.40)}\qquad \dot{V}_c(t, y(t), z(t)) \le -\epsilon z^2(t) + |h(\zeta(t))||z(t)| + [c + b\nu(\eta(t))]\dot{\eta}(t)$$

for almost all $t \in [0, \omega)$. Invoking the inequality $|h(\zeta)||z| \le \frac{1}{2}\epsilon z^2 + \frac{1}{2}\epsilon^{-1}h^2(\zeta)$, then integrating and invoking Assumption H, we have (for c sufficiently large)

$$\text{(3.41)}\quad \frac{1}{2}[cy^2(t) + z^2(t)] \le V_c(t, y(t), z(t)) \le V_c(\tau, y(\tau), z(\tau)) \\ + \frac{1}{2}\epsilon^{-1}\alpha_0(\|\zeta(\tau)\|) + \frac{1}{2}\epsilon^{-1}\alpha_1[\lambda(t) - \lambda(\tau)] \\ + c[\eta(t) - \eta(\tau)] + b\int_{\eta(\tau)}^{\eta(t)} \nu$$

for all $t, \tau \in [0, \omega)$, with $t \ge \tau$. A straightforward modification of the contradiction argument used previously in the proof of Lemma 3.4 establishes boundedness of $\eta(\cdot)$ (and hence, of $\lambda(\cdot)$). Boundedness of $\eta(\cdot)$ and $\lambda(\cdot)$, together with (3.41), imply boundedness of $y(\cdot)$ and $z(\cdot)$. That $\zeta(\cdot)$ is bounded is a consequence of the ISS property of $\Sigma_2 = (g, h, N)$. This establishes assertion (i) and assertion (ii) follows by monotonicity of $\lambda(\cdot)$. It remains to prove assertion (iii). With minor modification, the argument used in the proof of Lemma 3.4 applies to conclude that $x(\cdot)$ approaches the set $\{(y, z, \zeta, \lambda)|\ y = 0 = z\}$. In particular, $(y(t), z(t)) \to (0, 0)$ as $t \to \infty$, and so, by the convergent-input, convergent-state property of the ISS system $\Sigma_2 = (g, h, N)$, we may also conclude that $\zeta(t) \to 0$ as $t \to \infty$. □

Example. Linear minimum-phase systems of relative degree two. Let (A, B, C) define a linear, single-input u, single-output y minimum-phase system on $\mathbb{R}^{N+2}$ of relative degree two. Denoting its Markov parameters by $m_k := CA^{k-1}B$, then $m_1 = 0$, $m_2 \ne 0$, and the system has a representation (on $\mathbb{R}^2 \times \mathbb{R}^N$) of the form

$$\text{(3.42)}\qquad \ddot{y} = A_1 y + (m_3/m_2)\dot{y} + A_2\zeta + m_2 u, \quad \dot{\zeta} = A_3 y + A_4\zeta.$$

If we assume that $-m_3/m_2 > \delta > 0$ (that is, the system exhibits natural damping quantified by known δ), then we may identify (3.42) and (3.37) by setting

$$f : (p, y) \mapsto A_1 y,\ g : (y, \zeta) \mapsto A_3 y + A_4 \zeta,\ h : \zeta \mapsto A_2 \zeta, \quad b = m_2,\ d(\cdot) \equiv m_3/m_2.$$

By the relative-degree-two assumption, $b = m_2 \neq 0$ and, by the minimum-phase assumption, $\mathrm{spec}(A_4) \subset \mathbb{C}_-$. Defining $\phi \equiv \psi : |y| \mapsto \frac{1}{2}|y|$, we see that every relative-degree-two, minimum-phase system with $m_3/m_2 < -\delta$ is of class $(\mathcal{N}_{\delta,\phi}, \mathcal{P}_\psi)$ and we recover (modulo notation) the adaptive stabilizer proposed previously in [5]:

$$u(t) = \nu(\eta(t))y(t), \quad \eta(t) = \delta\lambda(t) + y^2(t), \quad \dot{\lambda}(t) = 2y^2(t),\ \lambda(0) = \lambda^0.$$

Remarks. We conclude with some observations on the servomechanism problem for dynamically perturbed planar systems. Akin to the Remarks in section 3.1.4, the adaptive servomechanism of section 3.2.2 can also be modified to incorporate dynamically perturbed systems, when the dynamic perturbations are generated by linear systems $\dot{\zeta} = A\zeta + By$, $w = C\zeta$, $\mathrm{spec}(A) \subset \mathbb{C}_-$. For such perturbations, the (modified) servomechanism assures convergence to zero of the tracking error and its derivative, convergence to a finite limit of the adapting parameter, and boundedness of the evolution $t \mapsto \zeta(t)$ of the perturbing system. For brevity, we omit full details here.

REFERENCES

[1] J. P. Aubin and A. Cellina, *Differential Inclusions*, Springer-Verlag, New York, 1984.

[2] C. I. Byrnes and C. F. Martin, *An integral-invariance principle for nonlinear systems*, IEEE Trans. Automat. Control, 40 (1995), pp. 983–994.

[3] C. I. Byrnes and J. C. Willems, *Adaptive stabilization of multivariable linear systems*, in Proc. 23rd IEEE Conference on Decision and Control, IEEE, New York, 1984, pp. 1574–1577.

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

[5] M. Corless and E. P. Ryan, *Adaptive control of a class of nonlinearly perturbed linear systems of relative degree two*, Systems Control Lett., 21 (1993), pp. 59–64.

[6] K. Deimling, *Nonlinear Functional Analysis*, Springer-Verlag, New York, 1983.

[7] K. Deimling, *Multivalued Differential Equations*, Walter de Gruyter, New York, 1992.

[8] A. F. Filippov, *Differential Equations with Discontinuous Righthand Sides*, Kluwer, Dordrecht, 1988.

[9] A. Ilchmann, *Non-Identifier-Based High-Gain Adaptive Control*, Springer-Verlag, New York, 1993.

[10] A. Ilchmann and E. P. Ryan, *Universal λ-tracking for nonlinearly-perturbed systems in the presence of noise*, Automatica, 30 (1994), pp. 337-346.

[11] J. P. LaSalle, *The Stability of Dynamical Systems*, SIAM Regional Conference Series in Applied Mathematics, SIAM, Philadelphia, 1976.

[12] P. D. Loewen, *Optimal Control via Nonsmooth Analysis*, American Mathematical Society, Providence, RI, 1993.

[13] B. Mårtensson, *Adaptive Stabilization*, Ph.D. thesis, Lund Institute of Technology, Sweden, 1986.

[14] B. Mårtensson, *Adaptive stabilization of multivariable linear systems*, Contemp. Math., 68 (1987), pp. 191–225.

[15] B. Mårtensson, *The order of a stabilizing regulator is sufficient* a priori *information for adaptive stabilization*, Systems Control Lett., 6 (1985), pp. 87–91.

[16] A. S. Morse, *New directions in parameter adaptive control*, in Proc. 23rd IEEE Conference on Decision and Control, IEEE, New York, 1984, pp. 1566–1568.

[17] R. D. Nussbaum, *Some remarks on a conjecture in parameter adaptive control*, Systems Control Lett., 3 (1983), pp. 243–246.

[18] L. Praly, G. Bastin, J.-B. Pomet, and Z.-P. Jiang, *Adaptive stabilization of nonlinear systems*, in Foundations of Adaptive Control, P. V. Kokotović, ed., Springer-Verlag, New York, 1991, pp. 347–433.

[19] E. P. Ryan, *Discontinuous feedback and universal adaptive stabilization*, in Control of Uncertain Systems, D. Hinrichsen and B. Mårtensson, eds., Birkhäuser, Boston, 1990, pp. 245–258.

[20] E. P. Ryan, *Universal $W^{1,\infty}$-tracking for a class of nonlinear systems*, Systems Control Lett., 18 (1992), pp. 201–210.

[21] E. P. Ryan, *A nonlinear universal servomechanism*, IEEE Trans. Automat. Control, 39 (1994), pp. 753–761.

[22] E. P. Ryan, *Universal stabilization of a class of nonlinear systems with homogeneous vector fields*, Systems Control Lett., 26 (1995), pp. 177–184.

[23] E. P. Ryan, *Adaptive stabilization of multi-input nonlinear systems*, Internat. J. Robust Nonlinear Control, 3 (1993), pp. 169–181.

[24] E. D. Sontag, *Smooth stabilization implies coprime factorization*, IEEE Trans. Automat. Control, 34 (1989), pp. 435–443.

[25] E. D. Sontag, *State-space and i/o stability for nonlinear systems*, in Feedback Control, Nonlinear Systems, and Complexity, Lecture Notes in Control and Inform. Sci., 202, B. A. Francis and A. R. Tannenbaum, eds., Springer-Verlag, New York, 1995, pp. 215–235.

[26] E. D. Sontag and Y. Wang, *On characterizations of the input-to-state stability property*, Systems Control Lett., 24 (1995), pp. 351–359.

[27] J. C. Willems and C. I. Byrnes, *Global adaptive stabilization in the absence of information on the sign of the high frequency gain*, in Lecture Notes in Control and Inform. Sci., 62, Springer-Verlag, New York, 1984, pp. 49–57.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 3, pp. 981–1004, May 1998

010

OPTIMAL BOUNDARY CONTROL OF THE STOKES FLUIDS WITH POINT VELOCITY OBSERVATIONS*

PUHONG YOU†, ZHONGHAI DING‡, AND JIANXIN ZHOU†

Abstract. This paper studies constrained linear-quadratic regulator (LQR) problems in distributed boundary control systems governed by the Stokes equation with point velocity observations. Although the objective function is not well defined, we are able to use hydrostatic potential theory and a variational inequality in a Banach space setting to derive a first-order optimality condition and then a characterization formula of the optimal control. Since matrix-valued singularities appear in the optimal control, a singularity decomposition formula is also established, with which the nature of the singularities is clearly exhibited. It is found that in general, the optimal control is not defined at observation points. A necessary and sufficient condition that the optimal control is defined at observation points is then proved.

Key words. LQR, Stokes fluid, distributed boundary control, point observation, hydrostatic potential, boundary integral equation, singularity decomposition

AMS subject classifications. 49N10, 49J20, 76D07, 76D10, 93C20, 65N38

PII. S0363012996300276

1. Introduction.

In this paper, we are concerned with the problems in boundary control of fluid flows. We consider the following constrained optimal boundary control problems in the systems governed by the Stokes equation with point velocity observations.

Let $\Omega \subset \mathbb{R}^3$ be a bounded domain with smooth boundary Γ, Γ_1 an open subset of Γ and $\Gamma_0 = \Gamma \setminus \Gamma_1$.

$$\text{(LQR)} \qquad \begin{cases} \min\limits_{\vec{u}\in\mathcal{U}} J(\vec{u}) = \sum\limits_{k=1}^{m} \mu_k |\vec{w}(P_k) - \vec{Z}_k|^2 + \gamma \int_{\Gamma_1} |\vec{u}(x)|^2 d\sigma_x, \\ \text{subject to (1.1)} \begin{cases} \nu\Delta\vec{w}(x) - \nabla p(x) = 0 & \text{in } \Omega, \\ \text{div } \vec{w}(x) = 0 & \text{in } \Omega, \\ \vec{\tau}(\vec{w})(x) = \vec{g}(x) & \text{on } \Gamma_0, \\ \vec{\tau}(\vec{w})(x) = \vec{u}(x) & \text{on } \Gamma_1, \end{cases} \end{cases}$$

where

$\vec{w}(x)$ is the velocity vector of the fluid at $x \in \Omega$;

$p(x)$ is the pressure of the fluid at $x \in \Omega$;

*Received by the editors June 10, 1996; accepted for publication (in revised form) March 21, 1997. http://www.siam.org/journals/sicon/36-3/30027.html

†Department of Mathematics, Texas A&M University, College Station, TX 77843 (pyou@math.tamu.edu, jzhou@math.tamu.edu). The research of these authors was supported in part by NSF grant DMS-9404380 and by an IRI Award of Texas A&M University.

‡Department of Aerospace Engineering, Texas A&M University, College Station, TX 77843. Present address: Department of Mathematical Sciences, University of Nevada-Las Vegas, Las Vegas, NV 89154-4020 (dingz@nevada.edu).

$\vec{\tau}(\vec{w})(x)$ is the surface stress of the fluid along Γ defined by

$$\vec{\tau}(\vec{w})(x) = (\tau_1(\vec{w})(x), \tau_2(\vec{w})(x), \tau_3(\vec{w})(x))^T,$$

$$\tau_i(\vec{w})(x) = \sum_{k=1}^{3} \left[\frac{\partial w_i(x)}{\partial x_k} + \frac{\partial w_k(x)}{\partial x_i} \right] n_k(x) - p(x) n_i(x);$$

$\vec{n}(x) = (n_1(x), n_2(x), n_3(x))$ is the unit outnormal vector of Γ at x;
$\vec{g}$ is a given (surface stress) Neumann boundary data (B.D.) on Γ_0;
$\vec{u}(x) \in \mathcal{U}$ is the (surface stress) Neumann boundary control on the surface Γ_1;
$\mathcal{U}$ is the admissible control set to be defined later for well-posedness of the problem and for applications;
$\gamma, \mu_k > 0, 1 \le k \le m$, are given weighting factors;
$P_k \in \Gamma, 1 \le k \le m$, are prescribed "observation points";
$Z_k \in \mathbb{R}^3, 1 \le k \le m$, are prescribed "target values" at P_k;
ν, a positive quantity, is the kinematic viscosity of the fluid. For simplicity, throughout this paper we assume that $\nu = 1$ and the density of the fluid is the constant one.

Let

$$M_0 = \{\vec{a} + \vec{b} \times \vec{x} \mid \vec{a}, \vec{b} \in \mathbb{R}^3\}, \tag{1.2}$$

which is the subspace of the rigid body motions in $\mathbb{R}^3$. Multiplying the Stokes equation by $\vec{a} + \vec{b} \times \vec{x} \in M_0$ and integration by parts yield the compatibility condition of the Stokes system, i.e.,

$$\int_\Gamma \vec{\tau}(\vec{w})(x) \cdot (\vec{a} + \vec{b} \times \vec{x}) d\sigma_x = 0$$

or

$$\vec{\tau}(\vec{w}) \perp M_0.$$

For $q \ge 1$, let A be a subspace of $(L^q(\Gamma))^3$ and denote

$$(L^q(\Gamma))^3_{\perp A} = \{\vec{f} \in (L^q(\Gamma))^3 | \vec{f} \perp A\}.$$

The Stokes equation (1.1) describes the steady state of an incompressible viscous fluid with low velocity in $\mathbb{R}^3$. It is a frequently used model in fluid mechanics and an interesting model in linear elastostatics due to its similarities. During the past years, considerable attention has been given to the problem of active control of fluid flows (see [1], [2], [7], [18], [19], and references therein). This interest is motivated by a number of potential applications, such as control of separation, combustion, fluid-structure interaction, and super maneuverable aircraft. In the study of those control problems and Navier–Stokes equations, the Stokes equations, which describe the slow steady flow of a viscous fluid, play an important role because of the needs in stability analysis, iterative computation of numerical solutions, boundary control, etc. The theoretical and numerical discussion of the Stokes equations on smooth or Lipschitz domains can be found from [14], [16], [17], [22], [25], [26], [27].

Our objective in this paper is to find the optimal surface stress $\vec{u}(x)$ on Γ_1, which yields a desired velocity distribution $\vec{w}(x)$, such that (s.t.) at observation points $P_k, 1 \le k \le m$, the observation values $\vec{w}(P_k)$ are as close as possible to the

target values Z_k with a least possible control cost $\int_{\Gamma_1} |\vec{u}(x)|^2 d\sigma_x$, which arise from the contemporary fluid control problems in the fluid mechanics.

Notice that point observations are assumed in the problem setting, because they are much easier to be realized in applications than distributed observations. They can be used in modeling contemporary "smart sensors."

Sensors can be used in boundary control systems (BCS) governed by partial differential equations (PDE) to provide information on the state as a feedback to the systems. According to the space-measure of the data that sensors can detect, sensors can be divided into two types, point sensors and distributed sensors. Point sensors are much more realistic and easier to design than distributed sensors. In contemporary "smart materials," piezoelectric or fiber-optic sensors (called *smart sensors*) can be embedded to measure deformation, temperature, strain, pressure, etc. Each smart sensor detects only the *average* of the data in between the sensor, and its size can be less than 10^{-6}m [29], [30], [24]. So in any sense, they should be treated as point sensors. As a matter of fact, so far *distributed sensors* have not been used in any real applications, to the best of our knowledge. However, once point observations on the boundary are used in a BCS, singularities will appear, and very often the system becomes ill-posed. Mathematically and numerically, it becomes very tough to handle. On the other hand, when point observations are used in the problem setting, the state variable has to be continuous, so the regularity of the state variable stronger than the one in the case of distributed observations is required. The fact is that in the literature of related optimal control theory, starting from the classic book [23] by J. L. Lions until recent papers [3], [4] by E. Casas and others, distributed observations are always assumed and the optimal controls are characterized by an adjoint system. The system is then solved numerically, typically by a finite-element method, which cannot efficiently tackle the singularity in the optimal control along the boundary.

On the other hand, since it is important in the optimal control theory to obtain a state-feedback characterization of the optimal control, with the bound constraints in the system, the Lagrange–Kuhn–Tucker approach is not desirable because theoretically it cannot provide us with a state-feedback characterization of the optimal control, which is important in our regularity/singularity analysis of the optimal control, and numerically it leads to a numerical algorithm to solve an optimization problem with a huge number of inequality constraints. A refinement of the boundary will double the number of the inequality constraints, so the numerical algorithm will be sensitive to the partition number of the boundary. Since the BCS is governed by a PDE system in $\mathbb{R}^3$, the partition number of the boundary can be very large and any numerical algorithm sensitive to the partition number of the boundary may fail to carry out numerical computation or provide reliable numerical solutions.

Recently in the study of a linear quadratic BCS governed by the Laplace equation with point observations, the potential theory and boundary integral equations (BIE) have been applied in [20], [10], [11], [12] to derive a characterization of the optimal control in terms of the optimal state directly and therefore bypass the adjoint system. This approach shows certain important advantages over others. It provides rather explicit information on the control and the state, and it is amenable to direct numerical computation through a boundary element method (BEM), which can efficiently tackle the singularities in the optimal control along the boundary.

In [10], [11], [9] several regularity results are obtained. The optimal control is characterized directly in terms of the optimal state. The exact nature of the singularities in the optimal control is exhibited through a decomposition formula. Based on

the characterization formula, numerical algorithms are also developed to approximate the optimal control. Their insensitivity to the discretization of the boundary and fast uniform convergences are mathematically verified in [12], [31].

The case with the Stokes system is much more complicated than the one with the Laplace equation due to the fact that the fundamental solution of the Stokes system is matrix-valued and has rougher singular behaviors. In this paper, we assume that the control is active on a part of the surface and the control variable is bounded by two vector-valued functions. A Banach space setting has been used in our approach. We first prove a necessary and sufficient condition for a variational inequality problem (VIP) which leads to a first-order optimality condition of our original optimization problem. A characterization of the optimal control and its singularity decomposition formula are then established. Our approach can be easily adopted to handle other cases, and it shows the essence of the characterization of the optimal control, through which gradient-related numerical algorithms can be designed to approximate the optimal control.

The organization of this paper is as follows. In the rest of section 1, we introduce some basic definitions and known regularity results that are required in the later development. In section 2, we first prove an existence theorem for an orthogonal projection. Next we derive a characterization result for a variational inequality which serves as a first-order optimality condition to our LQR problem. Then a state-feedback characterization of the optimal control is established. Section 3 will be devoted to study regularity/singularity of the optimal control. Since the optimal control contains a singular term, we first derive a singularity decomposition formula for the optimal control, with which we find that in general the optimal control is not defined at observation points. A necessary and sufficient condition that the optimal control is defined at observation points is then established. Some other regularities of the optimal control will also be studied in this section. Based on our characterization formulas, in a subsequent paper, we design a conditioned gradient projection method (CGPM) to approximate the optimal control. Numerical analysis for its (uniform) convergence and (uniform) convergence rate are presented there. We show that CGPM converges uniformly subexponentially, i.e., faster than any integer power of $\frac{1}{n}$. Therefore CGPM is insensitive to discretization of the boundary. The insensitivity of our numerical algorithm to discretization of boundary is a significant advantage over other numerical algorithms. Since the fundamental solution of the Stokes system is matrix valued with a very rough singular behavior, numerical analysis is also much more complicated than the case with scalar-valued fundamental solution, e.g., the Laplacian equation.

Let us now briefly recall some hydrostatic potential theory, BEM, and some known regularity results. Throughout this paper, for a sequence of elements in $\mathbb{R}^n$, we use a superscript to denote sequential index and a subscript to denote components, e.g., $\{x^k\} \subset \mathbb{R}^n$ and $x^k = (x_1^k, \ldots, x_n^k)$. We may also use $\vec{x}^k$ to emphasize that x^k is a vector. We may write $\vec{w}(x, \vec{u})$ to indicate that the velocity $\vec{w}$ depends also on $\vec{u}$. Unless stated otherwise, we assume $p > 2, q > 1$ with $\frac{1}{p} + \frac{1}{q} = 1$, $|\cdot|$ is the Euclidean norm in $\mathbb{R}^n$, and $\|\cdot\|$ is the norm in $(L^h(\Gamma))^n (h \geq 1)$.

Let $\{E(x,\xi), \vec{e}(x,\xi)\} = \{[E_{ij}(x,\xi)]_{3\times 3}, [e_i(x,\xi)]_{3\times 1}\}$ be the fundamental solution of the Stokes systems, i.e.,

$$\begin{cases} \Delta_x E(x,\xi) - \nabla_x \vec{e}(x,\xi) = -\delta(x-\xi) I_3, \\ \qquad\qquad \operatorname{div}_x E(x,\xi) = 0, \end{cases} \tag{1.3}$$

where $\delta(x-\xi)$ is the unit Dirac delta function at $x=\xi$ and I_3 is the 3×3 identity matrix. It is known [22] that

$$E_{ij}(x,\xi)=\frac{1}{8\pi}\left(\frac{\delta_{ij}}{|x-\xi|}+\frac{(x_i-\xi_i)(x_j-\xi_j)}{|x-\xi|^3}\right),\quad 1\le i,j\le 3,$$
$$e_i(x,\xi)=\frac{1}{4\pi}\frac{x_i-\xi_i}{|x-\xi|^3},$$

where $\delta_{i,j}$ is the Kronecker symbol.

Remark 1.1. The significant difference between the case with point observations and the case with distributed observations is as follows. For a given vector $\vec{V}\in\mathbb{R}^3$ the function

$$x\mapsto\sum_{k=1}^{m}\mu_k E(P_k,x)\vec{V} \tag{1.4}$$

has a singularity of order $O(\frac{1}{|x-P_k|})$ at $x=P_k$; however, it may oscillate between $-\infty$ and $+\infty$ as $x\to P_k$, so it is very tough to deal with, whereas the function

$$x\mapsto\int_{\Gamma_0}E(\xi,x)\vec{V}d\sigma_\xi \tag{1.5}$$

is well defined and continuous.

On the other hand, if $E(P_k,x)$ in (1.4) and (1.5) is replaced by the fundamental solution of the Laplace equation, in this case, $E(P_k,x)$ becomes scalar-valued, then (1.4) has the same order $O(\frac{1}{|x-P_k|})$ of singularity at $x=P_k$, but the limit as $x\to P_k$ exists (including $-\infty$ or $+\infty$). So the singularity can be easily handled. □

It is then known that the solution $(\vec{w},p)$ of the Stokes equation (1.1) has a simple-layer representation

$$\vec{w}(x)=\int_\Gamma E(x,\xi)\vec{\eta}(\xi)d\sigma_\xi+\vec{a}+\vec{b}\times\vec{x}\quad\forall\, x\in\overline{\Omega}, \tag{1.6}$$

$$p(x)=\int_\Gamma\vec{e}(x,\xi)\cdot\vec{\eta}(\xi)d\sigma_\xi+a\quad\forall\, x\in\Omega \tag{1.7}$$

for some constants $\vec{a},\vec{b}\in\mathbb{R}^3$ and $a\in\mathbb{R}$. $\vec{\eta}$ is called the layer density and $\vec{a}+\vec{b}\times\vec{x}$ represents a rigid body motion. By the jump property of the layer potentials, we obtain the BIE

$$\vec{\tau}(\vec{w})(x)=\frac{1}{2}\vec{\eta}(x)+\text{p.v.}\int_\Gamma T(x,\xi)\vec{\eta}(\xi)d\sigma_\xi\quad\forall\, x\in\Gamma, \tag{1.8}$$

where p.v. stands for principle value and

$$T(x,\xi)=[\vec{\tau}_x(E_1)(x,\xi),\vec{\tau}_x(E_2)(x,\xi),\vec{\tau}_x(E_3)(x,\xi)]=[T_{ij}(x,\xi)]_{3\times 3},$$
$$T_{ij}(x,\xi)=-\frac{3}{4\pi}\frac{(x_i-\xi_i)(x_j-\xi_j)}{|x-\xi|^5}(x-\xi)\cdot\vec{n}_x.$$

With a given Neumann B.D., the layer density $\vec{\eta}$ can be solved from the above BIE (1.8). Once the layer density is known, the solution $(\vec{w}(x),p(x))$ can be computed from (1.6) and (1.7). The velocity solution $\vec{w}(x)$ is unique only up to a rigid body motion, and the pressure solution $p(x)$ is unique up to a constant.

In BEM, the boundary $\Gamma = \Gamma_1 \bigcup \Gamma_0$ is divided into N elements with nodal points x_i. Assume that the layer density $\vec{\eta}(x)$ is piecewise smooth, e.g., piecewise constant, piecewise linear, etc. Then the BIE (1.8) becomes a linear algebraic system. This system can be solved for $\vec{\eta}(x_i)$ and then $(\vec{w}(x), p(x))$ can be computed from a discretized version of (1.6) and (1.7) for any $x \in \overline{\Omega}$.

For each $\vec{f} \in (L^2(\Gamma))^3$ and $x \in \mathbb{R}^3$, we define the simple layer potential of velocity $\mathcal{S}_v(\vec{f})$ by

$$\mathcal{S}_v(\vec{f})(x) = \int_\Gamma E(x,\xi)\vec{f}(\xi)d\sigma_\xi.$$

For each $\vec{f} \in (L^2(\Gamma))^3$ and $x \in \Gamma$, we define the boundary operators $\mathcal{K}$ and $\mathcal{K}^*$ by

$$\mathcal{K}(\vec{f})(x) = \text{p.v.} \int_\Gamma Q(x,\xi)\vec{f}(\xi)d\sigma_\xi,$$
$$\mathcal{K}^*(\vec{f})(x) = \text{p.v.} \int_\Gamma T(x,\xi)\vec{f}(\xi)d\sigma_\xi,$$

where

$$Q(x,\xi) = [\vec{\tau}_\xi(E_1)(x,\xi), \vec{\tau}_\xi(E_2)(x,\xi), \vec{\tau}_\xi(E_3)(x,\xi)] = [Q_{ij}(x,\xi)]_{3\times 3},$$
$$Q_{ij}(x,\xi) = \frac{3}{4\pi}\frac{(x_i-\xi_i)(x_j-\xi_j)}{|x-\xi|^5}(x-\xi)\cdot\vec{n}_\xi.$$

Next we collect some regularity results on $\mathcal{S}_v, \mathcal{K}$, and $\mathcal{K}^*$ into a lemma. Let

$$N = \ker\left(\frac{1}{2}I + \mathcal{K}^*\right),$$

which represents the set of all layer densities corresponding to the zero Neumann B.D., with which the Stokes system has only a rigid body motion. Hence we have

$$M_0 = \mathcal{S}_v(N) = \ker\left(\frac{1}{2}I + \mathcal{K}\right). \tag{1.9}$$

LEMMA 1.1. *Let $\Omega \subset \mathbb{R}^3$ be a bounded, simply connected domain with smooth boundary Γ.*

(a) $\mathcal{S}_v : (L^p(\Gamma))^3 \mapsto (C^{0,\alpha}(\mathbb{R}^3))^3$ *is a bounded linear operator for $p > 2$ and $0 < \alpha < \frac{p-2}{p}$.*

(b) *For any $1 \le p < +\infty$, $\mathcal{K}$ ($\mathcal{K}^*$) : $(L^p(\Gamma))^3 \mapsto (L^p(\Gamma))^3$ is a bounded linear operator and $\mathcal{K}$ ($\mathcal{K}^*$) is the adjoint of $\mathcal{K}^*$ ($\mathcal{K}$).*

(c) *For $p > 2$ and $0 < \alpha < \frac{p-2}{p}$, $\mathcal{K} : (L^p(\Gamma))^3 \mapsto (C^{0,\alpha}(\Gamma))^3$ is a bounded linear operator.*

(d) *For $1 < p < \infty$*

(1) $(\frac{1}{2}I + \mathcal{K}^*) : (L^p(\Gamma))^3_{\perp M_0} \mapsto (L^p(\Gamma))^3_{\perp M_0}$ *is invertible,*

(2) $(\frac{1}{2}I + \mathcal{K}) : (L^p(\Gamma))^3_{\perp M_0} \mapsto (L^p(\Gamma))^3_{\perp N}$ *is invertible.*

(e) *For $1 < q < 2$ and $s < \frac{2q}{2-q}$, $\mathcal{K} : (L^q(\Gamma))^3 \mapsto (L^s(\Gamma))^3$ is a bounded linear operator. Therefore $\mathcal{K}\circ\mathcal{K} : (L^q(\Gamma))^3 \mapsto (C^{0,\alpha}(\Gamma))^3$ for every $q > 1$ and $0 < \alpha < \frac{q-1}{q}$;*

(f) $(\frac{1}{2}I + \mathcal{K}) : (C(\Gamma))^3_{\perp M_0} \mapsto (C(\Gamma))^3_{\perp N}$ *is invertible.*

Proof. (a)–(d) can be found from [5], [6], [8], [13], [14], [22], [28].

To prove (e), since $\Gamma \subset \mathbb{R}^3$ is a compact set, it suffices to prove (e) for $q < s < \frac{2q}{2-q}$. Then we have $\frac{1}{q} > \frac{1}{s} > \frac{1}{q} - \frac{1}{2} = \frac{1}{2} + \frac{1}{q} - 1$. There exists an $\varepsilon \in (0,1)$, s.t. $\frac{1}{s} = \frac{1}{2-\varepsilon} + \frac{1}{q} - 1$. Let $r = 2-\varepsilon,\ \alpha = \frac{r'}{s'},\ \beta = \frac{q'}{s'}$, where r', q', s' are the conjugates of r, q, s, respectively. It can be verified that $1 < r < 2$ and

$$\frac{1}{\alpha} + \frac{1}{\beta} = 1, \quad \left(1 - \frac{q}{s}\right) s' = \frac{q}{\alpha}, \quad \left(1 - \frac{r}{s}\right) s' = \frac{r}{\beta}, \quad \frac{1}{\alpha} \cdot \frac{s}{s'} = \frac{s-q}{q}, \quad \frac{1}{\beta} \cdot \frac{s}{s'} = \frac{s-r}{r}.$$

Note

$$|Q_{ij}(x,\xi)| \le \frac{C}{|x-\xi|}, \quad 1 \le i, j \le 3, \tag{1.10}$$

and

$$\left(\int_\Gamma \frac{1}{|x-\xi|^r} \, d\sigma_\xi \right) < M < \infty \quad \forall x \in \Gamma,$$

where M is a constant independent of $x \in \Gamma$. Let $h(x) = \mathcal{K}(\vec{f})(x)$. Applying Hölder's inequality twice, we get

$$\begin{aligned}
|h(x)|^s &\le C^s \left(\int_\Gamma \frac{1}{|x-\xi|} |\vec{f}(\xi)| \, d\sigma_\xi \right)^s \\
&\le C^s \left(\int_\Gamma \left(\frac{1}{|x-\xi|} \right)^{\frac{r}{s}} |\vec{f}(\xi)|^{\frac{q}{s}} \left(\frac{1}{|x-\xi|} \right)^{1-\frac{r}{s}} |\vec{f}(\xi)|^{1-\frac{q}{s}} \, d\sigma_\xi \right)^s \\
&\le C^s \left(\int_\Gamma \frac{1}{|x-\xi|^r} |\vec{f}(\xi)|^q \, d\sigma_\xi \right) \left(\int_\Gamma \left(\frac{1}{|x-\xi|} \right)^{\frac{r}{\beta}} |\vec{f}(\xi)|^{\frac{q}{\alpha}} \, d\sigma_\xi \right)^{\frac{s}{s'}} \\
&\le C^s \left(\int_\Gamma \frac{1}{|x-\xi|^r} |\vec{f}(\xi)|^q \, d\sigma_\xi \right) \left(\int_\Gamma \frac{1}{|x-\xi|^r} \, d\sigma_\xi \right)^{\frac{s-r}{r}} \left(\int_\Gamma |\vec{f}(\xi)|^q \, d\sigma_\xi \right)^{\frac{s-q}{q}} \\
&\le C^s M^{s-r} \left(\int_\Gamma \frac{1}{|x-\xi|^r} |\vec{f}(\xi)|^q \, d\sigma_\xi \right) \cdot \|\vec{f}\|_q^{s-q}.
\end{aligned}$$

Thus

$$\begin{aligned}
\|h\|_{L^s(\Gamma)} &= \left(\int_\Gamma |h(x)|^s \, d\sigma_x \right)^{\frac{1}{s}} \\
&\le C M^{\frac{s-r}{s}} \left(\int_\Gamma \int_\Gamma \frac{1}{|x-\xi|^r} |\vec{f}(\xi)|^q \, d\sigma_\xi d\sigma_x \right)^{\frac{1}{s}} \cdot \|\vec{f}\|_q^{\frac{s-q}{s}} \\
&\le C M \|\vec{f}\|_q.
\end{aligned}$$

This proves the first part of (e). The second part follows from (c).

To prove (f), by (1.10), $Q_{ij}(x,\xi)$ is weakly singular for $1 \le i, j \le 3$. Thus $\mathcal{K}$ is an integral operator with a weakly singular kernel. By Theorem 2.22 in [21], $\mathcal{K}$ is a compact operator from $(C(\Gamma))^3$ to $(C(\Gamma))^3$. The rest follows from the Fredholm alternative (see [21, p. 44]. □

For a given Neumann B.D. $\vec{g} \in (L^p(\Gamma_0))^3$, we extend our control bound constraints $Bl, Bu \in (L^p(\Gamma_1))^3$ to the entire boundary Γ by

$$Bl(x) = \begin{cases} Bl(x), & x \in \Gamma_1, \\ \vec{g}(x), & x \in \Gamma_0, \end{cases} \quad \text{and} \quad Bu(x) = \begin{cases} Bu(x), & x \in \Gamma_1, \\ \vec{g}(x), & x \in \Gamma_0, \end{cases}$$

with

$$Bl(x) \le -\vec{B} < \vec{B} \le Bu(x) \quad \forall\, x \in \Gamma_1,$$

where $\vec{B} > 0$ is a constant vector depending on $\vec{g}$ and will be specified later. Define the feasible control set

$$\mathcal{U} = \left\{ \vec{u} \in (L^p(\Gamma))^3 \mid Bl(x) \le \vec{u}(x) \le Bu(x)\ \forall\, x \in \Gamma \text{ and } \vec{u} \perp M_0 \right\}, \tag{1.11}$$

where $\vec{u} \perp M_0$ stands for the compatibility condition of the Neumann B.D. in the Stokes system (1.1). It is clear that $\mathcal{U}$ is a closed bounded convex set in $(L^p(\Gamma))^3$.

According to Lemma 1.1 (a), for each given Neumann B.D. $\vec{u} \in \mathcal{U}$, the Stokes system (1.1) has a solution $\vec{w}$ in $(C(\overline{\Omega}))^3$ unique up to a vector $\vec{a} + \vec{b} \times \vec{x} \in M_0$, i.e.,

$$\vec{w}(x, \vec{u}) = \mathcal{S}_v \circ \left(\frac{1}{2} I + \mathcal{K}^* \right)^{-1} (\vec{u})(x) + \vec{a} + \vec{b} \times \vec{x}, \quad x \in \Omega, \tag{1.12}$$

$$= \vec{w}_0(x, \vec{u}) + \vec{a} + \vec{b} \times \vec{x}, \quad x \in \Omega, \tag{1.13}$$

where

$$\vec{w}_0(x, \vec{u}) = \mathcal{S}_v \circ \left(\frac{1}{2} I + \mathcal{K}^* \right)^{-1} (\vec{u})(x). \tag{1.14}$$

That is, for each given $\vec{u}$, the velocity state variable $\vec{w}$ is multiple valued, so the objective function $J(\vec{u})$ is not well defined. However, among all these velocity solutions, there is a unique solution $\vec{w}$ s.t.

$$\sum_{k=1}^{m} \mu_k |\vec{w}(P_k) - \vec{Z}_k|^2 = \min_{\vec{h} \in M_0} \sum_{k=1}^{m} \mu_k |\vec{w}_0(P_k) + \vec{h}(P_k) - \vec{Z}_k|^2. \tag{1.15}$$

A direct calculation yields that $\vec{w}(x) = \vec{w}_0(x) + \vec{a} + \vec{b} \times \vec{x}$ must satisfy

$$\begin{cases} \displaystyle\sum_{k=1}^{m} \mu_k (\vec{w}_0(P_k) + \vec{a} + \vec{b} \times \vec{P}_k - \vec{Z}_k) = 0, \\ \displaystyle\sum_{k=1}^{m} \mu_k (\vec{w}_0(P_k) + \vec{a} + \vec{b} \times \vec{P}_k - \vec{Z}_k) \times \vec{P}_k = 0. \end{cases} \tag{1.16}$$

Since such a $\vec{w}$ is unique and continuous, the point observations $\vec{w}(P_k)$ in our LQR problem setting make sense and the LQR problem is well posed.

From (1.14) and Lemma 1.1, we know

$$|\vec{w}(x, \vec{u}) - \vec{a}_u - \vec{b}_u \times \vec{x}| = |\vec{w}_0(x, \vec{u})| \le C \|\vec{u}\|_{L^p(\Gamma))^3}, \tag{1.17}$$

where C is a constant depending only on Γ. Let us observe (1.16). If we notice that $\vec{w}_0(x, \vec{u})$ is linear in $\vec{u}$, then we have the following lemma.

LEMMA 1.2. *Let $\vec{a}_0, \vec{b}_0 \in \mathbb{R}^3$ be the unique solution to*

$$\begin{cases} \vec{a}_0 \left(\displaystyle\sum_{k=1}^{m} \mu_k \right) + \vec{b}_0 \times \left(\displaystyle\sum_{k=1}^{n} \mu_k \vec{P}_k \right) = \displaystyle\sum_{k=1}^{m} \mu_k \vec{Z}_k \\ \vec{a}_0 \times \left(\displaystyle\sum_{k=1}^{m} \mu_k \vec{P}_k \right) + \displaystyle\sum_{k=1}^{m} \mu_k (\vec{b}_0 \times \vec{P}_k) \times \vec{P}_k = \displaystyle\sum_{k=1}^{m} \mu_k \vec{Z}_k \times \vec{P}_k. \end{cases}$$

Then for $\vec{u}_1, \vec{u}_2 \in \mathcal{U}$ *and* $t_1, t_2 \in \mathbb{R}$,

$$\vec{w}(x, t_1\vec{u}_1 + t_2\vec{u}_2) = t_1\vec{w}(x, \vec{u}_1) + t_2\vec{w}(x, \vec{u}_2) + (1 - t_1 - t_2)(\vec{a}_0 + \vec{b}_0 \times \vec{x}) \tag{1.18}$$

and

$$|\vec{w}(x, \vec{u}_1) - \vec{w}(x, \vec{u}_2)| \le C\|\vec{u}_1 - \vec{u}_2\|_{(L^p(\Gamma))^3}, \tag{1.19}$$

where C *is a constant depending only on* Γ.

2. Characterization of the optimal control. We establish an optimality condition of the LQR problem through a VIP. The characterization of the optimal control is then derived from the optimality condition.

In optimal control theory it is important to obtain a state-feedback characterization of the optimal control; i.e., the optimal control is stated as an explicit function of the optimal state. So the optimal control can be determined by a physical measurement of the optimal state. Our efforts are devoted to deriving such a result.

For each $\vec{f} \in (L^1(\Gamma))^3$, we define the vector-valued truncation function

$$\left[\vec{f}\right]_{Bl}^{Bu} = \left\{ [f_i(x)]_{Bl_i(x)}^{Bu_i(x)} = \left\{ \begin{array}{lll} Bu_i(x) & \text{if} & f_i(x) \ge Bu_i(x) \\ f_i(x) & \text{if} & Bl_i(x) < f_i(x) < Bu_i(x) \\ Bl_i(x) & \text{if} & f_i(x) \le Bl_i(x) \end{array} \right\} \right.$$

Let $\langle \cdot, \cdot \rangle$ be the pairing on $((L^q(\Gamma))^3, (L^p(\Gamma))^3)$. Since our feasible control set $\mathcal{U}$ defined in (1.11) is a convex closed bounded set in $(L^p(\Gamma))^3$, it is known that $\vec{u}^*$ is an optimal control of the LQR problem if

$$\langle \nabla J(\vec{u}^*), \vec{u} - \vec{u}^* \rangle \ge 0 \quad \forall\, \vec{u} \in \mathcal{U}. \tag{2.1}$$

For any $\alpha > 0$, (2.1) is equivalent to

$$\langle \vec{u}^* - (\vec{u}^* - \alpha \nabla J(\vec{u}^*)), \vec{u} - \vec{u}^* \rangle \ge 0 \quad \forall\, \vec{u} \in \mathcal{U}. \tag{2.2}$$

To derive an optimality condition, we need to find a characterization of a solution to the above variational inequality.

THEOREM 2.1. *For each* $f \in (L^q(\Gamma))^3$, u^f *is a solution to the variational inequality*

$$\langle u^f - f, u - u^f \rangle \ge 0 \quad \forall\, u \in \mathcal{U} \tag{VIP}$$

if and only if

$$u^f = [f + z^f]_{Bl}^{Bu}, \tag{2.3}$$

where $z^f \in M_0$ *such that* $[f + z^f]_{Bl}^{Bu} \perp M_0$ *(refer to Theorem* 2.2 *for the existence of such a* z^f*).*

Moreover, (2.3) *is well defined in the sense that if* z^1 *and* z^2 *are two vectors in* M_0 *s.t.*

$$[f + z^1]_{Bl}^{Bu} \perp M_0 \quad \textit{and} \quad [f + z^2]_{Bl}^{Bu} \perp M_0;$$

then

$$[f(x) + z^1(x)]_{Bl}^{Bu} = [f(x) + z^2(x)]_{Bl}^{Bu} \quad \textit{almost everywhere (a.e.)}\ x \in \Gamma. \tag{2.4}$$

Proof. By Theorem 2.2, there exists $z^f \in M_0$ s.t. $[f+z^f]_{Bl}^{Bu} \perp M_0$. Let $u^f = [f+z^f]_{Bl}^{Bu}$. We have for each $u \in \mathcal{U}$,

$$\begin{aligned}
&\langle u^f - f, u - u^f\rangle \\
&= \langle u^f - (f+z^f), u - u^f\rangle \\
&= \sum_{i=1}^{3} \int_\Gamma \left\{[f_i(x) + z_i^f(x)]_{Bl_i}^{Bu_i} - (f_i(x) + z_i^f(x))\right\} \left\{u_i(x) - [f_i(x) + z_i^f(x)]_{Bl_i}^{Bu_i}\right\} d\sigma_x \\
&\geq 0,
\end{aligned}$$

where the last inequality holds since each integrand, the product of two terms, is nonnegative.

Next we assume that u^f is a solution to the VIP; i.e.,

$$\langle u^f - f, u - u^f\rangle \geq 0 \quad \forall\, u \in \mathcal{U}.$$

Taking $u = [f+z^f]_{Bl}^{Bu}$, which is in $\mathcal{U}$, we obtain

$$\langle u^f - f, [f+z^f]_{Bl}^{Bu} - u^f\rangle \geq 0. \tag{2.5}$$

By the first part, we have

$$\langle [f+z^f]_{Bl}^{Bu} - f, u - [f+z^f]_{Bl}^{Bu}\rangle \geq 0 \quad \forall\, u \in \mathcal{U}. \tag{2.6}$$

Taking $u = u^f$ in (2.6) yields

$$\langle [f+z^f]_{Bl}^{Bu} - f, u^f - [f+z^f]_{Bl}^{Bu}\rangle \geq 0. \tag{2.7}$$

Combining (2.5) with (2.7) gives us

$$\langle u^f - [f+z^f]_{Bl}^{Bu}, u^f - [f+z^f]_{Bl}^{Bu}\rangle \leq 0. \tag{2.8}$$

Thus

$$u^f = [f+z^f]_{Bl}^{Bu}.$$

The proof of the second part of the theorem follows directly from taking $z^f = z^1$ and $u^f = [f+z^2]_{Bl}^{Bu}$ in (2.8). ∎

In a Hilbert space setting, the above theorem is called a characterization of projection. When $\mathcal{U}$ is a convex closed subset of a Hilbert space H, for each $f \in H$, u_f is a solution to the VIP if and only if

$$u_f = P_{\mathcal{U}}(f);$$

i.e., u_f is the projection of f on $\mathcal{U}$. This characterization is used to derive a first-order optimality condition for convex inequality constrained optimal control problems. However, this result is not valid in general Banach spaces. Instead we prove a characterization of truncation, which is a special case of a projection. Note that in a Hilbert space setting, a projection maps a point in the space into a subset of the same space. However, our truncation is a projection that maps a point in $(L^q(\Gamma))^3$ into a subset of $(L^p(\Gamma))^3 (p > 2, \frac{1}{p} + \frac{1}{q} = 1)$. It crosses spaces. This characterization gives a connection between the truncation and the solution to VIP, in our case, an optimality condition

in terms of the gradient. That is, by our characterization of truncation, $\vec{u}^* \in \mathcal{U}$ is a solution to the VIP (2.2) if and only if

$$\vec{u}^* = [\vec{u}^* - \alpha \nabla J(\vec{u}^*) + \vec{z}^*]_{Bl}^{Bu}, \tag{2.9}$$

where $\vec{z}^* \in M_0$ is defined in Theorem 2.2 s.t.

$$[\vec{u}^* - \alpha \nabla J(\vec{u}^*) + \vec{z}^*]_{Bl}^{Bu} \perp M_0.$$

To prove the existence of a rigid body motion z^f in (2.3), we establish the following existence theorem for an orthogonal projection, which is given in a very general case and plays a key role in establishing the optimality condition. It can be used to solve LQR problems governed by PDEs, e.g., the Laplacian, the Stokes, the linear elastostatics, etc., where the PDE has multiple solutions for given Neumann-type boundary data satisfying a certain orthogonality condition.

THEOREM 2.2. *Let Γ be a bounded closed set in $\mathbb{R}^n$ and $\Gamma_0 \subset \Gamma$ be a subset s.t. meas $(\Gamma_1) > 0$, where $\Gamma_1 = \Gamma \setminus \Gamma_0$. Let $\vec{g} \in (L^p(\Gamma_0))^n$ and $\vec{B}l, \vec{B}u \in (L^p(\Gamma))^n$ $(p \geq 2)$ be given s.t.*

$$\vec{B}l(x) < -\vec{B} < \vec{B} < \vec{B}u(x) \quad (a.e.) \quad \forall\, x \in \Gamma_1,$$

where $\vec{B} = (B, \ldots, B)$ is given by (2.17) and

$$\vec{B}l(x) = \vec{g}(x) = \vec{B}u(x) \quad \forall\, x \in \Gamma_0.$$

Assume that M_0 is an m-dimensional subspace in $(L^q(\Gamma))^n$ $(q \leq 2, \frac{1}{p} + \frac{1}{q} = 1)$ and $M_1 = \{\vec{z}|_{\Gamma_1} \mid \vec{z} \in M_0\}$. Then a necessary and sufficient condition that for each $\vec{f} \in (L^1(\Gamma))^n$ there exists $\vec{z}_f \in M_0$ s.t.

$$\left[\vec{f}(x) + \vec{z}_f(x)\right]_{Bl}^{Bu} \perp M_0 \tag{2.10}$$

is that

$$\vec{g} \perp M_1^c = \{\vec{z}|_{\Gamma_0} \mid \vec{z} \in M_0, \vec{z}|_{\Gamma_1} = 0\}. \tag{2.11}$$

Moreover the set of all solutions $\vec{z}_f$ in (2.10) is locally uniformly bounded in the sense that for each given $\vec{f} \in (L^1(\Gamma))^n$ there exist $r_0 > 0$ and $b > 0$ s.t. for any $\vec{h} \in (L^1(\Gamma))^n$ with $\|\vec{f} - \vec{h}\| \leq r_0$ and for any $\vec{z}_h \in M_0$ with

$$\left[\vec{h}(x) + \vec{z}_h(x)\right]_{Bl}^{Bu} \perp M_0$$

we have

$$\|\vec{z}_h\| \leq b. \tag{2.12}$$

Proof. Case 1. $\dim(M_1) = \dim(M_0)$, i.e., $M_1^c = \{0\}$. Let $y = (\vec{y}^1, \ldots, \vec{y}^m)$ be an orthonormal basis in M_1 and M_0. To prove the first part of the theorem, we have to show that for each $\vec{f} \in (L^1(\Gamma))^n$, there exists $C^f = (c_1^f, \ldots, c_m^f) \in \mathbb{R}^m$ s.t.

$$\left\langle \left[\vec{f}(x) + \sum_{i=1}^{m} c_i^f \vec{y}^i(x)\right]_{Bl}^{Bu}, \vec{y}^j \right\rangle_{\Gamma} = 0 \quad \forall\, j = 1, \ldots, m.$$

For each $\vec{f} \in (L^1(\Gamma))^n$, we define a map $T_f : \mathbb{R}^m \mapsto \mathbb{R}^m$, for $C = (c_1, \ldots, c_m) \in \mathbb{R}^m$, by

$$T_f(C) = \left\{ \left\langle \left[\vec{f}(x) + \sum_{i=1}^{m} c_i \vec{y}^i(x) \right]_{Bl}^{Bu}, \vec{y}^j \right\rangle_{\Gamma} \right\}_{j=1,\cdots,m}. \tag{2.13}$$

Then to prove the first part, it suffices to show that for each $\vec{f} \in (L^1(\Gamma))^n$, there exists $C_f \in \mathbb{R}^m$ s.t.

$$T_f(C_f) = 0.$$

It is easy to check that for any $\vec{f}, \vec{h} \in (L^1(\Gamma))^n$ and $C_1, C_2 \in \mathbb{R}^m$, there exist two constants γ_1, γ_2 depending only on Γ and the basis y s.t.

$$|T_f(C_1) - T_h(C_2)| \le \gamma_1 |\vec{f} - \vec{h}|_{L^1} + \gamma_2 |C_1 - C_2|. \tag{2.14}$$

So $C \mapsto T_f(C)$ is a bounded (depends on Bl and Bu) Lipschitz continuous map. To show that T_f has a zero, we prove that there exists a constant $R > 0$ s.t. when $C \in \mathbb{R}^m$ and $|C| > R$, we have

$$T_f(C) \cdot C > 0. \tag{2.15}$$

Once (2.15) is verified, we have

$$\begin{aligned} |C - T_f(C)|^2 &= |C|^2 - 2T_f(C) \cdot C + |T_f(C)|^2 \\ &< |C|^2 + |T_f(C)|^2 \qquad\qquad \forall\, C \in \mathbb{R}^m,\ |C| > R. \end{aligned}$$

By Altman's fixed-point theorem [15], the map $C \mapsto C - T_f(C)$ has a fixed point $C^f \in B_R$ (B_R is the ball of radius R at the origin), i.e.,

$$T_f(C^f) = 0.$$

So it remains to verify (2.15). Define

$$D = \left\{ C = (c_1, \ldots, c_m) \in \mathbb{R}^m \mid \sum_{i=1}^{m} c_i^2 = 1 \right\}.$$

It suffices to show that there exists $R > 0$ s.t. for $t > R$,

$$T_f(tC) \cdot C > 0 \quad \forall\, C \in D.$$

In the following, we prove that for each given $\vec{f} \in (L^1(\Gamma))^n$ and $C \in D$, there exist $r_0 > 0$ and $R > 0$ s.t. when $t > R$, for any $\vec{h} \in (L^1(\Gamma))^n$ with $\|\vec{f} - \vec{h}\|_{L^1} \le r_0$, we have

$$T_h(tC) \cdot C > 0 \quad \forall\, C \in D.$$

So the second part of the theorem also follows. For each $C \in D$, we denote

$$\vec{y}^C(x) = \sum_{i=1}^{m} c_i \vec{y}^i(x).$$

It is obvious that

$$\int_{\Gamma_1} |\vec{y}^C(x)| d\sigma_x$$

is continuous in C and positive on the compact set D. Hence

$$m_y = \min_{C\in D} \left\{ \int_{\Gamma_1} | \vec{y}^C(x)| d\sigma_x \right\} > 0 \tag{2.16}$$

and we set

$$B = \frac{\max_{C\in D} \int_{\Gamma_0} |\vec{g}(x)\cdot \vec{y}^C(x)| d\sigma_x}{m_y}. \tag{2.17}$$

For any given $\varepsilon > 0$, we assume

$$Bl_i(x) \le -B-\varepsilon, \quad Bu_i(x) \ge B+\varepsilon \quad \forall\, x\in\Gamma_1, \ \ i=1,\dots,n.$$

For each $C\in D, t>0$,

$$\begin{aligned} T_f(tC)\cdot C &= \sum_{j=1}^m \left(\int_\Gamma \left[\vec{f}(x) + \sum_{i=1}^m tc_i \vec{y}^i(x) \right]_{Bl}^{Bu} \cdot \vec{y}^j(x) d\sigma_x \right) c_j \\ &= \int_\Gamma [\vec{f}(x) + t\vec{y}^C(x)]_{Bl}^{Bu} \cdot \vec{y}^C(x) d\sigma_x \\ &= \int_{\Gamma_1} \left[\vec{f}(x) + t\vec{y}^C(x) \right]_{Bl}^{Bu} \cdot (\vec{y}^C(x)) d\sigma_x + \int_{\Gamma_0} \vec{g}(x)\cdot \vec{y}^C(x) d\sigma_x \\ &= \sum_{i=1}^n I_i^C(t) + \int_{\Gamma_0} \vec{g}(x)\cdot \vec{y}^C(x) d\sigma_x, \end{aligned}$$

where for $i=1,\dots,n$,

$$I_i^C(t) = \int_\Gamma [f_i(x) + t y_i^C(x)]_{Bl_i(x)}^{Bu_i(x)} y_i^C(x) d\sigma_x.$$

Let

$$\Gamma_i^{C+} = \{x\in\Gamma_1 \mid y_i^C(x) > 0\} \quad \text{and} \quad \Gamma_i^{C-} = \{x\in\Gamma_1 \mid y_i^C(x) < 0\}.$$

We have

$$\begin{aligned} \lim_{t\to+\infty} I_i^C &= \int_{\Gamma_i^{C+}} Bu_i(x)\cdot y_i^C(x) d\sigma_x + \int_{\Gamma_i^{C-}} Bl_i(x)\cdot y_i^C(x) d\sigma_x \\ &\ge (B+\varepsilon) \int_{\Gamma_1} |y_i^C(x)| d\sigma_x. \end{aligned}$$

Thus

$$\begin{aligned} \lim_{t\to+\infty} T_f(tC)\cdot C &\ge (B+\varepsilon) \sum_{i=1}^n \int_{\Gamma_1} |y_i^C(x)| d\sigma_x + \int_{\Gamma_0} \vec{g}(x)\cdot \vec{y}^C(x) d\sigma_x \\ &\ge (B+\varepsilon) \int_{\Gamma_1} |y^C(x)| d\sigma_x + \int_{\Gamma_0} \vec{g}(x)\cdot \vec{y}^C(x) d\sigma_x \\ &\ge \varepsilon m_y, \end{aligned}$$

where m_y given by (2.16) is independent of C. From (2.14), we see that $T_f(C)\cdot C$ is continuous in both $\vec{f}$ and C; therefore, there exist $R^C>0$, r_C and $\delta_C>0$, as $t>R^C$, $\|\vec{h}-\vec{f}\|_{L^1}\le r^C$, and $|C'-C|<\delta_C$. We have

$$T_h(tC')\cdot C'\ge\frac{1}{2}\varepsilon m_y>0.$$

Since D is compact, there exist $C_1,\dots,C_s\in D$ and $\delta_1,\dots,\delta_s$ s.t.

$$D\subset\cup_{k=1}^s B_{\delta_k}(C_k).$$

Let

$$R^0=\max\{R^{C_1},\dots,R^{C_s}\}\quad\text{and}\quad r_0=\min\{r^{C_1},\dots,r^{C_s}\}.$$

When $t>R^0$, for all $\vec{h}\in(L^1(\Gamma))^n$ with $\|\vec{h}-\vec{f}\|_{L^1}\le r_0$, we have

$$T_h(tC)\cdot C\ge\frac{1}{2}\varepsilon m_y>0\quad\forall\, C\in D.$$

So we need only to take

$$\vec{B}=(B,\dots,B)$$

and

$$\vec{B}l<-\vec{B}<\vec{B}<\vec{B}u\quad\text{a.e. on }\Gamma_1.$$

Case 2. $m_1=\dim(M_1)<\dim(M_0)=m$. Let $y=(\vec{y}^1,\dots,\vec{y}^m)$ be an orthonormal basis in M_0, where $(\vec{y}^1,\dots,\vec{y}^{m_1})$ is a basis in M_1 with

$$\vec{y}^i|_{\Gamma_0}=0,\ (i=1,\dots,m_1)\quad\text{and}\quad\vec{y}^j|_{\Gamma_1}=0,\ (j=m_1+1,\dots,m). \tag{2.18}$$

By the proof in Case 1, for each $\vec{f}\in(L^1(\Gamma))^n$, there exists $C^f=(c_1^f,\dots,c_{m_1}^f)\in\mathbb{R}^{m_1}$ s.t.

$$\left\langle\left[\vec{f}(x)+\sum_{i=1}^{m_1}c_i^f\vec{y}^i(x)\right]_{Bl}^{Bu},\vec{y}^j\right\rangle_{\Gamma_1}=0\quad\forall\, j=1,\dots,m_1.$$

Then for any $c_{m_1+1}^f,\dots,c_m^f\in\mathbb{R}$, by (2.18), we have

$$\begin{aligned}\left\langle\left[\vec{f}(x)+\sum_{i=1}^{m}c_i^f\vec{y}^i(x)\right]_{Bl}^{Bu},\vec{y}^j\right\rangle_{\Gamma}&=\langle\vec{g}(x),\vec{y}^j\rangle_{\Gamma_0}+\left\langle\left[\vec{f}(x)+\sum_{i=1}^{m_1}c_i^f\vec{y}^i(x)\right]_{Bl}^{Bu},\vec{y}^j\right\rangle_{\Gamma_1}\\&=0\quad\forall j=1,\dots,m_1.\end{aligned}$$

On the other hand, when $j>m_1$, for any $c_1,\dots,c_m\in\mathbb{R}$, by (2.18), we have

$$\left\langle\left[\vec{f}(x)+\sum_{i=1}^{m}c_i\vec{y}^i(x)\right]_{Bl}^{Bu},\vec{y}^j\right\rangle_{\Gamma}=\langle\vec{g}(x),\vec{y}^j\rangle_{\Gamma_0}.$$

Therefore

$$\left\langle\left[\vec{f}(x)+\sum_{i=1}^{m}c_i\vec{y}^i(x)\right]_{Bl}^{Bu},\vec{y}^j\right\rangle_{\Gamma}=0,\quad j>m_1,$$

if and only if

$$\langle \vec{g}(x), \vec{y}^j \rangle_{\Gamma_0} = 0, \quad j > m_1;$$

i.e., (2.11) is satisfied. The proof is complete. □

Remark 2.1. In the above theorem,

(1) when rigid body motion is considered,

$$M_0 = \{\vec{a} + \vec{b} \times \vec{x} \mid \vec{a}, \vec{b} \in \mathbb{R}^3\},$$

we have $\dim(M_0) = \dim(M_1) = 6$, so all the conditions in the theorem are satisfied. So for each $\vec{f} \in (L^1(\Gamma))^3$ there is $\vec{a}_f + \vec{b}_f \times \vec{x} \in M_0$ s.t.

$$\left[\vec{f} + \vec{a}_f + \vec{b}_f \times \vec{x}\right]_{Bl}^{Bu} \perp M_0;$$

(2) if

$$Bl(x) \equiv -\infty \quad \text{or} \quad Bu(x) \equiv +\infty \quad \text{on } \Gamma_1$$

the conclusion still holds for each $\vec{f} \in (L^l(\Gamma))^n$ $(l \geq 1)$ and M_0 an m-dimensional subspace of $(L^q(\Gamma))^n$, where $q \geq 1$, $\frac{1}{h} + \frac{1}{q} = 1$, and $h = \min\{l, p\}$. When $h = 1$, $q = +\infty$;

(3) the vector C in (2.13) represents the rigid body motion in our case. From the above theorem, we can see that the solution C_f such that $T_f(C_f) = 0$ is not unique.

The following error estimate contains a uniqueness result, which will also be used in proving the uniform convergence rate of an algorithm in a subsequent paper.

THEOREM 2.3. *Let us maintain all the assumptions in Theorem* 2.2. *Let* $\vec{f}, \vec{h}$ *be given in* $(L^1(\Gamma))^n$, C_f, C_h *be, respectively, two zeros of* T_f *and* T_h *defined by* (2.13). *If*

$$meas\,(\Gamma_{C_f}) + meas\,(\Gamma_{C_h}) > 0,$$

where

$$meas\,(\Gamma_{C_f}) = \sum_{i=1}^{n} meas\,\{x \in \Gamma \mid Bl_i(x) < f_i(x) + y_i^{C_f}(x) < Bu_i(x)\},$$

$$meas\,(\Gamma_{C_h}) = \sum_{i=1}^{n} meas\,\{x \in \Gamma \mid Bl_i(x) < h_i(x) + y_i^{C_h}(x) < Bu_i(x)\},$$

$$y^{C_f}(x) = \sum_{i=1}^{m} c_i^f y^i(x) \quad \textit{and} \quad y^{C_h} = \sum_{i=1}^{m} c_i^h y^i(x),$$

then

$$|C_f - C_h| \leq \gamma \|\vec{f} - \vec{h}\|_{(L^1(\Gamma))^n}, \tag{2.19}$$

where the constant γ *is independent of* C_f *and* C_h.

Proof. We may assume that

$$\text{meas}\,(\Gamma_{C_f}) > 0.$$

For $T_f(C)$, we denote

$$\Gamma_C^k = \{x \in \Gamma \mid Bl_k(x) < f_k(x) + y_k^C(x) < Bu_k(x)\},$$

where

$$y_k^C(x) = \sum_{i=1}^{m} c_i y_k^i(x).$$

Write

$$\text{meas}\,(\Gamma_C) = \sum_{k=1}^{n} \text{meas}\,(\Gamma_C^k).$$

Since $T_f(C)$ is Lipschitz continuous in C, a direct calculation leads to the Fréchet derivative

$$T_f'(C) = \left[\sum_{k=1}^{n} \langle y_i^k, y_j^k \rangle_{\Gamma_C^k}\right]_{m\times m} \quad \text{a.e. } C \in \mathbb{R}^m,$$

a Gram-matrix, which is symmetric positive semidefinite; i.e., for any nonzero vector $b = (b_1, \ldots, b_m) \in \mathbb{R}^m$,

$$(b_1, \ldots, b_m) T_f'(C)(b_1, \ldots, b_m)^T = \sum_{k=1}^{n} \left\langle \sum_{i=1}^{m} b_i y_i^k, \sum_{i=1}^{m} b_i y_i^k \right\rangle_{\Gamma_C^k} \geq 0,$$

where "$>$" holds strictly if

$$\text{meas}\,(\Gamma_C) > 0,$$

because $\{\vec{y}_1, \ldots, \vec{y}_m\}$ is linearly independent.

On the other hand, we have

$$\left[\sum_{k=1}^{n} \langle y_i^k, y_j^k \rangle_{\Gamma_C^k}\right]_{m\times m} + \left[\sum_{k=1}^{n} \langle y_i^k, y_j^k \rangle_{\Gamma\setminus\Gamma_C^k}\right]_{m\times m} = \left[\sum_{k=1}^{n} \langle y_i^k, y_j^k \rangle_{\Gamma}\right]_{m\times m} = I_{m\times m},$$

where the Gram-matrix

$$\left[\sum_{k=1}^{n} \langle y_i^k, y_j^k \rangle_{\Gamma\setminus\Gamma_C^k}\right]_{m\times m}$$

is also symmetric positive semidefinite. Therefore

$$0 \leq |T_f'(C)| \leq 1 \quad \text{a.e. } C \in \mathbb{R}^m,$$

where "$<$" holds strictly in the first inequality if meas $(\Gamma_C) > 0$ and "$<$" holds strictly in the second inequality if meas $(\Gamma \setminus \Gamma_C) > 0$. Next, for given f, h in $(L^1(\Gamma))^n$ and two zeros C_f, C_h of T_f and T_h, respectively, we let

$$C_t = tC_h + (1-t)C_f, \quad t \in (0,1).$$

Since $T_f(C)$ is Lipschitz continuous in C, once meas $(\Gamma_{C_f}) > 0$, there exists $\varepsilon > 0$ s.t.

$$\text{meas}\,(\Gamma_{C_t}) > 0 \quad \forall\, 0 < t < \varepsilon.$$

It follows that $T_f'(C_t)$ is a symmetric positive definite matrix with

$$0 < |T_f'(C_t)| \le 1 \quad \text{a.e. } 0 < t < \varepsilon.$$

Therefore $\int_0^1 T_f'(C_t)dt$ defines a symmetric positive definite matrix with

$$0 < \left| \int_0^1 T_f'(C_t)dt \right| \le 1.$$

For any $0 < \mu < 1$, we have

$$0 < \left| I - \mu \int_0^1 T_f'(C_t)dt \right| = (1 - \lambda_f) < 1$$

for some $0 < \lambda_f < 1$. Taking

$$C_f - \mu T_f(C_f) = C_f \quad \text{and} \quad C_h - \mu T_h(C_h) = C_h$$

into account, we arrive at

$$\begin{aligned} |C_f - C_h| &= |C_f - C_h - \mu(T_f(C_f) - T_h(C_h))| \\ &= |C_f - C_h - \mu(T_f(C_f) - T_f(C_h) + T_f(C_h) - T_h(C_h))| \quad \text{(use (2.14))} \\ &\le \left| I - \mu \int_0^1 T_f'(C_t)dt \right| |C_f - C_h| + \mu\gamma_1 \|f - h\|_1 \\ &= (1 - \lambda_f)|C_f - C_h| + \mu\gamma_1 \|f - h\|_1. \end{aligned}$$

Consequently we have

$$|C_f - C_h| \le \frac{\gamma_1 \mu}{\lambda_f} \|f - h\|_{(L^1(\Gamma))^n},$$

and the proof is complete. □

As a direct consequence of Theorem 2.3, we obtain the following uniqueness result.

COROLLARY 2.4. *Let us maintain all the assumptions in Theorem* 2.2. *For given* $\vec{f} \in (L^1(\Gamma))^n$, *if* C_f *is a zero of* T_f *with*

$$\textit{meas}\,(\Gamma_{C_f}) > 0,$$

then C_f *is the unique zero of* T_f. □

Now we present a state-feedback characterization of the optimal control.

THEOREM 2.5. *Let* $\Omega \subset \mathbb{R}^3$ *be a bounded domain with smooth boundary* Γ. *The LQR problem has a unique optimal control* $\vec{u}^* \in \mathcal{U}$ *and a unique optimal velocity state* $\vec{w}^* \in (C(\Gamma))^3$ *s.t.*

$$\text{(2.20)} \qquad \begin{cases} \displaystyle\sum_{k=1}^{M} \mu_k(\vec{w}^*(P_k) - \vec{Z}_k) = 0, \\ \displaystyle\sum_{k=1}^{M} \mu_k(\vec{w}^*(P_k) - \vec{Z}_k) \times \vec{P}_k = 0. \end{cases}$$

and

$$\vec{u}^*(x) = \left[-\frac{1}{\gamma} \left(\frac{1}{2} I + \mathcal{K} \right)^{-1} \left(\sum_{k=1}^{m} \mu_k E(P_k, \cdot)(\vec{w}^*(P_k) - \vec{Z}_k) \right)(x) + \vec{a} + \vec{b} \times \vec{x} \right]_{Bl}^{Bu}, \quad \forall\, x \in \Gamma, \tag{2.21}$$

where $\vec{a} + \vec{b} \times \vec{x}$ *is defined in Theorem* 2.2 *s.t.* $\vec{u}^* \perp M_0$ *and* M_0 *is given in* (1.2).

Proof. Let $X = (L^p(\Gamma))^3_{\perp M_0}$. Since our objective function $J(\vec{u})$ is strictly convex and differentiable, and the feasible control set $\mathcal{U}$ is a closed bounded convex subset in the reflexive Banach space X, the existence and uniqueness of the optimal control are well established. Equation (2.20) is just a copy of (1.16). By our characterization of truncation, Theorem 2.1 with $\alpha = \frac{1}{2\gamma}$,

$$\vec{u}^*(x) = \left[\vec{u}^*(x) - \frac{1}{2\gamma} \nabla J(\vec{u})(x) + \vec{a} + \vec{b} \times \vec{x} \right]_{Bl}^{Bu} \quad \forall\, x \in \Gamma,$$

where $\vec{a} + \vec{b} \times \vec{x} \in M_0$ is defined in Theorem 2.2 s.t.

$$\left[\vec{u}^* - \frac{1}{2\gamma} \nabla J(\vec{u}) + \vec{a} + \vec{b} \times \vec{x} \right]_{Bl}^{Bu} \perp M_0.$$

To prove (2.21), we need only to show

$$\nabla J(\vec{u}) = 2 \left\{ \left(\frac{1}{2} I + \mathcal{K} \right)^{-1} \sum_{k=1}^{m} \mu_k E(P_k, \cdot)(\vec{w}(P_k, \vec{u}) - \vec{Z}_k) + \gamma \vec{u} \right\}. \tag{2.22}$$

Applying (1.9), i.e., $M_0 = \mathcal{S}_v(N)$ and (2.20), we get

$$\sum_{k=1}^{m} \mu_k E(P_k, \cdot)(\vec{w}(P_k, \vec{u}) - \vec{Z}_k) \in (L^q(\Gamma))^3_{\perp N}, \tag{2.23}$$

and then

$$\left(\frac{1}{2} I + \mathcal{K} \right)^{-1} \left\{ \sum_{k=1}^{m} \mu_k E(P_k, \cdot)(\vec{w}(P_k, \vec{u}) - \vec{Z}_k) \right\} \in (L^q(\Gamma))^3_{\perp M_0}. \tag{2.24}$$

Since $\nabla J(\vec{u})$ defines a bounded linear functional on X, for any $\vec{h} \in X$, taking (1.12) into account, we have

$$\begin{aligned}
&\langle \nabla J(\vec{u}), \vec{h} \rangle \\
&= 2 \sum_{k=1}^{m} \mu_k (\vec{w}(P_k, \vec{u}) - \vec{Z}_k) \mathcal{S}_v \left(\left(\frac{1}{2} I + \mathcal{K}^* \right)^{-1} \vec{h} \right)(P_k) + 2\gamma \langle \vec{u}, \vec{h} \rangle \\
&= 2 \sum_{k=1}^{m} \mu_k (\vec{w}(P_k, \vec{u}) - \vec{Z}_k) \int_\Gamma E(P_k, \xi) \left[\left(\frac{1}{2} I + \mathcal{K}^* \right)^{-1} \vec{h} \right](\xi) d\sigma_\xi + 2\gamma \langle \vec{u}, \vec{h} \rangle \\
&= 2 \int_\Gamma \left(\frac{1}{2} I + \mathcal{K} \right)^{-1} \left[\sum_{k=1}^{m} \mu_k E(P_k, \cdot)(\vec{w}(P_k, \vec{u}) - \vec{Z}_k) \right](\xi) \cdot \vec{h}(\xi) d\sigma_\xi + 2\gamma \langle \vec{u}, \vec{h} \rangle \\
&= 2 \left\langle \left[\left(\frac{1}{2} I + \mathcal{K} \right)^{-1} \sum_{k=1}^{m} \mu_k E(P_k, \cdot)(\vec{w}(P_k, \vec{u}) - \vec{Z}_k) \right] + \gamma \vec{u}(\cdot), \vec{h}(\cdot) \right\rangle.
\end{aligned}$$

So (2.22) is verified and the proof is complete. ∎

3. Regularities of the optimal control. It is clear that (2.21) is a feedback characterization of the optimal control. To obtain such a characterization, $\alpha = \frac{1}{2\gamma}$ in (2.9) is crucial. Later on we will see that $\alpha = \frac{1}{2\gamma}$ is also crucial in proving the uniform convergence of our numerical algorithms in a subsequent paper. Observe that when $Bl = -\infty$ and $Bu = +\infty$, it corresponds to the LQR problem without constraints on the control variable. The optimal solution, if it exists, becomes

$$\vec{u}^*(x) = -\frac{1}{\gamma}\left(\frac{1}{2}I + \mathcal{K}\right)^{-1}\left(\sum_{k=1}^{m}\mu_k E(P_k,\cdot)(\vec{w}^*(P_k) - \vec{Z}_k)\right)(x) + \vec{a} + \vec{b}\times\vec{x} \quad \forall\, x \in \Gamma,$$

where $\vec{a}+\vec{b}\times\vec{x}$ is defined in Theorem 2.2 s.t. $\vec{u}^* \perp M_0$ (see Remark 2.1). But according to Lemma 1.1 (d), such a solution $\vec{u}^*$ is only in $(L^q(\Gamma))^3$ $(q < 2)$, since $E(P_k,\cdot)$ is only in $(L^q(\Gamma))^3$. So it is reasonable to apply bound constraints Bl and Bu on the control variable $\vec{u}$. However, we notice that the optimal control still contains a singular term

$$\left(\frac{1}{2}I + \mathcal{K}\right)^{-1}\sum_{k=1}^{m}\mu_k E(P_k,\cdot)(\vec{w}(P_k,\vec{u}) - \vec{Z}_k)(x),$$

which is not computable at $x = P_k$. In order to carry out the truncation by Bl and Bu, we have to know the sign of this singular term. Hence we derive a singularity decomposition formula of (2.21), in which the singular term is expressed as continuous bounded terms plus a simple dominant singular term and a lower-order singular term. With the simple dominant singular term, the nature of the singularity is clearly exposed.

THEOREM 3.1. *For the optimal control $\vec{u}^*$ given in* (2.21), *let*

$$\vec{f}^*(x) = \sum_{k=1}^{m}\mu_k E(P_k,x)(\vec{w}^*(P_k) - \vec{Z}_k).$$

Then

$$\left(\frac{1}{2}I + \mathcal{K}\right)^{-1}\vec{f}^*(x) = 2\vec{f}^*(x) - 4\mathcal{K}\vec{f}^*(x) + 4\left(\frac{1}{2}I + \mathcal{K}\right)^{-1}\circ\mathcal{K}\circ\mathcal{K}\vec{f}^*(x) + \vec{a}_* + \vec{b}_*\times\vec{x}, \tag{3.1}$$

where in the singular part, the second term $4\mathcal{K}\vec{f}^*(x)$ is dominated by the first term $2\vec{f}^*(x)$, whose nature of singularity can be determined at each P_k and the regular term $4(\frac{1}{2}I + \mathcal{K})^{-1}\circ\mathcal{K}\circ\mathcal{K}\vec{f}^*(x)$ is continuous on Γ.

Proof. For given $\vec{g} \in (L^q(\Gamma))^3_{\perp N}$ with $q > 2 - \varepsilon(\Gamma)$, we have

$$\left(\frac{1}{2}I + \mathcal{K}\right)^{-1}\vec{g} = 2\vec{g} - 4\mathcal{K}\vec{g} + 4\left(\frac{1}{2}I + \mathcal{K}\right)^{-1}\circ\mathcal{K}\circ\mathcal{K}\vec{g} + \vec{a}_g + \vec{b}_g\times\vec{x}. \tag{3.2}$$

Let

$$\vec{f}^*(x) = \sum_{k=1}^{m}\mu_k E(P_k,x)(\vec{w}^*(P_k) - \vec{Z}_k).$$

By (2.23), $\vec{f}^* \in (L^q(\Gamma))^3_{\perp N}$ for every $q < 2$; thus (3.1) follows. The first part of Lemma 1.1 (e) states that the singularity in $2\vec{f}^*$ dominates the one in $4\mathcal{K}\vec{f}^*$, whereas the second part of Lemma 1.1 (e) and (f) imply that $(\frac{1}{2}I + \mathcal{K})^{-1}\circ\mathcal{K}\circ\mathcal{K}\vec{f}^*$ is continuous. ∎

The above singularity decomposition formula plays an important role in our singularity analysis and also in our numerical computation. It is used to prove the uniform convergence and to estimate the uniform convergence rate of our numerical algorithms in a subsequent paper.

Note that the fundamental velocity solution

$$E(\xi,x)=\left\{\frac{1}{8\pi}\left(\frac{\delta_{ij}}{|x-\xi|}+\frac{(x_i-\xi_i)(x_j-\xi_j)}{|x-\xi|^3}\right)\right\},\quad 1\le i,j\le 3,$$

is not defined when $\xi=P_k$ and $x\to P_k$, in the sense that when $x\to P_k$, some of the entries may oscillate between $-\infty$ and $+\infty$. So if we look at the simple dominant singular term in the singularity decomposition formula of the optimal control, we can see that in general, the optimal control $\vec{u}^*(x)$ is not defined at P_k, even with the truncation by Bl and Bu. This is a significant difference between systems with scalar-valued fundamental solution and with matrix-valued fundamental solution. For the formal case, e.g., the Laplacian, the optimal control is continuous at every point where Bl and Bu are continuous. Of course, if $Bl(P_k)=Bu(P_k)=\vec{g}(P_k)$, i.e., $P_k\in\Gamma_0$, which means the control is not active at P_k, then trivially $\vec{u}^*(P_k)=\vec{g}(P_k)$, a prescribed value. This is the case when a sensor is placed at P_k, then a control device cannot be put at the same point P_k. However, in general point observation case, the control may still be active at P_k. The above analysis then states that the optimal control is not defined at P_k unless some other conditions are posed. This is the nature of point observations. Notice that a distributed parameter control is assumed in our problem setting. Theoretically the values of the control variable at finite points will not affect the system. However, in numerical computation we can only evaluate the optimal control $\vec{u}^*$ at a finite number of points. The observation points P_k's usually are of the most interest. On the other hand, the optimal velocity state $\vec{w}^*$ is well defined and continuous at P_k whether $\vec{u}^*(P_k)$ is defined or not. So if one does want the optimal control $\vec{u}^*$ to be defined at P_k, when $Bl(P_k)=Bu(P_k)$, $k=1,\dots,m$, it is clear that $\vec{u}^*(P_k)$ is defined at each P_k. When $Bl(P_k)<Bu(P_k)$ for some $k=1,\dots,m$, then we have the following necessary and sufficient condition.

THEOREM 3.2. *Let $Bl(P_k)<Bu(P_k)$ for some $k=1,\dots,m$, then the optimal control $\vec{u}^*$ is well defined at the observation points P_k if and only if*

$$|(\vec{w}(P_k,\vec{u}^*)-\vec{Z}_k)_i|\le 2|(\vec{w}(P_k,\vec{u}^*)-\vec{Z}_k)_j|,\quad 1\le i\ne j\le 3, \tag{3.3}$$

where for each fixed k and i, the equality holds for at most one $j\ne i$ unless

$$\vec{w}(P_k,\vec{u}^*)=\vec{Z}_k.$$

When $\vec{u}^*$ is well defined at P_k, we have

$$(\vec{u}^*(P_k))_i=\begin{cases} Bl_i(P_k) & \text{if}\quad (\vec{w}(P_k,\vec{u}^*)-\vec{Z}_k)_i<0,\\ Bu_i(P_k) & \text{if}\quad (\vec{w}(P_k,\vec{u}^*)-\vec{Z}_k)_i>0.\end{cases} \tag{3.4}$$

Proof. If we observe the fundamental velocity solution, we can see that the proof follows from the following argument. For $x=(x_1,x_2,x_3)$ and $1\le i,j,k\le 3$,

$$\begin{aligned}\lim_{x\to 0}\bar{e}_i(x)&=\lim_{x\to 0}\left(c_i\left(\frac{1}{|x|}+\frac{x_i^2}{|x|^3}\right)+c_j\frac{x_ix_j}{|x|^3}+c_k\frac{x_ix_k}{|x|^3}\right)\\&=\lim_{x\to 0}\frac{1}{|x|^3}\left((c_ix_i^2+c_jx_ix_j+c_ix_j^2)+(c_ix_i^2+c_kx_ix_k+c_ix_k^2)\right)\end{aligned}$$

exists (including $\pm\infty$) if and only if

$$c_j^2 - 4c_i^2 \le 0 \quad \text{and} \quad c_k^2 - 4c_i^2 \le 0, \tag{3.5}$$

where at most one equality can hold unless $c_i = c_j = c_k = 0$. Notice that when (3.5) holds, $c_i = 0$ leads to $c_j = c_k = 0$. So if $c_i \neq 0$ and two equalities hold in (3.5), then

$$\bar{e}_i(x) = \frac{c_i}{|x|^3}\left((x_i \pm x_j)^2 + (x_i \pm x_k)^2\right).$$

We can make the limit either equal to zero by taking $x_i = \mp x_j = \mp x_k \to 0$ or equal to sign $(c_i)\infty$ by taking $x_i \neq \mp x_j$ or $x_i \neq \mp x_k$ and $x \to 0$. So the limit will not exist. When $\lim_{x\to 0} \bar{e}_i(x)$ exists and $c = (c_1, c_2, c_3) \neq 0$, we have

$$\lim_{x\to 0} \bar{e}_i(x) = \text{sign } (c_i)\infty. \qquad \square$$

With the above result and the singularity decomposition formula for the optimal control, the following continuous result can be easily verified.

THEOREM 3.3. *Let Bu and Bl be continuous on Γ_1. If for each $k = 1, \ldots, m$ either $Bl(P_k) = Bu(P_k)$ or the condition* (3.3) *holds strictly with $(\vec{w}(P_k, \vec{u}_p^0) - \vec{Z}_k)_i \neq 0$, then the optimal control $\vec{u}^*$ is continuous on Γ_1. So the equality in* (2.4) *holds for every point on Γ.*

From the state-feedback characterization (2.20), the control can be determined by a physical measurement of the state at a finite number of observation points $P_k, k = 1, \ldots, m$. The question is then asked, "will a small error in the measurement of the state cause a large deviation in the control?" Due to the appearance of the singular term in (2.20), in general the answer is yes; i.e., the state-feedback system is not stable. However, under certain conditions, we can prove that the state-feedback system is uniformly stable.

THEOREM 3.4. *Let $\vec{w}(P_k)$ be the exact velocity state at observation points and $\vec{u}_p$ be the control determined from* (2.20) *in terms of $\vec{w}(P_k)$. If for each $k = 1, \ldots, m$ either Bl and Bu are continuous and equal at P_k or Bu and Bl are locally bounded at P_k, the condition* (3.3) *holds strictly with $(\vec{w}(P_k, \vec{u}_p^0) - \vec{Z}_k)_i \neq 0$. Then the state-feedback system* (2.20) *is uniformly stable in the sense that for any $\varepsilon > 0$, there is $\delta > 0$ such that for any measurement $\vec{w}'(P_k)$ of $\vec{w}(P_k)$,*

$$|\vec{u}'(x) - \vec{u}(x)| < \varepsilon \quad \forall x \in \Gamma \qquad \textit{whenever } |\vec{w}'(P_k) - \vec{w}(P_k)| < \delta,$$

where $\vec{u}'$ is the control determined from (2.20) *in terms of $\vec{w}'(P_k)$.*

Proof. For each $\varepsilon > 0$. For each fixed $k = 1, \ldots, m$, if Bl and Bu are continuous and equal at P_k, there is $d_k' > 0$ s.t.

$$Bu(x) - Bl(x) < \varepsilon \quad \forall x \in \Gamma_1,\ |x - P_k| \le d_k'.$$

Since the control variable is bounded by Bl and Bu,

$$|\vec{u}'(x) - \vec{u}(x)| < \varepsilon \quad \forall x \in \Gamma_1,\ |x - P_k| \le d_k'.$$

If instead the condition (3.3) holds strictly with $(\vec{w}(P_k, \vec{u}_p^0) - \vec{Z}_k)_i \neq 0$, let $\delta_1 > 0$ be chosen so that when $|\vec{w}'(P_k) - \vec{w}(P_k)| < \delta_1$, condition (3.3) still holds strictly with $(\vec{w}'(P_k, \vec{u}_p^0) - \vec{Z}_k)_i \neq 0$. Due to the singular term in (2.20) and since Bu and Bl are

locally bounded at P_k, there is $d_k > 0$ such that when $x \in \Gamma$ and $|x - P_k| < d_k$ for some $k = 1, \ldots, m$, we have

$$\left(-\frac{1}{\gamma} \left(\frac{1}{2} I + \mathcal{K} \right)^{-1} \left[\sum_{k=1}^{m} \mu_k E(P_k, \cdot)(\vec{w}'(P_k) - \vec{Z}_k) \right](x) + \vec{a}' + \vec{b}' \times x \right)_i$$

either $> Bu(x)_i$

or $< Bl(x)_i$.

After the truncation by Bu and Bl, it follows that

$$\vec{u}'(x)_i = \vec{u}(x)_i = \text{ either } Bu(x)_i \text{ or } Bl(x)_i \quad \forall x \in \Gamma_1, |x - P_k| < d_k.$$

So if we define

$$\Gamma_+ = \{x \in \Gamma |\ |x - P_k| < \min\{d_k', d_k, k = 1, \ldots, m\} \text{ for some } k = 1, \ldots, m\},$$

then in either case we have

$$|\vec{u}'(x) - \vec{u}(x)| < \varepsilon \quad \forall x \in \Gamma_+.$$

Denote

$$\vec{F}(x) = -\frac{1}{\gamma} \left(\frac{1}{2} I + \mathcal{K} \right)^{-1} \left(\sum_{k=1}^{m} \mu_k E(P_k, \cdot)(\vec{w}(P_k) - \vec{Z}_k) \right)(x),$$

$$\vec{F}'(x) = -\frac{1}{\gamma} \left(\frac{1}{2} I + \mathcal{K} \right)^{-1} \left(\sum_{k=1}^{m} \mu_k E(P_k, \cdot)(\vec{w}'(P_k) - \vec{Z}_k) \right)(x)$$

and

$$\text{meas}\,(\Gamma_{C_F}) = \sum_{i=1}^{3} \text{meas}\,\{x \in \Gamma \mid Bl_i(x) < (\vec{F}(x) + \vec{a} + \vec{b} \times x)_i < Bu_i(x)\},$$

$$\text{meas}\,(\Gamma_{C_{F'}}) = \sum_{i=1}^{3} \text{meas}\,\{x \in \Gamma \mid Bl_i(x) < (\vec{F}'(x) + \vec{a}' + \vec{b}' \times x)_i < Bu_i(x)\}.$$

Since $\text{meas}\,(\Gamma_{C_F}) + \text{meas}\,(\Gamma_{C_{F'}}) = 0$ implies that

$$\vec{u}'(x)_i = \vec{u}(x)_i = \text{ either } Bu(x)_i \text{ or } Bl(x)_i \quad \forall\, x \in \Gamma,$$

there is nothing to prove. So we assume that $\text{meas}\,(\Gamma_{C_F}) + \text{meas}\,(\Gamma_{C_{F'}}) > 0$, and then Theorem 2.3 can be applied. For $x \in \Gamma_- = \Gamma \setminus \Gamma_+$, a compact set, by using (2.20) and triangle inequality, we obtain

$$\begin{aligned} |\vec{u}'(x) - \vec{u}(x)| &= \left| \left[\vec{F}(x) + \vec{a} + \vec{b} \times x \right]_{Bl}^{Bu} - \left[\vec{F}'(x) + \vec{a}' + \vec{b}' \times x \right]_{Bl}^{Bu} \right| \\ &\le \left| -\frac{1}{\gamma} \left(\frac{1}{2} I + \mathcal{K} \right)^{-1} \left[\sum_{k=1}^{m} \mu_k E(P_k, \cdot)(\vec{w}'(P_k) - \vec{w}(P_k)) \right](x) \right| \\ &\qquad + |\vec{a}' + \vec{b}' \times x - (\vec{a} + \vec{b} \times x)| \\ &\equiv |I_1(x)| + |I_2(x)|. \end{aligned}$$

Since the operator $(\frac{1}{2}I + \mathcal{K})^{-1}$ is linear and bounded, and the function $E(P_k, \cdot)$ is continuous and bounded on the compact set Γ_-, there is $\delta_2 > 0$ such that

$$|I_1(x)| < \frac{1}{2}\varepsilon \quad \forall x \in \Gamma_- \quad \text{when } |\vec{w}'(P_k) - \vec{w}(P_k)| < \delta_2.$$

As for $I_2(x)$, Theorem 2.3 yields

$$|(\vec{a}', \vec{b}') - (\vec{a}, \vec{b})| \le \gamma \|\vec{F}' - \vec{F}\|_{(L^1(\Gamma))^3},$$

where the constant γ depends only on Γ. Since there is constant C_0 independent of $\vec{w}'(P_k)$ such that

$$\|\vec{F}' - \vec{F}\|_{(L^1(\Gamma))^3} \le C_0 |\vec{w}'(P_k) - \vec{w}(P_k)|, \quad k = 1, \ldots, m,$$

there is $\delta_3 > 0$ such that

$$|I_2(x)| = |\vec{a}' + \vec{b}' \times x - (\vec{a} + \vec{b} \times x)| < \frac{1}{2}\varepsilon \quad \forall x \in \Gamma_- \quad \text{whenever } |\vec{w}'(P_k) - \vec{w}(P_k)| < \delta_3.$$

Finally, for $\delta = \min\{\delta_1, \delta_2, \delta_3\}$, we have

$$|\vec{u}'(x) - \vec{u}(x)| < \varepsilon \quad \forall x \in \Gamma \quad \text{whenever } |\vec{w}'(P_k) - \vec{w}(P_k)| < \delta \quad \text{for } k = 1, \ldots, m.$$

The proof is complete. □

As a final comment, it is worth indicating that although in the problem setting, the governing differential equation—the Stokes equation—is linear, the bound constraint on the control variable introduces a nontrivial nonlinearity into the system. This can be clearly seen in Theorem 2.2. Also, our approach can be adopted to deal with certain nonlinear boundary control problems.

REFERENCES

[1] H. T. Banks and K. Ito, *Structural actuator control of fluid/structure interactions*, in Proc. 33rd Conference on Decision and Control, Tampa, FL, 1994, pp. 283–288.

[2] J. A. Burns and Y. Ou, *Feedback control of the driven cavity problem using LQR designs*, in Proc. 33rd Conference on Decision and Control, Tampa, FL, 1994, pp. 289–294.

[3] E. Casas, *Control of an elliptic problem with pointwise state constraints*, SIAM J. Control Optim., 24 (1986), pp. 1309–1318.

[4] E. Casas, *Boundary control of semilinear elliptic equations with pointwise state constraints*, SIAM J. Control Optim., 31 (1993), pp. 993–1006.

[5] M. Christ, *Lectures on Singular Integral Operators*, CBMS Regional Conf. Ser. in Math. 77, AMS, Providence, RI, 1990.

[6] R. Coifman, A. McIntoch, and I. Mayer, *L'intégral de Cauchy definit un operateur borné sur L^2 pour les curbs Lipschitziennes*, Ann. of Math., 116 (1982), pp. 361–387.

[7] C. Cuvelier, *Optimal control of a system governed by the Navier-Stokes equations coupled with heat equation*, in New Developments in Differential Equations, W. Eckhaus, ed., North-Holland, Amsterdam, 1976, pp. 81–98.

[8] E. J. Dahlberg, C. E. Hening, and G. C. Verchota, *Boundary value problems for the systems of elastostatics in Lipschitz domains*, Duke Math. J., 57 (1988), pp. 795–818.

[9] Z. Ding, *Topics on Potential Theory on Lipschitz Domains and Boundary Control Problems*, Ph.D. dissertation, Department of Mathematics, Texas A&M University, College Station, TX, 1994.

[10] Z. Ding, L. Ji, and J. Zhou, *Constrained LQR problems in elliptic distributed control systems with point observations*, SIAM J. Control Optim., 34 (1996), pp. 264–294.

[11] Z. Ding and J. Zhou, *Constrained LQR problems governed by the potential equation on Lipschitz domain with point observations*, J. Math. Pures Appl., 74 (1995), pp. 317–344.

[12] Z. Ding and J. Zhou, *Constrained LQR problems in elliptic distributed control systems with point observations—convergence results*, Appl. Math. Optim., 36 (1997), pp. 173–201.
[13] E. B. Fabes, M. Jodeid, Jr., and N. M. Riviere, *Potential techniques for boundary value problems on C^1 domains*, Acta Math., 141 (1978), pp. 165–186.
[14] E. B. Fabes, C. E. Henig, and G. C. Verchota, *The Dirichlet problems for the Stokes system on Lipschitz domains*, Duke Math. J., 57 (1988), pp. 769–793.
[15] J. Dugundji and A. Granas, *Fixed Point Theory*, PWN-Polish Scientific Publishers, Warszawa, 1982.
[16] V. Girault and P. A. Raviart, *Finite Element Methods for Navier-Stokes Equations: Theory and Algorithms*, Springer-Verlag, New York, 1986.
[17] M. Gunzburger, *Finite Element Methods for Viscous Incompressible Flows: A Guide to Theory, Practice, and Algorithms*, Academic Press, New York, 1989.
[18] M. Gunzburger, L. Hou, and T. Svobodny, *Boundary velocity control of incompressible flow with an application to viscous drag reduction*, SIAM J. Control Optim., 30 (1992), pp. 167–181.
[19] K. Ito and S. Kang, *A dissipative feedback control synthesis for systems arising in fluid dynamics*, SIAM J. Control Optim., 32 (1994), pp. 831–854.
[20] L. Ji and G. Chen, *Point observation in linear quadratic elliptic distributed control systems*, in Proc. AMS Summer Conference on Control and Identification of Partial Differential Equations, SIAM, Philadelphia, 1993, pp. 155–170.
[21] R. Kress, *Linear Integral Equations*, Springer-Verlag, New York, 1989.
[22] O. A. Ladyzhenskaia, *The Mathematical Theory of Viscous Incompressible Flow*, Gordon and Breach, New York, 1963.
[23] J. L. Lions, *Contrôle optimal des systèms gouvernés par deséquations aux dérivées partielles*, Dunod, Gauthier-Villars, Paris, 1968.
[24] C. E. Lee and H. E. Taylor, *Fiber-optic Fabry-Perot temperature sensor using a low-coherence light source*, J. Lightwave Technology, 9 (1994), pp. 129–134.
[25] V. G. Maz'ya and S. M. Nikol'skij, *Analysis IV: Linear and Boundary Integral Equations*, Encyclopaedia Math. Sci. 27, Springer-Verlag, New York, 1991.
[26] R. Temam, *Navier-Stokes Equations*, North-Holland, New York, 1977.
[27] R. Temam, *Navier-Stokes Equations and Nonlinear Functional Analysis*, SIAM, Philadelphia, 1983.
[28] G.Verchota, *Layer Potentials and Boundary Value Problems for Laplace's Equation on Lipschitz Domains*, Ph.D. thesis, University of Minnesota, Minneapolis, MN, 1982.
[29] M. C. Wang, *Fiber-optic Fabry-Perot Temperature and Dynamic Sensor System Using a Low-Coherence LED Light Source*, Ph.D. dissertation, Texas A&M University, College Station, TX, 1995.
[30] M. T. Wlodarczyk and G. He, *A fiber-optic combustion pressure sensor system for automotive engine control*, Sensors, 11 (1994), pp. 35–42.
[31] P. You and J. Zhou, *Constrained LQR problems in elliptic distributed control systems with point observations—on convergence rates*, SIAM J. Control Optim., 35 (1997), pp. 1739–1754.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 3, pp. 1005–1019, May 1998

011

GLOBALLY AND SUPERLINEARLY CONVERGENT ALGORITHM FOR MINIMIZING A NORMAL MERIT FUNCTION*

ELIJAH POLAK† AND LIQUN QI‡

Abstract. In this paper we present two new concepts related to the solution of systems of nonsmooth equations (NE) and variational inequalities (VI). The first concept is that of a *normal merit function*, which summarizes the simple basic properties shared by various known merit functions. In general, normal merit functions are locally Lipschitz, but not differentiable. The second concept is that of a *Newtonian operator*, whose values generalize the concept of the Hessian for normal merit functions. These two concepts are then used to generalize the nonsmooth Newton method for solving the equation $\nabla f(x) = 0$, where f is a normal merit function with $f \in C^1$, to the case where f is only locally Lipschitz and the set-valued inclusion $0 \in \partial f(x)$ needs to be solved. Combining the resulting generalized Newton method with certain first-order methods, we obtain a globally and superlinearly convergent algorithm for minimizing normal merit functions.

Key words. normal merit function, generalized Newton method, first-order algorithm, global convergence, superlinear convergence

AMS subject classifications. 65H10, 90C99

PII. S0363012996310245

1. Introduction.

Throughout this paper, we let $F : \Re^n \to \Re^n$ be a locally Lipschitz continuous mapping.

The NE problem is to find a vector $x \in \Re^n$ such that

$$F(x) = 0. \tag{1.1}$$

Let S be a nonempty closed convex set in $\Re^n$. The VI problem is to find a vector $x \in S$ such that

$$\langle F(x), y - x\rangle \geq 0 \quad \text{for all } y \in S, \tag{1.2}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product in $\Re^n$. When $S = \Re^n$, VI reduces to NE. A comprehensive survey of VI is given in [11]. On the other hand, the VI problem (1.2) can be reformulated in the form of the NE problem (1.1) [19]. There are other problems, such as the nonlinear complementarity problem and the maximal monotone operator problem, which can be reformulated as an NE problem. A survey of NEs is given in [19]. In this paper, by "the problem," we mean either the NE problem or the VI problem.

We propose to solve the NE and VI problems by minimizing a *merit function*, $f : \Re^n \to \Re$, using a new extension of Newton's method. A natural merit function f for the NE problem is the norm function

$$f_n(x) := \frac{1}{2}\langle F(x), F(x)\rangle. \tag{1.3}$$

*Received by the editors March 29, 1996; accepted for publication (in revised form) April 1, 1997. http://www.siam.org/journals/sicon/36-3/30124.html

†Department of Electrical Engineering, University of California at Berkeley, Berkeley, CA 94720 (polak@optimum.eecs.berkeley.edu). This paper was partly written while this author was visiting The University of New South Wales; the visit was partly funded by the Australian Research Council. Additional support was provided by National Science Foundation grant ECS-93-02926.

‡School of Mathematics, The University of New South Wales, Sydney, New South Wales 2052, Australia (l.qi@unsw.edu.au). The research of this author was supported by the Australian Research Council.

In this case, with $f = f_n : \Re^n \to \Re$, the merit function satisfies the following four properties:

(i) $f(x) \geq 0$ for all $x \in \Re^n$;
(ii) $f(x^*) = 0$ if and only if x^* solves the problem (here, NE);
(iii) f is locally Lipschitz;
(iv) f is strictly differentiable with the derivative equal to zero at a solution x^* of the problem.

In general, by Rademacher's theorem, any locally Lipschitz continuous function $H : \Re^n \to \Re^m$ is differentiable almost everywhere, and the B-differential of H at x is given by

$$\partial_B H(x) := \{V \in \Re^{n\times m} \,|\, V = \lim_{x_k \to x} \nabla H(x_k)^T,\ x_k \in \Omega_H\},$$

where $\Omega_H := \{x \in \Re^n \,|\, H \text{ is differentiable at } x\}$ [24]. For any $x \in \Re^n$, $\partial_B H(x)$ is a nonempty compact set consisting of $n \times n$ matrices. On the other hand, the Clarke Jacobian of H [2] at x is defined by

$$\partial H(x) := \text{conv } \partial_B H(x).$$

We will use g to denote ∇f if it exists, and G to denote ∂f. Using this notation, property (iv) can be expressed as follows:

$$G(x^*) \equiv \partial f(x^*) = \partial_B f(x^*) = \{g(x^*)\} = \{0\}.$$

For the VI problem we can use the recently discovered D-gap function [10, 21, 29]. The D-gap function $f_d : \Re^n \to \Re$ is defined by

$$f_d(x) := h_\alpha(x) - h_\beta(x), \tag{1.4}$$

where α and β are arbitrary positive parameters such that $\alpha < \beta$ and h_α is the regularized gap function

$$h_\alpha(x) := \max_{y\in S} \left\{ \langle F(x), x - y\rangle - \frac{\alpha}{2}\|x - y\|^2 \right\}. \tag{1.5}$$

(The function h_β is defined similarly with α replaced by β.) The D-gap function f_d also satisfies properties (i)–(iv) [10, 21, 29].

We call a merit function for an NE or VI problem a *normal merit function* if it satisfies properties (i)–(iv) above, as well as the following fifth property:

(v) there exists an auxiliary function $p : \Re^n \times \Re^{n\times n} \to \Re^n$ such that for any $x \in \Re^n$ and any $a \in G(x)$ there is $V \in \partial F(x)$ such that

$$a = p(x, V). \tag{1.6}$$

For $f = f_n$, we may let

$$p(x, V) = V^T F(x). \tag{1.7}$$

In particular, when F is smooth, $a = \nabla f_n(x)$, $V = F_x(x)$, and (1.6) holds automatically.

Thus we see that the norm function is a normal merit function for the NE problem. In section 3, we will show that the D-gap function also satisfies property (v). Hence it is a normal merit function for the VI problem.

The results in this paper remain true if we replace the definition $G := \partial f$ by $G := \partial_B f$ and the $V \in \partial F(x)$ by a $V \in \partial_B F(x)$ in (1.6).

Note that most merit functions for the nonlinear complementarity problem can be classified as special cases of the norm function for NE or the D-gap function for VI. Many recently developed algorithms for the nonlinear complementarity problem are based on the norm function of the Fischer NE function or the Pang min NE function of the nonlinear complementarity problem [3, 5, 6, 7, 8, 9, 13, 14, 16, 19, 28], while the implicit Lagrangian of the nonlinear complementarity problem [17] can be regarded as a special case of the D-gap function for VI with $S = \Re^n_+$ [15, 16].

In this paper we will consider the problem of minimizing a normal merit function f for either NE and VI. Thus we need to solve

$$\min_{x \in \Re^n} \quad f(x). \tag{1.8}$$

When f is smooth, the associated stationarity problem for (1.8) is to find an x such that

$$g(x) = 0. \tag{1.9}$$

Note that (1.9) turns out to be a system of NE, since g, in general, is not differentiable. A class of superlinearly convergent generalized Newton methods for solving NE has been developed in [24], [26], and [19]. Suppose that g is locally Lipschitz continuous. The generalized Newton method for solving (1.9) is of the form

$$x_{k+1} = x_k - V_k^{-1} g(x_k), \tag{1.10}$$

where V_k is an $n \times n$ symmetric matrix, which belongs to either $\partial g(x_k)$ or $\partial_B g(x_k)$. Note that all matrices in $\partial g(x)$ are symmetric. If x^* is a solution of (1.9), g is semismooth at x^*, and all the matrices in $\partial g(x^*)$ or $\partial_B g(x^*)$ (dependent upon the choice of V_k in (1.10)) are nonsingular, then the generalized Newton method (1.10) is locally superlinearly convergent at x^*. Furthermore if g is strongly semismooth at x^*, this convergence is quadratic [24, 26, 19].

However, if f is the D-gap function f_d for VI, even when F is smooth, the above generalized Newton method has difficulties:

i) It depends on the Rademacher theorem for the construction of ∂g or $\partial_B g$. Since the Rademacher theorem does not apply in infinite dimensional spaces, the above generalized Newton method cannot be extended to infinite dimensional spaces.

ii) Even if g is locally Lipschitz continuous, as required by the Rademacher theorem, and the problem is finite dimensional, as in this paper, it is difficult to calculate a V in either $\partial g(x)$ or $\partial_B g(x)$ when $g = \nabla f_d$ since there are no exact calculus rules for the Clarke generalized Jacobian and the B-differential.

iii) It follows from the definitions of ∂g and $\partial_B g$ that the computation of a V in either $\partial g(x)$ or $\partial_B g(x)$ when $g = \nabla f_d$ involves the computation of the second derivative of F or the generalized Jacobian of ∇F. This is not practical even if $\nabla^2 F$ exists. When F is only continuously differentiable, the generalized Jacobians of ∇F, $\partial(\nabla F)$, and $\partial_B(\nabla F)$ do not exist at all.

In [27], some approximate substitutes called computable generalized Jacobians are introduced to overcome the above difficulties. Using these computable generalized Jacobians, a globally and superlinearly convergent algorithm for minimizing the D-gap function for the VI problem is presented in [27] when F is smooth and S is defined by some twice-smooth convex functions. In that algorithm, the local method is the generalized Newton method while the global part of that algorithm is a trust region method. Notably, the generalized Newton method for solving (1.9) with $g = \nabla f_d$

in [27] does not require that ∇F or g are locally Lipschitz continuous. The D-gap function is further studied in [15].

When F is only locally Lipschitz, the stationarity problem associated with (1.8) is to find an x such that

$$0 \in G(x). \tag{1.11}$$

So far, there are no generalized Newton methods which solve a system of linear equations as a subproblem, for solving such a set-valued inclusion problem.

In this paper, we will introduce a new operator, called a *Newtonian operator*. For any $x \in \Re^n$, the value of the Newtonian operator, $T(x)$, is a compact set of $n \times n$ symmetric matrices. We will call to the value of the Newtonian operator a *Newtonian*. The calculation of a Newtonian does not involve any second-order properties of F, and Newtonian operators exist even when F is only locally Lipschitz. We will construct a local, superlinearly convergent generalized Newton method, based on our Newtonian operator, for solving both (1.9) and (1.11). Then we will construct a general globally convergent model of some first-order algorithms for solving (1.9) and (1.11). Combining the generalized Newton method and the general model of first-order methods, we will present a general globally and superlinearly convergent algorithm for minimizing a normal merit function in this class.

We use $\|\cdot\|$ for the 2-norms and denote $\mathsf{N} = \{0, 1, 2, \ldots\}$.

2. Newtonian operators and a generalized Newton method. We now formally define the *Newtonian operator*.

DEFINITION 2.1. *Suppose that F is locally Lipschitz and f is a normal merit function. Let $G = \partial f$. Suppose that T is a set-valued operator mapping from $\Re^n$ to $\Re^{n\times n}$. We say that T is a* Newtonian operator *of f if for any x^* solving the problem, there is a neighborhood $U(x^*)$ of x^* such that*

i) *T is upper semicontinuous at x^*;*

ii) *for any $x \in U(x^*)$, $T(x)$ is a nonempty compact set of $n \times n$ symmetric matrices;*

iii) *there is an* auxiliary*, possibly set-valued,* operator *q mapping from $\Re^n \times \Re^{n\times n}$ to $\Re^{n\times n}$ such that for any $x \in U(x^*)$ and $W \in T(x)$ there is a $V \in \partial F(x)$ such that $W \in q(x, V)$;*

iv) *for $d \in \Re^n$ small enough, any $V \in \partial F(x^* + d)$ and any $W \in q(x^* + d, V)$ ($W = q(x^* + d, V)$ if the latter is single-valued),*

$$p(x^* + d, V) - Wd = o(\|d\|), \tag{2.1}$$

where p is the auxiliary function in the definition of the normal merit function. We call T a strong Newtonian operator of f if instead of (2.1) *we have*

$$p(x^* + d, V) - Wd = O(\|d\|^2). \tag{2.2}$$

If F is smooth, we may combine iii) and iv) in the above definition as

iii)$'$ for $d \in \Re^n$ small enough and any $W \in T(x^* + d)$,

$$g(x^* + d) - Wd = o(\|d\|).$$

Recall [26, 25] that a locally Lipschitz continuous function F is said to be *semismooth* at $x \in \Re^n$ if and only if F is directionally differentiable at x and for any $V \in \partial F(x + d)$,

$$F'(x; d) = Vd + o(\|d\|). \tag{2.3}$$

If F is semismooth at x, then for any $V \in \partial F(x+d)$,

$$F(x+d) = F(x) + Vd + o(\|d\|). \tag{2.4}$$

We say that F is *strongly semismooth* at x if F is semismooth at x and for any $V \in \partial F(x+d)$,

$$F'(x;d) = Vd + O(\|d\|^2). \tag{2.5}$$

If F is strongly semismooth at x, then for any $V \in \partial F(x+d)$,

$$F(x+d) = F(x) + Vd + O(\|d\|^2). \tag{2.6}$$

PROPOSITION 2.2. *Consider* (1.1). *Suppose that F is semismooth at all solutions of* (1.1). *Let T be defined by $T(x) = \{V^T V \mid V \in \partial F(x)\}$. Then T is a Newtonian operator of f_n, where f_n is the norm function of NE. If F is strongly semismooth at all solutions of* (1.1), *then T is a strong Newtonian operator of f_n.*

Proof. Clearly, properties i) and ii) of Definition 2.1 hold for T. Let $q(x,V) = V^T V$. Then property iii) holds for T. Let $V \in \partial F(x^*+d)$ and $W = q(x^*+d, V)$. Then by (1.7),

$$\begin{aligned} p(x^*+d,V) - Wd &= V^T F(x^*+d) - V^T V d \\ &= V^T[F(x^*+d) - Vd] \\ &= V^T[F(x^*+d) - F(x^*) - Vd] \quad (\text{since } F(x^*) = 0) \\ &= o(\|d\|) \qquad\qquad\qquad\qquad\qquad\qquad (\text{by (2.4)}). \end{aligned}$$

This shows that property iv) of Definition 2.1 holds for T. Hence T is a Newtonian operator of f_n. Similarly, when F is strongly semismooth, we can show that T is a strong Newtonian operator. This completes the proof. ∎

In the next section, we will show that the D-gap function has a Newtonian operator even when F is only locally Lipschitz. Note that Newtonian operators of a normal merit function are not unique. Suppose that T is a Newtonian operator of a normal merit function f, x_0 is a fixed point in $\Re^n$, and M is a fixed $n \times n$ symmetric matrix. Define $T_0(x) \equiv T(x)$ if $x \neq x_0$ and $T_0(x_0) = T(x_0) \cup \{M\}$. Then T_0 is also a Newtonian operator of f.

We now propose a generalized Newton method for solving (1.11) as follows:

$$x_{k+1} = x_k - W_k^{-1} p(x_k, V_k), \tag{2.7}$$

where $V_k \in \partial F(x_k)$ and $W_k \in q(x_k, V_k)$.

Note that, even when F is smooth, (2.7) is more general than (1.10).

For any $x_0 \in \Re^n$ and $\delta > 0$, let $N(x_0;\delta) = \{x \in \Re^n : \|x - x_0\| \leq \delta\}$. The proof of the following lemma follows directly [26] from the first two properties of a Newtonian operator.

LEMMA 2.3. *Suppose that x^* is a solution of the problem. If all $W \in T(x^*)$ are nonsingular, then there are $c > 0$ and $\delta > 0$ such that for all $x \in N(x^*;\delta)$, all $W \in T(x)$ are nonsingular and*

$$\|W^{-1}\| \leq c. \tag{2.8}$$

We now can state the superlinear convergence theorem for our generalized Newton method (2.7).

THEOREM 2.4. *Suppose that T is a Newtonian operator of a normal merit function f, x^* is a solution of the problem, and all $W \in T(x^*)$ are nonsingular. Then the generalized Newton method* (2.7) *converges to x^* superlinearly in a neighborhood of x^*. Furthermore if T is a strong Newtonian operator of f, then this convergence is quadratic.*

Proof. For x close to x^* enough, we have

$$\begin{aligned}\|x_{k+1} - x^*\| &= \|x_k - x^* - W_k^{-1} p(x_k, V_k)\| && \text{(by (2.7))}\\ &\le \|W_k^{-1}\| \cdot \|p(x_k, V_k) - W_k(x_k - x^*)\| \\ &\le c\|p(x_k, V_k) - W_k(x_k - x^*)\| && \text{(by (2.8))}\\ &= o(\|x_k - x^*\|) && \text{(by (2.1)).}\end{aligned}$$

This shows superlinear convergence. Quadratic convergence can be proved similarly under the condition that T is a strong Newtonian. This completes the proof. ∎

Combining Proposition 2.2 and Theorem 2.4 and noting the equivalence between nonsingularity of all matrices in $\partial F(x^*)$ and nonsingularity of all matrices in $T(x^*)$, we have the following corollary.

COROLLARY 2.5. *Suppose that F is locally Lipschitz on $\Re^n$ and semismooth at a solution x^* of* (1.1). *Suppose that all matrices in $\partial F(x^*)$ are nonsingular. Then in a neighborhood of x^*, the Gauss–Newton method*

$$x_{k+1} = x_k - (V_k^T V_k)^{-1} V_k^T F(x_k),$$

where $V_k \in \partial F(x_k)$, converges to x^* superlinearly. If F is strongly semismooth at x^*, then this convergence is quadratic.

If we use ∂_B instead of ∂ throughout our discussion, then we have the following result.

COROLLARY 2.6. *Suppose that F is locally Lipschitz on $\Re^n$ and semismooth at a solution x^* of* (1.1). *Suppose that all matrices in $\partial_B F(x^*)$ are nonsingular. Then in a neighborhood of x^*, the Gauss–Newton method*

$$x_{k+1} = x_k - (V_k^T V_k)^{-1} V_k^T F(x_k),$$

where $V_k \in \partial_B F(x_k)$, converges to x^* superlinearly. If F is strongly semismooth at x^*, then this convergence is quadratic.

3. The VI problem. We now consider the D-gap function f_d, defined by (1.4) and (1.5). By [29], it satisfies properties (i)–(iv) of a normal merit function.

Assume that F is locally Lipschitz on $\Re^n$. By [27], if F is differentiable at x, then

$$\begin{aligned} g(x) &\equiv \nabla f_d(x) \\ &= \nabla F(x)(y_\beta(x) - y_\alpha(x)) + (\beta - \alpha)x + (\alpha y_\alpha(x) - \beta y_\beta(x)),\end{aligned}$$

where $y_\alpha(x) = \Pi_S(x - \alpha^{-1}F(x))$, $y_\beta(x) = \Pi_S(x - \beta^{-1}F(x))$, and Π_S is the projection operator on the set S. From this, we have

$$G(x) = \partial f_d(x) = \{p(x, V) \mid V \in \partial F(x)\},$$

where

$$p(x, V) = (V^T - \alpha I)(x - y_\alpha(x)) - (V^T - \beta I)(x - y_\beta(x)).$$

Hence property (v) of a normal merit function is also satisfied. This shows that f_d is a normal merit function of the VI problem.

Throughout the rest of this paper, we assume that

$$S = \{y \in \Re^n \mid h_i(y) \leq 0, i = 1, \ldots, m\}, \tag{3.1}$$

where each h_i is twice-continuously differentiable and convex. In [27], a computable generalized Hessian of the D-gap function was constructed for such a set S when F is smooth. By Lemma 4.1 of [27], such a computable generalized Hessian is in fact a Newtonian operator of the D-gap function. We now extend this result to the case when F is only locally Lipschitz on $\Re^n$ and semismooth at solutions.

Let $x \in \Re^n$ and $\bar{y} = \Pi_S(x)$. Then $\bar{y}$ is the unique solution of the following nonlinear programming problem in y :

$$\begin{array}{ll} \min & \dfrac{1}{2}\|y - x\|^2 \\ \text{subject to} & h_i(y) \leq 0, \quad i = 1, \ldots, m. \end{array} \tag{3.2}$$

Let $\mathcal{M}(x)$ denote the set of multipliers $\lambda \in \Re^m$ that satisfy KKT optimality conditions for (3.2) at $\bar{y}$:

$$\begin{gathered} \bar{y} - x + \sum_{i=1}^{m} \lambda_i \nabla h_i(\bar{y}) = 0, \\ \lambda_i \geq 0, \quad h_i(\bar{y}) \leq 0, \quad \lambda_i h_i(\bar{y}) = 0, \quad i = 1, \ldots, m. \end{gathered} \tag{3.3}$$

For any $y \in \Re^n$, we will denote the active set by

$$I(y) = \{i \mid h_i(y) = 0\}.$$

For a nonnegative vector $d \in \Re^m$, we let supp(d), called the support of d, be the subset of $\{1, \ldots, m\}$ consisting of the indices i for which $d_i > 0$. Let $\mathcal{C}(x)$ be a family of subsets of $\{1, \ldots, m\}$ defined as follows: $J \in \mathcal{C}(x)$ if and only if

$$\text{supp}\,(\lambda) \subseteq J \subseteq I(\bar{y}) \tag{3.4}$$

for some $\lambda \in \mathcal{M}(x)$. We say that the linear independence constraint qualification (LICQ) holds at $\bar{y}$ if the family of vectors

$$\{\nabla h_i(\bar{y}) \mid i \in I(\bar{y})\}$$

are linearly independent. We say that the constant rank constraint qualification (CRCQ) holds at $\bar{y}$ if there exists a neighborhood $N(\bar{y})$ of $\bar{y}$ such that for every set $J \subseteq I(\bar{y})$, the family of gradient vectors

$$\{\nabla h_i(y) \mid i \in J\}$$

has the same rank (which depends on J) for all vectors $y \in N(\bar{y})$ [12, 20]. Clearly, CRCQ is weaker than LICQ.

Let $\mathcal{B}(x)$ be the subfamily of $\mathcal{C}(x)$ such that $J \in \mathcal{B}(x)$ if and only if $J \in \mathcal{C}(x)$ and the vectors

$$\{\nabla h_i(\bar{y}) \mid i \in J\} \tag{3.5}$$

are linearly independent. Here, we allow the empty index set to be a member of $\mathcal{B}(x)$ and $\mathcal{C}(x)$.

For any $x \in \Re^n$ where CRCQ holds, we define

$$\Lambda_S(x) := \{P(x;J) \mid J \in \mathcal{B}(x)\}, \tag{3.6}$$

where $P(x;\emptyset) = I$ and if $J \neq \emptyset$,

$$P(x;J) = C^{-1} - C^{-1}D\left(D^TC^{-1}D\right)^{-1}D^TC^{-1}, \tag{3.7}$$

with

$$C \equiv C(x;J) \equiv I + \sum_{i=1}^{m} \lambda_i(x;J)\nabla^2 h_i(\Pi_S(x)), \quad D \equiv D(x;J) \equiv \nabla h_J(\Pi_S(x)), \tag{3.8}$$

and $\lambda = \lambda(x;J)$ is the multiplier used in (3.4) for the definition of J. Notice that the matrix $C = C(x,J)$ does not depend on the multipliers under CRCQ, at least as it operates on critical directions [20]. Since the vectors in (3.5) are linearly independent, by (3.3), $\lambda = \lambda(x;J)$ is uniquely determined by x and J.

Summarizing Lemmas 2.1–2.3 of [27], we have the following lemma.

LEMMA 3.1. *If CRCQ holds at $\bar{y} = \Pi_S(x)$, then there exists a neighborhood $N(x)$ of x, and for each $J \in \mathcal{B}(x)$ there exists a function $y(\cdot;J)$, such that*

(i) *for all $J \in \mathcal{B}(x)$, both $y(\cdot;J)$ and $\lambda(\cdot;J)$ are continuously differentiable in $N(x)$; and for all $z \in N(x)$,*

(ii) *CRCQ holds at $\Pi_S(z)$;*

(iii) *$\mathcal{B}(z) \subseteq \mathcal{B}(x)$;*

(iv) *$\Pi_S(z) = y(z;J)$ for all $J \in \mathcal{B}(z)$;*

(v) *$P(z;J) = \nabla y(z;J)$ is symmetric positive semidefinite and $\|P(z;J)\| \leq 1$, for all $J \in \mathcal{B}(z)$.*

The following theorem follows from the above lemma.

THEOREM 3.2. *If CRCQ holds at $\bar{y} = \Pi_S(x)$, then for any $P \in \Lambda_S(x+d)$,*

$$\Pi_S(x+d) = \Pi_S(x) + Pd + o(\|d\|). \tag{3.9}$$

Furthermore if all $\nabla^2 h_i$ are locally Lipschitz, then

$$\Pi_S(x+d) = \Pi_S(x) + Pd + O(\|d\|^2).$$

Proof. By Lemma 3.1, for $\|d\|$ small enough and $P \in \Lambda_S(x+d)$,

$$\Pi_S(x+d) - \Pi_S(x) - Pd = y(x+d;J) - y(x;J) - \nabla y(x;J)d$$

for some $J \in \mathcal{B}(x)$. Hence, we have our first conclusion.

If all $\nabla^2 h_i$ are locally Lipschitz, then by (3.7) and (3.8), $P(\cdot;J)$ is locally Lipschitz. The second conclusion follows now. □

We now can state the main result in this section.

THEOREM 3.3. *Consider* (1.2). *Suppose that F is semismooth and CRCQ holds at all solutions of* (1.2). *Let T be defined by $T(x) = \{q(x,V) \mid V \in \partial F(x)\}$, where*

$$\begin{aligned} & q(x,V) \\ & = (V^T - \alpha I)(I - \Lambda_S(x - \alpha^{-1}F(x))(I - \alpha^{-1}V)) \\ & -(V^T - \beta I)(I - \Lambda_S(x - \beta^{-1}F(x))(I - \beta^{-1}V)). \end{aligned}$$

Then T is a Newtonian operator of f_d, where f_d is the D-gap function of the VI problem. Furthermore if all $\nabla^2 h_i$ are locally Lipschitz and F is strongly semismooth at all solutions of (1.2)*, then T is a strong Newtonian operator of f_d.*

Proof. Suppose that x^* is a solution of (1.2). Then

$$x^* = \Pi_S(x^* - \alpha^{-1}F(x^*)) = \Pi_S(x^* - \beta^{-1}F(x^*)). \tag{3.10}$$

By Lemma 3.1, Λ_S is upper semicontinuous at x^*. Since ∂F is a closed operator, T is upper semicontinuous at x^*. This proves that property i) of the definition of a Newtonian operator holds for T.

By Lemma 3.1 and (3.10), for x in a neighborhood of x^*, $\Lambda_S(x - \alpha^{-1}F(x))$ and $\Lambda_S(x - \beta^{-1}F(x))$, are nonempty, compact sets of $n \times n$ symmetric matrices. Then, property ii) of the definition of a Newtonian operator follows, for T, from the expression of $q(x, V)$.

By the definition of $q(x, V)$, property iii) of the definition of a Newtonian operator holds.

Let $V \in \partial F(x^* + d)$ and $W \in q(x^* + d, V)$. Then there are $P_\alpha \in \Lambda_S(x^* + d - \alpha^{-1}F(x^* + d))$ and $P_\beta \in \Lambda_S(x^* + d - \beta^{-1}F(x^* + d))$ such that

$$W = (V^T - \alpha I)(I - P_\alpha(I - \alpha^{-1}V)) - (V^T - \beta I)(I - P_\beta(I - \beta^{-1}V)).$$

For $\|d\|$ small enough, we have

$$\begin{aligned}
&p(x^* + d, V) - Wd \\
&= (V^T - \alpha I)[x^* + d - \Pi_S(x^* + d - \alpha^{-1}F(x^* + d)) - (I - P_\alpha(I - \alpha^{-1}V))d] \\
&-(V^T - \beta I)[x^* + d - \Pi_S(x^* + d - \beta^{-1}F(x^* + d)) - (I - P_\beta(I - \beta^{-1}V))d] \\
&= (V^T - \alpha I)[\Pi_S(x^* - \alpha^{-1}F(x^*)) - \Pi_S(x^* + d - \alpha^{-1}F(x^* + d)) + P_\alpha(d - \alpha^{-1}Vd)] \\
&-(V^T - \beta I)[\Pi_S(x^* - \beta^{-1}F(x^*)) - \Pi_S(x^* + d - \beta^{-1}F(x^* + d)) + P_\beta(d - \beta^{-1}Vd)] \\
&\quad \text{(by (3.10))} \\
&= (V^T - \alpha I)[\alpha^{-1}P_\alpha(F(x^* + d) - F(x^*) - Vd) + o(\|\alpha^{-1}(F(x^* + d) - F(x^*)) - d\|)] \\
&-(V^T - \beta I)[\beta^{-1}P_\beta(F(x^* + d) - F(x^*) - Vd) + o(\|\beta^{-1}(F(x^* + d) - F(x^*)) - d\|)] \\
&\quad \text{(by (3.9))} \\
&= (V^T - \alpha I)[P_\alpha \cdot o(\|d\|) + o(\|F(x^* + d) - F(x^*)\|) + o(\|d\|)] \\
&-(V^T - \beta I)[P_\beta \cdot o(\|d\|) + o(\|F(x^* + d) - F(x^*)\|) + o(\|d\|)] \\
&\quad \text{(by (2.4))} \\
&= (V^T - \alpha I) \cdot o(\|d\|) - (V^T - \beta I) \cdot o(\|d\|) \\
&\quad \text{(by Lemma 3.1)} \\
&= o(\|d\|),
\end{aligned}$$

where the last equality is due to local boundedness of ∂F. This shows that property iv) of Definition 2.1 holds for T. Hence T is a Newtonian operator of f_d. Similarly we can prove the strong conclusion from the strong condition. This completes the proof. □

Note that f_d will reduce to $(\alpha^{-1} - \beta^{-1})f_n$ if $S = \Re^n$. Hence we may think the generalized Newton method (2.7) applied to f_d as a generalized version of the Gauss–Newton method for solving the VI problem.

4. First-order methods. Our globalization algorithm for the local Newton methods that we have described depends on the use of first-order, unconstrained optimization methods that are locally uniformly cost decreasing. We define this property as follows.

DEFINITION 4.1. *Consider the unconstrained optimization problem*

$$\min_{x \in \Re^n} f(x), \tag{4.1}$$

where f is at least locally Lipschitz continuous. We will say that an algorithm iteration map $A : \Re^n \to 2^{\Re^n}$ is locally uniformly cost decreasing, with respect to (4.1)*, if*

(i) *for any $x \in \Re^n$ such that $0 \notin \partial f(x)$, there exists a $\rho_x > 0$ and a $\delta_x > 0$ such that*

$$f(x'') - f(x') \leq -\delta_x \tag{4.2}$$

for all $x' \in N(x; \rho_x)$ and all $x'' \in A(x')$; and

(ii) *for all x such that $0 \in \partial f(x)$, $A(x) = \{x\}$.*

The most important property of algorithms that are locally uniformly cost decreasing, with respect to problem (4.1), is given below.

THEOREM 4.2. *Suppose that the algorithm iteration map $A : \Re^n \to 2^{\Re^n}$ is locally uniformly cost decreasing with respect to problem* (4.1) *and that, given an $x_0 \in \Re^n$, the sequence $\{x_i\}_{i=0}^{\infty}$ is constructed according to the rule*

$$x_{i+1} \in A(x_i), \quad i = 0, 1, 2, \ldots. \tag{4.3}$$

Then every accumulation point $\hat{x}$ of $\{x_i\}_{i=0}^{\infty}$ is stationary (i.e., $0 \in \partial f(\hat{x})$).

Proof. Without loss of generality, we can assume that $0 \notin \partial f(x_i)$ for all $i \in \mathsf{N}$. Hence it follows from (4.2) that $f(x_{i+1}) < f(x_i)$ for all $i \in \mathsf{N}$, and hence, if $\{x_i\}_{i=0}^{\infty}$ has an accumulation point $\hat{x}$, we must have that $f(x_i) \to f(\hat{x})$, as $i \to \infty$.

For the sake of contradiction, suppose that $0 \notin \partial f(\hat{x})$. Then, since $\hat{x}$ is an accumulation point of $\{x_i\}_{i=0}^{\infty}$, it follows from (4.2) that there exists a $\hat{\delta} > 0$ and a subsequence $\{x_{i_k}\}_{i=0}^{\infty}$ of $\{x_i\}_{i=0}^{\infty}$, such that $x_{i_k} \to \hat{x}$ as $k \to \infty$, and

$$f(x_{i_{k+1}}) - f(x_{i_k}) \leq -\hat{\delta} \tag{4.4}$$

for all $k \in \mathsf{N}$. Since this implies that $f(x_i) \to -\infty$ as $i \to \infty$, we have a contradiction, which completes our proof. □

When the function f in (4.1) is continuously differentiable, we have at least two choices of an algorithm iteration map $A : \Re^n \to 2^{\Re^n}$, which is locally uniformly cost decreasing, with respect to problem (4.1). The first is the iteration map of Armijo gradient method [1]; the other is a class of trust region methods.

THEOREM 4.3. *Suppose that the function f in* (4.1) *is continuously differentiable. Consider the Armijo gradient method iteration map $A : \Re^n \to 2^{\Re^n}$, for problem* (4.1)*, defined by*

$$A(x) := \{x - \lambda(x)\nabla f(x)\}, \tag{4.5}$$

where, for fixed $\alpha, \beta \in (0,1)$,

$$\lambda(x) := \max_{k \in \mathsf{N}} \{\beta^k | f(x - \beta^k \nabla f(x)) - f(x) + \beta^k \alpha \|\nabla f(x)\|^2 \leq 0\}. \tag{4.6}$$

The map A is locally uniformly cost decreasing, with respect to problem (4.1).

Proof. Suppose that $x^* \in \Re^n$ is such that $\nabla f(x) \neq 0$. Let $\varepsilon > 0$ be such that $\alpha + \varepsilon < 1$. Then there exists a $k^* \in \mathsf{N}$ such that

$$f(x^* - \beta^{k^*} \nabla f(x^*)) - f(x^*) \leq -\beta^{k^*}(\alpha + \varepsilon)\|\nabla f(x^*)\|^2. \tag{4.7}$$

Rearranging the terms in (4.7), we obtain that

$$f(x^* - \beta^{k^*} \nabla f(x^*)) + \beta^{k^*} \alpha \|\nabla f(x^*)\|^2 \leq -\beta^{k^*} \varepsilon \|\nabla f(x^*)\|^2. \tag{4.8}$$

It now follows from the continuity of f and of ∇f that there exists a $\rho^* > 0$ such that for all $x \in N(x^*; \rho^*)$, $\lambda(x) \geq \beta^{k^*}$ and $\|\nabla f(x)\|^2 \geq 1/2 \ \|\nabla f(x^*)\|^2$. Hence for all $x \in N(x^*; \rho^*)$,

$$f(A(x)) - f(x) \leq -1/2 \ \beta^{k^*} \alpha \|\nabla f(x^*)\|^2 := -\delta^*, \tag{4.9}$$

which completes our proof. □

Referring to [23], we find that when the function f in (4.1) is continuously differentiable, the following trust region algorithm iteration map is uniformly cost decreasing, with respect to problem (4.1).

Let H be a symmetric $n \times n$ matrix-valued function defined on $\Re^n$, let $\phi : \Re^n \to \Re$ be defined by

$$\phi(h) := \langle \nabla f(x), h \rangle + 1/2 \ \langle h, H(x)h \rangle, \tag{4.10}$$

let $\alpha, \beta \in (0,1)$, $\sigma_{\max} > 0$ be parameters, let $\Delta : \Re^n \times \mathsf{N} \to \Re^n$ be defined by

$$\Delta(x,k) := \min_{\lambda} \{\phi(-\lambda \nabla f(x)) | \lambda \|\nabla f(x)\| \leq \beta^k \sigma_{\max}\}, \tag{4.11}$$

for any $x \in \Re^n$, $k \in \mathsf{N}$, let $G(x,k) \subset \Re^n$ be defined by

$$G(x,k) := \{h \in \Re^n | \phi(h) \leq \Delta(x,k)\}, \tag{4.12}$$

and let

$$(D(x), K(x)) := \arg\max\{\beta^k | k \in \mathsf{N}, h \in G(x,k), f(x+h) - f(x) \leq \alpha\phi(h)\}. \tag{4.13}$$

Then we define the set-valued trust region algorithm iteration map $A : \Re^n \to 2^{\Re^n}$ as follows:

$$A(x) := \{x + h | h \in D(x)\}. \tag{4.14}$$

We infer from Theorem 1.2.29 in [23] the following result.

THEOREM 4.4. *Suppose that the function f in* (4.1) *is continuously differentiable and that there exists a* $c \in (0, \infty)$ *such that* $||H(x)|| \leq c$ *for all* $x \in \Re^n$. *Then the trust region algorithm iteration map A, defined by* (4.14), *is locally uniformly cost decreasing, with respect to* (4.1).

When the function f in (4.1) is only locally Lipschitz continuous, we can use algorithms based on *augmented convergent direction finding maps* that were introduced in [22]. These were defined as follows.

DEFINITION 4.5. *Let $f : \Re^n \to \Re$ be locally Lipschitz continuous. We shall say that $\bar{G}f : \Re^n \to 2^{\Re^{n+1}}$ is an* augmented convergent direction finding *(a.c.d.f.) map for f if*

(a) *$\bar{G}f$ is continuous (i.e., both upper and lower semicontinuous) and $\bar{G}f(x)$ is convex for all $x \in \Re^n$.*

(b) *for any $x \in \Re^n$, if $\bar{\xi} = (\xi^0, \xi) \in \Re^{n+1}$, where $\xi \in \Re^n$, is an element of $\bar{G}f(x)$, then $\xi^0 \geq 0$.*

(c) *for any $x \in \Re^n$, a point $\bar{\xi} = (0, \xi)$ is an element of $\bar{G}f(x)$ if and only if $\xi \in \partial f(x)$, where $\partial f(x)$ denotes the Clarke generalized gradient.*

In [22], we find the following result.

THEOREM 4.6. *Suppose that $f : \Re^n \to \Re$ is locally Lipschitz continuous and $\bar{G}f$ is an a.c.d.f. map for f. Then for any $x \in \Re^n$,*

(a) *$0 \in \partial f(x)$ if and only if $0 \in \bar{G}f(x)$.*

(b) *The functions $\Theta : \Re^n \to \Re$ and $\bar{h} : \Re^n \to \Re^{n+1}$ defined by*

$$\Theta(x) := \min\{\xi^0 + 1/2 \ \|\xi\|^2 | \bar{\xi} = (\xi^0, \xi) \in \bar{G}f(x)\}, \tag{4.15}$$

where $\xi^0 \in \Re$ and $\xi \in \Re^n$, and

$$\bar{h}(x) := \operatorname{argmin} \{\xi^0 + 1/2 \ ||\xi||^2 | \bar{\xi} \in \bar{G}f(x)\} \tag{4.16}$$

are both continuous. Furthermore $\Theta(x) = 0$ if and only if $0 \in \partial f(x)$.

(c) *Writing $\bar{h}(x) = (h^0(x), h(x))$, with $h(x) \in \Re^n$ we have*

$$d_0 f(x; h(x)) \leq -\Theta(x) \quad \forall x \in \Re^n, \tag{4.17}$$

where $d_0 f(x; h)$ denotes the Clarke generalized directional derivative.

The following algorithm map is a generalization of the Armijo gradient method for differentiable functions. See [22, section 5] for a proof of the following result.

THEOREM 4.7. *Suppose that the function f in* (4.1) *is locally Lipschitz continuous and that $\bar{G}f$ is an a.c.d.f. for f. Consider the generalized Armijo gradient method iteration map $A : \Re^n \to 2^{\Re^n}$, for problem* (4.1)*, defined by*

$$A(x) := \{x + \lambda(x)h(x)\}, \tag{4.18}$$

where $h(x)$ is defined by (4.16) *and, for fixed $\alpha, \beta \in (0, 1)$,*

$$\lambda(x) := \max_{k \in \mathsf{N}} \{\beta^k | f(x + \beta^k h(x)) - f(x) + \beta^k \alpha \Theta(x) \leq 0\}, \tag{4.19}$$

with $\Theta(x)$ defined by (4.15)*. Then the map A is locally uniformly cost decreasing with respect to problem* (4.1)*.*

5. A globally and superlinearly convergent algorithm. We now consider problem (1.8) again, where f is a normal merit function for the NE or VI problem, with F locally Lipschitz on $\Re^n$. Suppose that there is an algorithm iteration map $A : \Re^n \to 2^{\Re^n}$, which is locally uniformly cost decreasing, with respect to problem (1.8). Suppose that T is a Newtonian operator of f, with an auxiliary operator q. Then we have the following algorithm.

ALGORITHM 5.1. Let $x_0 \in \Re^n$ and $\epsilon \in (0, 1)$.

For any $k \in \mathsf{N}$, let

$$\hat{x}_k = x_k - W_k^{-1} p(x_k, V_k), \tag{5.1}$$

where $V_k \in \partial F(x_k)$ and $W_k \in q(x_k, V_k)$. If

$$f(\hat{x}_k) \leq \epsilon f(x_k), \tag{5.2}$$

let $x_{k+1} = \hat{x}_k$. Otherwise let $x_{k+1} \in A(x_k)$.

If $x_{k+1} = \hat{x}_k$, we call this a Newton step. To prove the convergence theorem for this algorithm, we need the following lemma.

LEMMA 5.2. *Suppose that x^* is a solution of the problem. If all $W \in T(x^*)$ are positive definite, then there are $c_1 > c_2 > 0$ and $\delta > 0$ such that for all $x \in N(x^*; \delta)$, all $W \in T(x)$ are positive definite and*

$$c_2\|x - x^*\|^2 \le f(x) \le c_1\|x - x^*\|^2. \tag{5.3}$$

Proof. Since all $W \in T(x^*)$ are positive definite, by the four properties of a Newtonian operator, there are $c_1 > 5c_2 > 0$ and $\delta > 0$ such that for all $x \in N(x^*; \delta)$, all $V \in \partial F(x)$ and all $W \in q(x, V)$,

$$2.5c_2\|x - x^*\|^2 \le (x - x^*)^T W(x - x^*) \le 0.5c_1\|x - x^*\|^2 \tag{5.4}$$

and

$$\|p(x, V) - W(x - x^*)\| \le 0.5c_2\|x - x^*\|. \tag{5.5}$$

Now, by the Lebourg mean-value theorem (MVT) (Theorem 2.3.7 of [2]), for any $x \in N(x^*; \delta)$,

$$\begin{aligned} f(x) &= f(x) - f(x^*) && \text{(since } f(x^*) = 0\text{)} \\ &= p(x^* + t(x - x^*), V)^T(x - x^*) && \text{(by MVT)} \\ &\le t(x - x^*)^T W(x - x^*) + 0.5c_2\|x - x^*\|^2 && \text{(by (5.5))} \\ &\le 0.5tc_1\|x - x^*\|^2 + 0.5c_2\|x - x^*\|^2 && \text{(by (5.4))} \\ &\le c_1\|x - x^*\|^2, \end{aligned}$$

where $0 \le t \le 1, V \in \partial F(x^* + t(x - x^*))$ and $W \in q(x^* + t(x - x^*), V)$. Similarly, for any $x \in N(x^*; \delta)$,

$$\begin{aligned} f(x) &= f(x) - f(x^*) \qquad \text{(since } f(x^*) = 0\text{)} \\ &= \lim_{N\to\infty} \sum_{j=1}^{N} \left[f\left(x^* + \frac{j}{N}(x - x^*)\right) - f\left(x^* + \frac{j-1}{N}(x - x^*)\right)\right] \\ &= \lim_{N\to\infty} \sum_{j=1}^{N} \frac{1}{N} p\left(x^* + \frac{j - 1 + t_j}{N}(x - x^*), V_j\right)^T (x - x^*) \qquad \text{(by MVT)} \\ &\ge \lim_{N\to\infty} \sum_{j=1}^{N} \left[\frac{j - 1 + t_j}{N^2}(x - x^*)^T W_j(x - x^*) - \frac{0.5c_2(j - 1 + t_j)}{N^2}\|x - x^*\|^2 \right] \\ &\quad \text{(by (5.5))} \\ &\ge \lim_{N\to\infty} \sum_{j=1}^{N} \left[\frac{j-1}{N^2}(x - x^*)^T W_j(x - x^*) - \frac{0.5c_2 j}{N^2}\|x - x^*\|^2 \right] \\ &\ge \lim_{N\to\infty} \sum_{j=1}^{N} \left[\frac{2.5c_2(j-1)}{N^2} - \frac{0.5c_2 j}{N^2} \right] \|x - x^*\|^2 \qquad \text{(by (5.4))} \\ &\ge \lim_{N\to\infty} \frac{[1.25N(N-1) - 0.25N(N+1)]c_2}{N^2}\|x - x^*\|^2 \\ &= c_2\|x - x^*\|^2, \end{aligned}$$

where $0 \le t_j \le 1, V_j \in \partial F(x^* + \frac{j-1+t_j}{N}(x - x^*))$, and $W_j \in q(x^* + \frac{j-1+t_j}{N}(x - x^*), V_j)$. This proves the lemma. ∎

For the NE problem with T defined by $T(x) = \{V^T V | V \in \partial F(x)\}$, all $W \in T(x^*)$ are positive definite as long as they are nonsingular. For the VI problem, the positive definiteness of matrices in $T(x^*)$ was established by Theorem 3.1 of [27].

We now can state the convergence theorem for Algorithm 5.1.

THEOREM 5.3. *Every accumulation point $\hat{x}$ of $\{x_k\}_{k=0}^{\infty}$, generated by Algorithm* 5.1*, is a stationary point of* (1.8)*, i.e., $0 \in \partial f(\hat{x})$. Function f has the same value on these accumulation points.*

If the Newton step is used infinitely many times, then every accumulation point of $\{x_k\}_{k=0}^{\infty}$ is a solution of the original problem. If at one of such accumulation point, say x^*, all matrices in $T(x^*)$ are positive definite, then the whole sequence $\{x_k\}_{k=0}^{\infty}$ converges to x^* superlinearly with $x_{k+1} = \hat{x}_k$ for all large k. Furthermore if T is a strong Newtonian operator of f, the convergence is quadratic.

Proof. Notice that

$$f(x_{k+1}) \leq f(x_k) \tag{5.6}$$

for all $k \in \mathsf{N}$.

If the Newton step is not used for all large k, then the conclusions follow from Theorem 4.2 and the nonincreasing property (5.6).

If the Newton step is used infinitely many times, then

$$\lim_{k \to \infty} f(x_k) = 0$$

by (5.2) and (5.6). Then every accumulation point of $\{x_k\}_{k=0}^{\infty}$ is a solution of the original problem. Suppose that x^* is such an accumulation point, hence a solution to the original problem. Suppose that all matrices in $T(x^*)$ are positive definite. Let x_k be close to x^*. Then by the proof of Theorem 2.4, we have

$$\|\hat{x}_k - x^*\| = o(\|x_k - x^*\|).$$

By Lemma 5.2, this implies,

$$f(\hat{x}^k) = o(f(x_k)).$$

This step and the following steps have to be Newton steps and the remaining conclusions follow from Theorem 2.4. □

This algorithm may end with a stationary point of (1.8) which is not a solution of the original problem. See page 152 of [4] for description of this phenomenon. An effort has been made in [18] to overcome this difficulty in the case of smooth nonlinear equations. However, if F is strongly monotone and f is the norm function for the NE problem or the D-gap function for the VI problem, any stationary point of f is the unique solution of the original problem [29, 27]. In this case, our algorithm converges to this unique solution of the original problem globally and superlinearly.

Acknowledgments. We are thankful to Houyuan Jiang for his help in preparing the manuscript of this paper and to the associate editor, two referees, and Defeng Sun for their helpful comments.

REFERENCES

[1] L. ARMIJO, *Minimization of functions having Lipschitz continuous first partial derivatives*, Pacific J. Math., 16 (1966), pp. 1–3.

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

[3] T. DE LUCA, F. FACCHINEI, AND C. KANZOW, *A semismooth equation approach to the solution of nonlinear complementarity problems*, Math. Programming, 75 (1996), pp. 407–439.

[4] J. E. Dennis and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice–Hall, Englewood Cliffs, NJ, 1983.
[5] F. Facchinei, A. Fischer, and C. Kanzow, *Inexact Newton methods for semismooth equations with applications to variational inequality problems*, in Nonlinear Optimization and Applications, G. Di Pillo and F. Giannessi, eds., Plenum Press, New York, 1996, pp. 125–139.
[6] F. Facchinei, A. Fischer, and C. Kanzow, *A semismooth Newton method for variational inequalities: The case of box constraints*, in Complementarity and Variational Problems—State of the Art, M. C. Ferris and J.-S. Pang, eds., SIAM, Philadelphia, 1997, pp. 76–90.
[7] F. Facchinei, A. Fischer, and C. Kanzow, *Regularity properties of a semismooth reformulation of variational inequalities*, SIAM J. Optim., 8 (1998), to appear.
[8] F. Facchinei and C. Kanzow, *A nonsmooth inexact Newton method for the solution of large-scale nonlinear complementarity problems*, Math. Programming, 76 (1997), pp. 493–512.
[9] A. Fischer, *An NCP-function and its use for the solution of complementarity problems*, in Recent Advances in Nonsmooth Optimization, D.-Z. Du, L. Qi, and R. S. Wormersley, eds., World Scientific Publishers, Singapore, 1995, pp. 88–105.
[10] M. Fukushima, *Merit functions for variational inequality and complementarity problems*, in Nonlinear Optimization and Applications, G. Di Pillo and F. Giannessi, eds., Plenum Press, New York, 1996, pp. 155–170.
[11] P. T. Harker and J. S. Pang, *Finite-dimensional variational inequality and nonlinear complementarity problems: A survey of theory, algorithms and applications*, Math. Programming, 48 (1990), pp. 161–220.
[12] R. Janin, *Directional derivative of the marginal function in nonlinear programming*, Math. Programming Study, 21 (1984), pp. 110–126.
[13] H. Jiang and L. Qi, *A new nonsmooth equations approach to nonlinear complementarities*, SIAM J. Control Optim., 35 (1997), pp. 178–193.
[14] C. Kanzow, *Nonlinear complementarity as unconstrained optimization*, J. Optim. Theory Appl., 88 (1996), pp. 139–155.
[15] C. Kanzow and M. Fukushima, *Theoretical and numerical investigation of the D-gap function for box constrained variational inequalities*, Math. Programming, to appear.
[16] Z.-Q. Luo and P. Tseng, *A new class of merit functions for the nonlinear complementarity problems*, in Complementarity and Variational Problems—State of the Art, M. C. Ferris and J.-S. Pang, eds., SIAM, Philadelphia, 1997, pp. 204–225.
[17] O. L. Mangasarian and V. Solodov, *Nonlinear complementarity as unconstrained and constrained minimization*, Math. Programming, 62 (1993), pp. 227–297.
[18] J. L. Nazareth and L. Qi, *Globalization of Newton's method for solving nonlinear equations*, J. Numerical Algebra Appl., 90 (1996), pp. 653–673.
[19] J. S. Pang and L. Qi, *Nonsmooth equations: Motivation and algorithms*, SIAM J. Optim., 3 (1993), pp. 443–465.
[20] J. S. Pang and D. Ralph, *Piecewise smoothness, local invertibility, and parametric analysis of normal maps*, Math. Oper. Res., 21 (1996), pp. 401–426.
[21] J. M. Peng, *Equivalence of variational inequality problems to unconstrained optimization*, Math. Programming, 78 (1997), pp. 347–355.
[22] E. Polak, *On the mathematical foundations of nondifferentiable optimization in engineering design*, SIAM Rev., 29 (1987), pp. 21–91.
[23] E. Polak, *Optimization: Algorithms and Consistent Approximations*, Springer-Verlag, New York, 1997.
[24] L. Qi, *Convergence analysis of some algorithms for solving nonsmooth equations*, Math. Oper. Res., 18 (1993), pp. 227–244.
[25] L. Qi and H. Jiang, *Semismooth Karush-Kuhn-Tucker equations and convergence analysis of Newton methods and quasi-Newton methods for solving these equations*, Math. Oper. Res., 22 (1997), pp. 301–325.
[26] L. Qi and J. Sun, *A nonsmooth version of Newton's method*, Math. Programming, 58 (1993), pp. 353–367.
[27] D. Sun, M. Fukushima, and L. Qi, *A computable generalized Hessian of the D-gap function and Newton-type methods for variational inequality problem*, in Complementarity and Variational Problems—State of the Art, M.C. Ferris and J.-S. Pang, eds., SIAM, Philadelphia, 1997, pp. 452–473.
[28] N. Yamashita and M. Fukushima, *Modified Newton methods for solving a semismooth reformulation of monotone complementarity problems*, Math. Programming, 76 (1997), pp. 469–491.
[29] N. Yamashita, K. Taji, and M. Fukushima, *Unconstrained optimization reformulations of variational inequality problems*, J. Optim. Theory Appl., 92 (1997), pp. 439–456.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 3, pp. 1020–1047, May 1998

012

ERGODIC BOUNDARY/POINT CONTROL OF STOCHASTIC SEMILINEAR SYSTEMS*

T. E. DUNCAN†, B. MASLOWSKI‡, AND B. PASIK-DUNCAN†

Abstract. A controlled Markov process in a Hilbert space and an ergodic cost functional are given for a control problem that is solved where the process is a solution of a parameter-dependent semilinear stochastic differential equation and the control can occur only on the boundary or at discrete points in the domain. The linear term of the semilinear differential equation is the infinitesimal generator of an analytic semigroup. The noise for the stochastic differential equation can be distributed, boundary and point. Some ergodic properties of the controlled Markov process are shown to be uniform in the control and the parameter. The existence of an optimal control is verified to solve the ergodic control problem. The optimal cost is shown to depend continuously on the system parameter.

Key words. ergodic control, stochastic semilinear equations, Markov processes in Hilbert spaces, invariant measures, boundary control

AMS subject classifications. 93E20, 93C20, 60H15

PII. S0363012996303190

1. Introduction.

An ergodic control problem for a stochastic process in a Hilbert space H is formulated and solved where the process is a solution of a parameter-dependent semilinear stochastic differential equation in H. The problem in the general setting is motivated by ergodic control problems for processes governed by stochastic partial differential equations (SPDEs) with control and noise occurring in the boundary conditions or at discrete points in the domain.

For example, consider the stochastic parabolic equation

$$\frac{\partial v}{\partial t}(t,\xi) = Lv(t,\xi) + F(\alpha, v(t,\xi)) + n(t,\xi) \tag{1.1}$$

for $(t,\xi) \in \mathbb{R}_+ \times (0,1)$ with initial and boundary conditions

$$v(0,\xi) = v_0(\xi), \tag{1.2}$$

$$\frac{\partial v}{\partial \xi}(t,0) = h_1(\alpha, v(t,\cdot), u(v(t,\cdot))) + \eta_1(t), \tag{1.3}$$

$$\frac{\partial v}{\partial \xi}(t,1) = h_2(\alpha, v(t,\cdot), u(v(t,\cdot))) + \eta_2(t), \tag{1.4}$$

where n denotes a space-dependent Gaussian noise that is white in time, η_1 and η_2 are one-dimensional standard Wiener processes, and these three processes are mutually independent. Furthermore,

$$Lv = a(\xi)\frac{\partial^2}{\partial \xi^2}v + b(\xi)\frac{\partial}{\partial \xi}v + c(\xi)v$$

*Received by the editors May 8, 1996; accepted for publication (in revised form) April 1, 1997. This research was partially supported by National Science Foundation grants DMS 9305936 and DMS 9623439, the Alexander von Humboldt Foundation, and GAČR grant 201/95/0629.

http://www.siam.org/journals/sicon/36-3/30319.html

†Department of Mathematics, University of Kansas, Lawrence, KS 66045-2142 (duncan@math.ukans.edu).

‡Institute of Mathematics, Czech Academy of Sciences, Prague, Czech Republic (maslow@cesnet.cz).

is a second-order uniformly elliptic operator, where $a, b, c \in C^\infty([0,1])$, $a > 0$, $c < 0$, $F : \mathcal{A} \times \mathbb{R} \to \mathbb{R}$, $h_i : \mathcal{A} \times H \times \mathcal{K} \to \mathbb{R}$, $i = 1, 2$, where $H = L^2(0,1)$, $\mathcal{A} \subset \mathbb{R}^{d_1}$, and $\mathcal{K} \subset \mathbb{R}^k$ are compact. The control problem is to minimize the ergodic cost functional

$$J(x, u, \alpha) = \limsup_{T \to \infty} \mathbb{E} \frac{1}{T} \int_0^T c(v(t), u(v(t)))dt$$

over the set of Markov controls $\mathcal{U} = \{u : H \to \mathcal{K} \mid u \text{ is Borel measurable}\}$, where $c : H \times \mathcal{K} \to \mathbb{R}$. The $\alpha \in \mathcal{A}$ in (1.1)–(1.4) represents a parameter.

The equations (1.1), (1.3), and (1.4) are only formal because the noise terms n, η_1, and η_2 are not well-defined stochastic processes (random fields). A standard approach for the rigorous treatment of the problem is to rewrite (1.1) as a controlled stochastic differential equation in the Hilbert space H, and to define the noise terms using Wiener processes with infinite-dimensional state spaces and the solution to the equation as a mild solution, using the semigroup theory (cf. [10, 27]).

In the present paper, this general framework is used. The controlled Markov process is defined by a Hilbert space-valued stochastic differential equation ((2.1) below). The linear term of the equation is the infinitesimal generator of an analytic semigroup. The general setting allows us to cover, as special cases, stochastic boundary/point control problems like the above example (see Examples 7.1 and 7.2). The noise for the stochastic differential equation can be distributed, boundary and point. The parameter-dependence occurs in the distributed and the boundary or the point drift terms. The control occurs only in the boundary or point drift term. The fact that the control is not distributed would seem to allow for more physically meaningful models. The noise is allowed to occur in both distributed and discrete forms to ensure more flexibility of the models. Since the H-valued Markov process depends on the control and the parameter, it is shown that some ergodic properties of the process are uniform in these quantities. For the solution of an ergodic control problem the existence of an optimal control is verified. It is shown that the optimal cost depends continuously on the system parameter.

Continuity of the optimal cost on the parameter is an important step in solving the adaptive control problem when the parameter is unknown. This verification is important to show the optimality of an adaptive control defined by means of a family of strongly consistent estimators of the unknown parameter α. In the case when the control and noise are distributed, the existence of an optimal control has been proven in [13], while the continuity of the optimal cost is new for this case.

The continuity of the optimal cost follows readily from the continuous dependence of the invariant measures for the controlled Markov process on the parameter α, uniform in the controls, in the norm of total variation of measures. This result can be of some independent interest and it may be interesting to note that even in some very simple cases the situation for Hilbert space–valued processes is significantly different from the finite-dimensional case. For example, consider the linear stochastic heat equation (without control)

$$\frac{\partial w}{\partial t}(t, \xi) = \alpha \frac{\partial^2 w}{\partial \xi^2}(t, \xi) + n(t, \xi), \quad (t, \xi) \in \mathbb{R}_+ \times (0, 1),$$

with initial and boundary conditions $w(0, \xi) = w_0(\xi)$, $w(t, 0) = w(t, 1) = 0$, where $\alpha \in [1/2, 1]$ and n is a space-time white noise. It is well known (see, e.g., [28]) that for each value of α, the probability laws (in the state space $H = L_2(0,1)$)

of the solutions converge in the norm of total variation to the Gaussian invariant measure $\mu(\alpha) = N(0, Q(\alpha))$, where $Q(\alpha) = \alpha^{-1}Q$, $Q = \int_0^\infty S(2t)dt$, and $S(\cdot)$ is the semigroup generated by the operator of the second derivative on $(0, 1)$, with zero Dirichlet boundary conditions.

However, by the dichotomy result for Gaussian measures it is easy to see that the invariant measures $\mu(\alpha)$ are singular for different values of $\alpha \in [1/2, 1]$, so there is no continuous dependence on α in the norm of total variation (see Remark 4.11 for some comparison between the finite- and infinite-dimensional state spaces).

A brief outline of the paper is given now. In section 2 the control problem is formulated and the basic assumptions are made and explained. The controlled process is the unique, weak, mild solution of the stochastic differential equation and induces a Markov process in H. Some estimates are made of this process, and an approximation of the transition probability function for the Markov process solution of the stochastic differential equation by transition functions of the solutions of the stochastic differential equation with bounded drifts is given, where the approximation is uniform in the control and the parameter. In section 3 the existence and uniqueness of the mild (backward) Kolmogorov equation for the controlled Markov process are verified. An estimate of the derivative of the mild solution of the Kolmogorov equation is given. In section 4 the results of section 3 are used to verify a uniform version of the strong Feller property and the strong (i.e., variation norm) continuity of the transition measures with respect to the parameter that is uniform in the control. The invariant measures of the controlled Markov process are shown to be continuous with respect to the parameter in the variation norm topology that is uniform in the control. In section 5 some tightness properties are verified. Initially it is shown that a "tightness" on balls implies tightness. A Lyapunov-type condition is shown to imply the tightness for the family of invariant measures depending on the parameter and the control. Section 6 contains the main results of the paper: the existence of an optimal control for a fixed parameter and the continuous dependence of the optimal cost on the parameter are verified using the results proven in sections 2, 3, and 4. In section 7 two examples are given that satisfy the assumptions that are made for the control problem: in Example 7.1 the control problem (1.1)–(1.4) is treated, and Example 7.2 contains a similar control problem, where the control and noise occur at given discrete points in the domain rather than on the boundary.

A brief description and a comparison of some previous results on these topics are given now. Similar results for the existence and the uniqueness of the weak, mild solutions to stochastic differential equations with only distributed noise and control are given in [10, 17, 18]. Some results for the existence and the uniqueness of mild solutions for semilinear stochastic equations with boundary or point noise are given in [11, 22, 27]. In [27] an existence result for the invariant measures is given. The methods to obtain the mild solution of the Kolmogorov equation are similar to the methods used in [6, 8, 9] for a fixed stochastic equation without parameter dependency. The approach to verifying the existence of an optimal control uses a standard procedure (see, e.g., [25, 32] for a finite-dimensional process and [13] for an infinite-dimensional process). There seems to be a fairly limited amount of work on infinite-time horizon control problems in infinite-dimensional spaces. Some work is devoted to discounted cost functionals. For this latter problem the existence of an optimal stationary control is shown in [4], and the stationary Hamilton–Jacobi–Bellman equation is investigated in [7, 20]. It seems that the ergodic control problem is only considered in [13], where a distributed control is used.

2. Preliminaries. Consider a controlled, infinite-dimensional process $(X(t),\, t \geq 0)$ that satisfies the stochastic differential equation

$$\begin{aligned} \mathrm{d}X(t) + AX(t)\mathrm{d}t &= (f(\alpha, X(t)) + Bh(\alpha, X(t), u(X(t))))\mathrm{d}t + B\mathrm{d}V(t) + Q^{1/2}\mathrm{d}W(t), \\ X(0) &= x, \end{aligned} \tag{2.1}$$

where $X(0), X(t) \in H$, H is a separable, infinite-dimensional Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and norm $|\cdot|$, $\alpha \in \mathcal{A} \subset \mathbb{R}^d$ is a parameter and $\mathcal{A}$ is compact, U is a separable Hilbert space with inner product $\langle \cdot, \cdot \rangle_U$ and norm $|\cdot|_U$, $\mathcal{K}$ is a compact product of intervals in $\mathbb{R}^k$, $-A : \mathrm{Dom}(-A) \to H$ is the infinitesimal generator of an analytic semigroup $(S(t),\, t \geq 0)$ such that $A^{-1} \in \mathcal{L}(H)$, which is often denoted $A > 0$,

$$\begin{aligned} f &: \mathcal{A} \times H \to H, \\ h &: \mathcal{A} \times H \times \mathcal{K} \to U \end{aligned}$$

are Borel measurable functions, $B \in \mathcal{L}(U, D_A^{\varepsilon - 1})$, the family of bounded linear operators from U to $D_A^{\varepsilon-1}$, where $\varepsilon \in (0, 1]$ is given and D_A^δ for $\delta \geq 0$ is the domain of the fractional power A^δ with the topology induced by the graph norm $|x|_{D_A^\delta} = |A^\delta x|$, while for $\delta < 0$ it is a completion of H in the norm $|\cdot|_{D_A^\delta}$. It is assumed that $Q \in \mathcal{L}(H)$ is positive and self-adjoint and $(V(t),\, t \geq 0)$ and $(W(t),\, t \geq 0)$ are independent, standard cylindrical Wiener processes in the spaces U and H, respectively, that are defined on a filtered, complete probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), \mathbb{P})$. The family of controls, $\mathcal{U}$, is

$$\mathcal{U} = \{u : H \to \mathcal{K} \mid u \text{ is Borel measurable}\}.$$

The control problem is to minimize, over $u \in \mathcal{U}$, the ergodic cost functional

$$J(x, u, \alpha) = \limsup_{T \to \infty} \mathbb{E} \frac{1}{T} \int_0^T c(X(s), u(X(s)))\mathrm{d}s, \tag{2.2}$$

where $c : H \times \mathcal{K} \to \mathbb{R}_+$ is bounded and Borel measurable.

The following assumptions, (A1)–(A7), are used selectively in this paper.

(A1) There exist a $\gamma \in (0, 1/2]$ and a $\Delta \in (0, 1/2]$ such that $B \in \mathcal{L}_2(U, D_A^{\gamma - 1/2})$ and $Q^{1/2} \in \mathcal{L}_2(H, D_A^{\Delta - 1/2})$, where $\mathcal{L}_2(\cdot, \cdot)$ is the family of Hilbert–Schmidt operators.

(A2) For each $\alpha \in \mathcal{A}$ the function $h(\alpha, \cdot, \cdot) : H \times \mathcal{K} \to U$ is continuous and $f(\alpha, \cdot) : H \to H$ is Lipschitz continuous on the bounded subsets of H, and there are constants k, k_f, k_h, and $\tilde{k}(\alpha)$ such that $|f(\alpha, x)| \leq k + k_f|x|$, $|h(\alpha, x, u)|_U \leq k + k_h|x|$, and $|h(\alpha, x, u)|_U \leq \tilde{k}(\alpha)$ for all $x \in H$, $u \in \mathcal{K}$, and $\alpha \in \mathcal{A}$.

By (A1) and the analyticity of $-A$, the composition $S(r)B$ is well defined for $r > 0$, and furthermore, $S(r)B \in \mathcal{L}_2(U, H)$, $S(r)Q^{1/2} \in \mathcal{L}_2(H)$, and

$$\int_0^t |S(r)B|^2_{\mathcal{L}_2(U,H)} dr + \int_0^t |S(r)Q^{1/2}|^2_{\mathcal{L}_2(H)} dr < \infty$$

for $t > 0$. Therefore, the family of operators $(Q_t, t \geq 0)$

$$Q_t = \int_0^t S(r)BB^*S^*(r)\mathrm{d}r + \int_0^t S(r)QS^*(r)\mathrm{d}r \tag{2.3}$$

is well defined and $Q_t \in \mathcal{L}_2(H)$ for each $t \geq 0$.

(A3) The following are satisfied:

$$\mathcal{R}(\tilde{S}(t)) \subset \mathcal{R}(Q_t^{1/2}), \qquad |Q_t^{-1/2} S(t) A^{1-\varepsilon}|_{\mathcal{L}(H)} \leq \frac{c}{t^\beta}$$

for $t \in (0, T]$ for some $T > 0$, $c > 0$, and $\beta < 1$, where $(\tilde{S}(t),\ t \geq 0)$ is the restriction of $(S(t),\ t \geq 0)$ to the space $D_A^{1-\varepsilon}$ and $\mathcal{R}(\cdot)$ is the range.

(A4) There is a continuous, increasing function $\omega : \mathbb{R}_+ \to \mathbb{R}_+$ with $\omega(0) = 0$ such that

$$|f(\alpha, x) - f(\beta, x)| + |h(\alpha, x, u) - h(\beta, x, u)|_U \leq \omega(|\alpha - \beta|)(1 + |x|)$$

for all $\alpha, \beta \in \mathcal{A}$, $x \in H$, and $u \in \mathcal{K}$.

(A5) For each $u \in \mathcal{U}$ and $\alpha \in \mathcal{A}$ there is an invariant measure $\mu(\alpha, u)$ for the process $(X(t),\ t \geq 0)$ that satisfies (2.1), and the family of measures $(\mu(\alpha, u),\ \alpha \in \mathcal{A},\ u \in \mathcal{U})$ is tight.

(A6) The function $c : H \times \mathcal{K} \to \mathbb{R}_+$ given in (2.2) is bounded and Borel measurable and $c(x, \cdot) : \mathcal{K} \to \mathbb{R}_+$ is continuous for each $x \in H$.

(A7) The set $h(\alpha, x, \mathcal{K}) \times c(x, \mathcal{K}) \subset U \times \mathbb{R}_+$ is convex for each $\alpha \in \mathcal{A}$ and $x \in H$.

Some comments on the above assumptions (A1)–(A7) are given now. Assumption (A1) is a standard condition guaranteeing that the solution of the linear version of the equation (2.1) (i.e., with $f = 0$ and $h = 0$) is an H-valued stochastic process (otherwise it is only a cylindrical process; see, e.g., [12]). Note that (A1) implies that the above-defined operators Q_t are trace class operators on H. They are covariance operators of the (Gaussian) probability distribution of the solution to the linear equation. (Some discussion on the verification of (A1) is contained, for example, in [12, 27]; (A1) is also verified in Examples 7.1, 7.2 of the present paper.)

The assumption (A2) is used to verify that there exists a unique, weak, mild solution to the equation (2.1) (below in this section).

The assumption (A3) is used in section 3 to prove some suitable smoothing properties of the mild backward Kolmogorov equation corresponding to the stochastic equation (2.1), which is needed to show the ergodicity of the solutions to (2.1) and some continuity properties of the transition probability kernels. The assumption is also rather standard in the context of the perturbation methods; for instance, for $\epsilon = 1$ the results of section 3 have been proven in [9, 10]. A class of examples in which (A3) can be easily verified is given also in section 3 (Proposition 3.4).

The assumption (A4) is a continuous dependence of the coefficients of the equation (2.1) on the parameter α. It is used for the verification of the results that are related to the continuous dependence of the optimal cost on the parameter.

The assumption (A5) is a kind of stability assumption that is usually needed in ergodic control problems. In section 5 (A5) is verified in terms of some more explicit conditions on the coefficients of equation (2.1) (Lyapunov-type conditions).

The assumptions (A7) and (A8) are typical conditions that are used in the ergodic control theory ((A7) is sometimes called the Roxin-type condition) and they are used to establish the existence of an optimal control for the given control problem.

Consider the following two stochastic differential equations:

$$\begin{aligned} \mathrm{d}Z(t) + AZ(t)\mathrm{d}t &= B\mathrm{d}V(t) + Q^{1/2}\mathrm{d}W(t), \\ Z(0) &= x, \end{aligned} \tag{2.4}$$

and

$$\begin{aligned} \mathrm{d}X(t) + AX(t)\mathrm{d}t &= f(\alpha, X(t))\mathrm{d}t + B\mathrm{d}V(t) + Q^{1/2}\mathrm{d}W(t), \\ X(0) &= x. \end{aligned} \tag{2.5}$$

Under the assumptions (A1) and (A2) it is easy to verify that each of the equations (2.4) and (2.5) has one and only one mild solution on the probability space $(\Omega, \mathcal{F}, P)$, that is, the solutions to the integral equations

$$Z(t) = S(t)x + \int_0^t S(t-r)B\mathrm{d}V(r) + \int_0^t S(t-r)Q^{1/2}\mathrm{d}W(r), \qquad t \geq 0, \tag{2.6}$$

and

$$\begin{aligned} X(t) = S(t)x &+ \int_0^t S(t-r)f(\alpha, X(t))\mathrm{d}t + \int_0^t S(t-r)B\mathrm{d}V(r) \\ &+ \int_0^t S(t-r)Q^{1/2}\mathrm{d}W(r), \quad t \geq 0. \end{aligned} \tag{2.7}$$

These solutions are D_A^δ-valued processes that belong to $C([0,T], L^p(\Omega, H)) \cap C((0,T], L^p(\Omega, D_A^\delta))$ for any $p \geq 1$, $T > 0$, and $\delta \in [0, \min(\varepsilon, \Delta, \gamma))$ (cf. [27]). Furthermore, the processes $(X(t),\, t \geq 0)$ and $(Z(t),\, t \geq 0)$ have D_A^δ-continuous versions (cf. [11, 30]), and in H they induce two Markov processes in the usual way.

Let $P_\alpha : \mathbb{R}_+ \times H \times \mathcal{B}(H) \to [0,1]$ be the transition probability function for $(X(t),\, t \geq 0)$ in (2.7), that is,

$$P_\alpha(t, x, \Gamma) = \mathbb{P}_x(X(t) \in \Gamma), \tag{2.8}$$

and let $(T(t),\, t \geq 0)$ be the Markov transition semigroup for $(Z(t),\, t \geq 0)$ in (2.6), that is,

$$T_t\varphi(x) = \mathbb{E}_x\varphi(Z(t)), \tag{2.9}$$

where $x \in H$ stands for the initial value of $X(\cdot)$, $t \geq 0$, and $\varphi \in \mathcal{M}(H)$, the bounded, Borel measurable functions on H. It is clear that

$$T_t 1_\Gamma(x) = N(S(t)x, Q_t)(\Gamma),$$

where $t \geq 0$, $\Gamma \in \mathcal{B}(H)$, $x \in H$, and Q_t is given by (2.3) so it is self-adjoint, nonnegative, and nuclear, and $N(S_t x, Q_t)$ is the Gaussian measure on H with mean $S_t x$ and covariance Q_t.

Let $\xi_T^{\alpha,u}$ be the random variable as follows:

$$\xi_T^{\alpha,u} = \int_0^T \langle h(\alpha, X(t), u(X(t))), \mathrm{d}V(t)\rangle_U - \frac{1}{2}\int_0^T |h(\alpha, X(t), u(X(t)))|_U^2 \mathrm{d}t \tag{2.10}$$

for $\alpha \in \mathcal{A}$, $u \in \mathcal{U}$, and $T > 0$, where $(X(t),\, t \in [0,T])$ is the solution of (2.7). A weak solution of (2.1) is constructed following the standard procedure of an absolutely continuous change of probability measure (cf. [10, 15, 17, 23]). For control problems, the method was initiated in [1, 14]. Note that $\mathbb{E}\exp(\xi_T^{\alpha,u}) = 1$ by (A2). There is a probability measure $\mathbb{P}_x^{\alpha,u}$ on $\mathcal{F}$ such that the restriction of $\mathbb{P}_x^{\alpha,u}$ to $\mathcal{F}_T$ is given by

$$\mathbb{P}_x^{\alpha,u}(\mathrm{d}\omega) = \exp(\xi_T^{\alpha,u})\mathbb{P}(d\omega), \tag{2.11}$$

the process $(V^*(t),\, t \geq 0)$ given by

$$V^*(t) = V(t) - \int_0^t h(\alpha, X(s), u(X(s)))\mathrm{d}s$$

is a cylindrical Wiener process on U, and using $\mathbb{P}_x^{\alpha,u}$ and the solution of (2.7), it follows that

(2.12)
$$\begin{aligned} X(t) = \; & S(t)x + \int_0^t S(t-r)f(\alpha, X(r))\mathrm{d}t + \int_0^t S(t-r)Bh(\alpha, X(r), u(X(r)))\mathrm{d}r \\ & + \int_0^t S(t-r)B\mathrm{d}V^*(r) + \int_0^t S(t-r)Q^{1/2}\mathrm{d}W(r). \end{aligned}$$

So there is a weak solution to (2.1) which is weakly unique and induces a Markov process on H whose Markov transition semigroup is denoted as

$$P_t^{\alpha,u}\varphi(x) = \mathbb{E}_x^{\alpha,u}\varphi(X(t)) \tag{2.13}$$

for $t \geq 0$ and $\varphi \in \mathcal{M}(H)$, where $\mathbb{E}_x^{\alpha,u}$ is the expectation using the probability measure $\mathbb{P}_x^{\alpha,u}$ and

$$P^{\alpha,u}(t, x, \Gamma) = P_t^{\alpha,u}1_\Gamma(x) \tag{2.14}$$

for $t \geq 0$, $\Gamma \in \mathcal{B}(H)$, and $x \in H$ is the corresponding transition probability function.

In the remainder of the section, three technical lemmas are given that are useful in what follows. Initially, Proposition 2.2 of [27] is given as the following lemma.

LEMMA 2.1. *If* (A1) *and* (A2) *are satisfied and* $\delta \in [0, \min(\gamma, \Delta, \varepsilon))$, $p > \max((\Delta - \delta)^{-1}, (\gamma - \delta)^{-1}, (\varepsilon - \delta)^{-1})$, *and* $x \in H$, *then for each* $T > 0$ *there is a constant* $C = \hat{C}(T, p, \delta)$ *such that*

$$\mathbb{E}|A^\delta X(T)|^p \leq C(1 + |x|^p), \tag{2.15}$$

where $(X(t),\, t \geq 0)$ *is the solution of* (2.7). *If* $\delta = 0$ *then* C *does not depend on* T *from finite intervals.*

The following two lemmas reduce some of the subsequent proofs to the case where f and h are uniformly bounded.

LEMMA 2.2. *If* (A1) *and* (A2) *are satisfied, then for each* $T > 0$ *and* $R > 0$

$$\lim_{N\to\infty} \inf \mathbb{P}_x^{\alpha,u}\left(\sup_{t\in[0,T]} |X(t)| \leq N \right) = 1, \tag{2.16}$$

where the infimum is taken over $\alpha \in \mathcal{A}$, $u \in \mathcal{U}$, *and* $x \in H$ *with* $|x| \leq R$.

Proof. Recall the equation (2.12) for $(X(t),\, t \geq 0)$. Let $\Omega_{x,N}^{\alpha,u} \subset \Omega$ be given by

$$\Omega_{x,N}^{\alpha,u} = \left\{ \sup_{t\in[0,T]} \left(\left| \int_0^t S(t-r)B\mathrm{d}V^*(r) \right| + \left| \int_0^t S(t-r)Q^{1/2}\mathrm{d}W(r) \right| \right) \leq N \right\}. \tag{2.17}$$

By a maximal inequality (Lemma 2.2 of [30])

$$\begin{aligned} & \mathbb{E}_x^{\alpha,u} \sup_{t\in[0,T]} \left(\left| \int_0^t S(t-r)B\mathrm{d}V^*(r) \right|^2 + \left| \int_0^t S(t-r)Q^{1/2}\mathrm{d}W(r) \right|^2 \right) \\ & \qquad \leq \int_0^T |S(r)B|^2_{\mathcal{L}_2(U,H)}\mathrm{d}r + \int_0^T |S(r)Q^{1/2}|^2_{\mathcal{L}_2(H)}\mathrm{d}r \leq M \end{aligned} \tag{2.18}$$

for some M that does not depend on α, x, and u by (A1) and the analyticity of

$(S(t),\, t \geq 0)$. Thus

(2.19)
$$\mathbb{P}_x^{\alpha,u}\left(\sup_{t\in[0,T]}\left(\left|\int_0^t S(t-r)B\mathrm{d}V^*(r)\right| + \left|\int_0^t S(t-r)Q^{1/2}\mathrm{d}W(r)\right|\right) \geq N\right) \leq \frac{2M}{N^2}.$$

By (A1), (A2), and (2.17) on the set $\Omega_{x,N}^{\alpha,u}$ it follows that

$$|X(t)| \leq c_1|x| + c_2 + c_3\int_0^t \frac{|X(s)|}{(t-s)^{1-\varepsilon}}\mathrm{d}s + N$$

for $t \in [0,T]$, where the constants c_1, c_2, and c_3 depend only on T. The generalized Gronwall lemma (Theorem 7.1 of [21]) implies that

$$|X(t)| \leq c_4|x| + c_5 + N \tag{2.20}$$

for $t \in [0,T]$ on $\Omega_{x,N}^{\alpha,u}$, where c_4 and c_5 only depend on T. Since $\mathbb{P}_x^{\alpha,u}(\Omega_{x,N}^{\alpha,u}) \geq 1 - 2M/N^2$ the equality (2.16) follows. □

By (A2) it follows that there is a sequence $(f_m, h_m,\, m \in \mathbb{N})$ such that for each $m \in \mathbb{N}$

$$(f_m(\alpha,x), h_m(\alpha,x,u)) = (f(\alpha,x), h(\alpha,x,u)) \tag{2.21}$$

for $\alpha \in \mathcal{A}$, $u \in \mathcal{K}$, $x \in H$ with $|x| \leq m$ and

$$|f_m| + |h_m|_U \leq M_m, \tag{2.22}$$

where M_m is a constant depending only on m, $f_m(\alpha,\cdot)$ is Lipschitz continuous, and $h_m(\alpha,\cdot,\cdot)$ is continuous for each $m \in \mathbb{N}$ and

$$|f_m(\alpha,x) - f_m(\beta,x)| + |h_m(\alpha,x,u) - h_m(\beta,x,u)|_U \leq \tilde{\omega}_m(|\alpha-\beta|) \tag{2.23}$$

for $\alpha,\beta \in \mathcal{A}$, $x \in H$, and $u \in \mathcal{K}$, where $\tilde{\omega}_m$ has the same properties as ω in (A4) for each $m \in \mathbb{N}$. It is clear that if f and h are replaced by f_m and h_m, respectively, in (2.1), then the same procedure gives a unique weak solution inducing a Markov process on H.

LEMMA 2.3. *Let $P_m^{\alpha,u} : \mathbb{R}_+ \times H \times \mathcal{B}(H)$ be the transition probability function for the Markov process that is the solution of* (2.1) *with f and h replaced by f_m and h_m, respectively, which are described above. If* (A1) *and* (A2) *are satisfied then*

$$\lim_{m\to\infty} \|P_m^{\alpha,u}(t,x,\cdot) - P^{\alpha,u}(t,x,\cdot)\| = 0 \tag{2.24}$$

uniformly in $\alpha \in \mathcal{A}$, $u \in \mathcal{U}$, and x from bounded sets in H where $\|\cdot\|$ is the variation norm.

The proof of this lemma follows easily from Lemma 2.2 and the local uniqueness theorem for stochastic integrals. Let $(X_m(t),\, t \geq 0)$ be the solution of (2.5) with f replaced by f_m. It easily follows that

$$\Omega_N = \left\{\sup_{\substack{t\in[0,T]\\ \alpha\in\mathcal{A},\ |x|\leq R}} |X_m(t)| \leq N\right\} = \left\{\sup_{\substack{t\in[0,T]\\ \alpha\in\mathcal{A},\ |x|\leq R}} |X(t)| \leq N\right\}$$

for $m \geq N > R > 0$ because the trajectories of $(X_m(t),\, t \geq 0)$ and $(X(t),\, t \geq 0)$ coincide for $t \in [0,T]$ on Ω_N. Lemma 2.2 implies that $\mathbb{P}_x^{\alpha,u}(\Omega_N) \to 1$ as $N \to \infty$ uniformly in $\alpha \in \mathcal{A}$, $u \in \mathcal{U}$, $x \in H$ with $|x| < R$. Defining $\mathbb{P}_{x,m}^{\alpha,u}$ in the same way as $\mathbb{P}_x^{\alpha,u}$ by replacing h by h_m, it follows that the probabilities $\mathbb{P}_{x,m}^{\alpha,u}$ and $\mathbb{P}_x^{\alpha,u}$ restricted to Ω_N coincide for $m \geq N$. Given $\varepsilon > 0$, choose $N \geq 0$ such that $\mathbb{P}_{x,m}^{\alpha,u}(\Omega_N) \geq 1 - \varepsilon$ for $\alpha \in \mathcal{A}$, $u \in \mathcal{U}$, $|x| \leq R$, and $m \geq N$. It follows that

$$|P_m^{\alpha,u}(T,x,\Gamma) - P^{\alpha,u}(T,x,\Gamma)| < \varepsilon$$

for each $\Gamma \in \mathcal{B}(H)$, $|x| \leq R$, $\alpha \in \mathcal{A}$, $u \in \mathcal{U}$, and $m \geq N$.

3. The mild Kolmogorov equation. In this section a version of the mild Kolmogorov equation is considered. The existence and the uniqueness of the solution of this equation is verified, as is an estimate on the derivatives which is important to establish a uniform version of the strong Feller property. Many of the results of this section are verified similarly to the verifications that are used for a single Hilbert space (cf. [6, 8, 9]), so some details are omitted. Recall the definitions of $(T_t,\, t \geq 0)$ in (2.9) and $(P_t^{\alpha,u},\, t \geq 0)$ in (2.13). Let D_x be the Fréchet derivative on H and let $\mathcal{U}_c = \{u \in \mathcal{U} : u \in C(H,\mathcal{K})\}$. Let

$$\mathcal{H} = \left\{ \psi \mid \psi : (0,T] \to C_b^1(H),\ D_x\psi : (0,T] \to C_b(H, D_{A^*}^{1-\varepsilon}),\right.$$
$$\left. |\psi|_{\mathcal{H}} := \sup_{\substack{t\in(0,T)\\ x\in H}} (t^\beta |\psi(t,x)| + t^\beta |D_x\psi(t,x)|_{D_{A^*}^{1-\varepsilon}}) < \infty \right\},$$

where $\beta \in (0,1)$ is given in (A3), which is assumed to be satisfied throughout this section.

PROPOSITION 3.1. *Let $\varphi \in C_b(H)$ and $n(t,x) = T_t\varphi(x)$ for $t \geq 0$ and $x \in H$. Then $n \in \mathcal{H}$ and*

$$|D_x n(t,x)|_{D_{A^*}^{1-\varepsilon}} \leq \frac{c}{t^\beta} \sup|\varphi| \tag{3.1}$$

for $t \in (0,T]$ and $x \in H$, where the constant c does not depend on φ.

Proof. By the absolute continuity of measures it follows that

$$\begin{aligned} n(t,x) &= \int \varphi(y) N(S(t)x, Q_t)(\mathrm{d}y) \\ &= \int \varphi(y) \exp\left[\langle \Gamma_t x, Q_t^{-(1/2)} y\rangle - \frac{1}{2}|\Gamma_t x|^2\right] N(0,Q_t)(\mathrm{d}y), \end{aligned} \tag{3.2}$$

where $\Gamma_t = Q_t^{-(1/2)} S_t \in \mathcal{L}(H)$ by (A3). Applying D_x to (3.2) it follows (cf. [10]) that

$$D_x n(t,x)h = \int \langle \Gamma_t h, Q_t^{-(1/2)} y\rangle \varphi(S(t)x + y) N(0,Q_t)(\mathrm{d}y)$$

for $h \in H$, so that (A3) yields

$$\begin{aligned} \sup_{|h|\leq 1} |D_x n(t,x)(A^{1-\varepsilon}h)| &\leq c_1 \sup_{|h|\leq 1} \int |\langle \Gamma_t A^{1-\varepsilon} h, Q_t^{-(1/2)} y\rangle| N(0,Q_t)(\mathrm{d}y) \sup|\varphi| \\ &\leq c_2 \sup|\varphi|\, |\Gamma_t A^{1-\varepsilon}|_{\mathcal{L}(H)} \leq \frac{c_3}{t^\beta} \sup|\varphi| \end{aligned} \tag{3.3}$$

for $t \in (0,T]$, where $c_i, i = 1,2,3$, are constants independent of φ and t. The inequality (3.1) follows because $(D_A^{\varepsilon-1})' = D_{A^*}^{1-\varepsilon}$. □

Consider the mild Kolmogorov equation of the form

(3.4)
$$v(t,x) = T_t\varphi(x) + \int_0^t T_{t-s}(\langle D_x v(s,\cdot), f(\alpha,\cdot)\rangle + \langle D_x v(s,\cdot), Bh(\alpha,\cdot,u(\cdot)\rangle)(x)\mathrm{d}s$$

for $t \geq 0$, where $\varphi \in C_b(H)$, $\langle\cdot,\cdot\rangle$ is used for the duality between the corresponding domains of the fractional powers of A and A^* as well as the inner product on H, and for notational convenience, the dependence of v on α and u is suppressed. The solution $v(t,x)$ of (3.4) is shown to be $P_t^{\alpha,u}\varphi(x)$.

PROPOSITION 3.2. *If* (A1)–(A3) *are satisfied,* $u \in \mathcal{U}_c$, $\varphi \in C_b(H)$, *and* $|f|$ *and* $|h|_U$ *are bounded independent of* $\alpha \in \mathcal{A}$ *and* $u \in \mathcal{U}_c$, *then the equation* (3.4) *has one and only one solution* $v(t,x) = P_t^{\alpha,u}\varphi(x)$ *in* $\mathcal{H}$ *that satisfies*

$$|D_x v(t,x)|_{D_{A^*}^{1-\varepsilon}} \leq \frac{\tilde{c}}{t^\beta}\sup|\varphi| \tag{3.5}$$

for $t \in (0,T]$, *where the constant* $\tilde{c}$ *does not depend on* $\varphi, u \in \mathcal{U}_c$ *or* $\alpha \in \mathcal{A}$.

Proof. To verify the existence and uniqueness of the solution of (3.4), the Banach fixed point theorem is used for the Banach space $(\mathcal{H}, |\cdot|_{\mathcal{H}})$. Define the mapping $\Phi : \mathcal{H} \to \mathcal{H}$ as follows:

$$\Phi v(t,x) = T_t\varphi(x) + \int_0^t T_{t-s}\psi(D_x v(s,\cdot))(x)\mathrm{d}s \tag{3.6}$$

for $t \in (0,T]$, where

$$\psi(D_x v(s,\cdot)) = \langle f(\alpha,\cdot), D_x v(s,\cdot)\rangle + \langle D_x v(s,\cdot), Bh(\alpha,\cdot,u(\cdot))\rangle \tag{3.7}$$

and the dependence of ψ on $\alpha \in \mathcal{A}$ and $u \in \mathcal{U}_c$ is suppressed. For $v_1, v_2 \in \mathcal{H}$ it follows that

(3.8)
$$\begin{aligned}|\Phi(v_1) - \Phi(v_2)|_{\mathcal{H}} = \sup_{\substack{t\in(0,T)\\ x\in H}} [t^\beta | \int_0^t T_{t-s}(\psi(D_x v_1(s,\cdot)) - \psi(D_x v_2(s,\cdot)))(x)\mathrm{d}s| \\ + t^\beta \int_0^t |(A^*)^{1-\varepsilon} D_x T_{t-s}(\psi(D_x v_1(s,\cdot)) - \psi(D_x v_2(s,\cdot)))(x)|\mathrm{d}s].\end{aligned}$$

Note that

$$\begin{aligned}&|\psi(D_x v_1(s,\cdot)) - \psi(D_x v_2(s,\cdot))| \leq c_1|D_x v_1(s,\cdot) - D_x v_2(s,\cdot)| \\ &\quad + |(A^*)^{1-\varepsilon}(D_x v_1(s,\cdot) - D_x v_2(s,\cdot))| \leq c_2|D_x v_1(s,\cdot) - D_x v_2(s,\cdot)|_{D_{A^*}^{1-\varepsilon}}\end{aligned}$$

for suitable constants c_1 and c_2. Applying this inequality to (3.8) yields

$$\begin{aligned}|\Phi(v_1) - \Phi(v_2)|_{\mathcal{H}} \leq\ & c_2 \int_0^t t^\beta \sup_{s,x} |D_x v_1(s,x) - D_x v_2(s,x)|_{D_{A^*}^{1-\varepsilon}}\mathrm{d}s \\ &+ c_2 c \int_0^t \frac{t^\beta}{(t-s)^\beta} \sup_{s,x} |D_x v_1(s,x) - D_x v_2(s,x)|_{D_{A^*}^{1-\varepsilon}}\mathrm{d}s \\ \leq\ & c_3|v_1 - v_2|_{\mathcal{H}} \left(\int_0^t \frac{t^\beta}{s^\beta}\mathrm{d}s + t^\beta \int_0^t \frac{\mathrm{d}s}{(t-s)^\beta s^\beta}\right).\end{aligned}$$

Thus Φ is a contraction for $t > 0$ sufficiently small. The fact that $\Phi(\mathcal{H}) \subset \mathcal{H}$ is verified similarly. Therefore, for $T > 0$ sufficiently small, there is a unique solution of (3.4). For arbitrary $T > 0$ the interval $[0, T]$ is subdivided into a finite number of small intervals.

To verify (3.5) it follows by (3.1) that

(3.9)

$$\sup_x |D_x v(t,x)|_{D^{1-\varepsilon}_{A^*}} \le \sup_x |D_x T_t \varphi(x)|_{D^{1-\varepsilon}_{A^*}} + \sup_x \int_0^t |D_x T_{t-s} \psi(D_x v(s,\cdot))(x)|_{D^{1-\varepsilon}_{A^*}} \mathrm{d}s$$
$$\le ct^{-\beta} \sup|\varphi| + c_4 \int_0^t \sup_x |D_x v(s,x)|_{D^{1-\varepsilon}_{A^*}} \frac{\mathrm{d}s}{(t-s)^\beta}$$

for $t \in (0, T)$ and c_4 is a constant. Applying the generalized Gronwall lemma (Theorem 7.1 of [21]) to the function $\lambda(t) = \sup_x |D_x v(t,x)|_{D^{1-\varepsilon}_{A^*}}$, it follows that

$$\sup_{x \in H} |D_x v(t,x)|_{D^{1-\varepsilon}_{A^*}} \le \frac{c_5}{t^\beta} \sup|\varphi| \tag{3.10}$$

for $t \in (0, T]$, where the constant c_5 does not depend on $t \in (0, T]$, $\varphi \in C_b(H)$, $\alpha \in \mathcal{A}$, and $u \in \mathcal{U}_c$, though it may depend on the bounds for $|f|$ and $|h|_U$. While it remains to show that $v(t,x)$ is $P^{\alpha,u}_t \varphi(x)$, this verification is identical to the proof of (Theorem 4 of [6]) and is omitted. Only note that (A1) implies that $B \in \mathcal{L}_2(U, D_A^{-1})$ and $Q^{1/2} \in \mathcal{L}_2(H, D_A^{-1})$, which is used here. □

Proposition 3.2 is essential in the following result, which gives a strong Feller property that is uniform for $u \in \mathcal{U}_c$. It is improved in the next section.

LEMMA 3.3. *Let $t > 0$ and $y \in H$ be fixed. If* (A1)–(A2) *are satisfied then there is a function $\tilde{\omega} : \mathbb{R}_+ \to \mathbb{R}_+$ that is increasing and continuous with $\tilde{\omega}(0) = 0$ such that*

$$\|P^{\alpha,u}(t,x,\cdot) - P^{\alpha,u}(t,y,\cdot)\| \le \tilde{\omega}(|x-y|) \tag{3.11}$$

for all $\alpha \in \mathcal{A}$, $u \in \mathcal{U}_c$, and $x \in H$, where $\|\cdot\|$ is the variation norm.

Proof. If $|f|$ and $|h|_U$ are bounded then by (3.5) it follows that

$$\|P^{\alpha,u}(t,x,\cdot) - P^{\alpha,u}(t,y,\cdot)\| = \sup_{\varphi \in C_b,\ |\varphi| \le 1} |P^{\alpha,u}_t \varphi(x) - P^{\alpha,u}_t \varphi(y)| \le \frac{\tilde{c}}{t^\beta} |x-y|_{D_A^{\varepsilon-1}}$$

for $x \in H$, which easily implies (3.11) ($\tilde{c}$ may depend on the bounds for $|f|$ and $|h|_U$). If $|f|$ and $|h|_U$ are not bounded, then use Lemma 2.3 to approximate $P^{\alpha,u}(t,x,\cdot)$ and $P^{\alpha,u}(t,y,\cdot)$ by $P^{\alpha,u}_k(t,x,\cdot)$ and $P^{\alpha,u}_k(t,y,\cdot)$, respectively, uniformly with respect to $\alpha \in \mathcal{A}$, $u \in \mathcal{U}_c$, and x from bounded sets in H. □

This section is concluded with a simple result which can be useful in some cases to verify (A3).

PROPOSITION 3.4. *If $\varepsilon > 1/2$ and $Q^{1/2} = A^{-\eta}\Gamma$, where $\eta \in [0, \varepsilon - 1/2)$ and $\Gamma, \Gamma^{-1} \in \mathcal{L}(H)$, then* (A3) *is satisfied.*

Proof. Let

$${}^1Q_t = \int_0^t S(r) B B^* S^*(r) \mathrm{d}r$$

and

$${}^2Q_t = \int_0^t S(r) Q S^*(r) \mathrm{d}r.$$

It is clear that $Q_t = {}^1Q_t + {}^2Q_t$ and ${}^1Q$ and ${}^2Q$ are nonnegative and self-adjoint. It suffices to verify (A3) with Q_t replaced by ${}^2Q_t$. By the minimum energy principle (cf. Remark B9 of [10]) it follows that

$$|{}^2Q_t^{-(1/2)}S(t)y| \le \left(\int_0^t |u(s)|^2 \mathrm{d}s\right)^{1/2}, \tag{3.12}$$

where $u \in L^2([0,t],H)$ is arbitrary such that the solution $(z(s),\ s\in[0,t])$ of

$$\dot z + Az = Q^{1/2}u, \qquad z(0) = y \tag{3.13}$$

satisfies $z(t) = 0$. The existence of such a function is a necessary condition for ${}^2Q_t^{-(1/2)}S(t) \in \mathcal{L}(H)$. For $x \in D_A^{1-\varepsilon}$ define $\tilde u(r) = -(1/t)Q^{-1/2}S(r)A^{1-\varepsilon}x$. Clearly $\tilde u \in L^2([0,t],H)$ and u gives $z(t) = 0$ if $y = A^{1-\varepsilon}x$. Thus

$$|{}^2Q_t^{-(1/2)}S(t)A^{1-\varepsilon}x| \le \left(\int_0^t |\tilde u(r)|^2 \mathrm{d}r\right)^{1/2} \le |x|\frac{\tilde c}{t^{\eta+(3/2)-\varepsilon}}$$

for a constant $\tilde c$, so (A3) is satisfied with $\beta = \eta + (3/2) - \varepsilon < 1$, □

4. The continuous dependence of some measures on a parameter. In this section, the verifications are made for the continuous dependence of $P^{\alpha,u}(t,x,\cdot)$ on the parameter α and the uniform strong Feller property, which yield (under the tightness condition (A5)) the uniform continuity of the invariant measures with respect to the parameter $\alpha \in \mathcal{A}$. This last result is used in section 6 to prove continuity of the optimal cost for the control problem (2.1), (2.2).

LEMMA 4.1. *If* (A1) *and* (A2) *are satisfied then for each* $t > 0$, $\alpha \in \mathcal{A}$, *and* $x \in H$

$$\lim_{u_n \to u} \|P^{\alpha,u_n}(t,x,\cdot) - P^{\alpha,u}(t,x,\cdot)\| = 0, \tag{4.1}$$

where $u_n \in \mathcal{U}$ *for all* $n \in \mathbb{N}$ *and* $u_n \to u$ *pointwise.*

Proof. By Lemma 2.3 it can be assumed that $|f|$ and $|h|_U$ are bounded uniformly in $\alpha \in \mathcal{A}$ and $u \in \mathcal{U}$. It easily follows that

$$\begin{aligned} \|P^{\alpha,u_n}(t,x,\cdot) - P^{\alpha,u}(t,x,\cdot)\| &= \sup_{\varphi\in C_b,\|\varphi|\le 1} |P_t^{\alpha,u_n}\varphi(x) - P_t^{\alpha,u}\varphi(x)| \\ &\le |\mathbb{E}\varphi(X(t))\exp(\xi_t^{\alpha,u_n}) - \mathbb{E}\varphi(X(t))\exp(\xi_t^{\alpha,u})| \le \mathbb{E}|\exp(\xi_t^{\alpha,u_n}) - \exp(\xi_t^{\alpha,u})|, \end{aligned} \tag{4.2}$$

where $(X(t),\ t \ge 0)$ satisfies (2.5). Since $\mathbb{E}\exp(2\xi_t^{\alpha,u_n}) \le \exp(t\sup|h|)$ the sequence $(\exp(\xi_t^{\alpha,u_n}),\ n\in\mathbb{N})$ is uniformly integrable, so for every $\varepsilon > 0$ there is an $R > 0$ such that

$$\mathbb{E}|\exp(\xi_t^{\alpha,u_n}) - \exp(\xi_t^{\alpha,u})| \le \mathrm{e}^R\mathbb{E}|\xi_t^{\alpha,u_n} - \xi_t^{\alpha,u}| + \varepsilon.$$

From (A2), the boundedness of $|h|_U$, and the dominated convergence theorem, (4.2) is verified. □

The following result is a uniform version of the strong Feller property.

LEMMA 4.2. *If* (A1)–(A3) *are satisfied, then for each* $t > 0$, $y \in H$, *there is a continuous, increasing function* $\tilde\omega : \mathbb{R}_+ \to \mathbb{R}_+$ *with* $\tilde\omega(0) = 0$ *such that*

$$\|P^{\alpha,u}(t,x,\cdot) - P^{\alpha,u}(t,y,\cdot)\| \le \tilde\omega(|x-y|)$$

for all $\alpha \in \mathcal{A}$, $u \in \mathcal{U}$, *and* $x \in H$.

Proof. Take the function $\tilde{\omega}$ from Lemma 3.3 (for the fixed $t > 0$ and $y \in H$) and let $\mathcal{U}' \in \mathcal{U}$ be the set of controls satisfying

$$\sup_{u \in \mathcal{U}'} \|P^{\alpha,u}(t,x,\cdot) - P^{\alpha,u}(t,y,\cdot)\| \leq \tilde{\omega}(|x-y|)$$

for all $x \in H$ and $\alpha \in \mathcal{A}$. By Lemma 3.3, $\mathcal{U}_c \subset \mathcal{U}'$ and by Lemma 4.1, $\mathcal{U}'$ is closed with respect to pointwise convergence. Since the families of Baire and Borel functions $H \to \mathcal{K}$ coincide (cf. [24, Theorem 2.31.IX]) it follows that $\mathcal{U}' = \mathcal{U}$.

PROPOSITION 4.3. *Denote by $\mathcal{P}(H)$ the space of probability measures on the Borel subsets of H endowed with the metric of total variation of measures. If* (A1)–(A4) *are satisfied, then for each $T > 0$ the function*

$$\eta : \mathcal{A} \to \mathcal{P}(H)$$

given by

$$\eta(\alpha) := P^{\alpha,u}(T,x,\cdot) \tag{4.3}$$

is continuous uniformly in $u \in \mathcal{U}$ and $x \in K$ for each compact set $K \subset H$.

Proof. By Lemma 2.3 it can be assumed that $|f|$ and $|h|_U$ are bounded and

$$|f(\alpha,x) - f(\beta,x)| + |h(\alpha,x,u) - h(\beta,x,u)|_U \leq \omega(|\alpha-\beta|) \tag{4.4}$$

for $x \in H$ and $\alpha, \beta \in \mathcal{A}$. Initially, the uniform continuity of (4.3) is verified for $u \in \mathcal{U}_c$. For $v_{\alpha,u}(t,x) = P_t^{\alpha,u}\varphi(x)$ for $x \in H$ and $\varphi \in C_b(H)$ it follows by Proposition 3.2 that

$$v_{\alpha,u}(t,x) = T_t\varphi(x) + \int_0^t T_{t-s}(\psi_{\alpha,u}(D_x v_{\alpha,u}(s,\cdot)))(x)\mathrm{d}s \tag{4.5}$$

for $t \in [0,T]$, where

$$\psi_{\alpha,u}(D_x v_{\alpha,u}(s,\cdot)) = \langle D_x v_{\alpha,u}(s,\cdot), f(\cdot)\rangle + \langle D_x v_{\alpha,u}(s,\cdot), Bh(\alpha,\cdot,u(\cdot))\rangle \tag{4.6}$$

and

$$|D_x v_{\alpha,u}(t,\cdot)|_{D_{A^*}^{1-\varepsilon}} \leq ct^{-\beta}\sup|\varphi| \tag{4.7}$$

for $t \in (0,T]$, where $c > 0$ does not depend on $t \in (0,T]$, $\alpha \in \mathcal{A}$, $u \in \mathcal{U}_c$, and $\varphi \in C_b(H)$. By Proposition 3.1 it follows that

$$\begin{aligned}
&\sup_x |v_{\alpha,u}(t,x) - v_{\alpha_0,u}(t,x)| + \sup_x |D_x v_{\alpha,u}(t,x) - D_x v_{\alpha_0,u}(t,x)|_{D_{A^*}^{1-\varepsilon}} \\
&\quad \leq \sup_x \int_0^t |T_{t-s}(\psi_{\alpha,u}(D_x v_{\alpha,u}(s,\cdot)) - \psi_{\alpha_0,u}(D_x v_{\alpha_0,u}(s,\cdot)))(x)|\mathrm{d}s \\
&\quad + \sup_x \int_0^t |D_x T_{t-s}(\psi_{\alpha,u}(D_x v_{\alpha,u}(s,\cdot)) - \psi_{\alpha_0,u}(D_x v_{\alpha_0,u}(s,\cdot)))(x)|_{D_{A^*}^{1-\varepsilon}}\mathrm{d}s.
\end{aligned} \tag{4.8}$$

By (4.4) and (4.7) it follows that

$$\begin{aligned}
&\sup_x |\psi_{\alpha,u}(D_x v_{\alpha,u}(s,x)) - \psi_{\alpha_0,u}(D_x v_{\alpha_0,u}(s,x))| \\
&\leq c_1 \sup_x |D_x v_{\alpha,u}(s,x) - D_x v_{\alpha_0,u}(s,x)|_{D_{A^*}^{1-\varepsilon}} + c_2 s^{-\beta}\omega(|\alpha-\alpha_0|)
\end{aligned} \tag{4.9}$$

for some constants c_1, c_2 depending only on the bounds for $|f|$, $|h|_U$, and $|B|_{\mathcal{L}(U,D_A^{\varepsilon-1})}$. Let $\lambda_{\alpha,u}(\cdot)$ be the left-hand side of (4.8). By (4.8) and (4.9) it follows that

$$\lambda_{\alpha,u}(t) \le \int_0^t \frac{k_1}{(t-s)^\beta}\lambda_{\alpha,u}(s)\, ds + \omega(|\alpha-\alpha_0|)\int_0^t \frac{k_2}{(t-s)^\beta s^\beta}\mathrm{d}s \tag{4.10}$$

for $t \in (0,T]$ for some constants k_1 and k_2. By the generalized Gronwall lemma (Theorem 7.1 of [21]) it follows that

$$\lambda_{\alpha,u}(t) \le k_3\omega(|\alpha-\alpha_0|)$$

for $t \in (0,T]$, so

$$\|P^{\alpha,u}(T,x,\cdot) - P^{\alpha_0,u}(T,x,\cdot)\| = \sup_{|\varphi|\le 1} |P_T^{\alpha,u}\varphi(x) - P_T^{\alpha_0,u}\varphi(x)| \le k_4\omega(|\alpha-\alpha_0|)$$

for some constants k_3 and k_4 that are independent of $x \in H$ and $u \in \mathcal{U}_c$. The last estimate is extended to $u \in \mathcal{U}$ using Lemma 4.1 by the same argument as in the proof of Lemma 4.2.

The following result is a version of the Itô formula that is applicable to functions of the solution of (2.1).

PROPOSITION 4.4. *If* (A1) *and* (A2) *are satisfied,* $g \in C^2(H)$, $D_x g(x) \in D_{A^*}^{1-\varepsilon}$ *for* $x \in H$, $D_x g : H \to D_{A^*}^{1-\varepsilon}$ *is continuous,* $D_{xx} g : H \to \mathcal{L}(D_A^{-\delta}, D_{A^*}^{\delta}) \cap \mathcal{L}(H, D_{A^*}^{1-\varepsilon})$ *is continuous for* $\delta = \max((1/2)-\Delta, (1/2)-\gamma)$, *where* D_{xx} *is the second Fréchet derivative,* $\langle A\cdot, D_x g(\cdot)\rangle : D_A^1 \to \mathbb{R}$ *can be extended to a continuous function* $\Phi : H \to \mathbb{R}$, *and*

$$|\Phi(x)| + |g(x)| + |D_x g(x)|_{D_{A^*}^{1-\varepsilon}} + |D_{xx} g(x)|_{\mathcal{L}(D_A^{-\delta}, D_{A^*}^{\delta})} \le \hat{k}(1+|x|^p) \tag{4.11}$$

for $x \in H$ *and some* $\hat{k}$ *and* $p > 0$, *then the following equality is satisfied:*

$$\begin{aligned}\mathbb{E}_x^{\alpha,u} g(X(t)) - g(x) = \ & \mathbb{E}_x^{\alpha,u}\int_0^t (-\Phi(X(s)) + \langle f(\alpha, X(s)), D_x g(X(s))\rangle \\ & + \langle h(\alpha, X(s), u(X(s))), B^* D_x g(X(s))\rangle_U \\ & + \frac{1}{2}\mathrm{tr}[(A^*)^{1/2-\gamma} D_{xx} g(X(s)) BB^* (A^*)^{\gamma-1/2}] \\ & + \frac{1}{2}\mathrm{tr}[(A^*)^{1/2-\Delta} D_{xx} g(X(s)) Q (A^*)^{\Delta-1/2}])\mathrm{d}s\end{aligned} \tag{4.12}$$

for $t \ge 0$, $\alpha \in \mathcal{A}$, $u \in \mathcal{U}$, *and* $x \in H$.

Proof. Fix $\alpha \in \mathcal{A}$ and $u \in \mathcal{U}$. Choose a sequence of functions $(h_n, n \in \mathbb{N})$ such that $h_n : H \to U$ is globally Lipschitz continuous and $h_n \to h$ pointwise as $n \to \infty$ and h is bounded by the constant $\tilde{k}(\alpha)$ from (A2). It follows as in Proposition 3.4 of [12] and Proposition 1.5 of [27] that

$$\begin{aligned}\mathbb{E} g(X(t))\exp(\xi_{n,t}) - g(x) = \mathbb{E}\int_0^t & (-\Phi(X(s)) \\ & + \langle f(\alpha, X(s)), D_x g(X(s))\rangle + \langle h_n(X(s)), B^* D_x g(X(s))\rangle_U \\ & + \frac{1}{2}\mathrm{tr}[(A^*)^{1/2-\gamma} D_{xx} g(X(s)) BB^* (A^*)^{\gamma-1/2}] \\ & + \frac{1}{2}\mathrm{tr}[(A^*)^{1/2-\Delta} D_{xx} g(X(s)) Q (A^*)^{\Delta-1/2}])\exp(\xi_{n,s})\mathrm{d}s\end{aligned} \tag{4.13}$$

for $t \geq 0$, where

$$\xi_{n,s} = \int_0^t \langle h_n(\alpha, X(r), u(X(r))), \mathrm{d}V(r)\rangle_U - \frac{1}{2}\int_0^t |h_n(\alpha, X(r), u(X(r)))|_U^2 \mathrm{d}r.$$

The remainder of the proof investigates the particular terms above as $n \to \infty$. As in the proof of Lemma 4.1, it can be shown that

$$\lim_{n\to\infty} \int_0^t \mathbb{E}|\exp(\xi_{n,s}) - \exp(\xi_s^{\alpha,u})|\mathrm{d}s = 0,$$

so there is a subsequence such that $\exp(\xi_{n_k,t}) \to \exp(\xi_t^{\alpha,u})$ on $[0,T]\times\Omega$, $\lambda\times\mathbb{P}$ almost everywhere, where λ is the Lebesgue measure on $\mathbb{R}$. It remains to verify the uniform integrability of the terms on the right-hand side of (4.13). It can be assumed that p in (4.11) is sufficiently large. Thus, for example,

$$\begin{aligned}|\langle f(\alpha, X(s)), D_x g(X(s))\rangle \exp(\xi_{n,s})|^2 &\leq c_1(k + k_f|X(s)|)^2\hat{k}^2(1+|X(s)|^p)^2\exp(2\xi_{n,s})\\ &\leq c_2 + c_3|X(s)|^{2p+2}\exp(2\xi_{n,s}).\end{aligned}$$

By Lemma 2.1 it follows that

$$\sup_{\substack{n\in\mathbb{N}\\ s\in[0,t]}} \mathbb{E}|X(s)|^{2p+2}\exp(2\xi_{n,s}) < \infty.$$

The uniform integrability of the other terms in (4.13) is verified in a similar way. □

REMARK. If the operator A^{-1} is compact and $D_{xx}g(x)BB^*$ can be extended to a nuclear operator on H for all $x \in H$, then

$$\mathrm{tr}[(A^*)^{1/2-\gamma}D_{xx}g(x)BB^*(A^*)^{\gamma-1/2}] = \mathrm{tr}D_{xx}g(x)BB^*$$

(Theorem iii.8.2 of [19]) and the analogous equality is satisfied for the last term on the right-hand side of (4.12). The equality (4.12) then has the usual form, which is called the Itô formula.

Choose and fix $\alpha_1 \in \mathcal{A}$ and let $\eta = P_{\alpha_1}(1,0,\cdot)$ (recall (2.8)). Note that by [27] and (A3) all of the transition functions $P_\alpha(t,x,\cdot)$, $\alpha \in \mathcal{A}$, $t > 0$, and $x \in H$ are equivalent. The following lemma is Lemma 3 of [13].

LEMMA 4.5. *Let $\varphi : H \to U$ and $G : H \to U$ be bounded, Borel measurable functions and let $(G_n, n \in \mathbb{N})$ be a sequence of bounded, Borel measurable functions that converge to G in $\sigma(L^\infty(H,\eta,H), L^1(H,\eta,H))$ (i.e., in the weak* topology of $L^\infty(H,\eta,H)$). If* (A1)–(A3) *are satisfied, then*

$$\lim_{n\to\infty} \mathbb{E}\left(\int_0^t \langle \varphi(X(s)), G_n(X(s)) - G(X(s))\rangle_U \mathrm{d}s\right)^2 = 0, \tag{4.14}$$

where $(X(t), t \geq 0)$ satisfies (2.5) *and $\alpha \in \mathcal{A}$, $x \in H$ are arbitrary.*

The following result is a technical lemma which will play an important role in the proofs of the existence of an optimal control and the uniformly continuous dependence of the invariant measures on the parameter α.

PROPOSITION 4.6. *Let $(\alpha_n, n \in \mathbb{N})$ be a sequence in $\mathcal{A}$ that converges to $\alpha_0 \in \mathcal{A}$ and let $(h(\alpha_0,\cdot,u_n(\cdot)), n \in \mathbb{N})$ be a sequence that converges to $h(\alpha_0,\cdot,u(\cdot))$ in $\sigma(L^\infty(H,\eta,H), L^1(H,\eta,H))$. If* (A1)–(A3) *are satisfied then*

$$\lim_{n\to\infty} P_t^{\alpha_n,u_n}\varphi(x) = P_t^{\alpha_0,u}\varphi(x) \tag{4.15}$$

for each $\varphi \in \mathcal{M}(H)$, $x \in H$, and $t > 0$.

Proof. It easily follows that

(4.16)
$$\begin{aligned}|P_t^{\alpha_n,u_n}\varphi(x) - P_t^{\alpha_0,u}\varphi(x)| &\le |P_t^{\alpha_n,u_n}\varphi(x) - P_t^{\alpha_0,u_n}\varphi(x)| + |P_t^{\alpha_0,u_n}\varphi(x) - P_t^{\alpha_0,u}\varphi(x)| \\ &\le \sup|\varphi|\|P^{\alpha_n,u_n}(t,x,\cdot) - P^{\alpha_o,u_n}(t,x,\cdot)\| \\ &\quad + |P_t^{\alpha_0,u_n}\varphi(x) - P_t^{\alpha_0,u}\varphi(x)|.\end{aligned}$$

By Proposition 4.3 the first term on the right-hand side of (4.16) tends to zero as $n \to \infty$, so it suffices to show that for any subsequence $(u_{n_k},\, k \in \mathbb{N})$

$$\lim_{k\to\infty} \mathbb{E}\varphi(X(t))\exp(\xi_t^{\alpha_0,u_{n_k}}) = \mathbb{E}\varphi(X(t))\exp(\xi_t^{\alpha_0,u}), \tag{4.17}$$

where $(X(t),\, t \ge 0)$ is a solution of (2.1) with $\alpha = \alpha_0$. The sequence $(\exp(\xi_t^{\alpha_0,u_k}),\, k \in \mathbb{N})$ is bounded in $L^1(\Omega,\mathbb{P})$, so there is a subsequence denoted as the full sequence and a $Z \in L^1(\Omega,\mathbb{P})$ such that

$$\lim_{n\to\infty} \exp(\xi_t^{\alpha_0,u_n}) = Z \tag{4.18}$$

in $\sigma(L^1(\Omega,\mathbb{P}), L^\infty(\Omega,\mathbb{P}))$. Since φ is bounded, the equality (4.17) follows if $Z = \exp(\xi_t^{\alpha_0,u})$. Let $g = \bar{g}(\langle x,e_1\rangle,\dots,\langle x,e_n\rangle)$, where $(e_i,\, i \in \mathbb{N})$ is a complete orthonormal basis in H such that $e_i \in D(A^*)$, $\bar{g} \in C_0^\infty(\mathbb{R}^n)$ is arbitrary, and $n \in \mathbb{N}$. By Proposition 4.4 it follows that

$$\begin{aligned}g(X(r)) - \int_0^r &(\langle -A^*D_xg(X(s)), X(s)\rangle + \langle D_xg(X(s)), f(\alpha_0,X(s))\rangle \\ &+ \langle B^*D_xg(X(s)), h(\alpha_0,X(s),u_n(X(s)))\rangle_U \\ &+ \frac{1}{2}\mathrm{tr}[(A^*)^{1/2-\gamma}D_{xx}g(X(s))BB^*(A^*)^{\gamma-1/2}] \\ &+ \frac{1}{2}\mathrm{tr}[(A^*)^{1/2-\Delta}D_{xx}g(X(s))Q(A^*)^{\Delta-1/2}])\mathrm{d}s\end{aligned} \tag{4.19}$$

for $r \in [0,t]$ is a martingale with respect to $\mathbb{P}_x^{\alpha_0,u_n}$. Apply Lemma 4.1 with $\varphi(y) = B^*D_xg(y)$, $G_n(y) = h(\alpha_0,y,u_n(y))$, and $G(y) = h(\alpha_0,y,u(y))$ to obtain

$$\lim_{n\to\infty} \mathbb{E}\left(\int_0^t \langle B^*D_xg(X(s)), h(\alpha_0,X(s),u_n(X(s)) - h(\alpha_0,X(s),u(X(s)))\rangle_U\right)^2 \mathrm{d}s = 0$$

so there is a subsequence such that for each $\Gamma \in \mathcal{F}$

$$\begin{aligned}&\lim_{k\to\infty}\int_\Gamma\int_0^r \langle B^*D_xg(X(s)), h(\alpha_0,X(s),u_{n_k}(X(s)))\rangle_U \exp(\xi_t^{\alpha_0,u_{n_k}})\mathrm{d}\mathbb{P} \\ &\quad = \int_\Gamma\int_0^r \langle B^*D_xg(X(s)), h(\alpha_0,X(s),u(X(s)))\rangle_U Z\mathrm{d}\mathbb{P}\end{aligned} \tag{4.20}$$

for all $r \in [0,t]$ by (4.34) of [2] and (33) of [32]. It follows that (4.19) is a continuous martingale with respect to $Z\mathbb{P}(\mathrm{d}\omega)$, and by the weak uniqueness of the solutions of (2.1) it follows that $Z = \exp(\xi_t^{\alpha_0,u})$ (cf. [18]). □

REMARK 4.7. In the remainder of this section some continuity properties of the invariant measures corresponding to the solution of (2.1) are verified. One of the basic assumptions here is the tightness condition (A5). Using a Lyapunov condition, (A5) is verified in section 5. Furthermore, if (A1)–(A3) are satisfied, then for each $\alpha \in \mathcal{A}$ and $u \in \mathcal{U}$ the transition probabilities $(P^{\alpha,u}(t,x,\cdot),\ t>0,\ x\in H)$ are equivalent, which follows from the equivalence of the transition probabilities $(P_\alpha(t,x,\cdot),\ t>0,\ x\in H)$. This latter fact is an immediate consequence of the strong Feller property (a special case of Lemma 4.2) and irreducibility (Proposition 2.7 of [28]). From the equivalence of $(P^{\alpha,u}(t,x,\cdot),\ t>0,\ x\in H)$, it follows that for each $\alpha\in\mathcal{A}$ and $u\in\mathcal{U}$ the invariant measure $\mu(\alpha,u)$ is ergodic and unique (Proposition 2.5 of [31]).

The following lemma follows basically from Roxin [29] (cf. also the Appendix in [2]).

LEMMA 4.8. *Let $\alpha\in\mathcal{A}$ be fixed. If* (A2), (A6), *and* (A7) *are satisfied then*

$$\{(h(\alpha,\cdot,u(\cdot)),c(\cdot,u(\cdot))) : u\in\mathcal{U}\} \subset L^\infty(H,\eta,U\times\mathbb{R})$$

is compact in the $\sigma(L^\infty(H,\eta,U\times\mathbb{R}),L^1(H,\eta,U\times\mathbb{R}))$ topology.

PROPOSITION 4.9. *If* (A1)–(A7) *are satisfied then*

$$\lim_{\alpha\to\alpha_0}\sup_{u\in\mathcal{U}}\rho_*(\mu(\alpha,u),\mu(\alpha_0,u))=0 \tag{4.21}$$

and

$$\lim_{n\to\infty}\rho_*(\mu(\alpha_0,\hat{u}_n),\mu(\alpha_0,u_0))=0, \tag{4.22}$$

where μ is the invariant measure and ρ_* is a metric for the weak convergence of measures, $\alpha_0\in\mathcal{A}$, $u_0\in\mathcal{U}$ and $(\hat{u}_n,\ n\in\mathbb{N})$ is a sequence in $\mathcal{U}$ such that*

$$\lim_{n\to\infty} h(\alpha_0,\cdot,\hat{u}_n(\cdot)) = h(\alpha_0,\cdot,u_0(\cdot))$$

in the $\sigma(L^\infty(H,\eta,U),L^1(H,\eta,U))$ topology.

Proof. Let $(\alpha_n,\ n\in\mathbb{N})$ be a sequence in $\mathcal{A}$ that converges to α_0 and let $(u_n,\ n\in\mathbb{N})$ be a sequence in $\mathcal{U}$. By Lemma 4.8 there exist subsequences (again denoted $(\alpha_n,\ n\in\mathbb{N})$ and $(u_n,\ n\in\mathbb{N})$) so that $\alpha_n\to\alpha_0$ and $(h(\alpha_0,\cdot,u_n(\cdot)),\ n\in\mathbb{N})$ converges to $h(\alpha_0,\cdot,u(\cdot))$ for some $u\in\mathcal{U}$ in the $\sigma(L^\infty(H,\eta,U),L^1(H,\eta,U))$ topology. To verify (4.21) it is necessary to show that from any such sequences there are subsequences $(\alpha_{n_k},\ k\in\mathbb{N})$ and $(u_{n_k},\ k\in\mathbb{N})$ such that

$$\lim_{k\to\infty}\rho_*(\mu(\alpha_{n_k},u_{n_k}),\mu(\alpha_0,u_{n_k}))=0. \tag{4.23}$$

By the tightness condition (A5) there are measures ν_1 and ν_2 such that $\mu(\alpha_{n_k},u_{n_k})\to\nu_1$ and $\mu(\alpha_0,u_{n_k})\to\nu_2$ as $k\to\infty$ in the weak* topology. It is shown that ν_1 is an invariant measure for $P_t^{\alpha_0,u}$; that is, for each $\varphi\in C_b(H)$,

$$\int\varphi\mathrm{d}\nu_1 = \int P_t^{\alpha_0,u}\varphi\mathrm{d}\nu_1 \tag{4.24}$$

for $t\geq 0$. Again, for notational simplicity, let the subsequences be denoted as $(\alpha_n,\ n\in\mathbb{N})$ and $(u_n,\ n\in\mathbb{N})$. It easily follows that

$$
\begin{aligned}
\left|\int \varphi \mathrm{d}\nu_1 - \int P_t^{\alpha_0,u}\varphi \mathrm{d}\nu_1\right| &\le \left|\int \varphi \mathrm{d}\nu_1 - \int \varphi \mathrm{d}\mu(\alpha_n,u_n)\right| \\
&+ \left|\int \varphi \mathrm{d}\mu(\alpha_n,u_n) - \int P_t^{\alpha_n,u_n}\varphi \mathrm{d}\mu(\alpha_n,u_n)\right| \\
&+ \left|\int P_t^{\alpha_n,u_n}\varphi \mathrm{d}\mu(\alpha_n,u_n) - \int P_t^{\alpha_0,u}\varphi \mathrm{d}\mu(\alpha_n,u_n)\right| \\
&+ \left|\int P_t^{\alpha_0,u}\varphi \mathrm{d}\mu(\alpha_n,u_n) - \int P_t^{\alpha_0,u}\varphi \mathrm{d}\nu_1\right| \\
&:= I_n^1 + I_n^2 + I_n^3 + I_n^4.
\end{aligned} \tag{4.25}
$$

It follows that $I_n^1 + I_n^4 \to 0$ as $n \to \infty$ because $\mu(\alpha_n, u_n) \to \nu_1$ in the weak* topology and $I_n^2 \equiv 0$ because $\mu(\alpha_n, u_n)$ is $P_t^{\alpha_n,u_n}$ invariant. Furthermore,

$$
I_n^3 \le \int_K |P_t^{\alpha_n,u_n}\varphi - P_t^{\alpha_0,u}\varphi| \mathrm{d}\mu(\alpha_n,u_n) + 2\max|\varphi|\mu(\alpha_n,u_n)(H \setminus K) \tag{4.26}
$$

for any compact $K \subset H$. By Proposition 4.6 and Lemma 4.2, $P_t^{\alpha_n,u_n}\varphi \to P_t^{\alpha_0,u}\varphi$ uniformly on compact subsets of H, so this fact and (A5) imply that

$$
\lim_{n\to\infty} I_n^3 = 0.
$$

Therefore, (4.24) is satisfied. Since $\mu(\alpha_0, u)$ is the unique invariant measure for $P_t^{\alpha_0,u}$, $\nu_1 = \mu(\alpha_0, u)$. In the same way it follows that $\nu_2 = \mu(\alpha_0, u)$, which verifies (4.23) and thereby (4.21). To verify (4.22) note that given any sequence $(\mu(\alpha_0, \hat{u}_n),\ n \in \mathbb{N})$ there is a subsequence $(\mu(\alpha_0, \hat{u}_{n_k}),\ k \in \mathbb{N})$ converging to a measure ν_3 in the weak* topology. By analogy to (4.25) it can be shown that ν_3 is $P_t^{\alpha_0,u_0}$ invariant so $\nu_3 = \mu(\alpha_0, u_0)$. □

In Proposition 4.9, (4.21) gives the uniformly continuous dependence of invariant measures on the parameter α. Using Propositions 4.9 and 4.3 a strong version of (4.21) is obtained now.

PROPOSITION 4.10. *If* (A1)–(A7) *are satisfied then*

$$
\lim_{\alpha\to\alpha_0} \sup_{u\in\mathcal{U}} \|\mu(\alpha,u) - \mu(\alpha_0,u)\| = 0, \tag{4.27}
$$

where $\|\cdot\|$ *is the variation norm.*

Proof. It easily follows that

$$
\begin{aligned}
\sup_{u\in\mathcal{U}} \|\mu(\alpha,u) - \mu(\alpha_0,u)\| &= \sup_{u\in\mathcal{U}} \sup_{\substack{|\varphi|\le 1\\ \varphi\in C_b}} \left|\int_H \varphi \mathrm{d}\mu(\alpha,u) - \int_H \varphi \mathrm{d}\mu(\alpha_0,u)\right| \\
&= \sup_{u\in\mathcal{U}} \sup_{|\varphi|\le 1} \left|\int_H P_1^{\alpha,u}\varphi \mathrm{d}\mu(\alpha,u) - \int_H P_1^{\alpha_0,u}\varphi \mathrm{d}\mu(\alpha_0,u)\right| \\
&\le 2\sup_{\alpha,u} \mu(\alpha,u)(H \setminus K) + \int_K \sup_u \|P^{\alpha,u}(1,x,\cdot) - P^{\alpha_0,u}(1,x,\cdot)\| \mu(\alpha,u)(\mathrm{d}x) \\
&+ \sup_{u,\varphi} \left|\int_K P_1^{\alpha,u}\varphi \mathrm{d}\mu(\alpha,u) - \int_K P_1^{\alpha_0,u}\varphi \mathrm{d}\mu(\alpha_0,u)\right|
\end{aligned} \tag{4.28}
$$

for any compact set $K \subset H$. By (A5) the first term on the right-hand side of (4.28) can be made arbitrarily small by choosing a suitable compact set K, and by Proposition 4.3 the second term converges to zero almost surely as $\alpha \to \alpha_0$. Furthermore, by Lemma 4.2 the family of functions $(P_1^{\alpha_0,u}\varphi, |\varphi| \le 1,\ u \in \mathcal{U})$ is uniformly continuous on K, so for sequences $(u_n,\ n \in \mathbb{N})$ and $(\varphi_n,\ n \in \mathbb{N})$, where $u_n \in \mathcal{U}$ and $\varphi_n \in C_b$ for $n \in \mathbb{N}$, there are subsequences $(u_{n_k},\ k \in \mathbb{N})$ and $(\varphi_{n_k},\ k \in \mathbb{N})$ and a $\psi \in C_b(K)$ such that $P_1^{\alpha_0,u_{n_k}}\varphi_{n_k}(x) \to \psi(x)$, as $k \to \infty$, uniformly in $x \in K$. Now the third term on the right-hand side of (4.28) is shown to converge to zero.

$$
\begin{aligned}
&\left|\int_K P_1^{\alpha_0,u_{n_k}}\varphi_{n_k}\mathrm{d}\mu(\alpha,u_{n_k}) - \int_K P_1^{\alpha_0,u_{n_k}}\varphi_{n_k}\mathrm{d}\mu(\alpha_0,u_{n_k})\right| \\
&\qquad \le \int_K |P_1^{\alpha_0,u_{n_k}}\varphi_{n_k} - \psi|\mathrm{d}\mu(\alpha,u_{n_k}) + \left|\int_K \psi\mathrm{d}\mu(\alpha,u_{n_k}) - \int_K \psi\mathrm{d}\mu(\alpha_0,u_{n_k})\right| \\
&\qquad\quad + \int_K |\psi - P_1^{\alpha_0,u_{n_k}}\varphi_{n_k}|\mathrm{d}\mu(\alpha_0,u_{n_k}) \\
&\qquad := I_n^1 + I_n^2 + I_n^3.
\end{aligned} \tag{4.29}
$$

By the uniform convergence $P_1^{\alpha_0,u_{n_k}}\varphi_{n_k} \to \psi$ on K it follows that $I_n^1 + I_n^3 \to 0$ as $n \to \infty$, and by (4.21) it follows that $I_n^2 \to 0$ as $n \to \infty$. This proves that the last term on the right-hand side of (4.28) tends to zero as $\alpha \to \alpha_0$. □

REMARK 4.11. The strong continuous dependence of the invariant measures on a parameter in Proposition 4.10 can be of independent interest even for equations without control. If the parameter occurs linearly in the generator of even a very simple example of an Ornstein–Uhlenbeck process then the invariant measures may not depend continuously on α in the variation norm. For example, consider the stochastic differential equation

$$
\mathrm{d}X(t) + \alpha AX(t)\mathrm{d}t = \mathrm{d}W(t), \qquad X(0) = x, \tag{4.30}
$$

where A and $(W(t),\ t \ge 0)$ are the same as in (2.1) and $\alpha \in [1/2, 2]$. If

$$
\int_0^\infty |S(t)|^2_{\mathcal{L}_2(H)}\mathrm{d}t < \infty \tag{4.31}
$$

then (4.30) has a unique mild solution that is a continuous H-valued process. If (4.31) is satisfied and $\alpha \in [1/2, 2]$, then there is a unique invariant measure $\mu(\alpha)$ for the solution of (4.30), where $\mu(\alpha) = N(0, \alpha^{-1}\tilde{Q})$ and $\tilde{Q} = \int_0^\infty S(t)S^*(t)\mathrm{d}t$. It is easy to verify that the family of measures $(\mu(\alpha),\ \alpha \in [1/2, 2])$ is tight and $\mu(\alpha) \xrightarrow{w^*} \mu(1)$ as $\alpha \to 1$. However, the variation norm convergence $\mu(\alpha) \to \mu(1)$ occurs if and only if $\dim H < \infty$ because the operator $(\alpha^{-1}\tilde{Q})\tilde{Q}^{-1} - I = (\alpha^{-1} - 1)I$ is not Hilbert–Schmidt for $\alpha \ne 1$ and $\dim H = \infty$, and so $\mu(\alpha)$ and $\mu(1)$ are singular by the well-known dichotomy for Gaussian measures. This occurs even in the strong Feller case when the solution of (4.30) converges in law to the invariant measure in the variation norm for each fixed α. For a specific example of this, consider the linear SPDE

$$
\frac{\partial w}{\partial t}(t,\xi) = \alpha\frac{\partial^2 w}{\partial \xi^2}(t,\xi) + n(t,\xi), \tag{4.32}
$$

where $\alpha \in [1/2, 2]$, $(t, \xi) \in \mathbb{R}_+ \times (0, 1)$, $w(0, \xi) = w_0(\xi)$, $w(t, 0) = v(t, 1) = 0$, and $(n(t, \xi),\ t \ge 0,\ \xi \in [0, 1])$ is a space-time white noise which can be expressed as an equation of the form (4.30) for $H = L^2(0, 1)$ (cf. Example 7.1).

5. Existence and tightness of invariant measures. In this section some more explicit sufficient conditions for the validity of (A5) are given by means of some Lyapunov-type inequalities. Throughout this section it is assumed that

(T1) $$A^{-1} \text{ is compact.}$$

Since the semigroup $S(\cdot)$ generated by $-A$ is assumed to be analytic and exponentially stable, there exist some $M > 0$ and $\omega > 0$ such that

(T2) $$|S(t)|_{\mathcal{L}(D_A^{-\delta},H)} \le M\mathrm{e}^{-\omega t}t^{-\delta}$$

for all $t > 0$ and $\delta \le 0$. (The constants M and ω will play some role in the Lyapunov-type conditions given below.)

While in the other sections of this paper the negativity of $-A$ is assumed merely for convenience (because $A + \beta I$ can be used instead of A, and βI can be added to f), in this section it is essential.

Define $\mu_T^{\alpha,u}$ as follows:

$$\mu_T^{\alpha,u}(\cdot) = \frac{1}{T}\int_0^T P^{\alpha,u}(t,0,\cdot)\mathrm{d}t \tag{5.1}$$

for $\alpha \in \mathcal{A}$, $u \in \mathcal{U}$, and $T > 0$. Since the solution of (2.1) is Feller, to verify (A5) it suffices to show that the family of measures $(\mu_T^{\alpha,u},\ \alpha \in \mathcal{A},\ u \in \mathcal{U},\ T \ge 1)$ is tight. In the following proposition it is shown that the tightness of $(\mu_T^{\alpha,u},\ \alpha \in \mathcal{A},\ u \in \mathcal{U},\ T \ge 1)$ follows from a similar property, where compact sets are replaced by balls (5.2). Note that (5.2) does not guarantee the existence of an invariant measure in general (cf. [33]).

PROPOSITION 5.1. *If* (A1), (A2), *and* (T1) *are satisfied and*

$$\lim_{n\to\infty} \mu_T^{\alpha,u}(H \setminus B_n) = 0, \tag{5.2}$$

where the convergence is uniform in $\alpha \in \mathcal{A}$, $u \in \mathcal{U}$, *and* $T \ge 1$, *and* $B_n = \{x \in H : |x| \le n\}$, *then the family of measures* $(\mu_T^{\alpha,u},\ \alpha \in \mathcal{A},\ u \in \mathcal{U},\ T \ge 1)$ *is tight.*

Proof. The weak solution of (2.1) satisfies the equation

$$\begin{aligned} X(t) &= S(t)x + \int_0^t S(t-r)f(\alpha, X(r))\mathrm{d}r \\ &\quad + \int_0^t S(t-r)Bh(\alpha, X(r), u(X(r)))\mathrm{d}r + Z_1(t) + Z_2(t), \end{aligned} \tag{5.3}$$

where

$$Z_1(t) = \int_0^t S(t-r)B\mathrm{d}V^*(r)$$

and

$$Z_2(t) = \int_0^t S(t-r)Q^{1/2}\mathrm{d}W(r)$$

for $t \ge 0$. By (A1) and Lemma 2.2 of [30] it follows that

$$\mathbb{E}_x^{\alpha,u}|Z_1(t)|_\delta^2 + \mathbb{E}_x^{\alpha,u}|Z_2(t)|_\delta^2 \le M_1$$

for $t \in [0,T]$, where $T > 0$ is fixed and the constant M_1 (as well as $M_2, \dots, M_5$ below) does not depend on $\alpha \in \mathcal{A}$, $u \in \mathcal{U}$, and $x \in H$, and $|\cdot|_\delta$ is the D_A^δ norm and

$\delta \in (0, \min(\varepsilon, \gamma, \Delta))$ is fixed. It follows that

$$\mathbb{E}_x^{\alpha,u}|X(t)|_\delta \le M_2|x|t^{-\delta} + \int_0^t \frac{M_3}{(t-s)^{1-\varepsilon+\delta}}\mathbb{E}_x^{\alpha,u}|X(s)|_\delta \mathrm{d}s + M_4 \tag{5.4}$$

for $t \in (0,T]$ and $x \in H$, so the generalized Gronwall lemma (Theorem 7.1 of [21]) implies that

$$\mathbb{E}_x^{\alpha,u}|X(T)|_\delta \le M_5(1+|x|) \tag{5.5}$$

for $\alpha \in \mathcal{A}$, $u \in \mathcal{U}$, and $x \in H$. By the Chebyshev inequality it follows that

$$\sup_{|y|\le R} \mathbb{P}_y^{\alpha,u}(|X(t)|_\delta \ge n) \le \frac{1}{n} M_5(1+R) \tag{5.6}$$

for $n \in \mathbb{N}$, $R > 0$, $\alpha \in \mathcal{A}$, and $u \in \mathcal{U}$.

Let $K_n \subset H$ be given by

$$K_n = \mathrm{Cl}_H A^{-\delta} B_n \tag{5.7}$$

for $n \in \mathbb{N}$, where Cl_H is the closure in H. Since $A^{-\delta}$ is a compact operator, K_n is compact in H. It follows that

(5.8)

$$\begin{aligned}
\frac{1}{T}\int_1^T P^{\alpha,u}(t,0,H\setminus K_n)\mathrm{d}t &= \frac{1}{T}\int_1^T \int_H P^{\alpha,u}(1,y,H\setminus K_n)P^{\alpha,u}(t-1,0,\mathrm{d}y)\mathrm{d}t \\
&= \frac{T-1}{T}\int_H P^{\alpha,u}(1,y,H\setminus K_n)\mu_{T-1}^{\alpha,u}(\mathrm{d}y) \\
&\le \mu_{T-1}^{\alpha,u}(H\setminus B_R) + \mu_{T-1}^{\alpha,u}(B_R)\sup_{|y|\le R} P^{\alpha,u}(1,y,H\setminus K_n) \\
&\le \mu_{T-1}^{\alpha,u}(H\setminus B_R) + \frac{1}{n}M_5(1+R)
\end{aligned}$$

for each $R > 0$. By (5.2) the right-hand side tends to zero as $n \to \infty$ uniformly in $\alpha \in \mathcal{A}$, $u \in \mathcal{U}$, and $T \ge 1$. □

In Theorem 5.3 below, the condition (5.2) is verified by a Lyapunov functional that completes the verification of (A5). Let V be given by

$$V = 2\int_0^\infty S(r)S^*(r)\mathrm{d}r. \tag{5.9}$$

If (T2) is satisfied then $V \in \mathcal{L}(H)$ is well defined, $V = V^*$, and $V \ge 0$.

The following estimates are easily verified.

LEMMA 5.2. *For $\beta, \lambda \in \mathbb{R}_+$ with $\beta + \lambda < 1$, $V \in \mathcal{L}(D_A^{-\beta}, D_A^{\lambda})$ and the following inequality is satisfied:*

$$|V|_{\mathcal{L}(D_A^{-\beta},D_A^{\lambda})} \le 2M^2(2\omega)^{\beta+\lambda-1}\Gamma(1-\beta-\lambda), \tag{5.10}$$

where Γ is the gamma function, M and ω are given in (T2)*, and V is given by* (5.9)*. Furthermore, if A is self-adjoint, then $V = A^{-1}$ and*

$$|V|_{\mathcal{L}(D_A^{-\beta},D_A^{\lambda})} \le \omega^{\beta+\lambda-1} \tag{5.11}$$

and ω is the first eigenvalue of A.

THEOREM 5.3. *If* (A1), (A2), (T1), (T2) *are satisfied and either*

$$M^2\omega^{-1}k_f + 2^{1-\varepsilon}M^2\omega^{-\varepsilon}|B|_{\mathcal{L}(U,D_A^{\varepsilon-1})}\Gamma(\varepsilon)k_h < 1 \tag{5.12}$$

for A not self-adjoint or

$$\omega^{-1}k_f + |B|_{\mathcal{L}(U,D_A^{\varepsilon-1})}\omega^{-\varepsilon}k_h < 1 \tag{5.13}$$

for A self-adjoint, where M and ω are given in (T2) *and Γ is the gamma function, then the condition* (A5) *is satisfied. In particular,* (5.12) *and* (5.13) *are satisfied if* $|f|$ *and* $|h|_U$ *are bounded uniformly with respect to* $\alpha \in \mathcal{A}$ *and* $u \in \mathcal{U}$.

Proof. By Proposition 5.1 it suffices to verify (5.2). Use Proposition 4.4 with $g(x) = \langle Vx, x\rangle$ so that $D_x g(x) = Vx$, $D_{xx}g(x) = V$,

$$|V|_{\mathcal{L}(H,D_{A^*}^{1-\varepsilon})} \le 2M^2(2\omega)^{-\varepsilon}\Gamma(\varepsilon), \tag{5.14}$$

and $\langle Ax, D_x g(x)\rangle = |x|^2$ for $x \in D_A^1$, and by Lemma 5.2 $V \in \mathcal{L}(D_A^{\gamma-1/2}, D_{A^*}^{1/2-\gamma}) \cap \mathcal{L}(D_A^{\Delta-1/2}, D_{A^*}^{1/2-\Delta})$. Thus the assumptions of Proposition 4.4 are satisfied, and by (A2) and (5.14),

$$\begin{aligned}
&\mathbb{E}_x^{\alpha,u}\langle VX(t), X(t)\rangle - \langle Vx, x\rangle \\
&\le \mathbb{E}_x^{\alpha,u}\int_0^t (|X(s)|^2(-1 + M^2\omega^{-1}k_f + 2^{1-\varepsilon}M^2\omega^{-\varepsilon}|B|_{\mathcal{L}(U,D_A^{\varepsilon-1})}\Gamma(\varepsilon)k_h) \\
&\quad + c_1|X(s)| + c_2)\mathrm{d}s
\end{aligned} \tag{5.15}$$

for $t \ge 0$, where the constants c_1 and c_2 (as well as the constants c_3 and c_4 below) do not depend on $\alpha \in \mathcal{A}$ and $u \in \mathcal{U}$. Choosing r such that $M^2\omega^{-1}k_f + 2^{1-\varepsilon}M^2\omega^{-2}|B|_{\mathcal{L}(U,D_A^{\varepsilon-1})}\Gamma(\varepsilon)\cdot k_h < r < 1$, it follows that

$$\mathbb{E}_x^{\alpha,u}\langle VX(t), X(t)\rangle - \langle Vx, x\rangle \le \mathbb{E}_x^{\alpha,u}\int_0^t ((r-1)|X(t)|^2 + c_3)\mathrm{d}s$$

for $t \ge 0$, and since $V \ge 0$ it follows that

$$\sup_{t\ge 1}\frac{1}{t}\int_0^t \mathbb{E}_x^{\alpha,u}|X(s)|^2\mathrm{d}s \le \sup_{t\ge 1}\frac{\langle Vx, x\rangle}{t(1-r)} + \frac{c_3}{1-r} \le c_4. \tag{5.16}$$

By (5.16) and the Chebyshev inequality it follows that (5.2) is satisfied. If A is self-adjoint then (5.11) can be used instead of (5.10). □

6. The existence of an optimal control. Recall that the control problem is described by the system (2.1) and the cost functional

$$J(\alpha, u) = \limsup_{T\to\infty}\mathbb{E}_x^{\alpha,u}\frac{1}{T}\int_0^T c(X(s), u(X(s)))\mathrm{d}s, \tag{6.1}$$

and the optimal cost is $J^*(\alpha) = \inf_{u\in\mathcal{U}} J(\alpha, u)$. If (A1)–(A3), (A5), and (A6) are satisfied then the following equality is satisfied:

$$J(\alpha, u) = \int_H c(y, u(y))\mu(\alpha, u)(\mathrm{d}y) \tag{6.2}$$

(cf. Remark 4.7), so the cost $J(\alpha, u)$ does not depend on the initial condition $X(0) = x \in H$. In this section the existence of an optimal control for the control problem (2.1) and (6.1) with a fixed parameter $\alpha \in \mathcal{A}$ and the continuity of the optimal cost $J^* : \mathcal{A} \to \mathbb{R}$ are verified. In Lemma 6.1 and Theorem 6.2 the parameter is fixed, so it is suppressed for notational convenience.

Recall that $P(t, x, \Gamma)$ is given in (2.8) and $\eta = P(1, 0, \cdot)$.

LEMMA 6.1. *Let $(A_n,\, n \in \mathbb{N})$ be a sequence in $\mathcal{B}(H)$ such that $\eta(A_n) \to 0$ as $n \to \infty$. If* (A1)–(A3) *and* (A5) *are satisfied then*

$$\lim_{n\to\infty} \sup_{u\in\mathcal{U}} \mu(u)(A_n) = 0. \tag{6.3}$$

Proof. Since $P(1, \cdot, \cdot) : H \to \mathcal{P}(H)$ is continuous in the variation norm and $P(1, x, \cdot)$ and η are equivalent for each $x \in K$, where $K \subset H$ is compact, it follows that

$$\lim_{n\to\infty} \sup_{x\in K} P(1, x, A_n) = 0. \tag{6.4}$$

Since, for a fixed $\alpha \in \mathcal{A}$, $|h|_U$ is bounded it follows that

$$\begin{aligned} \sup_{u\in\mathcal{U},\ x\in K} P^u(1, x, A_n) &= \sup_{u\in\mathcal{U}} \sup_{x\in K} \mathbb{E}_x 1_{A_n}(X(1)) \exp(\xi_1^u) \\ &\le \sup_{x\in K} (P(1, x, A_n))^{1/2} \exp(\sup |h|^2). \end{aligned} \tag{6.5}$$

The right-hand side of this inequality tends to zero as $n \to \infty$ by (6.4).

Finally it follows that

$$\begin{aligned} \sup_{u\in\mathcal{U}} \mu(u)(A_n) &= \sup_{u\in\mathcal{U}} \int_H P^u(1, x, A_n)\mu(u)(\mathrm{d}x) \\ &\le \sup_{u\in\mathcal{U}} \mu(u)(H \setminus K) + \int_K \sup_{u\in\mathcal{U}} P^u(1, x, A_n)\mu(u)(\mathrm{d}x). \end{aligned}$$

By (6.5) and the tightness of the family of measures $(\mu(u),\, u \in \mathcal{U})$ the right-hand side of this inequality tends to zero as $n \to \infty$.

THEOREM 6.2. *If* (A1)–(A3) *and* (A5)–(A7) *are satisfied for each $\alpha \in \mathcal{A}$, then there is an optimal control for the control problem given by* (2.1) *and* (6.1).

Proof. Let $(u_n,\, n \in \mathbb{N})$ be a sequence in $\mathcal{U}$ such that there is a subsequence in $(u_n,\, n \in \mathbb{N})$ denoted as $(u_n,\, n \in \mathbb{N})$ for notational convenience, such that

$$\lim_{n\to\infty} (h(\cdot, u_n(\cdot)), c(\cdot, u_n(\cdot))) = (h(\cdot, u(\cdot)), c(\cdot, u(\cdot))) \tag{6.6}$$

in the $\sigma(L^\infty(H, \eta, U \times \mathbb{R}), L^1(H, \eta, U \times \mathbb{R}))$ topology. To verify that u is an optimal control it is necessary to prove that for any subsequence $(u_{n_k},\, k \in \mathbb{N})$,

$$\lim_{k\to\infty} J(u_{n_k}) = J(u). \tag{6.7}$$

As in Lemma 4.5 it follows that

$$\lim_{n\to\infty} \mathbb{E}\left(\int_0^t (c(X(s), u_n(X(s))) - c(X(s), u(X(s))))\mathrm{d}s \right)^2 = 0,$$

where $(X(t),\, t \geq 0)$ satisfies (2.5) with $X(0) = x \in H$ arbitrary (cf. Theorem 2 of [13]). As in the passage to the limit in the proof of Proposition 4.4 it follows that

$$\lim_{n\to\infty} \mathbb{E}_x^{u_n} \int_0^t c(X(s), u_n(X(s)))\mathrm{d}s = \mathbb{E}_x^u \int_0^t c(X(s), u(X(s)))\mathrm{d}s \tag{6.8}$$

for a subsequence again denoted by $(u_n,\, n \in \mathbb{N})$. By Egorov's theorem the convergence in (6.8) is uniform in x except possibly on a set of arbitrarily small η-measure. This fact and Lemma 6.1 imply that

(6.9)

$$\lim_{n\to\infty} \int_H \left| \mathbb{E}_x^{u_n} \int_0^t c(X(s), u_n(X(s)))\mathrm{d}s - \mathbb{E}_x^u \int_0^t c(X(s), u(X(s)))\mathrm{d}s \right| \mu(u_n)(\mathrm{d}x) = 0.$$

For each fixed $t > 0$ it follows that

(6.10)

$$\begin{aligned}
&|J(u_n) - J(u)| \\
&\quad = \left| \int_H c(y, u_n(y))\mu(u_n)(\mathrm{d}y) - \int_H c(y, u(y))\mu(u)(\mathrm{d}y) \right| \\
&\quad \leq \left| \frac{1}{t} \int_0^t \left[\int_H \mathbb{E}_y^{u_n} c(X(s), u_n(X(s)))\mu(u_n)(\mathrm{d}y) \right.\right. \\
&\qquad\qquad \left.\left. - \int_H \mathbb{E}_y^u c(X(s), u(X(s)))\mu(u)(\mathrm{d}y) \right] \mathrm{d}s \right| \\
&\quad \leq \frac{1}{t} \int_H \left| \mathbb{E}_y^{u_n} \int_0^t c(X(s), u_n(X(s)))\mathrm{d}s - \mathbb{E}_y^u \int_0^t c(X(s), u(X(s)))\mathrm{d}s \right| \mu(u_n)(\mathrm{d}y) \\
&\qquad + \frac{1}{t} \int_0^t \left| \int_H \mathbb{E}_y^u c(X(s), u(X(s)))\mu(u_n)(\mathrm{d}y) - \int_H \mathbb{E}_y^u c(X(s), u(X(s)))\mu(u)(\mathrm{d}y) \right| \mathrm{d}s \\
&\quad := I_n^1 + I_n^2.
\end{aligned}$$

By (6.9) it suffices to show that $I_n^2 \to 0$ as $n \to \infty$. Since $c(\cdot, u(\cdot))$ is bounded and Borel measurable, the strong Feller property (Lemma 4.2) implies that

$$\mathbb{E}_\cdot c(X(s), u(X(s))) : H \to \mathbb{R}$$

is continuous for each $s > 0$ where $\mathbb{E}_x c(X(s), u(X(s))) = P_s^u c(\cdot, u(\cdot))(x)$. So by (4.22) and the dominated convergence theorem, $I_n^2 \to 0$ as $n \to \infty$. □

THEOREM 6.3. *If* (A1)–(A7) *are satisfied then the optimal cost* $J^* : \mathcal{A} \to \mathbb{R}$ *is continuous.*

Proof. It follows that

$$\sup_{u\in\mathcal{U}} |J(\alpha, u) - J(\alpha_0, u)| \leq \sup |c| \sup_{u\in\mathcal{U}} \|\mu(u, \alpha) - \mu(u, \alpha_0)\|. \tag{6.11}$$

By Proposition 4.10 it follows that the right-hand side of this inequality tends to zero as $n \to \infty$. Given $\varepsilon > 0$ there is a $\delta > 0$ such that if $|\alpha - \alpha_0| < \delta$ then

$$\sup_{u\in\mathcal{U}} |J(\alpha, u) - J(\alpha_0, u)| < \varepsilon.$$

Let $u_\alpha \in \mathcal{U}$ be an optimal control for the control problem (2.1) and (6.1), that is, $J^*(\alpha) = J(\alpha, u_\alpha)$ for $\alpha \in \mathcal{A}$. Since $J(\alpha, u_{\alpha_0}) \geq J(\alpha, u_\alpha)$ it follows that $J(\alpha_0, u_{\alpha_0}) \geq J(\alpha, u_\alpha) - \varepsilon$. Since $J(\alpha_0, u_{\alpha_0}) \leq J(\alpha_0, u_\alpha)$ it follows that $J(\alpha_0, u_{\alpha_0}) \leq J(\alpha, u_\alpha) + \varepsilon$ for $\alpha \in \mathcal{A}$ and $|\alpha - \alpha_0| < \delta$. □

7. Some Examples.

EXAMPLE 7.1. Consider the scalar stochastic parabolic partial differential equation

$$\frac{\partial v}{\partial t}(t,\xi) = Lv(t,\xi) + F(\alpha, v(t,\xi)) + n(t,\xi) \tag{7.1}$$

for $(t,\xi) \in \mathbb{R}_+ \times (0,1)$ with the initial and boundary conditions

$$v(0,\xi) = v_0(\xi), \tag{7.2}$$

$$\frac{\partial v}{\partial \xi}(t,0) = h_1(\alpha, v(t,\cdot), u(v(t,\cdot))) + \dot{\beta}_1(t), \tag{7.3}$$

$$\frac{\partial v}{\partial \xi}(t,1) = h_2(\alpha, v(t,\cdot), u(v(t,\cdot))) + \dot{\beta}_2(t), \tag{7.4}$$

where n denotes a space-dependent Gaussian noise that is white in time, β_1 and β_2 are one-dimensional standard Wiener processes, and these three processes are mutually independent. Furthermore,

$$Lv = a(\xi)\frac{\partial^2}{\partial \xi^2}v + b(\xi)\frac{\partial}{\partial \xi}v + c(\xi),$$

where $a, b, c \in C^\infty([0,1])$, $a > 0$, $c < 0$, $F : \mathcal{A} \times \mathbb{R} \to \mathbb{R}$, $h_i : \mathcal{A} \times H \times \mathcal{K} \to \mathbb{R}$, $i = 1,2$, where $H = L^2(0,1)$, $\mathcal{A} \subset \mathbb{R}^{d_1}$ is compact, $\mathcal{K} \subset \mathbb{R}^k$ is a compact product of intervals, $F(\alpha,\cdot) : \mathbb{R} \to \mathbb{R}$ is Lipschitz continuous, $h_i(\alpha,\cdot,\cdot) : H \times \mathcal{K} \to \mathbb{R}$, $i = 1,2$, is continuous and bounded for each $\alpha \in \mathcal{A}$ with at most linear growth that is uniform with respect to $\alpha \in \mathcal{A}$, and

$$|F(\alpha,\xi) - F(\beta,\xi)| + \sum_{i=1}^{2} |h_i(\alpha,x,u) - h_i(\beta,x,u)| \ \le \omega(|\alpha-\beta|)(1+\max(|x|,|\xi|)) \tag{7.5}$$

for $\alpha, \beta \in \mathcal{A}$, $\xi \in \mathbb{R}$, $x \in H$, and $u \in \mathcal{K}$, where ω satisfies the properties in (A2). The system of equations (7.1)–(7.4) can be rewritten in the form of (2.1) in a natural way, where $H = L^2(0,1)$, $U = \mathbb{R}^2$, $A = -L$ with

$$\mathrm{Dom}(A) = \left\{ \varphi : \varphi \in H^2(0,1), \frac{\partial}{\partial \xi}\varphi(0) = \frac{\partial}{\partial \xi}\varphi(1) = 0 \right\},$$

$f(\alpha,x)(\xi) = F(\alpha, X(\xi))$, $x \in H$, $\xi \in (0,1)$, and $h = [h_1, h_2]$. The operator B is defined as $B = \hat{A}N$, where $N \in \mathcal{L}(\mathbb{R}^2, D_A^\varepsilon)$, $\varepsilon < 3/4$ is the Neumann map corresponding to the elliptic Neumann problem

$$Lz(\xi) = 0, \qquad \xi \in (0,1), \tag{7.6}$$

$$\frac{\partial z}{\partial \xi}(0) = g_1, \qquad \frac{\partial z}{\partial \xi}(1) = g_2 \tag{7.7}$$

for $g_1, g_2 \in \mathbb{R}$, and $\hat{A} \in \mathcal{L}(D_A^\varepsilon, D_A^{\varepsilon-1})$ is the isomorphic extension of the operator A to D_A^ε. (See [26] for the theory of Dirichlet and Neumann maps, [16] for the identification of D_A^ε with the corresponding Sobolev spaces, and [22] or [27] for the mathematical justification of the form (2.1) for the equations (7.1)–(7.4).) Thus it follows that $B \in \mathcal{L}(U, D_A^{\varepsilon-1})$ for $\varepsilon < 3/4$ in the present case. Now it is verified that (A1) and (A3) are satisfied, where $Q^{1/2} = A^{-\eta}\Gamma$ with $\eta \ge 0$ and $\Gamma, \Gamma^{-1} \in \mathcal{L}(H)$. Since $A^{-\delta}$ is Hilbert–Schmidt for $\delta > 1/4$ (cf. Example 6.1 of [12]) it follows that

$Q^{1/2} \in \mathcal{L}_2(H, D_A^{\Delta-1/2})$ for $\Delta < 1/4 + \eta$. Since the space U is finite-dimensional, $B \in \mathcal{L}_2(U, D_A^{\gamma-1/2})$ for $\gamma < \varepsilon - 1/2$ and γ is positive if $\varepsilon > 1/2$. To verify (A3) use Proposition 3.4, which shows that (A3) is satisfied if $\eta \in [0, \varepsilon - 1/2)$ if $\varepsilon \geq 1/2$. Thus the assumptions (A1) and (A3) are satisfied for $\eta \in [0, 1/4)$ so that ε, γ, and Δ can be chosen to satisfy $\varepsilon \in (\eta + 1/2, 3/4)$, $\gamma < \varepsilon - 1/2$, and $\Delta < 1/4 + \eta$. The assumptions (A2) and (A4) are satisfied by the conditions imposed on F, h_1, and h_2. The tightness condition (A5) can be verified using Theorem 5.3 (note that A^{-1} is compact in the present case). For example, if $|F|$, $|h_1|$, and $|h_2|$ are (uniformly) bounded then (A5) is satisfied. Thus the results of the paper (in particular, Proposition 4.10, Theorems 6.2 and 6.3) can be applied for any cost functional $c : H \times \mathcal{K} \to \mathbb{R}$ that satisfies (A6) and satisfies with h_1 and h_2 the convexity condition (A7). A simple example of a boundary input (7.3), (7.4) that satisfies all the above conditions is

$$\frac{\partial v}{\partial \xi}(t, 0) = u_1(v(t, \cdot)) + \dot{\beta}_1(t), \tag{7.8}$$

$$\frac{\partial v}{\partial \xi}(t, 1) = u_2(v(t, \cdot)) + \dot{\beta}_2(t), \tag{7.9}$$

where $(u_1, u_2) : H \to [-M, M]^2 = \mathcal{K}$.

EXAMPLE 7.2. Consider the stochastic parabolic partial differential equation with pointwise noise and control

$$\frac{\partial v}{\partial t}(t, \xi) = Lv(t, \xi) + F(\alpha, v(t, \xi)) + \sum_{i=1}^{N} [h_i(\alpha, v(t, \cdot), u(v(t, \cdot))) + \dot{\beta}_i(t)]\delta_{\xi_i} + n(t, \xi) \tag{7.10}$$

for $(t, \xi) \in \mathbb{R}_+ \times (0, 1)$ with initial and boundary conditions

$$v(0, \xi) = v_0(\xi), \tag{7.11}$$

$$v(t, 0) = 0, \tag{7.12}$$

$$v(t, 1) = 0 \tag{7.13}$$

for $(t, \xi) \in \mathbb{R}_+ \times (0, 1)$, where L, F, n, β_i, and h_i are the same as in Example 7.1, and δ_{ξ_i}, $i = 1, 2, \ldots, N$, are the Dirac distributions at the points $\xi_i \in (0, 1)$, $i = 1, 2, \ldots, N$. The equation (7.10) is given a precise interpretation by using (2.1) with H and f as in Example 7.1, $V(t) = (\beta_1(t), \ldots, \beta_N(t))$, $U = \mathbb{R}^N$, $h = (h_1, \ldots, h_N)$, and $A = -L$ with $\mathrm{Dom}(A) = H^2(0, 1) \cap H_0^1(0, 1)$. It is possible to use the Neumann boundary conditions in (7.12), (7.13) as well, so that $\mathrm{Dom}(A)$ would be the same as in Example 7.1. Since the domain is one-dimensional it follows by the Sobolev imbedding theorem that $\delta_{\xi_i} \in D_A^{\varepsilon-1}$ for $\varepsilon < 3/4$ (cf. Theorem 1.1 of [5]). It trivially follows that $B \in \mathcal{L}(\mathbb{R}^N, D_A^{\varepsilon-1})$ for $B\lambda = \sum_{i=1}^{N} \lambda_i \delta_{\xi_i}$, $\lambda = (\lambda_1, \ldots, \lambda_N)$. The verification of assumptions (A1)–(A7) in the present example is almost identical to the verifications in Example 7.1 because H and f are the same and U, h, and $V(t)$ are analogous (but the dimension is N instead of 2), $A^{-\delta}$ is Hilbert–Schmidt for $\delta > 1/4$, A^{-1} is compact, and it is again required that $\varepsilon < 3/4$. If the covariance Q of the distributed Wiener process can be expressed as $Q^{1/2} = A^{-\eta}\Gamma$ for $\Gamma, \Gamma^{-1} \in \mathcal{L}(H)$, $\eta \in [0, 1/4)$, then the assumptions (A1) and (A3) are satisfied. Given an $M > 0$ and the set of controls $\mathcal{U} = \{u : H \to [-M, M]^N |\ u \text{ is Borel mesurable}\}$ it is now possible to apply Proposition 4.10 and Theorems 6.2 and 6.3 with any cost functional $c : H \times [-M, M]^N \to \mathbb{R}$ that satisfies (A6) and, together with h, the convexity condition (A7).

REFERENCES

[1] V. E. Beneš, *Existence of optimal stochastic control laws*, SIAM J. Control, 9 (1971), pp. 446–472.

[2] T. Bielecki and L. Stettner, *On ergodic control problems for singularly perturbed Markov processes*, Appl. Math. Optim., 20 (1989), pp. 131–161.

[3] J. M. Bismut, *Théorie probabiliste du controle des diffusions*, Mem. Amer. Math. Soc., 167 (1976), pp. 1–130.

[4] V. S. Borkar and T. E. Govindan, *Optimal control for semilinear evolution equations*, Nonlinear Anal., 23 (1994), pp. 15–35.

[5] S. Chen and R. Triggiani, *Characterization of domains of fractional powers of certain operators arising in elastic systems*, J. Differential Equations, 88 (1990), pp. 279–293.

[6] A. Chojnowska-Michalik and B. Goldys, *Existence, uniqueness and invariant measures for stochastic semilinear equations on Hilbert spaces*, Probab. Theory Related Fields, 102 (1995), pp. 331–356.

[7] P. Chow and J. L. Menaldi, *Infinite-dimensional Hamilton-Jacobi-Bellman equations in Gauss-Sobolev spaces*, Nonlinear Anal., 29 (1997), pp. 415–426.

[8] G. Da Prato, M. Fuhrman, and J. Zabczyk, *Differentiability of Ornstein-Uhlenbeck Semigroups*, Preprint 1995/26, Scuola Normale Superiore, Pisa, Italy.

[9] G. Da Prato and J. Zabczyk, *Smoothing properties of transition semigroups in Hilbert spaces*, Stochastics Stochastics Rep., 35 (1991), pp. 63–77.

[10] G. Da Prato and J. Zabczyk, *Stochastic Equations in Infinite Dimensions*, Cambridge Univ. Press, Cambridge, UK, 1992.

[11] G. Da Prato and J. Zabczyk, *Evolution equations with white noise boundary conditions*, Stochastics Stochastics Rep., 42 (1993), pp. 167–182.

[12] T. E. Duncan, B. Maslowski, and B. Pasik-Duncan, *Adaptive boundary and point control of linear stochastic distributed parameter systems*, SIAM J. Control Optim., 32 (1994), pp. 648–672.

[13] T. E. Duncan, B. Pasik-Duncan, and L. Stettner, *On ergodic control of stochastic evolution equations*, Stochastic Anal. Appl., 15 (1997), pp. 723–750.

[14] T. E. Duncan and P. Varaiya, *On the solutions of a stochastic control system*, SIAM J. Control, 9 (1971), pp. 354–371.

[15] T. E. Duncan and P. Varaiya, *On the solutions of a stochastic control system* II, SIAM J. Control, 13 (1975), pp. 1077–1092.

[16] D. Fujiwara, *Concrete characterizations of the domains of fractional powers of some elliptic differential operators of the second order*, Proc. Japan Acad. Ser. A Math. Sci., 43 (1967), pp. 82–86.

[17] D. Gatarek and B. Goldys, *On solving stochastic evolutions by the change of drift with application to optimal control*, Proceedings SPDE's and Applications, G. Da Prato and L. Tubaro, eds., Pitman, Boston, 1992, pp. 180–190.

[18] D. Gatarek and B. Goldys, *On weak solutions of stochastic equations in Hilbert spaces*, Stochastics Stochastics Rep., 46 (1994), pp. 41–51.

[19] I. C. Gokhberg and M. G. Krein, *Introduction to the Theory of Linear Nonselfadjoint Operators*, Nauka, Moscow, 1965 (in Russian); AMS, Providence, RI, 1969 (in English).

[20] F. Gozzi and E. Rouy, *Regular solutions of second-order stationary hamilton-jacobi equations*, J. Differential Equations, to appear.

[21] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Springer-Verlag, New York, 1981.

[22] A. Ichikawa, *Stability of parabolic equations with boundary and pointwise noise*, in Stochastic Space-Time Models and Limit Theorems, D. Reidel, Dordrecht, the Netherlands, 1985, pp. 81–94.

[23] S. M. Kozlov, *Some questions of stochastic equations with partial derivatives*, Trudy Sem. Petrovsk., 4 (1978), pp. 147–172 (in Russian).

[24] K. Kuratowski, *Topology, Vol.* I, Academic Press, New York, London, 1966.

[25] H. I. Kushner, *Optimality conditions for the average cost per unit time problem with a diffusion model*, SIAM J. Control Optim., 16 (1978), pp. 330–346.

[26] I. L. Lions and E. Magenes, *Nonhomogeneous Boundary Value Problems and Applications* I, Springer-Verlag, Berlin, 1972.

[27] B. Maslowski, *Stability of semilinear equations with boundary and pointwise noise*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 22 (1995), pp. 55–93.

[28] B. Maslowski, *On probability distributions of solutions of semilinear stochastic evolution equations*, Stochastics Stochastics Rep., 45 (1993), pp. 11–44.

[29] E. Roxin, *The existence of optimal controls*, Michigan Math. J., 91 (1962), pp. 109–119.
[30] J. Seidler, *Da Prato-Zabczyk's maximal inequality revisited* I, Math. Bohem., 118 (1993), pp. 67–106.
[31] J. Seidler, *Ergodic behaviour of stochastic parabolic equations*, Czech. Math. J., 47 (1997), pp. 277–316.
[32] L. Stettner, *On the existence of optimal per unit time control for degenerate diffusion model*, Bull. Polish Acad. Sci. Math., 34 (1986), pp. 749–769.
[33] I. Vrkoč, *A dynamical system in a Hilbert space with a weakly attractive nonstationary point*, Math. Bohem., 118 (1993), pp. 401–423.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 3, pp. 1048–1072, May 1998

013

PROXIMAL ANALYSIS AND THE MINIMAL TIME FUNCTION*

PETER R. WOLENSKI† AND YU ZHUANG†

Abstract. Under general hypotheses on the target set S and the dynamics of the system, we show that the minimal time function $T_S(\cdot)$ is a proximal solution to the Hamilton–Jacobi equation. Uniqueness results are obtained with two different kinds of boundary conditions. A new propagation result is proven, and as an application, we give necessary and sufficient conditions for $T_S(\cdot)$ to be Lipschitz continuous near S. A Petrov-type modulus condition is also shown to be sufficient for continuity of $T_S(\cdot)$ near S.

Key words. minimal time function, proximal analysis, Hamilton–Jacobi equations, nonsmooth analysis, continuity of value functions

AMS subject classifications. 49L20, 35D05

PII. S0363012996299338

1. Introduction.

The minimal time control problem consists of a given closed set S (the "target set") and a control system in which the goal is to steer an initial point x to the target set along a trajectory of the system in minimal time. The minimal time value is denoted by $T_S(x)$, which could be $+\infty$ if no trajectory from x can reach S. We shall model the control system in this paper as a differential inclusion.

The function $x \mapsto T_S(x)$ is called the minimal time function. The first goal of this paper is to prove that $T_S(\cdot)$ is the unique proximal solution to the Hamilton–Jacobi (HJ) equation. This is Theorem 3.2 below, which holds under very mild hypotheses on F and S.

Solving HJ equations in some nonclassical sense has developed into an active research area with several different schools of thought participating. The viscosity solution method was pioneered by Crandall and Lions [20] and is closely linked with classical PDE theory. See [19] and [22] for historical references. Minimax solutions to HJ in the context of differential games were introduced in the Russian school by Subbotin [33], [34], and one of the important contributions of this approach is its extensive reliance on flow invariance. Clarke and Vinter [17] considered solutions using generalized gradients to construct verification functions, although these are not necessarily unique. More recently, proximal solutions to (HJ) appeared in Clarke and Ledyaev [11], where the various concepts were also unified. See the discussion in [12, Remark 9.4].

Characterizing $T_S(\cdot)$ as a solution to a HJ equation without controllability assumptions requires one to handle discontinuities that may appear, and perhaps infinite values as well. In problems formulated on a fixed time interval, Barron and Jensen [5], [6] initiated the study of lower semicontinuous vicosity solutions and found an appropriate boundary condition to maintain uniqueness results. See also Frankowska

*Received by the editors February 28, 1996; accepted for publication (in revised form) April 2, 1997.

http://www.siam.org/journals/sicon/36-3/29933.html

†Department of Mathematics, Louisiana State University, Baton Rouge, LA 70803 (wolenski@math.lsu.edu, zhuang@marais.math.lsu.edu). The research of the first author was supported in part by NSF grant DMS-9623406.

[24] and Clarke et al. [12]. However, the minimal time problem does not fit into the category of problems covered by these results.

HJ theory specifically tailored to handle $T_S(\cdot)$ has also received considerable attention. Viscosity approaches have been undertaken by Bardi [2], Evans and James [21], Staicu [32], and Bardi and Falcone [4]. Local controllability assumptions are made to guarantee that $T_S(\cdot)$ will be continuous in these papers. Bardi and Staicu [3] and Soravia [30] have also used viscosity-type methods without controllability, although the uniqueness result in [3] requires that S be the closure of its interior (and in particular, preclude the target to be a single point). Soravia [30] contains two uniqueness results, one of the type just mentioned and another that is equivalent to our Theorem 3.2 below. Various interesting data perturbations and envelope constructions are made in these papers, but we do not require any of those techniques here. Rather, the approach in this paper is based on flow invariance and mimics the broad outline sketched in [12]. See also Frankowska [24], in which invariance is prominent. Adaptations of the arguments in [24] to minimal time problems are made in Carja, Mignanego, and Pieri [8], where the target set is restricted to be the origin. One common feature in all these papers (except [30]) is that a boundary condition is required at the boundary of the reachable set (this was introduced in [2]), whereas our approach dispenses with any such condition. Although some of the arguments given in section 3 below are routine modifications of those in [12, section 9], we attempt, except in cases where the modifications are obvious, to give detailed proofs for both clarity and completeness.

The basic assumptions on the multifunction F that appears on the right-hand side of the differential inclusion are standard. Namely, F will be locally bounded, have convex values, and exhibit local Lipschitz behavior. We shall not make an a priori growth assumption on F, but an additional hypothesis is required to imply the lower semicontinuity of $T_S(\cdot)$. This additional hypothesis will depend on properties of both S and F, but is always satisfied if F should exhibit linear growth in its state variable.

In Theorem 3.2, the minimal time function is shown to be the unique solution to a proximal form of the HJ equation satisfying an analytic boundary condition in the form of a HJ inequality. This boundary condition is probably the most natural one under the circumstances and allows for the easiest proof of the theorem. In section 4, we give an alternative but equivalent formulation of Theorem 3.2 by introducing a geometric boundary condition. This second boundary condition is closer in spirit to the one introduced by Barron and Jensen [5], [6] (and used in [24], [12]), but is somewhat more difficult to state in the present circumstance. On the other hand, it has a meaningful interpretation directly in terms of the data. The main result in section 5 is Theorem 5.1, which is a characterization of the proximal subgradients of T_S. This result describes the propagation of the level sets of T_S and is a different approach to some work by Soravia [31] on front propagation. In particular, our result does not require assumptions that force the level sets to strictly enlarge at each point. We make one application of Theorem 5.1 in section 6 by giving necessary and sufficient conditions for T_S to be Lipschitz continuous near S, and sufficient conditions are also provided for continuity of an arbitrary modulus. Remarks in that section address the history of such results. Finally, some examples are given in section 7.

2. Preliminaries.

2.1. Background in nonsmooth analysis. We first review some concepts from nonsmooth analysis that are pertinent in this paper. For a complete treatment, see [10], [25], or [13].

For a closed subset of S of $\mathbb{R}^n$, the distance function to S is defined by

$$d_S(x) = \inf\{\|x - s\| : s \in S\}$$

for all $x \in \mathbb{R}^n$.

DEFINITION 2.1. *Suppose $S \subseteq \mathbb{R}^n$ is closed and $s \in S$. A vector $\zeta \in \mathbb{R}^n$ is a* proximal normal *to S at s provided there exists $r > 0$ so that $d_S(s + r\zeta) = r\|\zeta\|$. The set of all proximal normal vectors to S at s is denoted by $N_S^P(s)$.*

DEFINITION 2.2. *Suppose $\theta : \mathbb{R}^n \to (-\infty, \infty]$ is lower semicontinuous and $x \in \operatorname{dom}\theta := \{x' : \theta(x') < \infty\}$. A vector $\xi \in \mathbb{R}^n$ is a* proximal subgradient *of θ at x provided $(\xi, -1) \in N^P_{\operatorname{epi}\theta}(x, \theta(x))$, where $\operatorname{epi}\theta$ denotes the epigraph $\{(x, r) : x \in \operatorname{dom}\theta,\ r \geq \theta(x)\}$ of $\theta(\cdot)$, which is a closed subset of $\mathbb{R}^{n+1}$. The set (which could be empty) of all proximal subgradients of $\theta(\cdot)$ at x is denoted by $\partial_P\theta(x)$. If $x \notin \operatorname{dom}\theta$, then $\partial_P\theta(x) = \phi$ by definition.*

The following analytic descriptions of the proximal objects are often useful. See [25] or [13] for the proofs.

PROPOSITION 2.1.

(a) *Suppose S is closed and $s \in S$. Then $\zeta \in N_S^P(s)$ if and only if there exists $\sigma > 0$ and $\eta > 0$ such that $\langle \zeta, s' - s\rangle \leq \sigma\|s' - s\|^2$ for all $s' \in S \cap \{x + \eta B\}$.*

(b) *Suppose $\theta : \mathbb{R}^n \to (-\infty, \infty]$ is lower semicontinuous and $x \in \operatorname{dom}\theta$. Then $\xi \in \partial_P\theta(x)$ if and only if there exists $\sigma > 0$ and $\eta > 0$ such that*

$$\theta(y) - \langle \xi, y - x\rangle + \sigma\|y - x\|^2 \geq \theta(x)$$

for all $y \in x + \eta B$.

The only nonsmooth constructions in this paper are the proximal objects, but we mention that some of our results have natural formulations in terms of other normal cones and subgradients and can be proven by taking the appropiate limits. We shall not go into those details here, however.

2.2. Background in differential inclusions. Throughout this subsection, $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is a given multifunction (i.e., a set-valued map). Associated with F is the differential inclusion

$$\begin{aligned} &\dot{x}(t) \in F\big(x(t)\big) \quad \text{(almost everywhere) } t \in [0, T], \\ &x(0) = x. \end{aligned} \tag{2.1}$$

A solution to (2.1) is an absolutely continuous function $x(\cdot)$ defined on the interval $[0, T]$ with initial value $x(0) = x$, in which case we say that $x(\cdot)$ is a trajectory of F that originates from x. The notation $\dot{x}(t)$ refers to the derivative of $x(\cdot)$ at t and is the right derivative if $t = 0$. The set of endpoints of all such trajectories of F is denoted by $R_F^{(T)}(x)$ and is called the *reachable set* (from x and at time T). That is, $R_F^{(T)}(x) := \{x(T) : x(\cdot)\,\text{satisfies (2.1)}\}$. The notation $R_F^{(\leq T)}(x)$ signifies the set of all points reachable from x at a time less than or equal to T.

Basic hypotheses to be imposed on F in various combinations are the following.

(H1) For each $x \in \mathbb{R}^n$, $F(x) \neq \phi$ with the graph $\operatorname{gr} F := \{(x, v) : v \in F(x)\}$ closed in $\mathbb{R}^{2n}$, and for each compact set $K \subset \mathbb{R}^n$, there exists a constant $M > 0$ such that

$$\sup\big\{\|v\| : x \in K,\ v \in F(x)\big\} \leq M.$$

(H2) For each $x \in \mathbb{R}^n$, $F(x)$ is convex.

(H3) For each compact subset $K \subset \mathbb{R}^n$, there exists a constant $k > 0$ such that

$$F(x) \subseteq F(y) + k\|x - y\|B \quad \text{for all} \quad x, y \in K.$$

In (H3) and below, we use B to designate the *closed* unit ball. The open ball will be written as $\operatorname{int} B$ (the interior of B). The following two propositions contain some elementary properties of differential inclusions.

PROPOSITION 2.2. *Suppose $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is a multifunction and satisfies* (H1).

(a) *For each compact set $K \subset \mathbb{R}^n$ and $\varepsilon > 0$, there exists $\tau > 0$ such that*

$$R_F^{(\leq\tau)}(K) := \bigcup_{x \in K} R_F^{(\leq\tau)}(x) \subset K + \varepsilon B.$$

(b) *If $\{x(\cdot)\}$ is a trajectory of F on $[0, T)$ originating from x with $T < \infty$ and satisfies*

$$\liminf_{t \uparrow T} \|x(t)\| < \infty,$$

then the limit of $x(t)$ exists as $t \uparrow T$.

(c) *Suppose in addition that either* (H2) *or* (H3) *holds. Then there exists $T > 0$ such that* (2.1) *admits at least one solution.*

Proof. (a) See [38, Lemma 5.1]. (b) Let $x(\cdot)$ be a trajectory, and suppose K is compact such that $x(t) \in K$ infinitely often as $t \uparrow T$. Let M be as in (H1) associated to the compact set $K + B$, and let τ be chosen as in part (a) (where $\varepsilon = 1$). Now choose $t_0 \in [T - \tau, T)$ so that $x(t_0) \in K$, and by part (a) it follows that $x(t) \in K + B$ for all $t \in [t_0, T)$. By the choice of M, we have for all $t_0 \leq t < t' \leq T$ that

$$\|x(t') - x(t)\| \leq \int_t^{t'} \|\dot{x}(\tau)\|\, d\tau \leq M(t' - t). \tag{2.2}$$

Since T is assumed to be finite, it follows from (2.2) that the limit of $x(t)$ as $t \uparrow T$ exists. (c) This is well known (see [1]). □

The following proposition gives some further information regarding C^1 trajectories of a differential inclusion under (H1)–(H3).

PROPOSITION 2.3. *Suppose F satisfies* (H1)–(H3), *and let $K \subset \mathbb{R}^n$ be compact.*

(a) *There exists $T > 0$ such that associated to every $x \in K$ and $v \in F(x)$ is a C^1 trajectory $x(\cdot)$ defined on $[0, T]$ with $\dot{x}(0) = v$. Moreover, the modulus of continuity of $\dot{x}(\cdot)$ does not depend on the particular initial value $x \in K$.*

(b) *For each solution $x(\cdot)$ of* (2.1) *and $\varepsilon > 0$, there exists a C^1 solution $x_\varepsilon(\cdot)$ of* (2.1) *with $x_\varepsilon(t) \in x(t) + \varepsilon B$ for all $t \in [0, T]$. (In other words, C^1 trajectories are dense with respect to the sup norm in the set of all trajectories.)*

Proof. (a) For the construction of $x(\cdot)$, see [1, pp. 115–117]. See also [38, Lemma 5.3]. (b) See [23, Theorem 6], or [39, Theorem 3.1]. □

The following result is fundamental to differential inclusion theory and is referred to as "the compactness of trajectories" theorem. This nomenclature is slightly misleading because the result says more than just that a bounded set of solutions to (2.1) is relatively compact. Rather, a stronger conclusion holds in that approximate trajectories have subsequences that converge to a trajectory. See [9, Theorem 3.1.7] for the proof.

PROPOSITION 2.4. *Suppose F satisfies* (H1) *and* (H2) *and $\{y_i(\cdot)\}_i$ is a sequence of uniformly bounded absolutely continuous functions with $y_i(0) = x_i$, where $y_i(\cdot)$ is*

defined on an interval $[0,T_i]$ *such that* $T_i \to T > 0$ *and* $x_i \to x$ *as* $i \to \infty$. *Assume further that*

$$\dot{y}_i(t) \in F\big(y_i(t) + r_i(t)\big) + \varepsilon_i B \quad \text{a.e. } t \in I_i,$$

where $\varepsilon_i \downarrow 0$, $\{r_i(\cdot)\}$ *is a sequence of measurable functions converging uniformly to* 0 *as* $i \to \infty$, *and* I_i *is a measurable subset of* $[0,T_i]$ *such that the measure of* I_i *converges to* T. *Then there exists a trajectory* $x(\cdot)$ *of* F *on* $[0,T]$ *with* $x(0) = x$ *and such that a subsequence of* $y_i(\cdot)$ *converges to* $x(\cdot)$ *uniformly and the subsequence of the derivatives* $\dot{y}_i(\cdot)$ *converges to* $\dot{x}(\cdot)$ *weakly in* $L^1[0,T]$.

We shall require no hypothesis that limits the growth of F as $\|x\| \to \infty$, but the lack thereof necessitates the introduction of an "escape time." Even more important to our analysis is that the escape time will record the first instance when a trajectory leaves a given open set. In classical ODE theory (or where trajectories are uniquely defined in passing through any given state), escape times are generally defined depending only on the initial value. However, for differential inclusions in general, the definition we require relies upon the particular trajectory, and subsequently there may be many trajectories originating from a state x with different escape times. Nonetheless, our definition is consistent with the classical one in that the escape time gives a maximal (positive) interval of definition for the trajectory to stay in the open set. We use the notation U^c to denote $\mathbb{R}^n \backslash U$, the complement of a given set U.

DEFINITION 2.3. *Suppose* $U \subseteq \mathbb{R}^n$ *is open,* $x \in U$, *and* $x(\cdot)$ *is a trajectory of* F *with* $x(0) = x$ *and defined on the* half-open *interval* $[0,T)$, *where* $0 < T \leq \infty$. *Then* T *is an* escape time *from* U *(in which case we write* $T =: \mathrm{Esc}(x(\cdot);U)$*), provided at least one of the following conditions hold:*

(a) $T = \infty$ *and* $x(t) \in U$ *for all* $t \geq 0$,

(b) $x(t) \in U$ *for all* $t \in [0,T)$ *and* $\|x(t)\| \to \infty$ *as* $t \uparrow T$, *or*

(c) $T < \infty$, $x(t) \in U$ *for all* $t \in [0,T)$, *and* $d_{U^c}\big(x(t)\big) \to 0$ *as* $t \uparrow T$.

The next proposition says that under standard existence theory assumptions, any trajectory can be extended to a trajectory that has an escape time.

PROPOSITION 2.5. *Suppose* F *satisfies* (H1) *and either* (H2) *or* (H3), *and* $x(\cdot)$ *is a trajectory of* F *on* $[0,T)$ *with* $x(t) \in U$ *for all* $t \in [0,T)$. *If* T *is not an escape time from* U, *then* $x(\cdot)$ *can be extended to a trajectory* $\widetilde{x}(\cdot)$ *defined on a strictly larger interval* $[0,\widetilde{T})$, *and in which* $\widetilde{T} = \mathrm{Esc}(\widetilde{x}(\cdot);U)$.

Proof. Since T is not an escape time, we must have $T < \infty$, for otherwise (a) would hold. We also must have $\sup_{t\in[0,T)} \|x(t)\| < \infty$, for otherwise (b) would hold. Thus, by Proposition 2.2(b), there exists $y \in \mathbb{R}^n$ such that $x(t) \to y$ as $t \uparrow T$. Since (c) does not hold, we must have $y \in U$. Using standard existence theory for differential inclusions (which we quoted in Proposition 2.2(c)), the trajectory $x(\cdot)$ can be extended from y. Such an extension will remain in U on an interval $[0,T+\tau)$ for small $\tau > 0$ by Proposition 2.2(a). Now we take trajectory $\widetilde{x}(\cdot)$ and time $\widetilde{T}$ as a maximal element (in the sense of graph inclusion) over such extensions. It follows that $\widetilde{T} = \mathrm{Esc}(\widetilde{x}(\cdot);U)$, since otherwise the preceding considerations on $x(\cdot)$ and T could be applied to $\widetilde{x}(\cdot)$ and $\widetilde{T}$, which would violate the maximality. ∎

Suppose $U \subseteq \mathbb{R}^n$ is open and $x \in U$. The set of all trajectories of F originating from x that remain in U over a maximal interval is denoted by $\Upsilon_{(F,U)}(x)$. That is, $\Upsilon_{(F,U)}(x)$ consists of those trajectories $x(\cdot)$ of F defined on a half-open interval $[0,T)$ with $x(0) = x$ and for which $\mathrm{Esc}(x(\cdot);U)$ is defined with $T = \mathrm{Esc}(x(\cdot);U)$. By Proposition 2.5, the set $\Upsilon_{(F,U)}(x)$ is nonempty for each $x \in U$.

2.3. The minimal time function. Now suppose $S \subset \mathbb{R}^n$ is closed and $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is a multifunction. The minimal time function $T_S : \mathbb{R}^n \to [0, \infty]$ is defined as follows. If $x \notin S$, then

$$T_S(x) := \inf\big\{T : \text{there exists } x(\cdot) \text{ satisfying (2.1)} \\ \text{with } x(0) = x \text{ and } x(T) \in S\big\}. \tag{2.3}$$

If no trajectory of F originating from x can reach S in finite time, then the above infimum is taken over the empty set, and hence $T_S(x) = \infty$ in this case, which is the usual convention. If $x \in S$, then $T_S(x) = 0$ by definition, which is consistent with the above definition if we allow trajectories to be defined on the degenerate interval $[0, 0]$.

It turns out that (H1)–(H3) are not sufficient by themselves to give many of the desired properties of $T_S(\cdot)$. Lower semicontinuity, for example, is not assured (see Example 7.1), nor if the infimum of (2.3) is finite is it necessarily attained (see Example 7.2). There are also several instances below where we shall want to exclude certain kinds of trajectories from entering the discussion, and the following hypothesis serves all of these needs. Note that this assumption is not merely on F, but depends on both F and S.

(H4) For all $x \notin S$ and $x(\cdot) \in \Upsilon_{(F,\mathbb{R}^n)}(x)$, if

$$\mathrm{Esc}(x(\cdot); \mathbb{R}^n) < \infty$$

then

$$\mathrm{Esc}(\bar{x}(\cdot); S^c) < \mathrm{Esc}(x(\cdot); \mathbb{R}^n),$$

where $\bar{x}(\cdot)$ is a restriction of $x(\cdot)$.

Roughly speaking, if (H4) holds, then any trajectory of F escaping to infinity in finite time must pass through S.

Remark 2.1. There are natural hypotheses explicitly given on the data that guarantee that (H4) will hold. For example, if there exist constants c_1, c_2 such that

$$\sup_{v \in F(x)} \|v\| \le c_1 + c_2\|x\| \quad \text{for all} \quad x \in \mathbb{R}^n,$$

then $\mathrm{Esc}(x(\cdot); \mathbb{R}^n) = \infty$ for all trajectories $x(\cdot)$, and so (H4) holds trivially in this case. Other simple conditions pertaining only to the target set could be given; for example, if S^c is bounded or if $\{r_i\}$ is a sequence of numbers converging to $+\infty$ and S is given by

$$S = \bigcup_{i=1}^{\infty} \{x : \|x\| = r_i\},$$

then (H4) is satisfied.

The following two characterizations of $T_S(x)$ are immediate consequences of the definitions.

$$T_S(x) = \inf\big\{T \ge 0 : R^{(T)}(x) \cap S \ne \phi\big\},$$

and if (H4) holds and $x \notin S$,

$$T_S(x) = \inf\big\{\mathrm{Esc}(x(\cdot); S^c) : x(\cdot) \in \Upsilon_{F,S^c}(x)\big\}.$$

PROPOSITION 2.6. *Suppose* $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ *satisfies* (H1), (H2), *and* (H4). *If* $x \in S^c \cap \operatorname{dom} T_S$, *then there exists* $x(\cdot) \in \Upsilon_{(F,S^c)}(x)$ *with* $\operatorname{Esc}(x(\cdot), S^c) = T_S(x)$ *and* $x(T_S(x)) \in S$ *(that is, the infimum in* (2.3) *is attained). Furthermore,* $T_S(\cdot)$ *is lower semicontinuous on* $\mathbb{R}^n$.

Proof. Suppose $x \notin S$ and $T_S(x) < \infty$. Let $\{x_i(\cdot)\}$ be a minimizing sequence of (2.3), which means that $x_i(\cdot) \in \Upsilon_{(F,S^c)}(x)$ and $T_i := \operatorname{Esc}(x_i(\cdot); S^c) \to T_S(x)$ as $i \to \infty$ and $x_i(T_i) \in S$ for all i. Let

$$T := \inf\{t \in [0, T_S(x)] : \limsup_{i\to\infty} \|x_i(t)\| = \infty\}.$$

If the limsup is always finite, then we take $T = T_S(x)$ by convention. Note that $0 < T$ by Proposition 2.2(a), and for any $t < T$, the sequence $\{x_i(\cdot)\}$ is uniformly bounded on the interval $[0, t]$. Hence, using a diagonal process and the compactness of trajectories theorem, we may assume that $x_i(\cdot)$ converges uniformly to a trajectory $x(\cdot)$ on each compact interval of $[0, T)$. Since $T \le T_S(x)$, we have

$$\liminf_{t\uparrow T} \|x(t)\| < \infty,$$

since otherwise $\operatorname{Esc}(x(\cdot); \mathbb{R}^n) = T$ and (H4) would be violated. Hence, by Proposition 2.2(b), the limit $\lim_{t\uparrow T} x(t) =: x(T)$ exists. To prove that the infimum in (2.3) is attained, we show that $x(T) \in S$. We first claim that $T = T_S(x)$. Indeed, let $K := \{x(t) : t \in [0, T]\}$ and choose τ as in Proposition 2.2(a) associated to the compact set $K + B$ and $\varepsilon = 1$. If $T < T_S(x)$, then there exists t_0 and $0 < \tau_0 \le \tau$ with

$$0 < t_0 < T < t_0 + \tau_0 \le T_S(x).$$

Since $x_i(t_0) \to x(t_0)$ as $i \to \infty$, we have $x_i(t_0) \in K + B$ for all large i, and by Proposition 2.2(a), it follows that $x_i(t_0 + \tau_0) \in K + 2B$ for all large i. However, the definition of T as an infimum says that $\limsup_{i\to\infty} \|x_i(t_0+\tau_0)\| = \infty$, a contradiction. Hence $T = T_S(x)$ as claimed.

To see that $x(T) \in S$, let M be given as in (H1) associated to the compact set $K + 2B$. Now let $\eta > 0$ be small and choose t_1 such that

$$T - \min\{\tau, \eta/M\} < t_1 < T \quad \text{and} \tag{2.4}$$

$$\|x(T) - x(t_1)\| < \eta. \tag{2.5}$$

We now choose i large enough such that

$$\|x(t_1) - x_i(t_1)\| < \eta \quad \text{and} \tag{2.6}$$

$$(T_i - t_1) < \min\{\tau, \eta/M\}. \tag{2.7}$$

Note that (2.7) is possible in view of (2.4) and since $T_i \to T_S(x) = T$. Observe next that $x_i(t_1) \in K + B$, and thus $x_i(t) \in K + 2B$ for all $t \in [t_1, T_i]$ by Proposition 2.2(a), and consequently

$$\|\dot{x}_i(t)\| \le M \quad \text{a.e. } t \in [t_1, T_i] \tag{2.8}$$

by (H1). Since $x_i(T_i) \in S$, we have

$$\begin{aligned} d_S(x(T)) &\le \|x(T) - x_i(T_i)\| \\ &\le \|x(T) - x(t_1)\| + \|x(t_1) - x_i(t_1)\| + \|x_i(t_1) - x_i(T_i)\| \\ &\le 2\eta + \int_{t_1}^{T_i} \|\dot{x}_i(t)\|\, dt \\ &\le 2\eta + M(T_i - t_1) \\ &\le 3\eta, \end{aligned}$$

where we used (2.5) and (2.6) to deduce the third inequality, (2.8), the fourth, and (2.7), the last. This implies $x(T) \in S$ since η is arbitrary.

To prove lower semicontinuity, suppose $x_i \to x$, and we may assume without loss of generality that

$$0 < \lim_{i\to\infty} T_S(x_i) =: T < \infty.$$

For each $i = 1, 2\ldots$, let $x_i(\cdot) \in \Upsilon_{(F,S^c)}(x_i)$ satisfy $\mathrm{Esc}(x_i(\cdot), S^c) = T_S(x_i)$ and $x_i(T_S(x_i)) \in S$, which exists by part (a). We can proceed at this point in a manner completely analogous to the proof above and produce a trajectory $x(\cdot) \in \Upsilon_{(F,S^c)}(x)$ with $\mathrm{Esc}(x(\cdot); S^c) \le T$ and $x(T) \in S$. (The only difference between here and the above is in the initial values $x_i(0) = x_i$ of the trajectories $x_i(\cdot)$, but the estimates remain valid.) Since $T_S(x)$ is defined as an infimum, we have $T_S(x) \le \mathrm{Esc}(x(\cdot); S^c) \le T$, which proves that $T_S(\cdot)$ is lower semicontinuous. ∎

3. HJ theory.

3.1. Invariance. We shall apply invariance results to objects obtained through modifying the given data, thus these concepts are introduced in terms other than S and F. Moreover, in contrast to [12], we require our notions to be local.

DEFINITION 3.1. *Suppose $E \subseteq \mathbb{R}^n$ is nonempty, $U \subseteq \mathbb{R}^n$ is open, and $\Gamma : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is a multifunction.*

(a) *Then (Γ, E) is* weakly invariant *in U provided that for all $x \in E \cap U$, there exists a trajectory $x(\cdot) \in \Upsilon_{(\Gamma,U)}(x)$ that satisfies $x(t) \in E$ for all $t \in [0, \mathrm{Esc}(x(\cdot); U))$.*

(b) *(Γ, E) is called* strongly invariant *in U provided that for every $x \in E$, every trajectory $x(\cdot) \in \Upsilon_{(\Gamma,U)}(x)$ satisfies $x(t) \in E$ for all $t \in [0, \mathrm{Esc}(x(\cdot); U))$.*

The next proposition relates these concepts to the minimal time problem. Recall that the closed set $S \subset \mathbb{R}^n$ and the multifunction $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ are given. We write $-F \times \{1\}$ for the multifunction defined at $(x, r) \in \mathbb{R}^n \times \mathbb{R}$ as

$$\big(-F \times \{1\}\big)(x, r) := \{-v : v \in F(x)\} \times \{1\} \subset \mathbb{R}^{n+1}.$$

A similar notation is in effect for the multifunction $F \times \{-1\}$.

PROPOSITION 3.1. *Suppose F satisfies* (H1) *and* (H2), *and let $E := \mathrm{epi}\, T_S$.*

(a) *If* (H4) *holds, then $(F \times \{-1\}, E)$ is weakly invariant in $U := S^c \times \mathbb{R}$.*

(b) *$(-F \times \{1\}, E)$ is strongly invariant in $\mathbb{R}^{n+1}$.*

Proof. (a) Let $(x, r) \in E \cap U$, and hence $x \notin S$ and $T_S(x) \le r < \infty$. By Proposition 2.6, there exists $x(\cdot) \in \Upsilon_{(F,S^c)}(x)$ satisfying $\mathrm{Esc}(x(\cdot); S^c) = T_S(x)$. By the principle of optimality, we have

$$T_S\big(x(t)\big) = T_S(x) - t \le r - t \tag{3.1}$$

for all $t \in [0, T_S(x)]$. Define $z(t) := \big(x(t), r-t\big)$ for $t \in [0, T_S(x))$. Then $\mathrm{Esc}(z(\cdot); U) = T_S(x)$, and clearly $z(\cdot) \in \Upsilon_{(F\times\{-1\})}(x, r)$. Moreover, it follows immediately from (3.1) that $z(t) \in E$ for all $t \in [0, \mathrm{Esc}(z(\cdot); U))$, which yields (a). (b) Let $(x, r) \in E$ and suppose $z(\cdot) \in \Upsilon_{(-F\times\{1\},\mathbb{R}^{n+1})}(x, r)$. Then $z(\cdot)$ has the representation $z(t) = \big(y(t), r + t\big)$ for $t \in [0, T)$, where $y(\cdot) \in \Upsilon_{(-F,\mathbb{R}^n)}(x)$ and

$$T := \mathrm{Esc}\big(z(\cdot); \mathbb{R}^{n+1}\big) = \mathrm{Esc}\big(y(\cdot); \mathbb{R}^n\big).$$

Fix $t \in [0, T)$. We must show that $z(t) \in E$. For $t' \in [0, t]$, define $x(t') = y(t - t')$. It is clear that $x(\cdot)$ is a trajectory for F since $y(\cdot)$ is a trajectory for $-F$. Hence, by the principle of optimality, we have

$$T_S\big(x(t)\big) + t \geq T_S\big(x(0)\big).$$

Using this and the fact that $(x, r) \in E$, we conclude that

$$r + t \geq T_S(x) + t = T_S(y(0)) + t = T_S(x(t)) + t \geq T_S(x(0)) = T_S(y(t)),$$

which says that $z(t) = \big(y(t), r + t\big) \in E$. Hence (b) holds. □

The notions of invariance lead directly to comparison results between T_S and certain lower semicontinuous functions θ, as the next result shows.

PROPOSITION 3.2. *Suppose F satisfies* (H1) *and* (H2) *and $\theta : \mathbb{R}^n \to (-\infty, \infty]$ is lower semicontinuous and satisfies $\theta(s) = 0$ for all $s \in S$. Let $E := \operatorname{epi}\theta$ and $U := S^c \times \mathbb{R}$.*

(a) *Suppose* (H4) *is also satisfied. If $(F \times \{-1\}, E)$ is weakly invariant in U and $\theta(\cdot)$ is bounded below on $\mathbb{R}^n$, then $\theta(x) \geq T_S(x)$ for all $x \in \mathbb{R}^n$.*

(b) *If $(-F \times \{1\}, E)$ is strongly invariant in $\mathbb{R}^{n+1}$, then $\theta(x) \leq T_S(x)$ for all $x \in \mathbb{R}^n$.*

Proof. (a) Suppose $x \in \mathbb{R}^n$. The conclusion is trivial if $x \in S$ or if $\theta(x) = \infty$, so assume

$$x \in S^c \cap \operatorname{dom}\theta.$$

By weak invariance, there exists a $z(\cdot) \in \Upsilon_{(F\times\{-1\},U)}(x, \theta(x))$ that remains in E. Note that $z(\cdot)$ has the form $z(t) = \big(x(t), \theta(x) - t\big)$, where $x(\cdot) \in \Upsilon_{(F,S^c)}(x)$. By the nature of U, we have

$$T := \mathrm{Esc}(z(\cdot), U) = \mathrm{Esc}(x(\cdot), S^c). \tag{3.2}$$

Observe next that the statement "$z(\cdot)$ remains in E" is equivalent to

$$\theta(x(t)) \leq \theta(x) - t \quad \text{for all} \quad t \in [0, T). \tag{3.3}$$

Since we have assumed that θ is bounded below, it follows from (3.3) that $T < \infty$. Assumption (H4) in conjunction with (3.2) implies that $\inf_{t\in[0,T)} \|x(t)\| < \infty$, and so it follows from Proposition 2.2(b) that $x(t) \to y \in S$ as $t \uparrow T$. We simply set $x(T) := y$. The lower semicontinuity of θ implies that (3.3) holds for $t = T$ as well, and the boundary condition on θ says that $\theta(x(T)) = 0$. Hence we have $\theta(x) \geq T$. Finally, the definition of T_S as an infimum yields that $T \geq T_S(x)$, and we conclude that $\theta(x) \geq T \geq T_S(x)$, which is (a). (b) Suppose $x \in \mathbb{R}^n$. If $T_S(x) = \infty$ or $x \in S$, there is nothing to show, so assume $x \in S^c \cap \operatorname{dom} T_S$. Let $\eta > 0$. There exists $x(\cdot) \in \Upsilon_{(F,S^c)}(x)$ with $\mathrm{Esc}(x(\cdot); S^c) =: T < T_S(x) + \eta$ and $x(T) \in S$. Let

$$z(t) := \big(x(T - t), t\big),$$

which is a trajectory of $-F \times \{1\}$ originating from $\big(x(T), 0\big) \in E$. By the strong invariance assumption, the trajectory $z(\cdot)$ remains in E, and thus

$$t \geq \theta\big(x(T-t)\big) \quad \text{for all} \quad t \in [0,T]. \tag{3.4}$$

Setting $t = T$ in (3.4) says that

$$T_S(x) + \eta > T \geq \theta\big(x(0)\big) = \theta(x),$$

and letting $\eta \downarrow 0$ proves (b). □

3.2. HJ inequalities. Recall that the (minimized) Hamiltonian associated to a multifunction $\Gamma : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is the function $h_\Gamma : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^1$ given by

$$h_\Gamma(x,\zeta) = \min\big\{\langle v, \zeta\rangle : v \in \Gamma(x)\big\}.$$

It is clear that $\zeta \mapsto h_\Gamma(x,\zeta)$ is positively homogeneous of degree 1.

The following theorem is a local version of results contained in [12], and the proofs given here must differ only to take account of the local nature. We also attempt to fill in a few more details in the strong invariance proof. Note that in part (a), Γ must only be defined (that is, be nonempty) on $E \cap U$, as is pointed out in [12, Remark 2.1(c)]. We also point out that the global version of part (a) was originally proved by Veliov [36], [37], work that was overlooked in the writing of [12].

THEOREM 3.1. *Suppose Γ satisfies* (H1) *and* (H2), *$E \subseteq \mathbb{R}^n$ is closed, and $U \subseteq \mathbb{R}^n$ is open.*

(a) *Then (Γ, E) is weakly invariant in U if and only if $h_\Gamma(x,\zeta) \leq 0$ for all $x \in E \cap U$ and $\zeta \in N_E^P(x)$.*

(b) *In addition, suppose Γ satisfies* (H3). *Then (Γ, E) is strongly invariant in U if and only if $h_\Gamma(x,-\zeta) \geq 0$ for all $x \in E \cap U$ and $\zeta \in N_E^P(x)$.*

Proof. (a)(⇒) The proximal normal cones $N_E^P(x)$ and $N^P_{E\cap\{x+\varepsilon B\}}(x)$ coincide, and thus this direction follows as in [12, Theorem 2.2]. (⇐) Let $x \in E \cap U$. We first choose $\tau > 0$ as in Proposition 2.2(a) with $K = \{x\}$ and $\varepsilon > 0$ chosen so that $x + \varepsilon B \subset U$. Then the construction used in the proof of Theorem 2.1 of [12] is valid here and produces a trajectory $x(\cdot)$ on $[0,\tau]$ that remains in $E \cap U$. We now can choose a trajectory $x(\cdot)$ on a half-open interval $[0,T)$ that has maximal graph and remains in $E \cap U$. For such a maximally defined trajectory, we must have $T = \mathrm{Esc}(x(\cdot); U)$, and thus $x(\cdot)$ is as in Definition 3.1(a). (b)(⇒) Suppose (Γ, E) is strongly invariant in U, $x \in E$, $v \in \Gamma(x)$, and $\zeta \in N_E^P(x)$. By Proposition 2.3(a), there exists a C^1 trajectory $x(\cdot)$ of Γ on $[0,T]$ with $x(0) = x$ and $\dot{x}(0) = v$. For small $t > 0$, $x(t)$ remains in U also. By Proposition 2.1(a), there exists $\sigma > 0$ such that

$$\langle \zeta, x' - x\rangle \leq \sigma \|x' - x\|^2 \quad \text{for all} \quad x \in E. \tag{3.5}$$

The strong invariance assumption implies that $x(t) \in E$ for all t, and thus inserting these values into (3.5) gives

$$\langle \zeta, x(t) - x\rangle \leq \sigma \|x(t) - x\|^2 \quad \text{for all} \quad t \in [0,T]. \tag{3.6}$$

Upon dividing (3.6) by t and letting $t \downarrow 0$ yields

$$\langle \zeta, v\rangle \leq 0. \tag{3.7}$$

Taking the supremum of the left side in (3.7) over $v \in F(x)$ and multiplying through by -1 implies $h_\Gamma(x,-\zeta) \geq 0$. (⇐) Let $x \in E$ and $x(\cdot) \in \Upsilon_{(\Gamma,U)}(x)$. Fix $T < \mathrm{Esc}(x(\cdot); U)$,

and it suffices to show that $x(t) \in E$ for all $t \in [0,T]$. We claim that without loss of generality $x(\cdot)$ is C^1. Indeed, we have $x(t) + \varepsilon B \subset U$ for all $t \in [0,T]$ whenever $\varepsilon > 0$ is small, and there exists a C^1 trajectory $x_\varepsilon(\cdot)$ that is within $x(t) + \varepsilon B$ for all t by Proposition 2.3(b). If it is known that $x_\varepsilon(t) \in E$ for all t and ε, then since E is closed, it follows that $x(t) \in E$ by letting $\varepsilon \downarrow 0$.

Hence we may assume $x(\cdot)$ is C^1. Let $\varepsilon > 0$ such that

$$K := \bigcup_{t\in[0,T]} \{x(t) + \varepsilon B\} \subset U,$$

and let k be the Lipschitz constant for K as in (H3). We now proceed as in the proof of Theorem 3.1 in [12] and define $\widetilde{\Gamma} : [0,T] \times K :\rightrightarrows \mathbb{R}^{n+1}$ by

$$\widetilde{\Gamma}(t,y) := \{1\} \times \big\{v \in F(y) : \|v - \dot{x}(t)\| \le k\|y - x(t)\|\big\}.$$

It is easily seen that Γ satisfies (H1) and (H2) on $[0,T] \times K$. Set

$$\widetilde{E} := \mathbb{R} \times E \quad \text{and} \quad \widetilde{U} := \mathbb{R} \times \operatorname{int} K,$$

where $\operatorname{int} K$ is the interior of K. We claim that $(\widetilde{\Gamma}, \widetilde{E})$ is weak invariant in $\widetilde{U}$. To see this, note that

$$N^P_{\widetilde{E}}(t,y) = \{0\} \times N^P_E(y), \tag{3.8}$$

and observe that for each $\zeta \in N^P_E(y)$, we have by assumption and the definition of $\widetilde{\Gamma}$ that

$$0 \le h_\Gamma(y, -\zeta) \le -h_{\widetilde{\Gamma}}\big((t,y),(0,\zeta)\big). \tag{3.9}$$

Hence, by (3.8) and (3.9), we see that

$$h_{\widetilde{\Gamma}}\big(\tilde{y}, \tilde{\zeta}\big) \le 0$$

for all $\tilde{y} \in \widetilde{E} \cap \widetilde{U}$ and $\tilde{\zeta} \in N^P_{\widetilde{E}}(\tilde{y})$. By part (a), it then follows that $(\widetilde{\Gamma}, \widetilde{E})$ is weakly invariant in $\widetilde{U}$ as claimed. Hence there exists a trajectory $z(\cdot) \in \Upsilon_{(\widetilde{\Gamma},\widetilde{U})}(0,x)$ that remains in $\widetilde{E}$. It is obvious that $z(t)$ has the form $(t, y(t))$ for some trajectory $y(\cdot)$ of Γ, and it follows from Gronwall's inequality that $y(\cdot) = x(\cdot)$ (see [12, Proof of Theorem 2.2]). □

We next interpret these results in terms of state-augmented data and epigraphs of lower semicontinuous functions. The following proposition is the analogue of the monotone results of Theorem 7.4 (a) and (d) in [12], recast into terms relevant to minimal time problems.

PROPOSITION 3.3. *Suppose F satisfies* (H1) *and* (H2), $\theta : \mathbb{R}^n \to (-\infty, \infty]$ *is lower semicontinuous, and $E = \operatorname{epi}\theta$.*

(a) *$(F \times \{-1\}, E)$ is weakly invariant in $S^c \times \mathbb{R}$ if and only if*

$$1 + h_F(x,\xi) \le 0 \quad \textit{for all} \quad x \notin S \textit{ and } \xi \in \partial_P\theta(x).$$

(b) *Suppose F in addition satisfies* (H3). *Then $(-F \times \{1\}, E)$ is strongly invariant if and only if*

$$1 + h_F(x,\xi) \ge 0 \quad \textit{for all} \quad x \in \mathbb{R}^n \textit{ and } \xi \in \partial_P\theta(x).$$

Proof. (a) Let $(x,\xi)\in\mathbb{R}^{2n}$, $r\in\mathbb{R}$, and $\rho<0$ and note that

$$\begin{aligned} h_{(F\times\{-1\})}\big((x,r),(\xi,\rho)\big) &= \inf_{v\in F(x)}\big\{\langle v,\xi\rangle-\rho\big\} \\ &= -\rho\left(1+h_F\left(x,-\frac{\xi}{\rho}\right)\right). \end{aligned} \tag{3.10}$$

($\Rightarrow$) Suppose $x\notin S$ and $\xi\in\partial_P\theta(x)$. By Theorem 3.1(a), we have

$$h_{(F\times\{-1\})}\big((x',r),-\zeta\big)\le 0 \tag{3.11}$$

for all $(x',r)\in E$, $x'\notin S$, and $\zeta\in N_E^P(x,r)$. Using the values $(x',r)=(x,\theta(x))\in \operatorname{epi}\theta=E$ and $\zeta=(\xi,-1)$ (which is a valid choice of ζ by Definition 2.2), we see from (3.10) and (3.11) that

$$1+h_F(x,\xi)\le 0.$$

($\Leftarrow$) Let $(x,r)\in E\cap S^c\times\mathbb{R}$ and $\zeta=(\xi,\rho)\in N_E^P(x,r)$. By the nature of epigraphs, we have $\rho\le 0$. Let us assume first that $\rho<0$, from which it follows that $r=\theta(x)$. Since $N_E^P(x,\theta(x))$ is a cone, we have $(-\xi/\rho,-1)\in N_E^P(x,\theta(x))$, and consequently, $-\xi/\rho\in\partial_P\theta(x)$. By (3.10), we have

$$h_{(F\times\{-1\})}\big((x,\theta(x)),(\xi,\rho)\big)=-\rho\left(1+h_F\big(x,-\xi/\rho\big)\right)\ge 0, \tag{3.12}$$

where we deduced the inequality from $-\rho>0$ and our assumption that $(1+h_F)\ge 0$.

Now suppose $\rho=0$. It is easily checked that $(\xi,0)\in N_E^P(x,\theta(x))$ as well, and so by Rockafellar's horizontality theorem [27], there exist sequences $\{x_i\}$, $\{\xi_i\}$, and $\{\rho_i\}$ such that $x_i\to x$, $\theta(x_i)\to\theta(x)$, $\xi_i\to\xi$, $\rho_i<0$, and $\rho_i\uparrow 0$, and $-\xi_i/\rho_i\in\partial_P\theta(x_i)$. By (3.12) we have

$$-\rho_i\left(1+h_F\big(x_i,-\xi_i/\rho_i\big)\right)\ge 0$$

for all i, and letting $i\to\infty$ yields $h_F(x,\xi)\ge 0$, and hence

$$h_{(F\times\{-1\})}\big((x,r),(\xi,0)\big)=h_f(x,\xi)\ge 0. \tag{3.13}$$

In view of (3.12) and (3.13), it follows from Theorem 3.1(a) that $F\times\{-1\}, E$ is weakly invariant on $S^c\times\mathbb{R}$. (b) The analogue to (3.10) needed here is

$$h_{(-F\times\{1\})}\big((x,r),-(\xi,\rho)\big)=-\rho\big(1+h_F(x,-\xi/\rho)\big), \tag{3.14}$$

which holds for all $(x,\xi)\in\mathbb{R}^{2n}$, $r\in\mathbb{R}$, and $\rho<0$. The proof of the equivalence in (b) is virtually identical to the one of (a), where Theorem 3.1(b) is quoted instead of (a). ∎

3.3. The HJ equation. Again suppose $S\subseteq\mathbb{R}^n$ is closed. We now characterize $T_S(\cdot)$ as the solution to the HJ equation on S^c that satisfies certain boundary conditions.

THEOREM 3.2. *Suppose $F:\mathbb{R}^n\rightrightarrows\mathbb{R}^n$ satisfies* (H1)–(H3), *and $S\subset\mathbb{R}^n$ is closed such that* (H4) *holds. Then there exists a unique lower semicontinuous function $\theta:\mathbb{R}^n\to(-\infty,\infty]$ bounded below on $\mathbb{R}^n$ and satisfying the following.*

(HJ) *For each $x\notin S$ and $\xi\in\partial_P\theta(x)$, we have*

$$1+h_F(x,\xi)=0.$$

(ABC) *Each* $x \in S$ *satisfies* $\theta(x) = 0$ *and*

$$1 + h_F(x, \xi) \geq 0$$

whenever $\xi \in \partial_P \theta(x)$.
The unique such function is $\theta(\cdot) = T_S(\cdot)$.

Remark 3.1. (HJ) is the Hamilton–Jacobi equation as it applies to minimal time problems, but in which proximal subgradients have replaced the usual gradient. (ABC) is an analytic boundary condition. In the next section, we analyze (ABC) in more detail and show that it is equivalent (in the context of the theorem) to a geometric boundary condition (GBC).

Proof. It is obvious that $T_S(\cdot)$ is bounded below by zero and it is lower semicontinuous by Proposition 2.6. It equals zero on S by definition. Propositions 3.1(a) and 3.3(a) combine to imply that

$$1 + h_F(x, \xi) \leq 0 \quad \text{for all} \quad x \notin S \text{ and } \xi \in \partial_P T_S(x). \tag{3.15}$$

Similarly, Propositions 3.1(b) and 3.3(b) combine to imply that

$$1 + h_F(x, \xi) \geq 0 \quad \text{for all} \quad x \in \mathbb{R}^n \text{ and } \xi \in \partial_P T_S(x). \tag{3.16}$$

We conclude from (3.15) and (3.16) that both (HJ) and (ABC) hold for $\theta(\cdot) := T_S(\cdot)$.

To prove uniqueness, suppose $\theta(\cdot)$ is lower semicontinuous, bounded below, and satisfies (HJ) and (ABC). By Propositions 3.3(a) and 3.2(a), we conclude that

$$\theta(x) \geq T_S(x) \tag{3.17}$$

for all $x \in \mathbb{R}^n$. Similarly, by Propositions 3.3(b) and 3.2(b), we conclude that

$$\theta(x) \leq T_S(x) \tag{3.18}$$

for all $x \in \mathbb{R}^n$. Obviously, $\theta(\cdot) = T_S(\cdot)$ follows from (3.17) and (3.18). ∎

4. A geometric boundary condition. In this section we give an alternative to the boundary condition (ABC) of Theorem 3.2 that can be more easily interpreted in terms of the target set and the dynamics of the system. The two conditions are equivalent only in the context of Theorem 3.2. The new boundary condition we consider is the following.

(GBC) For $x \in S$, we have $\theta(x) = 0$, and if $v \in F(x)$ and $\zeta \in N_S^P(x)$ satisfy $\langle v, \zeta \rangle < 0$, then

$$\liminf_{\substack{x' \to x \\ \frac{x'-x}{\|x'-x\|} \to \frac{-v}{\|v\|}}} \theta(x') = 0.$$

Loosely speaking, (GBC) says that if a velocity vector v at a point $x \in S$ makes an obtuse angle with some proximal normal at x, then $\theta(x')$ must go to zero along a sequence approaching x tangentially in the direction $-v$.

THEOREM 4.1. *Suppose the boundary condition* (ABC) *in Theorem* 3.2 *is replaced by* (GBC). *Then the conclusions of Theorem* 3.2 *remain valid.*

Proof. (ABC)⇒(GBC). By the uniqueness assertion in Theorem 3.2, we must show that $T_S(\cdot)$ satisfies (GBC).

Suppose $x \in S$, $v \in F(x)$, and $\zeta \in N_S^P(x)$ satisfy $\langle \zeta, v\rangle < 0$. By Proposition 2.2(b), there exists a C^1 trajectory $y(\cdot)$ on $[0,T]$ of $-F$ originating from x with $\dot y(0) = -v$. Let $x(t) := y(T-t)$, $t \in [0,T]$, and note that $x(\cdot)$ is a trajectory for F with $x(T) = x \in S$. By the principle of optimality, we have

$$T_S(x(t)) \le T - t \quad \text{for all} \quad t \in [0,T],$$

and consequently $T_S(x(t)) \to 0$ as $t \uparrow T$. Moreover,

$$\frac{x(t)-x}{T-t} \to \dot x(T) = \dot y(0) = -v \quad \text{as} \quad t \uparrow T,$$

and thus $x' = x(t)$ as $t \uparrow T$ is an admissible set of values in the liminf in (GBC). It follows that $T_S(\cdot)$ satisfies (GBC).

(GBC)⇒(ABC). Now assume $\theta(\cdot)$ satisfies (GBC) and the conditions in Theorem 3.2, except possibly (ABC). We show, in fact, that (ABC) holds for $\theta(\cdot)$ as well.

Suppose $x \in S$ and $\zeta \in \partial_P \theta(x)$. Hence there exists $\sigma > 0$ and $\eta > 0$ such that

$$\theta(x') - \langle \zeta, x' - x\rangle + \sigma\|x' - x\|^2 \ge \theta(x) = 0 \tag{4.1}$$

for all $x' \in x + \eta B$. We require the following simple lemma.

LEMMA 4.1. *Suppose x, ζ, and $\theta(\cdot)$ are as above. Then $\zeta \in N_S^P(x)$.*

Proof. Since (4.1) holds in particular for all $x' \in S \cap \{x + \eta B\}$, and $\theta(x') = 0$ for such x', we have

$$\langle \zeta, x' - x\rangle \le \sigma\|x' - x\|^2 \tag{4.2}$$

for all $x' \in S \cap \{x + \eta B\}$. Thus by (4.2) and Proposition 2.1(a), we conclude that $\zeta \in N_S^P(x)$. □

Now let $v \in F(x)$ be arbitrary, and it suffices to show that

$$1 + \langle \zeta, v\rangle \ge 0. \tag{4.3}$$

If $\langle \zeta, v\rangle \ge 0$, then (4.3) is trivial, so assume that

$$-\delta := \langle \zeta, v\rangle < 0. \tag{4.4}$$

We have by Lemma 4.1 that $\zeta \in N_S^P(x)$, and by Definition 2.1, there exists $r > 0$ such that

$$U := x + r\|\zeta\| + r\|\zeta\|\operatorname{int} B = \{x' : \|x' - x - r\zeta\| < r\|\zeta\|\}$$

satisfies $\operatorname{cl} U \cap S = \{x\}$, where $\operatorname{cl} U$ signifies the closure of U. Let $K := \operatorname{cl} U + B$, and let M, k be as in (H1), (H3), respectively. We next invoke (GBC) and obtain a sequence $x_i \to x$ such that

$$\theta(x_i) \to 0 \quad \text{and} \quad \frac{x_i - x}{\|x_i - x\|} \to \frac{-v}{\|v\|}. \tag{4.5}$$

Let $v_i = \operatorname{proj}_{F(x_i)}(v)$ be the projection (that is, the closest element) of v in $F(x_i)$. Assumption (H3) says

$$\|v_i - v\| \le k\|x_i - x\| \quad \text{for all} \quad i. \tag{4.6}$$

By Proposition 2.3(a), there exists $T > 0$ such that for each i there is a trajectory $y_i(\cdot)$ of $-F$ with $y_i(0) = x_i$, $\dot{y}(0) = -v_i$, and

$$\|\dot{y}_i(t) + v_i\| \le m(t) \quad \text{for all} \quad i, \tag{4.7}$$

where $m(\cdot)$ is a modulus function ($m(t) \searrow 0$ as $t \downarrow 0$) that is independent of i. We want to show that $y_i(\cdot)$ stays in U if T is small and i is large enough, which is the content of the following lemma.

LEMMA 4.2. *By shrinking T if necessary (but not dependent on i), we have that $y_i(t) \in U$ for all $t \in [0,T]$ and all sufficiently large i.*

Proof. Let τ be as in Proposition 2.2(a) applied to the compact set $\operatorname{cl} U$ and $\varepsilon = 1$, and shrink T if necessary so that $T < \tau$ and satisfies $M^2T + 2r\|\zeta\|m(T) < r\delta$, where δ is as in (4.4). If i is large enough that $2k\|x_i - x\|\|\zeta\| < \delta$, then for all $t \in [0,T]$ we have

$$\begin{aligned}
\|y_i(t) - x - r\zeta\|^2 &= \left\| \int_0^t \dot{y}_i(t')\,dt' \right\|^2 - 2\left\langle r\zeta, \int_0^t \dot{y}_i(t')\,dt' \right\rangle + r^2\|\zeta\|^2 \\
&\le M^2t^2 - 2\left\langle r\zeta, \int_0^t \big(\dot{y}_i(t') + v_i\big)\,dt' \right\rangle \\
&\qquad + 2t\big\langle r\zeta, \big(v_i - v\big)\big\rangle + 2t\langle r\zeta, v\rangle + r^2\|\zeta\|^2 \\
&\le t\big[M^2t + 2r\|\zeta\|m(t) + 2rk\|x_i - x\|\|\zeta\| - 2r\delta\big] \\
&\qquad + r^2\|\zeta\|^2 \\
&< r^2\|\zeta\|^2.
\end{aligned}$$

The second inequality follows from (4.7), (4.6), and (4.4). Hence $y_i(t) \in U$ as claimed. ∎

Recall that $\theta(\cdot)$ satisfies (HJ) and, in particular,

$$1 + h_F(y,\zeta) \ge 0 \quad \text{for all} \quad y \notin S,\ \zeta \in \partial_P\theta(y). \tag{4.8}$$

(The inequality (4.8) is one-half of (HJ), and is actually all that is required in this part of the proof.) By Proposition 3.3(b), it follows that $(-F \times \{1\}, \operatorname{epi}\theta)$ is strongly invariant in U, and thus by Lemma 4.2 we conclude that

$$\theta(y_i(t)) \le \theta(y_i(0)) + t = \theta(x_i) + t \tag{4.9}$$

for all large i and $t \in [0,T]$. By the compactness of trajectories theorem, we may assume without loss of generality that $y_i(\cdot) \to y(\cdot)$ uniformly on $[0,T]$ for some trajectory $y(\cdot)$ of $-F$ originating from x. For fixed $t \in [0,T]$, we let $i \to \infty$ in (4.9) and recall that $\theta(\cdot)$ is lower semicontinuous and $\theta(x_i) \to 0$ (see (4.5)). It follows from (4.9) that

$$\theta(y(t)) \le t \quad \text{for all} \quad t \in [0,T]. \tag{4.10}$$

For t small enough that $y(t) \in x + \eta B$, we can substitute $x' = y(t)$ into (4.1), and using (4.10), conclude that

$$t - \langle \zeta, y(t) - x\rangle + \sigma\|y(t) - x\|^2 \ge 0. \tag{4.11}$$

Observe for $t \in (0, T]$ that

$$\begin{aligned} \left\| \frac{y(t) - x}{t} + v \right\| &= \lim_{i \to \infty} \left\| \frac{y_i(t) - x_i}{t} + v_i \right\| \\ &\leq \frac{1}{t} \int_0^t \|\dot{y}(t') + v_i\| \, dt' \\ &\leq m(t), \end{aligned} \tag{4.12}$$

where in the last inequality we used (4.7). Therefore, dividing (4.11) by t and letting $t \downarrow 0$, we finally see from (4.12) that (4.3) holds, which finishes the proof. □

5. Propagation. In this section we prove an apparently new result that in effect characterizes the proximal subgradients of $T_S(\cdot)$. The result is known in the special case where $F(x) = B$ for all x. In this case, time is parametrized by Euclidean distance, and $T_S = d_S$. Thus Theorem 5.1 generalizes Theorem 3.4 of [16] to allow for much more general F.

For $r \geq 0$, define

$$S(r) := \{x \in \mathbb{R}^n : T_S(x) \leq r\},$$

the r-level set of $T_S(\cdot)$.

THEOREM 5.1. *Suppose $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ satisfies* (H1)–(H3), *$S \subseteq \mathbb{R}^n$ is closed, and* (H4) *holds.*

(a) *For all $x \in S$, we have*

$$\partial_P T_S(x) = N_S^P(x) \cap \{\zeta \in \mathbb{R}^n : \min_{v \in F(x)} \langle v, \zeta \rangle \geq -1\}.$$

(b) *Whenever $r > 0$ and $T_S(x) = r$, then we have*

$$\partial_P T_S(x) = N_{S(r)}^P(x) \cap \{\zeta \in \mathbb{R}^n : \min_{v \in F(x)} \langle v, \zeta \rangle = -1\}.$$

Proof. (a) Suppose $x \in S$ and $\zeta \in \partial_P T_S(x)$. Since $T_S(\cdot)$ satisfies the conditions of Theorem 4.1, we have by Lemma 4.1 that $\zeta \in N_S^P(x)$. We also have that

$$\min_{v \in F(x)} \langle v, \zeta \rangle \geq -1 \tag{5.1}$$

holds, since this is precisely the boundary condition (ABC) satisfied by $T_S(\cdot)$, as stated in Theorem 3.2.

For the opposite inclusion, suppose now that ζ satisfies (5.1) and

$$\langle \zeta, x' - x \rangle \leq \sigma \|x' - x\|^2 \quad \text{for all} \quad x' \in S, \tag{5.2}$$

where $\sigma > 0$ and $\eta > 0$. Note that (5.2) is simply the statement that $\zeta \in N_S^P(x)$ (Proposition 2.1(a)). We will show that there exists $\sigma' > 0$ and $\eta' > 0$ so that

$$T_S(x') - \langle \zeta, x' - x \rangle + \sigma' \|x' - x\|^2 \geq 0 \tag{5.3}$$

for all $x' \in x + \eta' B$, which implies $\zeta \in \partial_P T_S(x)$ by Proposition 2.1(b).

For the purpose of obtaining a contradiction, suppose to the contrary, and that (5.3) fails. Then for each $i = 1, 2, \ldots$, there exists x_i such that $x_i \to x$ as $i \to \infty$, and

$$T_S(x_i) - \langle \zeta, x_i - x \rangle < -i \|x_i - x\|^2 \tag{5.4}$$

holds for all i. Observe that since (5.2) holds, we have $x_i \notin S$ for large i. Set $t_i := T_S(x_i)$ and first note that

$$t_i \le \|\zeta\|\|x_i - x\| \tag{5.5}$$

for all i, which is a consequence of the Cauchy–Schwarz inequality and (5.4).

Now let $z_i(\cdot)$ be a time optimal trajectory originating from x_i, which exists by Proposition 2.6. Thus $z_i(\cdot)$ is a trajectory of F defined on $[0, t_i]$ and satisfies $z_i(0) = x_i$ and $z_i(t_i) =: y_i \in S$. Since $x_i \to x$, we deduce from (5.5) that $t_i \to 0$ as $i \to \infty$, and subsequently that for all large i, the entire range of the trajectory $z_i(\cdot)$ remains in $x + \eta B$ by Proposition 2.2(a). We have

$$\begin{aligned} t_i - \langle \zeta, x_i - x\rangle &= t_i + \langle \zeta, y_i - x_i\rangle - \langle \zeta, y_i - x\rangle \\ &\ge t_i + \left\langle \zeta, \int_0^{t_i} \dot z_i(t')\,dt' \right\rangle - \sigma\|y_i - x\|^2, \end{aligned} \tag{5.6}$$

which holds for large i in lieu of (5.2). Next, let $v_i(\cdot)$ be the pointwise projection of $\dot z_i(\cdot)$ onto $F(x)$, which is a measurable function defined on $[0, t_i]$. That is, $v_i(t') := \operatorname{proj}_{F(x)}(\dot z_i(t')) \in F(x)$. Let $K := x + \eta B$ and choose M, k as in (H1), (H3), respectively. We seek appropriate bounds for the last two terms in (5.6).

First we estimate $\|y_i - x\|$. We have

$$\begin{aligned} \|y_i - x\| &\le \|y_i - x_i\| + \|x_i - x\| \\ &\le \int_0^{t_i} \|\dot z_i(t')\|\,dt' + \|x_i - x\| \\ &\le t_i M + \|x_i - x\| \\ &\le \big(1 + M\|\zeta\|\big)\|x_i - x\|, \end{aligned} \tag{5.7}$$

where we used (5.5) to deduce the last inequality.

Next we estimate the integral term in (5.6). Note that any $t' \in [0, t_i]$ satisfies

$$\begin{aligned} \|z_i(t') - x\| &\le \|z_i(t') - x_i\| + \|x_i - x\| \\ &\le \int_0^{t_i} \|\dot z(t')\|\,dt' + \|x_i - x\| \\ &\le M t_i + \|x_i - x\|. \end{aligned} \tag{5.8}$$

We have by (5.1) and the Lipschitz assumption on F that

$$\begin{aligned} \left\langle \zeta, \int_0^{t_i} \dot z_i(t')\,dt' \right\rangle &= \int_0^{t_i} \langle \zeta, v_i(t')\,dt'\rangle + \left\langle \zeta, \int_0^{t_i} \big(\dot z_i(t') - v_i(t')\big)\,dt' \right\rangle \\ &\ge -t_i - k\|\zeta\| \int_0^{t_i} \|z_i(t') - x\|\,dt' \\ &\ge -t_i - t_i k\|\zeta\|\big[M t_i + \|x_i - x\|\big] \\ &\ge -t_i - k\|\zeta\|^2[M\|\zeta\| + 1]\|x_i - x\|^2 \\ &=: -t_i - c\|x_i - x\|^2 \end{aligned} \tag{5.9}$$

for all i, where (5.8) was used to deduce the next to last inequality, and (5.5), the last inequality.

We now insert (5.7) and (5.9) into (5.6) and deduce that

$$t_i - \langle \zeta, x_i - x\rangle \geq -[c + \sigma(1 + M\|\zeta\|)]\|x_i - x\|^2.$$

This contradicts (5.4) whenever i is sufficiently large and finishes the proof. (b) The proof of (b) is very similar to that of (a). The only modifications needed are (1) the substitution of $\theta(x) = r$ rather than 0, and (2) the use of (HJ) rather than (ABC). The details are left to the reader. □

Remark 5.1. It may be noted that $S(r)$ is nothing more than $R_{-F}^{(\leq r)}(S)$, the set of points reachable from S in times less than or equal to r by trajectories of $-F$. Thus Theorem 5.1(b) gives some information regarding the existence of proximal normals to reachable sets. We plan to further exploit this in future work, but in the present paper, we give only one application of this result in the next section.

6. Regularity results. In this section, we give necessary and sufficient conditions for T_S to be Lipschitz continuous near S. A sufficient condition for T_S to be continuous with proscribed modulus of continuity is also derived. Throughout this entire section, we assume that F satisfies (H1) and (H2), $S \subset \mathbb{R}^n$ is compact, and (H4) holds. We often impose (H3) as well, but this will be stated explicitly.

Recall that a modulus function $m(\cdot) : [0, \infty) \to [0, \infty)$ is a nondecreasing continuous function satisfying $m(0) = 0$. For our purposes, the distinguishing property of a modulus function is its behavior near 0. To make this formal, we say that two modulus functions $m_1(\cdot)$ and $m_2(\cdot)$ are equivalent if there exists a constant $c > 0$ such that

$$0 < \liminf_{r\downarrow 0} \frac{m_1(cr)}{m_2(r)} \leq \limsup_{r\downarrow 0} \frac{m_1(cr)}{m_2(r)} < \infty.$$

This defines an equivalence relation among modulus functions, and an equivalence class is called a *modulus class*. Note that if $m_1(\cdot)$ is a modulus function and $m_2(r) := c_1 m(c_2 r)$ for some positive constants c_1, c_2, then m_1 and m_2 belong to the same modulus class.

Let $\mathcal{M}$ be a modulus class. We say that a function f is $\mathcal{M}$-continuous on a set $U \subseteq \mathbb{R}^n$ if there exists $m(\cdot) \in \mathcal{M}$ such that

$$|f(x) - f(y)| \leq m(\|x - y\|) \quad \text{for all} \quad x, y \in U.$$

The first proposition shows that the local $\mathcal{M}$-continuity of $T_S(\cdot)$ near S need only be checked with one of the points belonging to S. This result appears in Soravia [28, Lemma 4.1] for modulus functions of the form $m(r) = c\, r^\alpha$, $0 < \alpha \leq 1$ and $c > 0$.

PROPOSITION 6.1. *Suppose $\mathcal{M}$ is a modulus class and* (H3) *holds. The following are equivalent.*

(a) *There exists $\eta > 0$ such that $T_S(\cdot)$ is $\mathcal{M}$-continuous on $S + \eta B$.*

(b) *There exists $\eta > 0$ and $m(\cdot) \in \mathcal{M}$ such that*

$$T_S(x) \leq m\big(d_S(x)\big)$$

for all $x \in S + \eta B$.

Proof. (a)⇒(b). This is trivial. (b)⇒(a) Let $\eta > 0$ and $m(\cdot)$ be as in (b). By Proposition 2.2(a), there exists $\eta' > 0$ such that

$$R_F^{(\leq m(\eta'))}(S + \eta' B) \subset S + \eta B. \tag{6.1}$$

We show that $T_S(\cdot)$ is continuous on $S+\eta' B$ with modulus of continuity $m_1(\cdot) := m\big(e^{km(\eta')}(\cdot)\big)$, where k is the Lipschitz constant of F on $S+\eta B$. Since $m_1 \in \mathcal{M}$, this is sufficient to prove (a).

Let $x, y \in S+\eta' B$. Since $T_S(x) \le m(\eta') < \infty$, we have by Proposition 2.6 that there exists $x(\cdot) \in \Upsilon_{(F,S^c)}(x)$ with $T := T_S(x)$ and $x(T) \in S$. By a well-known fact regarding the dependence of the reachable set on initial values (cf. [1, p. 120] or [39, Corollary 7.2]), there exists a trajectory $y(\cdot)$ for F originating from y that satisfies

$$\|x(T)-y(T)\| \le e^{kT}\|x-y\|. \tag{6.2}$$

Note that $y(T) \in S+\eta B$ by (6.1). The principle of optimality gives

$$T_S(y) \le T + T_S(y(T)). \tag{6.3}$$

Thus, from (6.3), we have

$$\begin{aligned} T_S(y) - T_S(x) &\le T_S(y(T)) \\ &\le m\big(d_S(y(T))\big) \\ &\le m\big(\|x(T)-y(T)\|\big) \\ &\le m_1\big(\|x-y\|\big). \end{aligned} \tag{6.4}$$

We used (b) in deducing the second inequality, the monotonicity of $m(\cdot)$ and $x(T) \in S$ in the third, and the monotonicity again and (6.2) in the last. Switching the roles of x and y in (6.4) shows that $T_S(\cdot)$ is continuous in $S+\eta' B$ of modulus m_1. □

The following theorem gives necessary and sufficient conditions for Lipschitz continuity of T_S near S.

THEOREM 6.1. *Suppose* (H3) *holds. Then the following are equivalent.*

(a) *There exists $\eta > 0$ such that $T_S(\cdot)$ is Lipschitz continuous on $S+\eta B$.*

(b)
$$\sup\big\{\|\zeta\| :\ \zeta \in \partial_P T_S(s),\ s \in S\big\} < \infty.$$

(c) *There exist $\eta > 0$ and $\delta > 0$ such that $x \in S^c \cap \{S+\eta B\}$ and $\zeta \in x - \mathrm{proj}_S(x)$ imply*
$$h_F(x,\zeta) \le -\delta\|\zeta\|.$$

Proof. (a)⇒(b). Let $s \in S$ and $\zeta \in \partial_P T_S(s)$. From Proposition 2.1(b), there exists $\sigma > 0$ such that for all x near s we have

$$\langle \zeta, x-s\rangle \le T_S(x) + \sigma\|x-s\|^2. \tag{6.5}$$

Since there exists $\lambda > 0$ for which $T_S(x) \le \lambda\|x-s\|$ by (a), (b) follows immediately from (6.5).

(b)⇒(c). If (c) fails, then for each $i = 1, 2\ldots$, there exists $x_i \notin S$ with $d_S(x_i) \to 0$ and $\zeta_i := x_i - s_i \in x_i - \mathrm{proj}_S(x_i)$ such that

$$h_F(x_i,\zeta_i) \ge \frac{-1}{i}\|\zeta_i\| \tag{6.6}$$

for all i. Let k be chosen for the compact set $S+B$ as in (H3). Then $x \mapsto h_F(x,\zeta)$ is Lipschitz of rank $k\|\zeta\|$ on $S+B$. Therefore, by (6.6), we have for large i

$$\begin{aligned} h_F(s_i,\zeta_i) &\ge h_F(x_i,\zeta_i) - k\|x_i-s_i\|\|\zeta_i\| \\ &= -\left[\frac{1}{i} + k d_S(x_i)\right] d_S(x_i), \end{aligned}$$

since $d_S(x_i) = \|x_i - s_i\| = \|\zeta_i\|$. Let $\rho_i := \big[1/i + kd_S(x_i)\big]$. Rewriting the last display and using the positive homogeneity of $h_F(s,\cdot)$ gives

$$h_F\left(s_i, \frac{\zeta_i}{\rho_i d_S(x_i)}\right) \geq -1. \tag{6.7}$$

Since $N_S^P(s_i)$ is a cone, we also have

$$\frac{\zeta_i}{\rho_i d_S(x_i)} \in N_S^P(s_i). \tag{6.8}$$

Thus (6.7), (6.8), and Theorem 5.1(a) imply that

$$\frac{\zeta_i}{\rho_i d_S(x_i)} \in \partial_P T_S(s_i).$$

Let $\lambda > 0$ be the supremum value in (b), and by assumption (b) we have

$$\lambda \geq \frac{\|\zeta_i\|}{\rho_i d_S(x_i)} = \frac{1}{\rho_i}.$$

But $\rho_i \to 0$ as $i \to \infty$, which is a contradiction, and hence we conclude that (c) holds.

(c)⇒(a). Under the hypothesis (c), it is shown in [18, Corollary 3.1] that $T_S(x) \leq c\, d_S(x)$ for some $c > 0$ and for all x sufficiently close to S (this result also appears in Veliov [37]). Hence (a) follows from this and Proposition 6.1. □

Remark 6.1. From the above proof of (b)⇒(c), one can also obtain a relation for the constant δ appearing in (c). Namely, if $0 < \eta \leq 1$ is sufficiently small, then since ρ_i cannot be larger than λ, we can take $\delta = 1/\lambda + k$. Note, however, that the proof does not seem to provide an a priori estimate for η, but it implies that such a value must exist.

Remark 6.2. There is an extensive literature behind the equivalence of (a) and (c) in Theorem 6.1. The implication (c)⇒(a) was first proved by Petrov [26] with $S = \{0\}$ and was extended to arbitrarily closed sets S by Soravia [29]. The converse (a)⇒(c) was shown by Bardi and Falcone [4] in the case when the boundary of S was piecewise C^2. Cannarsa and Sinestrari [7] allowed for "proximally smooth" S (see [16]) in proving this implication, but also required state differentiability in the dynamics. More recently, Yue [40] proved the equivalence in considerable generality, although the dynamics were given an explicit control formulation, as was the dynamics in all of the above-mentioned papers. Veliov [37] goes further yet by allowing the multifunction F to be nonautonomous and to depend measurably on t. The equivalence of condition (b) in our theorem seems to be new, however.

Remark 6.3. The condition that the proximal subgradient of a lower semicontinuous function f is locally bounded on an open set U is equivalent to f locally Lipschitz on U. See [15]. Note that in Theorem 6.1 above, we only used a part of one direction of this equivalence, and only the "easy" direction at that. We emphasize that in Theorem 6.1(b), the boundedness of the proximals is posited only for points in S.

We now give a sufficient condition for $T_S(\cdot)$ to satisfy condition (b) in Proposition 6.1 for some C^2 modulus function $m(\cdot)$, by which we mean that $m(\cdot)$ is a modulus function and is twice continuously differentiable on $(0,\infty)$. The estimate in fact does not require (H3). However, if (H3) holds as well, then it follows (Corollary 6.1) from Proposition 6.1 that $T_S(\cdot)$ is $\mathcal{M}$-continuous near S, where $m(\cdot) \in \mathcal{M}$.

THEOREM 6.2. *Assume that $m(\cdot)$ is a C^2 modulus function with $m'(r) > 0$ for $r > 0$, and that*

$$h_F(x,\zeta) \le \frac{-1}{m'\big(d_S(x)\big)}\|\zeta\| \tag{6.9}$$

for all x near S and $\zeta \in x - \operatorname{proj}_S(x)$. Then $T_S(\cdot)$ satisfies

$$T_S(x) \le m\big(d_S(x)\big)$$

for all x near S.

We require the following technical lemma, which is a version of the chain rule.

LEMMA 6.1. *Suppose $m(\cdot)$ is as in the theorem, and $f : U \to (0,\infty)$ is lower semicontinuous on the open set U. Then $m \circ f$ is lower semicontinuous on U, and $\zeta \in \partial_P f(x)$ if and only if $m'(f(x))\zeta \in \partial_P(m \circ f)(x)$.*

Proof. Let $x \in U$ and $\zeta \in \partial_P f(x)$. There exists $\sigma > 0$ such that

$$f(y) \ge f(x) + \langle \zeta, y - x\rangle - \sigma\|y - x\|^2 \tag{6.10}$$

for all y near x. Since $m(\cdot)$ is strictly increasing, taking m on both sides of (6.10) preserves the inequality (both sides of (6.10) are positive if y is near enough to x), and we have

$$m\big(f(y)\big) \ge m\big(f(x) + \langle \zeta, y - x\rangle - \sigma\|y - x\|^2\big). \tag{6.11}$$

Also, m is C^2, and so there exists a $\sigma' > 0$ such that

$$m(z) \ge m\big(f(x)\big) + m'(f(x))\big(z - f(x)\big) - \sigma'\|z - f(x)\|^2 \tag{6.12}$$

for all z near $f(x)$. Setting $z = f(x) + \langle \zeta, y - x\rangle - \sigma\|y - x\|^2$, we obtain from (6.11) and (6.12) that

$$m\big(f(y)\big) \ge m\big(f(x)\big) + \big\langle m'\big(f(x)\big)\zeta, y - x\big\rangle - \sigma''\|y - x\|^2, \tag{6.13}$$

where $\sigma'' := m'\big(f(x)\big)\sigma + 2\sigma'\|\zeta\|^2$, which holds if y is sufficiently near x. This shows that $m'\big(f(x)\big)\zeta \in \partial_P\big(m \circ f\big)(x)$.

The converse follows since the inverse of $m(\cdot)$ has the same properties as $m(\cdot)$. ∎

We record another fact in the next proposition, which contains results taken from [16] and [14]. This gives the relationship between the vectors $\zeta \in x - \operatorname{proj}_S(x)$ and proximal subgradients of the distance function.

PROPOSITION 6.2. *Suppose $x \notin S$ and $\partial_P d_S(x) \ne \phi$. Then both $\operatorname{proj}_S(x)$ and $\partial_P d_S(x)$ are singletons and equal $\{s_x\}$, $\{(x - s_x)/\|x - s_x\|\}$, respectively. Moreover, the Petrov modulus condition* (6.9) *is equivalent to*

$$h_F(x,\zeta) \le \frac{-1}{m'\big(d_S(x)\big)} \quad \text{for all } \zeta \in \partial_P d_S(x). \tag{6.14}$$

Proof. See [14, Theorem 4.1] for the statements regarding the proximal subgradient and the projection. That (6.9) implies (6.14) is then obvious. The reverse implication follows by taking limits, since if $\zeta = x - s_x \in x - \operatorname{proj}_S(x)$, then $\zeta/\|\zeta\|$ is the single element belonging to $\partial_P d_S(s_x + \varepsilon\zeta)$, $0 < \varepsilon < 1$ (see [16]). ∎

Proof of Theorem 6.2. Let $U := \{S + \eta \operatorname{int} B\} \cap S^c$, where $\eta > 0$ is sufficiently small that (6.9) holds for $x \in S + \eta B$. We shall see that (6.9) and each of the following statements are equivalent:

$$h_F\big(x,\zeta\big) \le \frac{-1}{m'\big(d_S(x)\big)} \quad \text{for all } x \in U, \text{ for all } \zeta \in \partial_P d_S(x), \tag{6.15}$$

$$1 + h_F\big(x, m'\big(d_S(x)\big)\zeta\big) \le 0 \quad \text{for all } x \in U, \text{ for all } \zeta \in \partial_P d_S(x), \tag{6.16}$$

$$1 + h_F\big(x,\zeta\big) \le 0 \quad \text{for all } x \in U, \text{ for all } \zeta \in \partial_P\big(m \circ d_S\big)(x), \tag{6.17}$$

$$\big(F \times \{-1\}, \operatorname{epi}(m \circ d_S)\big) \quad \text{is weakly invariant in } U \times \mathbb{R}. \tag{6.18}$$

The equivalence of (6.9) and (6.15) is contained in Proposition 6.2; that of (6.15) and (6.16) is due only to a rearrangement of terms and the positive homogeneity of $h_F(x,\cdot)$; the equivalence of (6.16) and (6.17) follows from Lemma 6.1, and that of (6.17) and (6.18) is a consequence of Proposition 3.3(a).

By Proposition 2.2(a), there exists $0 < \eta' \le \eta$ such that

$$R_F^{(\le m(\eta'))}(S + \eta' B) \subseteq S + \frac{\eta}{2} B. \tag{6.19}$$

Let $x \in \{S + \eta' B\} \cap S^c$. Our assumption (6.9) has been shown to be equivalent to (6.18), and thus there exists a trajectory $\tilde{x}(\cdot)$ of $F \times \{-1\}$ originating from $\big(x, m\big(d_S(x)\big)\big)$ that remains in epi $\big(m \circ d_S\big)$. We can write $\tilde{x}(t) = \big(x(t), m(d_S(x)) - t\big)$ for $0 \le t < \operatorname{Esc}(\tilde{x}(\cdot); U \times \mathbb{R}) = \operatorname{Esc}(x(\cdot); U) =: T$, where $x(\cdot)$ is a trajectory for F. Since $\tilde{x}(\cdot)$ remains in epi $\big(m \circ d_S\big)$, we have

$$m\big(d_S(x)\big) - t \ge m\big(d_S(x(t))\big) \quad \text{for all } t \in [0, T). \tag{6.20}$$

Since $m \ge 0$, we must have $T \le m\big(d_S(x)\big) \le m(\eta')$, and so it follows from (6.19) that $\lim_{t \uparrow T} x(t) = x(T) \in S$ (that is, $x(\cdot)$ escapes by hitting S first, not by going to the boundary of $S + \eta B$). Letting $t \uparrow T$ in (6.20) yields

$$T_S(x) \le T \le m\big(d_S(x)\big),$$

which finishes the proof. □

The following is an immediate corollary.

COROLLARY 6.1. *Suppose* (H3) *holds in addition to the hypotheses of Theorem* 6.2. *Then* $T_S(\cdot)$ *is* $\mathcal{M}$*-continuous near* S.

Proof. This follows immediately from Theorem 6.2 and Proposition 6.1. □

Remark 6.4. The condition (6.15) reduces to the Petrov–Lipschitz condition in Theorem 6.1(c) if $\mathcal{M}$ contains the modulus function $m(r) = r$, and hence Theorem 6.2 generalizes the implication (c)⇒(a) in Theorem 6.1. The proof also provides an alternative to relying on results from [18], as was done in the proof of Theorem 6.1. We also mention that from (6.20), one can deduce the "rate of weak attainability" estimate derived by other means in [18].

Remark 6.5. If m in the previous theorem is taken as $m(r) = c\, r^\alpha$, where $c > 0$ and $0 < \alpha \le 1$, then Theorem 6.2 is a sufficient condition for α-Hölder continuity of T_S. Soravia [28] gave a sufficient condition for $\alpha = \frac{1}{2}$ under some special hypotheses, and Yue [40] considered any $0 < \alpha \le 1$. Thus Theorem 6.2 extends a result of Yue [40] to arbitrary (albeit C^2) moduli. Although the dynamics in [40] use the control formulation and the hypotheses there are stated using directional derivatives

rather than proximal subgradients, Theorem 6.2 and [40, Theorem 2.1] appear to be equivalent in the case of Hölder moduli.

Remark 6.6. As pointed out in Remark 5.1, the level set $S(r)$, $r > 0$, is the reachable set $R_{-F}^{(\leq r)}(S)$. If it is known that $T_S(\cdot)$ is continuous near S, then it follows immediately that S is contained in the interior of the reachable set associated with $-F$ up to time r. Thus Corollary 6.1 gives a sufficient condition for *small time local controllability*. The converse is also true: that is, small time local controllability of the system $-F$ implies the continuity of $T_S(\cdot)$ near S. There is a considerable literature devoted to local controllability (see Sussmann [35]), and there are systems which are controllable but violate any Petrov modulus condition. Thus a converse to Corollary 6.1 will not hold in general, although it does hold if the continuity is Lipschitz (Theorem 6.1). This can be explained by the fact that Lipschitz continuity is characterized by properties of its proximal subgradient, whereas continuity of a less restrictive modulus is not. We note, however, that a Petrov modulus condition as a sufficient condition for local controllability uses no additional structure of the control system beyond knowledge of certain admissible velocities at points near the target.

7. Examples.

Example 7.1. This is an example where F satisfies (H1)–(H3), but yet $T_S(\cdot)$ fails to be lower semicontinuous. Define $f : \mathbb{R}^2 \to \mathbb{R}^2$ by

$$f(x, y) = \{1\} \times \{1 + y^2\}.$$

Let $F(x, y) := \{f(x, y)\}$ and $S := \left\{\frac{\pi}{2}\right\} \times \mathbb{R}$. One can easily check that $T_S(0, 0) = \infty$, but $T_S(\varepsilon, 0) \to \frac{\pi}{2}$ as $\varepsilon \downarrow 0$.

Example 7.2. Here is an example showing that the existence of optimal trajectories cannot be assured unless additional hypotheses are added to (H1)–(H3). Again, we use state space $\mathbb{R}^2$, and F is obtained by modifying the previous example.

$$F(x, y) = \{1\} \times [0, 1 + y^2] \quad \text{and} \quad S := \left\{(x, y) : x > \frac{\pi}{2},\, y = \frac{1}{x - \frac{\pi}{2}}\right\}.$$

Then

$$R^{(T)}(0, 0) = \begin{cases} \{T\} \times [0, \tan T] & \text{if} \quad 0 \leq T < \frac{\pi}{2}, \\ \{T\} \times [0, \infty) & \text{if} \quad \frac{\pi}{2} \leq T. \end{cases}$$

Then one has $T_S(0, 0) = \frac{\pi}{2}$, but no trajectory reaches S from $(0, 0)$ in this time.

Acknowledgments. The authors wish to express their appreciation for the referees' comments, and their thanks to M. Bardi and P. Soravia for bringing to our attention some of the pertinent literature.

REFERENCES

[1] J. P. Aubin and A. Cellina, *Differential Inclusions*, Springer-Verlag, New York, 1984.
[2] M. Bardi, *A boundary value problem for the minimum-time function*, SIAM J. Control Optim., 27 (1989), pp. 776–785.
[3] M. Bardi and V. Staicu, *The Bellman equation for time-optimal control of noncontrollable nonlinear systems*, Acta Appl. Math., 31 (1993), pp. 201–223.
[4] M. Bardi and M. Falcone, *An approximation scheme for the minimum time function*, SIAM J. Control Optim., 28 (1990), pp. 950–965.

[5] E. N. Barron and R. Jensen, *Optimal control and semicontinuous viscosity solutions*, Proc. Amer. Math. Soc., 113 (1991), pp. 397–402.
[6] E. N. Barron and R. Jensen, *Semicontinuous viscosity solutions for Hamilton-Jacobi equations with convex Hamiltonians*, Comm. Partial Differential Equations, 15 (1990), pp. 1713–1742.
[7] P. Cannarsa and C. Sinestrari, *Convexity Properties of the Minimal Time Function*, preprint.
[8] O. Carja, F. Mignanego, and G. Pieri, *Lower semicontinuous solutions of the Bellman equation for the minimal time problem*, J. Optim. Theory Appl., 85 (1995), pp. 563–574.
[9] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Classics Appl. Math. 5, SIAM, Philadelphia, 1990. (Originally published by Wiley Interscience, New York, 1983.)
[10] F. H. Clarke, *Methods of Dynamic and Nonsmooth Optimization*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 57, SIAM, Philadelphia, 1989.
[11] F. H. Clarke and Yu. S. Ledyaev, *Mean value inequalities in Hilbert space*, Trans. Amer. Math. Soc., 344 (1994), pp. 307–324.
[12] F. H. Clarke, Yu. S. Ledyaev, R. J. Stern, and P. R. Wolenski, *Qualitative properties of trajectories of control systems: A survey*, J. Dynamical Control Systems, 1 (1995), pp. 1–48.
[13] F. H. Clarke, Yu. S. Ledyaev, R. J. Stern, and P. R. Wolenski, *Subdifferential and Nonlinear Analysis*, in Nonsmooth Analysis and Control Theory, Springer-Verlag, New York, 1988.
[14] F. H. Clarke, Yu. S. Ledyaev, and P. R. Wolenski, *Proximal analysis and minimization principles*, J. Math. Anal. Appl., 196 (1995), pp. 722–735.
[15] F. H. Clarke, R. J. Stern, and P. R. Wolenski, *Subgradient criteria for monotonicity, the Lipschitz condition, and convexity*, Canad. J. Math., 45 (1983), pp. 1167–1183.
[16] F. H. Clarke, R. J. Stern, and P. R. Wolenski, *Proximal smoothness and the lower C^2 property*, J. Convex Anal., 2 (1995), pp. 117–144.
[17] F. H. Clarke and R. B. Vinter, *Local optimality conditions and Lipschitzian solutions to the Hamilton–Jacobi equation*, SIAM J. Control Optim., 21 (1983), pp. 856–870.
[18] F. H. Clarke and P. R. Wolenski, *Control of systems to sets and their interiors*, J. Optim. Theory Appl., 88 (1996), pp. 3–23.
[19] M. G. Crandall, H. Ishii, and P.-L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc., 27 (1992), pp. 1–67.
[20] M. G. Crandall and P.-L. Lions, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.
[21] L. C. Evans and M. R. James, *The Hamilton–Jacobi–Bellman equation for time-optimal control*, SIAM J. Control Optim., 27 (1989), pp. 1477–1489.
[22] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.
[23] A. F. Filippov, *Classical solutions of differential equations with multivalued right-hand side*, SIAM J. Control Optim., 5 (1967), pp. 609–621.
[24] H. Frankowska, *Lower semicontinuous solutions of the Hamilton-Jacobi equation*, SIAM J. Control Optim., 31 (1993), pp. 257–272.
[25] P. D. Loewen, *Optimal Control via Nonsmooth Analysis*, CRM Proc. Lecture Notes 2, AMS, Providence, RI, 1993.
[26] N. N. Petrov, *On the Bellman function for the time optimal process problem*, J. Appl. Math. Mech., 34 (1970), pp. 785–791.
[27] R. T. Rockafellar, *Proximal subgradients, marginal values, and augmented Lagrangians in nonconvex optimization*, Math. Oper. Res., 6 (1981), pp. 424–436.
[28] P. Soravia, *Hölder continuity of the minimum-time function for C^1-manifold targets*, J. Optim. Theory Appl., 75 (1992), pp. 401–421.
[29] P. Soravia, *Pursuit-evasion problems and viscosity solutions of Isaacs equations*, SIAM J. Control Optim., 31 (1993), pp. 604–623.
[30] P. Soravia, *Discontinuous viscosity solutions to Dirichlet problems for Hamilton–Jacobi equations with convex Hamiltonians*, Comm. Partial Differential Equations, 18 (1993), pp. 1493–1514.
[31] P. Soravia, *Generalized motion of a front propagating along its normal direction: A differential games approach*, Nonlinear Anal., 22 (1994), pp. 1247–1262.
[32] V. Staicu, *Minimal time function and viscosity solutions*, J. Optim. Theory Appl., 60 (1989), pp. 81–91.
[33] A. I. Subbotin, *A generalization of the basic equation of the theory of differential games*, Soviet Math. Dokl., 22 (1980), pp. 358–362.

[34] A. I. Subbotin, *Generalization of the main equation of differential games*, J. Optim. Theory Appl., 43 (1984), pp. 103–133.
[35] H. J. Sussmann, *A general theorem on local controllability*, SIAM J. Control Optim., 25 (1987), pp. 158–194.
[36] V. Veliov, *Sufficient conditions for viability under imperfect measurement*, Set-Valued Anal., 1 (1993), pp. 305–317.
[37] V. Veliov, *Lipschitz Continuity of the Value Function in Optimal Control*, J. Optim. Theory Appl., 94 (1997), pp. 335–363.
[38] P. R. Wolenski, *A uniqueness theorem for differential inclusions*, J. Differential Equations, 84 (1990), pp. 165–182.
[39] P. R. Wolenski, *The exponential formula for the reachable set of a Lipschitz differential inclusion*, SIAM J. Control Optim., 28 (1990), pp. 1148–1161.
[40] R. Yue, *On the Properties of Bellman's Function in Time Optimal Control Problems*, preprint.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 3, pp. 1073–1085, May 1998

014

A CLASSICAL APPROACH TO UNIFORM NULL CONTROLLABILITY FOR ELASTIC BEAMS*

MIGUEL ANGEL MORELES†

Abstract. The Rayleigh beam equation is the formal limit of the Timoshenko beam equation as the *shear modulus* $K \to +\infty$. Following a method in W. Littman, *Ann. Scuola Norm. Sup. Pisa Cl. Sci.* (4), 5 (1978), pp. 567–580 and W. Littman, Proc. 24th IEEE CDC, Fort Lauderdale, FL, 1985, controllability is possible. That is, the evolution systems associated with the Rayleigh and Timoshenko equations can be driven to rest by applying appropriate controls at both ends of the beam. In this work we show that the process is uniform; more precisely, controllability of the Rayleigh system can be achieved by letting $K \to +\infty$ in the solution of the Timoshenko controllability problem.

Key words. null controllability, elastic beams, hyperbolic operators

AMS subject classifications. 93B, 35L, 34H

PII. S0363012996304882

1. Introduction.

Let us recall two problems in boundary controllability.

The null boundary controllability problem (NBCP). Suppose we have a well-posed initial-boundary value problem for an *evolution equation* $Lw = 0$ in a cylindrical domain $Q = \Omega \times [0, +\infty)$, where Ω is a bounded domain in $\mathbb{R}^n$. The NBCP deals with the following question: Given initial data in Ω at $t = 0$, can these data be supplemented with appropriate inhomogeneous time-dependent boundary data (boundary controls), prescribed on the lateral boundary of Q, such that the solution of the initial-boundary value problem will vanish for $t \geq T_0$?

The uniform null boundary controllability problem (UNBCP). Let $L_\varepsilon = L + \varepsilon M$, where L_ε and L are evolution operators with well-posed initial-boundary value problems as before. The UNBCP deals with the following questions: Is it possible to find ε_0 and T_0 with T_0 independent of $\varepsilon \leq \varepsilon_0$ such that one has null controllability for $T > T_0$ for all equations $L_\varepsilon w_\varepsilon = 0$, $\varepsilon \leq \varepsilon_0$, and for $Lw = 0$? And what happens, as $\varepsilon \to 0$ to the boundary controls which drive the system to rest?

We will be concerned with these two problems when L, L_ε are *elastic beam operators.*

In Russell [14] several models for the elastic beam are considered, in particular the model of Timoshenko described by the system

$$I_\rho \psi_{tt} - EI\psi_{xx} + K(\psi - w_x) = 0,$$
$$\rho w_{tt} + K(\psi - w_x)_x = 0$$

plus appropriate initial and boundary conditions.

Here $w(x,t)$ represents the vertical displacement of the *elastic axis* of the beam, and $\psi(x,t)$ is the rotation angle due to bending and shear.

The physical constants in the model are $\rho \equiv$ density, $EI \equiv$ flexural rigidity, $I_\rho \equiv$ rotary inertia, and $K \equiv$ shear modulus.

By formal differentiation the system can be uncoupled to obtain the Timoshenko equation

$$\rho w_{tt} - I_\rho w_{ttxx} + EI w_{xxxx} + \frac{\rho}{K}(I_\rho w_{tttt} - EI w_{ttxx}) = 0. \tag{1.1}$$

*Received by the editors June 7, 1996; accepted for publication (in revised form) April 10, 1997. http://www.siam.org/journals/sicon/36-3/30488.html

†Centro de Investigación en Matemáticas (CIMAT), Apdo. Postal 402, Guanajuato, GTO 36000, Mexico (moreles@fractal.cimat.mx).

In this equation $-I_\rho w_{ttxx}$ is the contribution of rotary inertia and the term due to shear is $\frac{\rho}{K}\left(I_\rho w_{tttt} - EIw_{ttxx}\right)$.

If the shear effect is neglected we are led to the Rayleigh equation

$$\rho w_{tt} - I_\rho w_{ttxx} + EIw_{xxxx} = 0. \tag{1.2}$$

It is observed that (1.1) is a perturbation of (1.2).

In our exposition we deal with the UNBCP for (1.1) and (1.2). Uniformity is studied when $K \to +\infty$. The outline is as follows.

In section 2 we present the main result; roughly speaking it says that the null-controlled problem for the Timoshenko equation converges to that for the Rayleigh equation. The problem is not new, and there are several related results in the literature covering also similar problems for plate systems. The last paragraph of the section, entitled Earlier Results, gives references as well as a brief discussion of these results.

Our proof is carried out in sections 3 and 4. The idea is to follow a well-known result by Littman [9, 10] to solve the NBCP.

Two perturbation problems arise from Littman's method. Section 3 studies convergence of the solution of a homogeneous Cauchy problem for the Timoshenko equation to that of Rayleigh. The Cauchy data, compactly supported, are given in the x-axis. Section 4 deals with the problem in the x-direction, now the Cauchy problems are nonhomogeneous and the Cauchy data, in the t-axis, are zero. It will become apparent that the former is a *singular perturbation* problem, whereas the latter is a *regular perturbation* problem.

A natural continuation to this work is to solve the UNBCP with controls in only one end of the beam. Mention of this, and other problems of related interest is in section 5.

To simplify the statements of results, we shall use the term *smooth* loosely, hoping that the notion of smooth will be clear from the context.

Also, all positive constants independent of the shear modulus K will be denoted by c. When precision is necessary, we distinguish between different constants by using subscripts.

2. Main result. Denote by L_K the Timoshenko operator

$$L_K = \rho\partial_t^2 - I_\rho\partial_t^2\partial_x^2 + EI\partial_x^4 + \frac{\rho}{K}\left(I_\rho\partial_t^4 - EI\partial_t^2\partial_x^2\right)$$

and by L_0 the Rayleigh operator

$$L_0 = \rho\partial_t^2 - I_\rho\partial_t^2\partial_x^2 + EI\partial_x^4.$$

For any $s \in \mathbb{R}$ consider the function (in $\mathbb{R}^n$)

$$\Lambda_s \equiv \Lambda_s(\xi) = \left(1 + |\xi|^2\right)^{\frac{s}{2}} \ . \tag{2.1}$$

Denote as customary by H^s the Sobolev space with norm

$$\|u\|_s^2 = \int \left(\Lambda_s(\xi)\,|\widehat{u}(\xi)|\right)^2 d\xi,$$

where $\widehat{u}(\xi) \equiv (\mathcal{F}u)(\xi)$ is the Fourier transform of u. For bounded domains Ω define $H^s(\Omega)$ as usual.

Let $\Omega = (-a, a)$, $Q = [-a, a] \times [0, +\infty)$, and let L be one of the operators above and let us say its order with respect to t is n, where n is either 2 or 4.

The initial-boundary value problem (IBVP) to consider is the following:

$$\begin{aligned} &Lw = 0, && (x,t) \in Q, \\ &\partial_t^j w(x,0) = w^j(x), && x \in \Omega, \qquad j = 0,1,\dots,n-1 \end{aligned} \tag{2.2}$$

with boundary conditions

$$B_- w(-a,t) = g_-(t), \quad B_+ w(a,t) = g_+(t),$$

where B_-, B_+ are differential boundary operators. We assume that the problem is *well posed.* Later we will see that uniqueness is enough.

For $b \geq 0$ denote by $\sum_b$ the strip $[-a,a] \times [b,+\infty)$. Finally, extend the Cauchy data in (2.2) as smoothly as possible to have compact support, and consider the Cauchy problem in the upper half-plane. Following Littman [9, 10], there exists $T_1 > 0$ and a function w, smooth away from zero, such that w satisfies this new Cauchy problem, and $w(x,t) \equiv 0$ in the strip $\sum_{T_1}$. An important observation is that Littman's proof is constructive and independent of the boundary conditions.

Let w_K be the Littman's solution for the Timoshenko equation and T_{1_K} be the control time. Similarly, w_0 and T_{1_0} for the Rayleigh equation.

The UNBCP is a consequence of the following theorem.

THEOREM 2.1. *Let $m \geq 0$. On the interval $(-a,a)$ let*

$$w^0 \in H^{m+3}, \quad w^1 \in H^{m+2}, \quad w^2, w^3 \in H^m. \tag{2.3}$$

(i) *There exists T_1 independent of K such that $w_K(x,t)$ and $w_0(x,t)$ both vanish in the strip $\sum_{T_1}$ (i.e., $T_1 = T_{1_K} = T_{1_0}$).*

(ii) *For bounded subsets of $\sum_0$ $\partial_x^m \partial_t^l w_K$ converges to $\partial_x^m \partial_t^l w_0$ as $K \to +\infty$ in the L^∞-norm for $l = 0, 1$.*

Several remarks are in order:

1. To solve the NBCP for the Rayleigh and Timoshenko equations we just need to read off appropriate boundary conditions. More precisely, the following must hold:
   Uniqueness assumption. If the IBVP has a solution, it is unique in *sufficiently large* function spaces. This will be made precise in the proof.
   Let c be $-a$ or a. To fulfill this assumption some admissible boundary conditions are
   (i) $w(c,t) = w_x(c,t) = 0$ (clamped end);
   (ii) $w(c,t) = w_{xx}(c,t) = 0$ (simply supported end);
   (iii) $w_{xx}(c,t) = I_\rho w_{ttx}(c,t) - EI w_{xxx}(c,t) = 0$ (free end).
   (see Russell [14].)
2. For clamped and simply supported ends we obtain *strong* convergence as implied by (ii) in the theorem. It will become apparent later that for a free end weaker convergence holds.
3. We shall see that for T_0 as in (3.14) if we restrict to $\sum_{T_0}$ we obtain uniform convergence in compacta.
4. Observe that in (2.3) we require more regularity than necessary to solve the IBVP. This is because of estimates (3.11) and (3.12) in section 3, which we have been unable to improve.
5. Thanks to Littman's method, our result is suitable for numerical implementation. A drawback is the need to impose controls in the whole boundary.

Earlier results. There are several results related to the problem of concern of this paper and with the more complex problem of plate systems. Our problem was

motivated by the works of Lions [7, 8] and Lagnese and Lions [6], where a complete treatment of boundary controllability is given. See also the monograph by Lagnese [3].

Komornik [2] proves that for a Reissner–Mindlin plate, the control time *is independent of* K. For the Timoshenko beam equation a similar result follows. It is not known if the solution to the control problem for the Reissner–Mindlin system (resp., for the Timoshenko system) converges to the solution to the control problem for the Kirchhoff system (resp., the Rayleigh system) as the shear modulus approaches infinity. For the beam problem, our result provides a positive answer.

The perturbation problem, i.e., convergence of solutions of the Timoshenko equation to that of Rayleigh, as well as for plate systems, is dealt in a more complex setting in the works of Lagnese and Leugering [4] and Lagnese, Leugering, and Schmidt [5] . Namely, they describe the dynamics of networks of interconnected Reissner–Mindlin thin plates and establish convergence to the corresponding dynamic model for networks of interconnected Kirchhoff plates.

The proof of Theorem 2.1 is in the next two sections.

3. Singular perturbation. Since we are interested in K large, we assume $K \geq K_0$ for K_0 such that

$$\frac{K_0}{\rho} >> \frac{EI}{I_\rho}. \tag{3.1}$$

In this section the main tool is the Fourier transform; hence it is convenient to introduce the notation

$$D_x \equiv -i\partial_x, \quad D_t \equiv -i\partial_t.$$

The Timoshenko and Rayleigh operators are written, respectively, in the form

$$P_K = D_t^4 - \left[\frac{K}{I_\rho} + \left(\frac{K}{\rho} + \frac{EI}{I_\rho}\right) D_x^2\right] D_t^2 + \frac{KEI}{\rho I_\rho} D_x^4,$$

$$P_0 = \left(1 + \frac{I_\rho}{\rho} D_x^2\right) D_t^2 - \frac{EI}{\rho} D_x^4.$$

The associated Cauchy problems are

$$\begin{gathered} P_K[u_K] = 0, \\ D_t^j u_K(x,0) = w^j(x), \quad j = 0,1,2,3, \end{gathered}$$

and

$$\begin{gathered} P_0[u_0] = 0, \\ u_0(x,0) = w^0(x), \quad D_t u_0(x,0) = w^1(x). \end{gathered}$$

Observe that in the t-direction, the Timoshenko equation is of order 4, whereas the Rayleigh equation is of order 2; there is a *loss* of two initial conditions. Hence, the Timoshenko equation is a singular perturbation of that of Rayleigh.

THEOREM 3.1. *For $m \geq 1$, assume*

$$w^0 \in H^{m+3}, w^1 \in H^{m+2}, w^2 \in H^m, w^3 \in H^{m-1}.$$

Then $D_x^j D_t^l u_K$ converges to $D_x^j D_t^l u_0$ when $K \to \infty$ for $l = 0, 1$; $j \leq m$ in the norm $L^\infty((-\infty,\infty) \times [\tau_0, \tau_1])$ with $0 \leq \tau_0 \leq \tau_1 < \infty$.

The main step of the proof is the theorem below. For the statement we need some notation.

Let U_K be the Fourier transform of u_K and U_0 be that of u_0. Decompose U_K in the form

$$U_K = U_0 + R.$$

Then R satisfies

$$\begin{aligned} &P_K(\xi, D_t)R(\xi,t) = F(\xi,t), \\ &D_t^j R(\xi,0) = R^j(\xi), \quad j = 0,1,2,3, \end{aligned} \tag{3.2}$$

where

$$\begin{aligned} &R^0(\xi) = R^1(\xi) = 0, \\ &R^j(\xi) = W^j(\xi) - D_t^j U_0(\xi,0), \quad j = 2,3, \end{aligned} \tag{3.3}$$

and

$$F(\xi,t) = -\left(D_t^4 U_0 - \frac{EI}{I_\rho}\xi^2 D_t^2 U_0\right).$$

Recall the function Λ_s in (2.1) and observe that $\Lambda_0 = 1$ and $\Lambda_{s+\sigma} = \Lambda_s \Lambda_\sigma$. Moreover $\Lambda_s < \Lambda_\sigma$ if $s < \sigma$.

THEOREM 3.2. *Let $R(\xi,t)$ be the solution of* (3.2). *Then*

$$|R(\xi,t)| \le \frac{c}{K}(1+t)^2\left(\Lambda_2\left|W^0\right| + \Lambda_1\left|W^1\right| + \Lambda_{-2}\left|W^2\right| + \Lambda_{-2}\left|W^3\right|\right) \tag{3.4}$$

and for $l = 1,2,3$

$$\left|D_t^l R(\xi,t)\right| \le \frac{c\,(1+t)}{\left(\sqrt{K}\right)^{2-l}}\left(\Lambda_{l+1}\left|W^0\right| + \Lambda_l\left|W^1\right| + \Lambda_{l-2}\left|W^2\right| + \Lambda_{l-3}\left|W^3\right|\right). \tag{3.5}$$

Proof. It is readily seen that

$$U_0(\xi,t) = W^0(\xi)\cos\lambda t + W^1(\xi)\frac{i}{\lambda}\sin\lambda t$$

with

$$\lambda \equiv \lambda(\xi) = \frac{\sqrt{\frac{EI}{\rho}}\,\xi^2}{\sqrt{1 + \frac{I_\rho}{\rho}\,\xi^2}}.$$

We have the estimates

$$\begin{aligned} |U_0(\xi,t)| &\le \left|W^0(\xi)\right| + t\left|W^1(\xi)\right|, \\ |D_t^n U_0(\xi,t)| &\le c\left(\Lambda_n\left|W^0(\xi)\right| + \Lambda_{n-1}\left|W^1(\xi)\right|\right), \quad n = 1,2,\ldots. \end{aligned} \tag{3.6}$$

Let

$$p_4(\xi,z) = P_K(\xi,z) = \sum_{i=0}^{4} a_i(\xi) z^{4-i};$$

notice that $a_0 = 1$.

Define

$$p_j(\xi,z)=\sum_{i=0}^{j}a_i(\xi)z^{j-i}$$

and

$$G_j(\xi,t)=\frac{1}{2\pi i}\int_C\frac{e^{izt}p_{3-j}(\xi,z)}{p_4(\xi,z)}dz\,,\qquad j=0,1,2,3, \tag{3.7}$$

where C is a simple curve in the z-plane which surrounds the zeros of $p_4(\xi,z)$.

By residues it follows that $G_j(\xi,t)$ satisfies

$$p_4(\xi,D_t)G_j=0$$

and initial data

$$D_t^iG_j(\xi,0)=\begin{cases}1, & i=j,\\ 0, & i\neq j.\end{cases}$$

Hence

$$R(\xi,t)=\sum_{j=0}^{3}R^j(\xi)G_j(\xi,t)+\int_0^tG_3(\xi,t-s)F(\xi,s)ds.$$

Since $R^0(\xi)=R^1(\xi)=0$ we need to bound the t-derivatives of

$$R(\xi,t)=R^2(\xi)G_2(\xi,t)+R^3(\xi)G_3(\xi,t)+\int_0^tG_3(\xi,t-s)F(\xi,s)ds; \tag{3.8}$$

$R_2(\xi)$ and $R_3(\xi)$ as in (3.3). Let us compute the function $G_2(\xi,t)$ explicitly.

The roots of the polynomial

$$P_K(\xi,z)=z^4-\left[\frac{K}{I_\rho}+\left(\frac{K}{\rho}+\frac{EI}{I_\rho}\right)\xi^2\right]z^2+\frac{KEI}{\rho I_\rho}\xi^4$$

are given by

$$z_4=-z_1=\tfrac{1}{\sqrt{2}}\sqrt{a+\sqrt{a^2-4b}}\quad\text{and}\quad z_3=-z_2=\tfrac{1}{\sqrt{2}}\sqrt{a-\sqrt{a^2-4b}},$$

where

$$a=\frac{K}{I_\rho}+\left(\frac{K}{\rho}+\frac{EI}{I_\rho}\right)\xi^2\quad\text{and}\quad b=\frac{KEI}{\rho I_\rho}\xi^4.$$

We obtain by residues in (3.7) that

$$G_2(\xi,t)=\frac{1}{(z_4)^2-(z_3)^2}\left(\cos z_4t-\cos z_3t\right).$$

By the properties of Λ_s we see also that

$$\frac{1}{(z_4)^2-(z_3)^2}\leq\frac{c}{K}\Lambda_{-2},\quad z_3\leq c\Lambda_1,\quad z_4\leq c\sqrt{K}\,\Lambda_1.$$

It follows that

$$|D_t^nG_2(\xi,t)|\leq c\left(\sqrt{K}\right)^{n-2}\Lambda_{n-2}\,,\qquad n=0,1,2,\dots\,. \tag{3.9}$$

Since $D_t G_3(\xi,t) = G_2(\xi,t)$ we have

$$|G_3(\xi,t)| \le c\frac{t}{K}\Lambda_{-2} \quad \text{and} \quad |D_t^n G_3(\xi,t)| \le c\left(\sqrt{K}\right)^{n-3}\Lambda_{n-3}. \tag{3.10}$$

On the other hand, from the estimates in (3.6) for $U_0(\xi,t)$ it follows that

$$|F(\xi,t)| \le c\left(\Lambda_4\left|W^0\right| + \Lambda_3\left|W^3\right|\right).$$

All of these estimates in the expression (3.8) prove the result. □

The estimates (3.4), (3.5), and Parseval's formula imply

$$\begin{aligned} \left|D_x^m\left(u_K(x,t) - u_0(x,t)\right)\right| &\le \frac{c}{K}(1+t)^2\left(\left\|w^0\right\|_{m+3} + \left\|w^1\right\|_{m+2}\right. \\ &\quad \left. + \left\|w^2\right\|_{m-1} + \left\|w^3\right\|_{m-1}\right) \end{aligned} \tag{3.11}$$

and for $l = 1,2,3$

$$\begin{aligned} \left|D_x^m D_t^l\left(u_K(x,t) - u_0(x,t)\right)\right| &\le \frac{c(1+t)}{\left(\sqrt{K}\right)^{2-l}}\left(\left\|w^0\right\|_{m+l+2} + \left\|w^1\right\|_{m+l+1}\right. \\ &\quad \left. + \left\|w^2\right\|_{m+l-1} + \left\|w^3\right\|_{m+l-2}\right). \end{aligned} \tag{3.12}$$

Thus Theorem 3.1 follows as claimed.

For later reference we deduce from (3.6) the estimates

$$\begin{aligned} \left|D_x^m u_0(x,t)\right| &\le c(1+t)\left(\left\|w^0\right\|_{m+1} + \left\|w^1\right\|_{m+1}\right), \\ \left|D_x^m D_t^l u_0(x,t)\right| &\le c\left(\left\|w^0\right\|_{m+l+1} + \left\|w^1\right\|_{m+l}\right), \quad l = 1,2,\dots\ . \end{aligned} \tag{3.13}$$

Remarks.

1. Notice that convergence is valid even at $t = 0$ for $D_x^j u_K$ and $D_x^j D_t u_K$. Also weaker convergence is obtained for second-order derivatives with respect to t as shown by the estimate (3.12).
2. If the Cauchy data are compactly supported, then u_K and u_0 are smooth away from zero. This fact is a consequence of the regularity of the fundamental solutions of the Timoshenko and Rayleigh operators. A brief study follows.

The Timoshenko operator. The *principal part* of the Timoshenko operator is

$$\begin{aligned} P_{K4} &= \frac{\rho I_\rho}{K}\partial_t^4 - \left(I_\rho + \frac{\rho EI}{K}\right)\partial_t^2\partial_x^2 + EI\partial_x^4 \\ &\equiv \frac{\rho I_\rho}{K}\left(\partial_t - \sqrt{\frac{K}{\rho}}\partial_x\right)\left(\partial_t + \sqrt{\frac{K}{\rho}}\partial_x\right)\left(\partial_t - \sqrt{\frac{EI}{I_\rho}}\partial_x\right)\left(\partial_t + \sqrt{\frac{EI}{I_\rho}}\partial_x\right). \end{aligned}$$

Observe that P_K is strictly hyperbolic with respect to both the x- and t-axes. Moreover, there are unique *fundamental solutions* G_+, G_R, G_L of P_K supported, respectively, in the cones

$$\left\{|x| \le \sqrt{\tfrac{K}{\rho}}t, \quad t \ge 0\right\}, \quad \left\{|t| \le \sqrt{\tfrac{EI}{I_\rho}}x, \quad x \ge 0\right\}, \quad \left\{|t| \le \sqrt{\tfrac{EI}{I_\rho}}|x|, \quad x \le 0\right\},$$

denoted, respectively, by Γ_+ , Γ_R , Γ_L . In particular G_+ is smooth in the cone

$$\Gamma = \left\{|x| < \sqrt{\frac{EI}{I_\rho}}t, \quad t > 0\right\}.$$

All these facts follow from the theory of hyperbolic operators extensively covered in the literature. A classical reference is Hörmander [1].

The Rayleigh operator. In the sense of Ortner and Wagner [12, 13] the Rayleigh operator is quasi-hyperbolic with respect to the t-axis. They prove that there is a unique fundamental solution s_0 of P_0 with the properties $s_0 = 0$ for $t < 0$ and $e^{-\sigma t} s_0 \in \mathcal{S}'$. Here $\mathcal{S}'$ is the space of *tempered distributions*. Moreover s_0 is also smooth in the cone Γ. For a constructive proof see Moreles [11].

With respect to the x-axis P_0 is hyperbolic with $\sqrt{\frac{I_\rho}{EI}}$ the maximum speed of propagation. As before, there are unique fundamental solutions G_{0R}, G_{0L} supported, respectively, in the cones Γ_R and Γ_L.

If the Cauchy data are supported in $[-a, a]$ let $t_0 > 3a\sqrt{\frac{I_\rho}{EI}}$. Define

$$T_0 = t_0 - a\sqrt{\frac{I_\rho}{EI}}. \tag{3.14}$$

Then u_K and u_0 are smooth in a neighborhood of $[-a, a] \times [T_0, +\infty)$ because of the smoothness of the fundamental solutions s_0, G_+ of the Rayleigh and Timoshenko operators in the cone Γ.

4. Regular perturbation. Consider the Timoshenko operator in the form

$$\begin{aligned} L_K(\partial_x, \partial_t) &= \partial_x^4 - \left(\frac{I_\rho}{EI} + \frac{\rho}{K}\right)\partial_t^2\partial_x^2 + \frac{\rho I_\rho}{EIK}\partial_t^4 + \frac{\rho}{EI}\partial_t^2 \\ &\equiv \prod_{j=1}^{4}(\partial_x - \mu_j\partial_t) + \frac{\rho}{EI}\partial_t^2, \end{aligned}$$

where

$$\mu_4 = -\mu_1 = \sqrt{\frac{I_\rho}{EI}}, \qquad \mu_3 = -\mu_2 = \sqrt{\frac{\rho}{K}},$$

and the Rayleigh operator

$$L_0(\partial_x, \partial_t) = \partial_x^4 - \frac{I_\rho}{EI}\partial_t^2\partial_x^2 + \frac{\rho}{EI}\partial_t^2.$$

In what follows we only consider $x \geq 0$ since the case $x \leq 0$ is similar. Choose $t_1 > t_0$. Consider a cutoff function of the t variable so that

$$\varphi(t) = \begin{cases} 1, & t \leq t_0, \\ 0, & t \geq t_1. \end{cases}$$

Let

$$T_1 = t_1 + a\sqrt{\frac{I_\rho}{EI}}.$$

Define

$$f_K(x,t) = L_K[\varphi u_K], \qquad f_0(x,t) = L_0[\varphi u_0].$$

Let v_K be the solution of the Cauchy problem

$$\begin{aligned} &L_K[v_K] = f_K, \\ &\partial_x^j v_K(0,t) = 0, \qquad j = 0, 1, 2, 3, \end{aligned}$$

and v_0 the solution of

$$\begin{aligned} &L_0[v_0] = f_0, \\ &\partial_x^j v_0(0,t) = 0, \qquad j = 0,1,2,3. \end{aligned}$$

Observe that f_K and f_0 vanish outside the strip $[t_0, t_1]$. Since L_K and L_0 are hyperbolic, it follows that v_K and v_0 are smooth and for $-a \le x \le a$ vanish in a neighborhood of $t = 0$ and of $t \ge T_1$.

We call the problem of convergence of $v_K \to v_0$ a regular perturbation problem because both the Rayleigh and Timoshenko operators are of order 4 in the x-direction.

Let $v = v_K - v_0$. Then v satisfies

$$\begin{aligned} &L_K[v] = f, \\ &\partial_x^j v(0,t) = 0, \end{aligned} \tag{4.1}$$

where

$$f = f_K - f_0 + \frac{\rho I_\rho}{EIK}\partial_t^4 v_0 + \frac{\rho}{K}\partial_t^2 v_0\,. \tag{4.2}$$

LEMMA 4.1. *The function in* (4.2) *satisfies*

$$\begin{aligned} \left|\partial_x^j f(x,t)\right| \;\le\; & \frac{c}{\sqrt{K}}\,(1+t)^2\left[1 + \left(\max_{1\le i\le 4}\left|\partial_t^i \varphi(t)\right|\right)\right] \\ & \cdot \left(\left\|w^0\right\|_{j+5} + \left\|w^1\right\|_{j+4} + \left\|w^2\right\|_{j+2} + \left\|w^3\right\|_{j+1}\right). \end{aligned}$$

Proof. Here

$$\begin{aligned} f_K \;=\; & -\left(\frac{I_\rho}{EI} - \frac{\rho}{K}\right)\left(\partial_t^2\varphi\partial_x^2 u_K + 2\partial_t\varphi\partial_t\partial_x^2 u_K\right) \\ & + \frac{\rho I_\rho}{EIK}\left[\left(\partial_t^4\varphi\right)u_K + 4\partial_t^3\varphi\partial_t u_K + 6\partial_t^2\varphi\partial_t^2 u_K + 4\partial_t\varphi\partial_t^3 u_K\right] \\ & + \frac{\rho}{EI}\left[\left(\partial_t^2\varphi\right)u_K + 2\partial_t\varphi\partial_t u_K\right] \end{aligned}$$

and

$$f_0 \;=\; -\left(\frac{I_\rho}{EI}\right)\left(\partial_t^2\varphi\partial_x^2 u_0 + 2\partial_t\varphi\partial_t\partial_x^2 u_0\right) + \frac{\rho}{EI}\left[\left(\partial_t^2\varphi\right)u_0 + 2\partial_t\varphi\partial_t u_0\right].$$

Thus

$$\begin{aligned} f_K - f_0 \;=\; & -\left(\frac{I_\rho}{EI}\right)\left[\left(\partial_t^2\varphi\right)\left(\partial_x^2 u_K - \partial_x^2 u_0\right) + \left(2\partial_t\varphi\right)\left(\partial_t\partial_x^2 u_K - \partial_t\partial_x^2 u_0\right)\right] \\ & + \left(\frac{\rho}{EI}\right)\left[\left(\partial_t^2\varphi\right)\left(u_K - u_0\right) + \left(2\partial_t\varphi\right)\left(\partial_t u_K - \partial_t u_0\right)\right] \\ & + \left(\frac{\rho}{K}\right)\left(\partial_t^2\varphi\partial_x^2 u_K + 2\partial_t\varphi\partial_t\partial_x^2 u_K\right) \\ & + \frac{\rho I_\rho}{EIK}\left[\left(\partial_t^4\varphi\right)u_K + 4\partial_t^3\varphi\partial_t u_K + 6\partial_t^2\varphi\partial_t^2 u_K + 4\partial_t\varphi\partial_t^3 u_K\right]. \end{aligned}$$

Hence

$$\begin{aligned} \left|f_K - f_0\right| \;\le\; & c\left(\max_{1\le i\le 4}\left|\partial_t^i\varphi(t)\right|\right)\cdot\left[\left|\partial_x^2 u_K - \partial_x^2 u_0\right| + \tfrac{1}{K}\left|\partial_x^2 u_K\right|\right. \\ & + \left|\partial_t\partial_x^2 u_K - \partial_t\partial_x^2 u_0\right| + \tfrac{1}{K}\left|\partial_t\partial_x^2 u_K\right| + \left|u_K - u_0\right| + \tfrac{1}{K}\left|u_K\right| \\ & \left. + \left|\partial_t u_K - \partial_t u_0\right| + \tfrac{1}{K}\left(\left|\partial_t u_K\right| + \left|\partial_t^2 u_K\right| + \left|\partial_t^3 u_K\right|\right)\right]. \end{aligned}$$

To estimate the right-hand side use (3.11) and (3.12) for the terms of the form $\left|\partial_t^l \partial_x^m u_K - \partial_t^l \partial_x^m u_0\right|$. For the terms with factor $1/K$ add (3.13) and the triangle inequality. To complete the proof notice that v_0 in (4.2) is smooth and independent of K. □

The behavior of v_K as $K \to \infty$ is summarized in the following result.

THEOREM 4.2. *Let l, m be nonnegative integers such that $l = 0, 1$; $m + l \leq 3$. Then $\partial_x^m \partial_t^l v_K$ converges to $\partial_x^m \partial_t^l v_0$ uniformly in compacta.*

Proof. The result is proven by reducing (4.1) to a system. Let us introduce the operators

$$L_i = \prod_{j \neq i} \left(\partial_x - \mu_j \partial_t\right), \qquad i = 1, 2, 3, 4\,.$$

Two observations are in order. First, the operators $\left(\partial_x - \mu_i \partial_t\right) L_i$ and L_K have the same principal part

$$\prod_{j=1}^{4} \left(\partial_x - \mu_j \partial_t\right).$$

Second, the operators L_i constitute a set of 4 linearly independent forms in the third-order derivatives. In particular we obtain by Lagrange's interpolation formula

$$\begin{aligned} \partial_x^m \partial_t^{3-m} &= \sum_{i=1}^{4} (\mu_i)^m \prod_{j \neq i} \frac{\partial_x - \mu_j \partial_t}{\mu_i - \mu_j} \\ &= \frac{1}{\mu} \left((\mu_1)^{m-1} L_1 - (\mu_2)^{m-1} L_2 - (\mu_3)^{m-1} L_3 + (\mu_4)^{m-1} L_4 \right), \end{aligned} \tag{4.3}$$

where

$$\mu = 2 \left(\frac{I_\rho}{EI} - \frac{\rho}{K} \right).$$

By (3.1) $\mu > 0$. Let us make the change of variables

$$\begin{aligned} v_i &= L_i v, \qquad i = 1, 2, 3, 4, \\ v_5 &= v_{tt}\,. \end{aligned}$$

We are led to the following IBVP:

$$\begin{aligned} v_{jx} &= \mu_i v_{jt} - \frac{\rho}{EI} v_5 + f, \qquad j = 1, 2, 3, 4, \\ v_{5x} &= \frac{1}{\mu} \left(v_1 - v_2 - v_3 + v_4 \right), \\ v_j(0, t) &= 0, \qquad j = 1, \ldots, 5, \end{aligned} \tag{4.4}$$

and the boundary conditions

$$v_j(x, t) = 0 \qquad \text{in a neighborhood of} \qquad t = T_0, t = T_1.$$

Let

$$c_l = \begin{cases} c, & l = 0, 1, \\ c\sqrt{K}, & l = 2. \end{cases}$$

We claim that

$$\left|\partial_x^{2-l}\partial_t^l v(x,t)\right| \le c_l \left(\int_0^a \|f(y,\cdot)\|^2\, dy\right)^{\frac{1}{2}}, \tag{4.5}$$

and for third-order derivatives

$$\left|\partial_x^{3-l}\partial_t^l v(x,t)\right| \le c_l \left(\int_0^a \|\partial_y f(y,\cdot)\|^2\, dy\right)^{\frac{1}{2}}. \tag{4.6}$$

Here $\|g\|^2 = \int_{T_0}^{T_1} g(t)^2 dt$.

First we bound the functions in the system (4.4). Multiply the equation corresponding to v_j by v_j to obtain

$$\frac{d}{dx}\sum_{j=1}^{5} \|v_j(x,\cdot)\|^2 \le c_\mu \sum_{j=1}^{5} \|v_j(x,\cdot)\|^2 + 4\|f(y,\cdot)\|^2,$$

where

$$c_\mu = 4\left(1 + \frac{\rho}{EI} + \frac{1}{\mu}\right).$$

Thus by Gronwall's lemma it follows that

$$\|v_j(x,\cdot)\|^2 \le 4\, e^{c_\mu x} \int_0^x \|f(y,\cdot)\|^2\, dy, \qquad (x,t) \in [0,a]\times[T_0,T_1],$$

but $x \le a$ and K is large; hence there is a constant c independent of K such that

$$\|v_j(x,\cdot)\|^2 \le c \int_0^a \|f(y,\cdot)\|^2\, dy, \qquad (x,t) \in [0,a]\times[T_0,T_1]. \tag{4.7}$$

Now consider the second-order derivatives of v, namely $\partial_x^m \partial_t^{2-m} v$ with $m = 0,1,2$. Writing

$$\partial_x^m \partial_t^{2-m} v(x,t) = \int_{T_0}^t \partial_x^m \partial_s^{3-m} v(x,s) ds$$

we have from (4.3)

$$\left|\partial_x^m \partial_t^{2-m} v(x,t)\right| \le \frac{1}{\mu}\sum_{i=1}^{4} |\mu_i|^{m-1} \int_{T_0}^t |v_i|,$$

and by Hölder's inequality and (4.7)

$$\left|\partial_x^m \partial_t^{2-m} v(x,t)\right| \le c\left(\int_0^a \|f(y,\cdot)\|^2\, dy\right)^{\frac{1}{2}}, \quad m = 1,2. \tag{4.8}$$

For $m = 0$ in notice that $\sqrt{\frac{K}{\rho}} = \max\{\frac{1}{|\mu_i|}\}$, thus

$$\left|\partial_t^2 v(x,t)\right| \le c\sqrt{K}\left(\int_0^a \|f(y,\cdot)\|^2\, dy\right)^{\frac{1}{2}}.$$

From the estimate (4.8) and Lemma 4.1 we obtain for $m = 1, 2$ that

$$\begin{aligned}\left|\partial_x^m \partial_t^{2-m} v(x,t)\right| &\leq \frac{c\sqrt{a}}{\sqrt{K}}(1+t)^2\left[1+\left(\max_{1\leq i\leq 4}\left|\partial_t^i\varphi(t)\right|\right)\right] \\ &\quad\cdot\left(\left\|w^0\right\|_5+\left\|w^1\right\|_4+\left\|w^2\right\|_2+\left\|w^3\right\|_1\right)\end{aligned} \tag{4.9}$$

and for $m = 0$

$$\begin{aligned}\left|\partial_t^2 v(x,t)\right| &\leq c\sqrt{a}(1+t)^2\left[1+\left(\max_{1\leq i\leq 4}\left|\partial_t^i\varphi(t)\right|\right)\right] \\ &\quad\cdot\left(\left\|w^0\right\|_5+\left\|w^1\right\|_4+\left\|w^2\right\|_2+\left\|w^3\right\|_1\right)\end{aligned}$$

For third-order derivatives, let $w = v_x$ satisfying $L_K[w] = \partial_x f$ and repeat the foregoing argument to conclude our claim.

From (4.9) we obtain uniform estimates for the second-order derivatives, consequently for $\partial_x v$, $\partial_t v$, v. The theorem then follows. □

Observe that from estimates (4.5) and (4.6), $\partial_x^m \partial_t^2 v$ is bounded in compacta. It follows that $\partial_x^m \partial_t^2 v_K$ converges to $\partial_x^m \partial_t^2 v_0$ in a weaker sense.

To conclude this section let us write for the Timoshenko solution the decomposition

$$w_K = u_K\varphi - v_K,$$

whereas for the Rayleigh solution

$$w_0 = u_0\varphi - v_0.$$

Then w_K (respectively, w_0) coincides with u_K (respectively, u_0) near $t = 0$, satisfies the equation $L_K w_K = 0$ (respectively, $L_0 w_0 = 0$), and is zero for $-a \leq x \leq a$, $t \geq T_1$.

As a corollary of Theorems 3.1 and 4.2 we obtain our main result, Theorem 2.1. Therefore, the null-controlled problem for the Timoshenko equation converges to that for the Rayleigh equation as asserted.

5. Concluding comments. In (1.2) if we assume further that there is no rotary inertia effect, the resulting equation is

$$\rho w_{tt} + EIw_{xxxx} = 0,$$

the so-called Bernoulli–Euler equation.

In essence Theorem 3.1 follows from the estimates (3.9) and (3.10) for the functions $G_2(\xi,t)$ and $G_3(\xi,t)$ in (3.7). Including dependence on the rotary inertia I_ρ we obtain the corresponding estimates

$$|D_t^n G_2(\xi,t)| \leq c\frac{I_\rho}{K}\left(\left(\sqrt{\frac{K}{I_\rho}}\right)^n \Lambda_n + \xi^{2n}\right), \qquad n = 0,1,2,\ldots,$$

and

$$|G_3(\xi,t)| \leq c\frac{I_\rho}{K}t, \quad |D_t^n G_3(\xi,t)| \leq c\frac{I_\rho}{K}\left(\left(\sqrt{\frac{K}{I_\rho}}\right)^{n-1} \Lambda_{n-1} + \xi^{2(n-1)}\right), n = 1,2,\ldots .$$

We may build on this to show that for Cauchy data in suitable Sobolev spaces, the solution of the Cauchy problem for the Timoshenko equation (1.1) converges to that

of the Bernoulli–Euler equation (with like Cauchy data) as $(1/K, I_\rho) \to (0,0)$. This solves a particular case of the singular perturbation problem proposed by Russell [14].

To close our exposition let us mention some problems of related interest, namely, the UNBCP with controls in only one end; a generalization to several space dimensions, e.g., plate equations; the Bernoulli–Euler beam and the corresponding UNBCPs.

These and other considerations are left for future investigations.

Acknowledgment. This work is a summary of the author's doctoral dissertation under the direction of Prof. Walter Littman. It is a pleasure to acknowledge his advice.

REFERENCES

[1] L. Hörmander, *The Analysis of Linear Partial Differential Operators* I, II, Springer-Verlag, New York, 1983.

[2] V. Komornik, *Contrôlabilité exacte en temps minimal de quelques modèles de plaques*, C. R. Acad. Sci. Paris Sér. I Math. 307 (1988), pp. 471–474.

[3] J. Lagnese, *Boundary Stabilization of Thin Plates*, SIAM Stud. Appl. Math. 10, SIAM, Philadelphia, 1989.

[4] J. Lagnese and G. Leugering, *Modeling of dynamic networks of thin elastic plates*, Math. Methods Appl. Sci., 16 (1993), pp. 379–407.

[5] J. Lagnese, G. Leugering, and E. J. P. G. Schmidt, *Modeling, Analysis and Control of Dynamic Elastic Multi-Link Structures*, Birkhäuser Boston, Cambridge, MA, 1994.

[6] J. Lagnese and J. L. Lions, *Modelling, Analysis and Control of Thin Plates*, Masson, Paris, 1988.

[7] J. L. Lions, *Exact controllability, stabilization and perturbation for distributed systems*, SIAM Rev., 30 (1988), pp. 1–68.

[8] J. L. Lions, *Controllabilite Exacte, Perturbation et Stabilization de Systemes Distribues, Vol. 2: Perturbations*, Collect. RMA 9, Masson, Paris, 1988.

[9] W. Littman, *Boundary control theory for hyperbolic and parabolic partial differential equations with constant coefficients*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 5 (1978), pp. 567–580.

[10] W. Littman, *Boundary control theory for beams and plates*, in Proc. 24th IEEE CDC, Fort Lauderdale, FL, 1985.

[11] M. A. Moreles, *Uniform Null Boundary Controllability for the Beam Equations of Rayleigh and Timoshenko*, Ph.D. thesis, University of Minnesota, Minnneapolis, MN, 1995.

[12] N. Ortner and P. Wagner, *Some new fundamental solutions*, Math. Methods Appl. Sci., 12 (1990), pp. 439–461.

[13] N. Ortner and P. Wagner, *On the fundamental solutions of the operators of S. Timoshenko and R. D. Mindlin*, Math. Methods Appl. Sci., 15 (1992), pp. 525–535.

[14] D. L. Russell, *Mathematical models for the elastic beam and their control-theoretic implications*, in Semigroups, Theory and Applications, Vol. II, H. Brézis, M. G. Crandall, and F. Kappel, eds., Longman, New York, pp. 177–216.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 3, pp. 1086–1098, May 1998

015

EXPONENTIAL DECAY OF ENERGY OF THE EULER–BERNOULLI BEAM WITH LOCALLY DISTRIBUTED KELVIN–VOIGT DAMPING*

KANGSHENG LIU† AND ZHUANGYI LIU ‡

Abstract. In this paper, we consider the longitudinal and transversal vibrations of the Euler–Bernoulli beam with Kelvin–Voigt damping distributed locally on any subinterval of the region occupied by the beam. We prove that the semigroup associated with the equation for the transversal motion of the beam is exponentially stable, although the semigroup associated with the equation for the longitudinal motion of the beam is not exponentially stable. Due to the locally distributed and unbounded nature of the damping, we use a frequency domain method and combine a contradiction argument with the multiplier technique to carry out a special analysis for the resolvent. We also show that the associated semigroups are not analytic.

Key words. exponential stability, semigroup, local Kelvin–Voigt damping

AMS subject classifications. 35B37, 35B40

PII. S0363012996310703

1. Introduction.

Consider a clamped elastic beam of length L. One segment of the beam is made of a viscoelastic material with Kelvin–Voigt constitutive relation. By the Kirchhoff hypothesis, neglecting the rotatory inertia, the longitudinal and transversal vibration of the beam can be described by the following equations and boundary-initial conditions:

$$\begin{cases} \rho\ddot{u} - (pu' + D_a\dot{u}')' = 0 \quad \text{in } (0,L)\times\mathbf{R}^+, \\ u(0,t) = u(L,t) = 0, \\ u(x,0) = u_0(x),\ \dot{u}(x,0) = u_1(x), \end{cases} \tag{1.1}$$

$$\begin{cases} \rho\ddot{w} + (qw'' + D_b\dot{w}'')'' = 0 \quad \text{in } (0,L)\times\mathbf{R}^+, \\ w(0,t) = w(L,t) = w'(0,t) = w'(L,t) = 0, \\ w(x,0) = w_0(x),\ \ \dot{w}(x,0) = w_1(x), \quad x\in(0,L), \end{cases} \tag{1.2}$$

where u and w represent the longitudinal and transveral displacement of the beam, respectively. The coefficient functions $\rho, p, q, D_a, D_b \in L^\infty(0,L)$, $\rho, p, q \geq c_0 > 0$, $D_a = a(x)\chi_{(\alpha,\beta)}$, $D_b = b(x)\chi_{(\alpha,\beta)}$, with $\chi_{(\alpha,\beta)}$ being the characteristic function of the interval (α,β), $0 < \alpha, \beta < L$, and $a(x), b(x) \geq c_0 > 0$.

Recent advances in material science have provided new means for the suppression of vibrations of elastic structures. One approach is to bond or embed patches made of "smart material" to the underlying structure as passive or active damper. Due to the presence of the patches, the material properties of the structure, such as the density, Young's moduli, and damping coefficients, are changed. In particular, jump discontinuities at the location of the edges of the patch are usually introduced into these properties. Equations (1.1) and (1.2) model these phenomena by taking the

*Received by the editors October 18, 1996; accepted for publication (in revised form) April 10, 1997.

http://www.siam.org/journals/sicon/36-3/31070.html

†Department of Applied Mathematics, Zhejiang University, Hangzhou, 310027, China (ksliu@pub.zjpta.net.cn). The research of this author was supported in part by the National Natural Science Foundation of China.

‡Department of Mathematics and Statistics, University of Minnesota, Duluth, MN 55812-2496 (zliu@d.umn.edu).

coefficients to have the above special forms. For more information on the modeling aspect, we refer the readers to [BSW].

Our main concern is the following question: Is the locally distributed Kelvin–Voigt damping (on any subinterval of $(0, L)$) strong enough to cause uniform exponential decay of the energy of the beam? That is, do there exist constants $M_\mu, \mu > 0$ such that the energies, $\mathcal{E}(t)$, associated with (1.1) and (1.2) satisfy the inequality

$$\mathcal{E}(t) \le M_\mu e^{-\mu t}\mathcal{E}(0), \quad t > 0? \tag{1.3}$$

The exponential stability of an elastic system with damping distributed either over the entire region or along the boundary of the region has been studied extensively during the past two decades; however, only a relatively small amount of attention was paid to the stability of the system with damping distributed locally inside the domain. Discussions related to this issue were initiated by Lagnese [La] in 1983, where the exact controllability of the wave equation with locally distributed control was considered (see [Li] for the equivalence between the exact controllability and the exponential stabilizability for a conservative system). It is known that, when viscous damping is only distributed on a subinterval of the domain, (1.3) holds for the Euler–Bernoulli beam equation (longitudinal and transveral motion) with both constant coefficients [CFNS] and variable coefficients [K1], [K2]. For a higher dimensional domain (1.3) depends on the geometric properties of the subregion, where viscous damping is applied. We refer to [CFNS] for a two-dimensional Schrödinger equation on a disk and a rectangle, to [Li] for a class of n-dimensional conservative partial differential equations (PDEs), and to [Zu1], [Zu2] for an n-dimensional semilinear wave equation. It should be pointed out that the operator corresponding to the viscous damping is bounded on the underlying space while the one corresponding to the Kelvin–Voigt damping is unbounded, and is not a lower-order perturbation of the elastic operator. The effect of such locally distributed damping on the energy decay is unknown. Recently, we were able to show that locally distributed Kelvin–Voigt damping ensures the asymptotic stability of a general second-order elastic system [CLL]. In this paper, we will show that when Kelvin–Voigt damping is distributed only on a subinterval of the domain, (1.3) holds for the transversal motion but not for the longitudinal motion of the Euler–Bernoulli beam equation.

Our approach is the frequency domain method (FDM). Roughly speaking, the FDM is based on the boundedness on the imaginary axis of the resolvent of a semigroup generator in order to establish the exponential stability of the C_0-semigroup on a Hilbert space (see [Ge], [Hu], [Pr]). This method has been applied successfully to several models with locally distributed viscous damping (see [CFNS], [R]). However, due to the unboundedness of the Kelvin–Voigt damping operator, the arguments used in [CFNS] and [R] are not valid here since their arguments particularly depend on the boundedness of the damping operator. To overcome this difficulty, we combine the FDM with the multiplier technique and carry out a special analysis of the resolvent.

This paper is organized as follows. In section 2, we prove the exponential decay of energy for (1.2). In section 3, we prove the nonexponential decay of energy for (1.1). Finally, in section 4, we show that the C_0-semigroup associated with (1.2) is not analytic.

2. Transversal motion. Let $H = L^2_\rho(0, L)$ with the norm

$$\|v\| = \left(\int_0^L \rho |v(x)|^2 dx \right)^{\frac{1}{2}}$$

and $V = H_0^2(0, L)$ with the norm

$$\|v\|_V = \left(\int_0^L q|v''(x)|dx \right)^{\frac{1}{2}}.$$

Define $\mathcal{H}_1 = V \times H$ with the norm $\|(w, v)\|_{\mathcal{H}_1} = (\|w\|_V^2 + \|v\|^2)^{\frac{1}{2}}$. Then $\mathcal{H}_1$ is a Hilbert space—the finite energy state space. Define in $\mathcal{H}_1$

$$D(\mathcal{A}_1) = \left\{ (w, v) \mid w, v \in V,\ -M \equiv qw'' + D_b v'' \in H^2(0, L) \right\} \tag{2.1}$$

and

$$\mathcal{A}_1(w, v) = \left(v, \frac{1}{\rho} M'' \right). \tag{2.2}$$

Thus, (1.2) can be rewritten as an abstract evolution equation on $\mathcal{H}_1$,

$$(\dot{w}(t), \dot{v}(t)) = \mathcal{A}_1(w(t), v(t)), \quad (w(0), v(0)) = (w_0, w_1). \tag{2.3}$$

It is known that $\mathcal{A}_1$ generates a C_0-semigroup of contractions on $\mathcal{H}_1$. (See [CLL].) Therefore, $(w(t), \dot{w}(t)) = e^{\mathcal{A}_1 t}(w_0, w_1)$ gives the mild solution of (1.2) for every $(w_0, w_1) \in \mathcal{H}_1$. Moreover, $\mathcal{A}_1^{-1}$ is a bounded operator on $\mathcal{H}_1$.

We assume that ρ and q are positive constants on $[0, \alpha)$ and $(\beta, L]$; $b(x), q \in C[\alpha, \beta]$; and $\rho \in C^{1,1}[\alpha, \beta]$.

LEMMA 2.1. *The imaginary axis, $i\mathbf{R}$, $\subset \rho(\mathcal{A}_1)$ the resolvent set of $\mathcal{A}_1$.*

Proof. It is easy to show that there is no point spectrum on the imaginary axis, i.e., $i\mathbf{R} \cap \sigma_p(\mathcal{A}_1) = \emptyset$. By Lemma 4.1 in [CLL], the conclusion of this lemma is true. □

THEOREM 2.2. *Under the above assumptions on the coefficients of* (1.2)*, the semigroup $e^{\mathcal{A}_1 t}$ is exponentially stable; i.e., there exist $\nu > 0, M_\nu \geq 1$ such that*

$$\|e^{\mathcal{A}_1 t}\| \leq M_\nu e^{-\nu t} \quad \forall\, t > 0. \tag{2.4}$$

Proof. We need only to verify the condition for a C_0-semigroup of contractions on a Hilbert space being exponentially stable (see [Hu], [Pr], or [Ge]), i.e.,

$$\sup \left\{ \|(\lambda - \mathcal{A}_1)^{-1}\| \mid \lambda \in i\mathbf{R} \right\} < +\infty. \tag{2.5}$$

Suppose (2.5) is not true. By the continuity of the resolvent and the resonance theorem, there exist $\lambda_n \in i\mathbf{R}, (w_n, v_n) \in D(\mathcal{A}_1)$, $n = 1, 2, \ldots$, such that

$$\|(w_n, v_n)\|_{\mathcal{H}_1} = 1, \quad |\lambda_n| \to \infty, \tag{2.6}$$

and

$$(\lambda_n - \mathcal{A}_1)(w_n, v_n) \equiv (f_n, g_n) \to 0 \ \text{ in } \ \mathcal{H}_1. \tag{2.7}$$

This implies

$$\lambda_n w_n - v_n = f_n \to 0 \ \text{ in } \ V, \tag{2.8}$$

$$\lambda_n v_n \rho - M_n'' = \rho g_n \to 0 \ \text{ in } \ L^2(0, L), \tag{2.9}$$

where $M_n = -(qw_n'' + D_b v_n'')$.

Define

$$J(v) = \int_x^L \int_s^L \rho v(\tau) d\tau ds \tag{2.10}$$

and

$$y_n = \frac{1}{\lambda_n}[M_n + J(g_n)]. \tag{2.11}$$

Comparing (2.9) and (2.11) we have

$$y_n'' = \rho v_n. \tag{2.12}$$

The rest of the proof depends on the following two lemmas. Let $\omega_n = \sqrt{|\lambda_n|}$.

LEMMA 2.3. *The function y_n defined above has the following properties:*

$$y_n \to 0 \quad \textit{in} \ \ H^4(\alpha, \beta), \tag{2.13}$$

$$\lambda_n y_n \to 0 \quad \textit{in} \ \ L^2(\alpha, \beta), \tag{2.14}$$

$$\omega_n y_n \to 0 \quad \textit{in} \ \ H^2(\alpha, \beta). \tag{2.15}$$

Proof. From (2.7),

$$\mathrm{Re}\langle(\lambda_n - \mathcal{A}_1)(w_n, v_n), (w_n, v_n)\rangle_{\mathcal{H}_1} = \int_\alpha^\beta D_b |v_n''|^2 dx \to 0. \tag{2.16}$$

Therefore, from (2.8) we have

$$M_n \ \to \ 0 \quad \text{in} \ \ L^2(\alpha, \beta) \tag{2.17}$$

and

$$\frac{1}{\lambda_n}\|\xi v_n\|_V = \mathcal{O}(1) \tag{2.18}$$

for every $\xi \in C^\infty[0, L]$.

Equations (2.9), (2.17), and (2.18) imply that

$$\int_\alpha^\beta \xi \rho |v_n|^2 dx \to 0 \ \ \forall\, \xi \in C^\infty[0, L], \ \ \text{Supp}\, \xi \subset (\alpha, \beta). \tag{2.19}$$

Applying the interpolation theorem involving compact subdomains [A, Theorem 4.23], we find that (2.16) and (2.19) imply

$$v_n \ \to \ 0 \quad \text{in} \ \ H^2(\alpha, \beta). \tag{2.20}$$

Thus, (2.12) yields

$$\int_\alpha^\beta |y_n''''|^2 dx \ \to \ 0. \tag{2.21}$$

On the other hand, (2.14) follows from

$$\lambda_n y_n = J(g_n) - \frac{q + \lambda_n D_b}{\lambda_n} v_n'' - \frac{q}{\lambda_n} f_n'' \to 0 \quad \text{in} \ \ L^2(\alpha, \beta). \tag{2.22}$$

Since $|\lambda_n| \to \infty$, we obtain that $y_n \to 0$ in $L^2(\alpha, \beta)$. This, combined with (2.21), yields (2.13). From the interpolation inequality [A, Theorem 4.17], we also have (2.15). □

LEMMA 2.4. *The functions $w_n \in H^4(0,\alpha), H^4(\beta, L), n = 1, 2, \ldots$, have the following properties:*

$$\omega_n^4 \left(|w_n(\alpha)|^2 + |w_n'(\alpha)|^2 + |w_n(\beta)|^2 + |w_n'(\beta)|^2\right) \to 0, \tag{2.23}$$

$$\alpha q(0)|w_n''(\alpha^-)|^2 + (L-\beta)q(L)|w_n''(\beta^+)|^2 \to 2, \tag{2.24}$$

$$\omega_n^{-1} w_n'''(\alpha^-), \ \ \omega_n^{-1} w_n'''(\beta^+) \to 0. \tag{2.25}$$

Proof. Since $w_n, v_n \in V \subset H^2(0, L)$, Sobolev's embedding theorem implies that they are also in $C^1[0, L]$. By (2.8) and (2.20),

$$\lambda_n w_n \to 0 \ \ \text{in} \ \ H^2(\alpha, \beta). \tag{2.26}$$

Thus, $\lambda_n w_n$ converges to zero in $C^1[\alpha, \beta]$, which immediately leads to (2.23).

Note that $M_n = -q w_n''$ on $(0, \alpha) \cup (\beta, L)$, $q \equiv q(0)$ on $[0, \alpha)$, and $q \equiv q(L)$ on $(\beta, L]$. From the definition of the domain of $\mathcal{A}_1$, we know $w_n \in H^4(0, \alpha)$, $w_n \in H^4(\beta, L)$. It follows from (2.11) that

$$q(0)w_n''(\alpha^-) = (J(g_n) - \lambda_n y_n)(\alpha), \quad q(L)w_n''(\beta^+) = (J(g_n) - \lambda_n y_n)(\beta), \tag{2.27}$$

$$q(0)w_n'''(\alpha^-) = (J(g_n) - \lambda_n y_n)'(\alpha), \quad q(L)w_n'''(\beta^+) = (J(g_n) - \lambda_n y_n)'(\beta). \tag{2.28}$$

Dividing (2.28) by ω_n we obtain (2.25) by using (2.15) in the previous lemma.

In order to prove (2.24), we substitute (2.8) into (2.9) to get

$$\lambda_n^2 \rho w_n - M_n'' = \rho(g_n + \lambda_n f_n) \quad \text{for} \ \ x \in (0, L). \tag{2.29}$$

We multiply the above equation by $\overline{w}_n$, then integrate by parts on $(0, L)$. This leads to

$$\|\lambda_n w_n\|^2 - \|w_n\|_V^2 \to 0. \tag{2.30}$$

Here, we have used (2.6), (2.8), (2.9), and (2.16). Since $\|w_n\|_V^2 + \|v_n\|^2 = 1$ and $\lambda_n w_n - v_n$ also converges to zero in $L^2(0, L)$, (2.30) implies that both $\|\lambda_n w_n\|^2$ and $\|w_n\|_V^2$ must converge to $\frac{1}{2}$ as $n \to \infty$. This further leads to

$$\lim_{n\to\infty} \left(\int_0^\alpha + \int_\beta^L\right) \rho|\lambda_n w_n|^2 dx = \lim_{n\to\infty} \left(\int_0^\alpha + \int_\beta^L\right) q|w_n''|^2 dx = \frac{1}{2} \tag{2.31}$$

when (2.26) is taken into account.

On the intervals $(0, \alpha)$ and (β, L), (2.29) becomes

$$\lambda_n^2 \rho w_n + q w_n'''' = \rho(g_n + \lambda_n f_n). \tag{2.32}$$

We multiply the above equation by $x\overline{w_n'}$, integrate on $(0, \alpha)$, and then take the real part. Hence,

$$\text{Re} \int_0^\alpha \lambda_n^2 \rho w_n x \overline{w_n'} dx + \text{Re} \int_0^\alpha q w_n'''' x \overline{w_n'} dx = \text{Re} \int_0^\alpha \rho(g_n + \lambda_n f_n) x \overline{w_n'} dx. \tag{2.33}$$

It is easy to see that the term on the right-hand side of (2.33) converges to zero. After a straightforward calculation (integration by parts), the two terms on the left-hand side of (2.33) are

$$\mathrm{Re}\int_0^\alpha \lambda_n^2 \rho w_n x\overline{w_n'}dx = \frac{-\rho}{2}\omega_n^4 \alpha |w_n(\alpha)|^2 + \frac{1}{2}\int_0^\alpha \rho|\lambda_n w_n|^2 dx, \tag{2.34}$$

$$\begin{aligned}\mathrm{Re}\int_0^\alpha q w_n'''' x\overline{w_n'}dx &= q(0)\mathrm{Re}(\alpha w_n'''(\alpha^-) - w_n''(\alpha^-))\overline{w_n'}(\alpha)\\ &\quad + \frac{3}{2}\int_0^\alpha q|w_n''|^2 dx - \frac{\alpha}{2}q(0)|w_n''(\alpha^-)|^2.\end{aligned} \tag{2.35}$$

After substituting these terms into (2.33) and applying (2.23), (2.25), and (2.27), we have

$$\frac{1}{2}\int_0^\alpha \rho|\lambda_n w_n|^2 dx + \frac{3}{2}\int_0^\alpha q|w_n''|^2 dx - \frac{\alpha}{2}q(0)|w_n''(\alpha^-)|^2 \to 0. \tag{2.36}$$

Similarly, we can multiply (2.32) by $(L-x)\overline{w_n'}$ and integrate on (β, L) to get

$$\frac{1}{2}\int_\beta^L \rho|\lambda_n w_n|^2 dx + \frac{3}{2}\int_\beta^L q|w_n''|^2 dx - \frac{1}{2}(L-\beta)q(L)|w_n''(\beta^+)|^2 \to 0. \tag{2.37}$$

Finally, we subtract (2.37) from (2.36) and use (2.31) to obtain (2.24). ∎

In what follows, we will show that

$$|w''(\alpha^-)|^2 + |w''(\beta^+)|^2 \to 0 \tag{2.38}$$

to get a contradiction to (2.24). Denote by

$$\phi_n = \omega_n\left(\frac{\rho}{q}\right)^{\frac{1}{4}}, \quad F_n = \frac{\rho}{q}(g_n + \lambda_n f_n), \quad D = \frac{d}{dx}.$$

Then (2.32) can be rewritten as

$$(D - i\phi_n)(D + i\phi_n)(D^2 - \phi_n^2)w_n = F_n. \tag{2.39}$$

On the interval $(0, \alpha)$, by solving the first-order linear equation, we have

$$(D + i\phi_n)(D^2 - \phi_n^2)w_n = C_1 e^{i\phi_n(x-\alpha)} + \int_\alpha^x e^{i\phi_n(x-s)}F_n(s)ds; \tag{2.40}$$

$$\begin{aligned}(D^2 - \phi_n^2)w_n &= C_2 e^{-i\phi_n(x-\alpha)} + \frac{C_1}{\phi_n}\sin\phi_n(x-\alpha)\\ &\quad + \int_\alpha^x \frac{1}{\phi_n}\sin\phi_n(x-s)F_n(s)ds,\end{aligned} \tag{2.41}$$

where

$$C_1 = w_n'''(\alpha^-) + i\phi_n w_n''(\alpha^-) - \phi_n^2 w_n'(\alpha) - i\phi_n^3 w_n(\alpha), \tag{2.42}$$

$$C_2 = w_n''(\alpha^-) - \phi_n^2 w_n(\alpha); \tag{2.43}$$

and

$$
\begin{aligned}
(D-\phi_n)w_n = &\frac{C_2}{\phi_n(1-i)}\left[e^{-i\phi_n(x-\alpha)}-e^{-\phi_n(x-i\alpha)}\right]\\
&+\frac{C_1}{2\phi_n^2}\left[\sin\phi_n(x-\alpha)-\cos\phi_n(x-\alpha)\right]+\frac{C_1}{2\phi_n^2}e^{-\phi_n x}(\sin\phi_n\alpha+\cos\phi_n\alpha)\\
&+\int_0^x\int_\alpha^\tau e^{-\phi_n(x-\tau)}\frac{1}{\phi_n}\sin\phi_n(\tau-s)F_n(s)dsd\tau. \qquad (2.44)
\end{aligned}
$$

To obtain (2.44), we have used the boundary conditions $w_n(0)=w_n'(0)=0$. Multiplying (2.44) by $2\phi_n$ and taking $x=\alpha$, we have

$$
\begin{aligned}
2\phi_n w_n'(\alpha)-2\phi_n^2 w_n(\alpha) &= (1+i)C_2(1-e^{-\phi_n\alpha}e^{i\phi_n\alpha})-\frac{C_1}{\phi_n}\\
&+\frac{e^{-\phi_n\alpha}}{\phi_n}C_1(\sin\phi_n\alpha+\cos\phi_n\alpha)+2\int_0^\alpha\int_\alpha^\tau e^{-\phi_n(\alpha-\tau)}\sin\phi_n(\tau-s)F_n(s)dsd\tau. \qquad (2.45)
\end{aligned}
$$

We substitute (2.42) and (2.43) into (2.45) and let $n\to\infty$. By the results in Lemma 2.4, (2.45) yields

$$
\lim_{n\to\infty} w_n''(\alpha^-) = -2\lim_{n\to\infty}\int_0^\alpha\int_\alpha^\tau e^{-\phi_n(\alpha-\tau)}\sin\phi_n(\tau-s)F_n(s)dsd\tau. \qquad (2.46)
$$

We argue that the above limit is zero by the following estimates:

$$
\begin{aligned}
\left|\int_0^\alpha\int_\alpha^\tau e^{-\phi_n(\alpha-\tau)}\sin\phi_n(\tau-s)g_n(s)dsd\tau\right| &\le \int_0^\alpha\int_0^\alpha |g_n(s)|dsd\tau\\
&\le \alpha^{\frac{3}{2}}\left(\int_0^\alpha |g_n(s)|^2ds\right)^{\frac{1}{2}}\to 0 \qquad (2.47)
\end{aligned}
$$

and

$$
\begin{aligned}
&\left|\int_0^\alpha\int_\alpha^\tau e^{-\phi_n(\alpha-\tau)}\sin\phi_n(\tau-s)\lambda_n f_n(s)dsd\tau\right|\\
&=\left|\lambda_n\int_0^\alpha\left(\int_0^s e^{-\phi_n(\alpha-\tau)}\sin\phi_n(\tau-s)d\tau\right)f_n(s)ds\right|\\
&=\left|\frac{\lambda_n e^{-\phi_n\alpha}}{2\phi_n}\int_0^\alpha(\cos\phi_n s+\sin\phi_n s-e^{\phi_n s})f_n(s)ds\right|\\
&\le\left(\frac{q}{\rho}\right)^{\frac{1}{2}}\left(\alpha\phi_n e^{-\phi_n\alpha}+\frac{1}{2}-\frac{1}{2}e^{-\phi_n\alpha}\right)\max_{s\in[0,\alpha]}|f_n(s)|\to 0, \qquad (2.48)
\end{aligned}
$$

where we have used the fact that $g_n\to 0$ in $L^2(0,L)$, $f_n\to 0$ in $V\hookrightarrow C^1[0,L]$, and $\phi_n\to+\infty$. Thus we have proved

$$
w_n''(\alpha^-)\to 0 \quad \text{as} \quad n\to\infty. \qquad (2.49)
$$

Let $x'=L-x$, which maps the interval (β,L) onto $(0,L-\beta)$. Repeating the argument after (2.39), we can also prove

$$
w_n''(\beta^+)\to 0 \quad \text{as} \quad n\to\infty. \qquad (2.50)
$$

Equations (2.49) and (2.50) give us the promised contradiction. Thus, the energy of (1.2), $\mathcal{E}_1(t)=\frac{1}{2}\|(w(t),\dot w(t))\|_{\mathcal{H}_1}$, satisfies the inequality (1.3). □

3. Longitudinal motion. In this section, we study the longitudinal motion of the beam described in (1.1). Note that this equation also models the longitudinal vibration of an elastic rod.

Let $H = L^2_\rho(0, L)$ with the norm

$$\|v\| = \left(\int_0^L \rho |v(x)|^2 dx \right)^{\frac{1}{2}}$$

and $V = H^1_0(0, L)$ with the norm

$$\|v\|_V = \left(\int_0^L p |v'(x)| dx \right)^{\frac{1}{2}}.$$

Define $\mathcal{H}_2 = V \times H$ with the norm $\|(u, v)\|_{\mathcal{H}_2} = (\|u\|^2_V + \|v\|^2)^{\frac{1}{2}}$. Then $\mathcal{H}_2$ is a Hilbert space—the finite energy state space. Define in $\mathcal{H}_2$

$$D(\mathcal{A}_2) = \left\{ (u, v) \mid u, v \in V,\ S \equiv pu' + D_a v' \in H^1(0, L) \right\} \tag{3.1}$$

and

$$\mathcal{A}_2(u, v) = \left(v, \frac{1}{\rho} S' \right). \tag{3.2}$$

Thus (1.1) can be rewritten as an abstract evolution equation on $\mathcal{H}_2$:

$$(\dot{u}(t), \dot{v}(t)) = \mathcal{A}_2(u(t), v(t)), \quad (u(0), v(0)) = (u_0, u_1). \tag{3.3}$$

It is known that $\mathcal{A}_2$ generates a C_0-semigroup of contractions on $\mathcal{H}_2$ (see [CLL]). Therefore, $(u(t), \dot{u}(t)) = e^{\mathcal{A}_2 t}(u_0, u_1)$ gives the mild solution of (1.1) for every $(u_0, u_1) \in \mathcal{H}_2$. Moreover, $\mathcal{A}_2^{-1}$ is a bounded operator on $\mathcal{H}_2$.

We assume that the beam is homogeneous on the segments $(0, \alpha), (\alpha, \beta), (\beta, L)$, i.e., the coefficients in (1.1) are

$$\begin{cases} \rho = \rho_1 + (\rho_2 - \rho_1)\chi_{(\alpha,\beta)}, \\ p = p_1 + (p_2 - p_1)\chi_{(\alpha,\beta)}, \\ D_a = a\chi_{(\alpha,\beta)}, \quad \text{damping coefficient}, \end{cases} \tag{3.4}$$

with $\rho_1, \rho_2, p_1, p_2, a$ being positive constants. It is known that $e^{\mathcal{A}_2 t}$ is strongly stable (see [CLL]).

THEOREM 3.1. *The semigroup $e^{\mathcal{A}_2 t}$ for* (1.1) *with coefficients in* (3.4) *is not exponentially stable.*

Proof. We will show that $\|(\lambda - \mathcal{A}_2)^{-1}\|$ is unbounded on the imaginary axis. Let $\gamma_1^2 = \sqrt{\frac{p_1}{\rho_1}}$ and $\lambda = \lambda_n = i\omega_n = i\frac{2n\pi\gamma_1}{L-\beta}$, $n = 1, 2, \ldots$. Define

$$f = f(x, n) = \begin{cases} 0 & \text{in } (0, \beta), \\ \frac{1}{\omega_n} \sin \frac{\omega_n (x-\beta)}{\gamma_1} & \text{in } (\beta, L), \end{cases} \in V,$$

$$g = g(x, n) = \begin{cases} 0 & \text{in } (0, \beta), \\ \cos \frac{\omega_n (x-\beta)}{\gamma_1} & \text{in } (\beta, L), \end{cases} \in H.$$

We solve the resolvent equation $(\lambda-\mathcal{A}_2)(u,v)=(f,g),\ (u,v)\in D(\mathcal{A}_2)$ in the intervals $(0,\alpha),(\alpha,\beta),(\beta,L)$, respectively.

For $x\in(0,\alpha)$, we have

$$\begin{cases} \lambda u - v = 0, \\ \lambda v - \gamma_1^2 u'' = 0, \\ u(0) = 0. \end{cases} \tag{3.5}$$

It is easy to see that

$$u(x) = c_1 \sin \frac{\omega_n}{\gamma_1} x. \tag{3.6}$$

For $x\in(\alpha,\beta)$, we have

$$\begin{cases} \lambda u - v = 0, \\ \lambda v - (\gamma_2^2 + a_2\lambda)u'' = 0, \end{cases} \tag{3.7}$$

where $\gamma_2^2 = \sqrt{\frac{p_2}{\rho_2}}$ and $a_2 = \frac{a}{\rho_2}$. The solution of (3.7) is

$$u(x) = c_2 e^{\mu x} + c_3 e^{-\mu x} \tag{3.8}$$

with

$$\mu = \frac{\omega_n}{(\gamma_2^4 + a_2^2\omega_n^2)^{\frac{1}{4}}} \left(\cos\frac{\theta}{2} + i\sin\frac{\theta}{2}\right), \tag{3.9}$$

$$\cos\theta = \frac{-\gamma_2^2}{(\gamma_2^4 + a_2^2\omega_n^2)^{\frac{1}{2}}} \to 0, \quad \sin\theta = \frac{a_2\omega_n}{(\gamma_2^4 + a_2^2\omega_n^2)^{\frac{1}{2}}} \to 1 \quad \text{as } n\to\infty. \tag{3.10}$$

By the continuity conditions at $x=\alpha$, i.e.,

$$\begin{cases} u(\alpha^-) = u(\alpha^+), \\ \gamma_1^2 u'(\alpha^-) = (\gamma_2^2 + ia_2\omega_n)u'(\alpha^+), \end{cases} \tag{3.11}$$

we solve c_2, c_3 to get

$$u(x) = c_1 \left[\sin\frac{\omega_n\alpha}{\gamma_1}\cosh\mu(x-\alpha) + \frac{\gamma_1\omega_n}{(\gamma_2^2 + ia_2\omega_n)\mu}\cos\frac{\omega_n\alpha}{\gamma_1}\sinh\mu(x-\alpha)\right]. \tag{3.12}$$

Therefore,

$$u(\beta^-) = c_1 \left[\sin\frac{\omega_n\alpha}{\gamma_1}\cosh\mu(\beta-\alpha) + \frac{\gamma_1\omega_n}{(\gamma_2^2 + ia_2\omega_n)\mu}\cos\frac{\omega_n\alpha}{\gamma_1}\sinh\mu(\beta-\alpha)\right], \tag{3.13}$$

$$u'(\beta^-) = c_1 \left[\mu\sin\frac{\omega_n\alpha}{\gamma_1}\sinh\mu(\beta-\alpha) + \frac{\gamma_1\omega_n}{(\gamma_2^2 + ia_2\omega_n)}\cos\frac{\omega_n\alpha}{\gamma_1}\cosh\mu(\beta-\alpha)\right]. \tag{3.14}$$

For $x\in(\beta,L)$, we have

$$\begin{cases} \lambda u - v = f, \\ \lambda v - \gamma_1^2 u'' = g, \\ u(L) = 0. \end{cases} \tag{3.15}$$

Let

$$w_{\pm}(x) = \frac{1}{2}[v(x) \pm \gamma_1 u'(x)].$$

Then (3.15) can be transformed into a first-order, diagonal, nonhomogeneous system:

$$\begin{pmatrix} w_+ \\ w_- \end{pmatrix}' = \begin{pmatrix} \frac{\lambda}{\gamma_1} & 0 \\ 0 & -\frac{\lambda}{\gamma_1} \end{pmatrix} \begin{pmatrix} w_+ \\ w_- \end{pmatrix} + \begin{pmatrix} 0 \\ \frac{1}{\gamma_1} \end{pmatrix} g \equiv A \begin{pmatrix} w_+ \\ w_- \end{pmatrix} + \begin{pmatrix} 0 \\ \frac{1}{\gamma_1} \end{pmatrix} g. \tag{3.16}$$

Using the boundary condition $0 = v(L) = w_+(L) + w_-(L)$, we obtain the solution

$$\begin{pmatrix} w_+(x) \\ w_-(x) \end{pmatrix} = w_+(L) e^{(x-L)A} \begin{pmatrix} -1 \\ 1 \end{pmatrix} + \int_x^L e^{(x-\tau)A} \begin{pmatrix} \frac{1}{\gamma_1} \\ 0 \end{pmatrix} g(\tau) d\tau. \tag{3.17}$$

Since

$$\begin{aligned} v(x) &= w_+(x) + w_-(x) \\ &= w_+(L) \left[-e^{i\frac{2n\pi(x-L)}{L-\beta}} + e^{-i\frac{2n\pi(x-L)}{L-\beta}} \right] - \frac{1}{\gamma_1} \int_x^L e^{i\frac{2n\pi(x-\tau)}{L-\beta}} g(\tau) d\tau \\ &= -2iw_+(L) \sin \frac{2n\pi(x-L)}{L-\beta} - \frac{1}{\gamma_1} \int_x^L e^{i\frac{2n\pi(x-\tau)}{L-\beta}} \cos \frac{2n\pi(x-\beta)}{L-\beta} d\tau, \end{aligned} \tag{3.18}$$

we obtain that

$$v(\beta^+) = -\frac{L-\beta}{2\gamma_1}. \tag{3.19}$$

Furthermore, from $\lambda u(\beta^+) = f(\beta^+) + v(\beta^+) = v(\beta^+)$,

$$u(\beta^+) = i\frac{L-\beta}{2\gamma_1\omega_n}. \tag{3.20}$$

Similarly, from $\gamma_1 u'(x) = w_+(x) - w_-(x)$, we can also get

$$\gamma_1^2 u'(\beta^+) = -2\gamma_1 w_+(L) - \frac{1}{2}(L-\beta). \tag{3.21}$$

Applying the continuity conditions at $x = \beta$, i.e.,

$$\begin{cases} u(\beta^+) = u(\beta^-), \\ \gamma_1^2 u'(\beta^+) = (\gamma_2^2 + ia_2\omega_n) u'(\beta^-), \end{cases} \tag{3.22}$$

from (3.13), (3.14), (3.20), and (3.21), we can solve c_1 and obtain

$$\begin{aligned} &-2\gamma_1 w_+(L) - \frac{1}{2}(L-\beta) \\ &= i(L-\beta)\mu \frac{\gamma_2^2 + ia_2\omega_n}{2\gamma_1\omega_n} \cdot \frac{(\gamma_2^2 + ia_2\omega_n)\mu \sin \frac{\omega_n\alpha}{\gamma_1} + \gamma_1\omega_n \cos \frac{\omega_n\alpha}{\gamma_1} \coth \mu(\beta-\alpha)}{(\gamma_2^2 + ia_2\omega_n)\mu \sin \frac{\omega_n\alpha}{\gamma_1} \coth \mu(\beta-\alpha) + \gamma_1\omega_n \cos \frac{\omega_n\alpha}{\gamma_1}}. \end{aligned} \tag{3.23}$$

It follows from the definition of ω_n and μ that

$$|\coth \mu(\beta-\alpha)| \to 1, \quad |\mu| \to \infty \ \text{ as } \ n \to \infty. \tag{3.24}$$

If $\frac{\omega_n\alpha}{n\pi\gamma_1} = \frac{2\alpha}{L-\beta}$ is a rational number, there exists a subsequence of $\{\omega_n\}$, still denoted by ω_n, such that $\sin\frac{\omega_n\alpha}{\gamma_1} \equiv 0$. If $\frac{2\alpha}{L-\beta}$ is an irrational number, there exists a subsequence of $\{\omega_n\}$, still denoted by ω_n, such that $\sin\frac{\omega_n\alpha}{\gamma_1} \geq \epsilon_0 > 0$. In both cases, (3.23)–(3.24) imply

$$|w_+(L)| \to \infty \text{ as } n \to \infty. \tag{3.25}$$

This further leads to

$$\begin{aligned}\int_\beta^L |v(x)|^2 dx &= \int_\beta^L |w_+(x) + w_-(x)|^2 dx \\ &\geq 4|w_+(L)|^2 \int_\beta^L \sin^2\frac{2n\pi(x-L)}{L-\beta} dx - \frac{(L-\beta)^2}{2\gamma_1} \\ &= 2(L-\beta)|w_+(L)|^2 - \frac{(L-\beta)^2}{2\gamma_1} \to \infty \quad \text{as } n \to \infty. \end{aligned} \tag{3.26}$$

Finally, by $\|(f,g)\|^2_{\mathcal{H}_2} = \rho_1(L-\beta)$ and

$$\begin{aligned}\|(i\omega_n - \mathcal{A}_2)^{-1}(f,g)\|^2_{\mathcal{H}_2} &= \|(u,v)\|^2_{\mathcal{H}_2} \\ &\geq \int_\beta^L |v(x)|^2 dx \to \infty \quad \text{as } n \to \infty \end{aligned} \tag{3.27}$$

we conclude that

$$\sup_{\omega\in\mathbf{R}} \|(i\omega - \mathcal{A}_2)^{-1}\| = \infty. \tag{3.28}$$

Thus, $e^{\mathcal{A}_2 t}$ is not exponentially stable. □

Since the energy of (1.1) is $\mathcal{E}_2(t) = \frac{1}{2}\|(u(t), \dot{u}(t))\|^2_{\mathcal{H}_2} = \frac{1}{2}\|e^{\mathcal{A}_2 t}(u_0, u_1)\|^2_{\mathcal{H}_2}$, inequality (1.3) fails to hold for $\mathcal{E}_2(t)$. We believe that the lack of exponential stability here is due to the discontinuities in the damping coefficient D_a and the high order of damping operator. Some waves outside the interval (α, β) are strongly reflected at α, β. We don't know whether the conclusion in Theorem 3.1 is still true if the D_a is smooth at α and β.

4. Lack of analyticity. It is known that when the Kelvin–Voigt damping is globally distributed over the beam (i.e., $\alpha = 0, \beta = L$ in (1.1) and (1.2)), the corresponding semigroup is analytic (e.g., [CLL]). In this section, we will demonstrate that this is not true when the Kelvin–Voigt damping is only distributed locally. From Theorem 3.1, we already know that $e^{\mathcal{A}_2 t}$ is not analytic.

THEOREM 4.1. *$e^{\mathcal{A}_1 t}$ is not an analytic semigroup.*

Proof. From the analytic semigroup theory [Pa, Theorem 2.5.2], an exponential semigroup $e^{\mathcal{A}_1 t}$ is analytic if and only if

$$\sup\{\|\omega(i\omega - \mathcal{A}_1)^{-1}\| \mid \omega \in \mathbf{R}\} < +\infty. \tag{4.1}$$

We will show that (4.1) is not true. Choose $\gamma > 0$ such that $a = \gamma/\pi < \alpha$. Let

$$x_n = \frac{\pi}{4\gamma n}, \quad \widetilde{x}_n = \frac{\pi}{\gamma} - \frac{\pi}{4\gamma n}, \quad \phi_n(x) = \sin\gamma n x, \tag{4.2}$$

and construct a sequence of functions

$$u_n(x) = \begin{cases} \xi_n(x), & 0 \le x < x_n, \\ \phi_n(x), & x_n \le x < \widetilde{x}_n, \\ \eta_n(x), & \widetilde{x}_n \le x < a, \\ 0, & a \le x \le L, \end{cases} \tag{4.3}$$

where

$$\xi_n(x) = x^4 \sum_{k=0}^{3} b_{n,k} \frac{x^k}{k!}, \quad \eta_n(x) = (x-a)^4 \sum_{k=0}^{3} c_{n,k} \frac{(x-a)^k}{k!}. \tag{4.4}$$

The coefficients $b_{n,k}, c_{n,k}$ are uniquely determined by the smoothly connected conditions

$$\xi_n^{(k)}(x_n) = \phi_n^{(k)}(x_n), \quad \eta_n^{(k)}(\widetilde{x}_n) = \phi_n^{(k)}(\widetilde{x}_n), \quad k = 0, 1, 2, 3. \tag{4.5}$$

Since

$$|\phi_n^{(k)}(x_n)| = |\phi_n^{(k)}(\widetilde{x}_n)| = \frac{\sqrt{2}}{2}(\gamma n)^k, \tag{4.6}$$

we can directly verify that

$$b_{n,k} = \frac{d^k}{dx^k}\left(\frac{\xi_n(x)}{x^4}\right)\Big|_{x=x_n} = \mathcal{O}(n^{4+k}), \qquad (n \to \infty), \tag{4.7}$$

$$c_{n,k} = \frac{d^k}{dx^k}\left(\frac{\eta_n(x)}{(x-a)^4}\right)\Big|_{x=\widetilde{x}_n} = \mathcal{O}(n^{4+k}), \qquad (n \to \infty), \tag{4.8}$$

for $k = 0, 1, 2, 3$. This further leads to

$$\xi_n^{(k)}(x) = \mathcal{O}(n^k) \quad \forall\, x \in [0, x_n] \quad (n \to \infty), \tag{4.9}$$

$$\eta_n^{(k)}(x) = \mathcal{O}(n^k) \quad \forall\, x \in [\widetilde{x}_n, a] \quad (n \to \infty), \tag{4.10}$$

for $k = 0, 1, 2, 3, 4$. It is easy to see that $u_n \in H_0^4(0, L)$ and supp $u_n \subset (0, \alpha)$. Now, let

$$\omega_n = \sqrt{\frac{q(0)}{\rho(0)}}(\gamma n)^2, \quad v_n = i\omega_n u_n. \tag{4.11}$$

Then $(u_n, v_n) \in D(\mathcal{A}_1)$ and

$$(i\omega_n - \mathcal{A}_1)(u_n, v_n) = \left(0, -\omega_n^2 u_n + \frac{q(0)}{\rho(0)} u_n^{(4)}\right). \tag{4.12}$$

Since

$$-\omega_n^2 u_n + \frac{q(0)}{\rho(0)} u_n^{(4)} = \begin{cases} -\omega_n^2 \xi_n + \frac{q(0)}{\rho(0)} \xi_n^{(4)}, & 0 \le x < x_n, \\ 0, & x_n \le x < \widetilde{x}_n, \\ -\omega_n^2 \eta_n + \frac{q(0)}{\rho(0)} \eta_n^{(4)}, & \widetilde{x}_n \le x < a, \\ 0, & a \le x \le L, \end{cases} \tag{4.13}$$

by (4.9) and (4.10) we have

$$\left\| \frac{1}{\omega_n^2}(i\omega_n - \mathcal{A}_1)(u_n, v_n) \right\|_{\mathcal{H}_2} \to 0. \tag{4.14}$$

On the other hand,

$$\begin{aligned} \left\| \frac{1}{\omega_n}(u_n, v_n) \right\|_{\mathcal{H}_2} &\geq \left(\int_0^{\alpha} \rho |u_n|^2 dx \right)^{\frac{1}{2}} \\ &\geq \left(\int_{x_n}^{\widetilde{x}_n} \rho |\phi_n|^2 dx \right)^{\frac{1}{2}} \\ &= \left(\int_{x_n}^{\widetilde{x}_n} \rho \sin^2 \gamma n x dx \right)^{\frac{1}{2}} \to \left(\frac{\rho \pi}{2\gamma} \right)^{\frac{1}{2}}. \end{aligned} \tag{4.15}$$

Equations (4.14) and (4.15) imply that (4.1) is not true. □

Remark. This proof does not depend on the boundary conditions or the type of damping. Thus, we conclude that the semigroups associated with linear beam equations with local damping are never analytic.

REFERENCES

[A] R. A. Adams, *Sobolev Space*, Academic Press, New York, 1975.

[BSW] H. T. Banks, R. C. Smith, and Y. Wang, *Modeling aspects for piezoelectirc patch activation of shells, plates and beams*, Quart. Appl. Math., LIII (1995), pp. 353–381.

[CFNS] G. Chen, S. A. Fulling, F. J. Narcowich, and S. Sun, *Exponential decay of energy of evolution equation with locally distributed damping*, SIAM J. Appl. Math., 51 (1991), pp. 226–301.

[CLL] S. Chen, K. Liu, and Z. Liu, *Spectrum character and stability for the elastic systems with global/local Kelvin–Voigt damping*, SIAM J. Appl. Math., to appear.

[Ge] L. M. Gearhart, *Spectral theory for contraction semigroups on Hilbert space*, Trans. Amer. Math. Soc., 236 (1978), pp. 385–394.

[Hu] F. L. Huang, *Characteristic condition for exponential stability of linear dynamical systems in Hilbert spaces*, Ann. Differential Equations, 1 (1985), pp. 43–56.

[K1] J. U. Kim, *Exponential decay of the energy of a one-dimensional nonhomogeneous medium*, SIAM J. Control Optim., 292 (1991), pp. 368–380.

[K2] J. U. Kim, *Exact internal controllability of a one-dimensional aeroelastic plate*, Appl. Math. Optim., 24 (1991), pp. 99–111.

[La] J. Lagnese, *Control of wave process with distributed controls supported on a subregion*, SIAM J. Control Optim., 21 (1983), pp. 68–85.

[Li] K. S. Liu, *Locally distributed control and damping for the conservative systems*, SIAM J. Control Optim., 35 (1997), pp. 1574–1590.

[Pa] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.

[Pr] J. Prüss, *On the spectrum of C_0-semigroups*, Trans. Amer. Math. Soc., 284 (1984), pp. 847–857.

[R] M. Renardy, *On the type of certain C_0-semigroups*, Comm. Partial Differential Equations, 18 (1993), pp. 1299–1307.

[Zu1] E. Zuazua, *Exponential decay for the semilinear wave equation with localized damping*, Comm. Partial Differential Equations, 15 (1990), pp. 205–235.

[Zu2] E. Zuazua, *Exponential decay for the semilinear wave equation with localized damping in unbounded domains*, J. Math. Pures Appl., 70 (1991), pp. 513–529.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 3, pp. 1099–1132, May 1998

016

OPTIMAL CONTROL OF INFLATION: A CENTRAL BANK PROBLEM*

MARIA B. CHIAROLLA† AND ULRICH G. HAUSSMANN‡

Abstract. This paper models the action of the central bank on the dynamics of the nominal interest rate with the aim of controlling inflation. The problem is set up as a two-dimensional *bounded variation control problem*; it is shown that its variational formulation leads to a stochastic differential game with stopping times between the conservative and the expansionist tendencies of the bank.

Key words. central bank, inflation, bounded variation stochastic control, weak variational inequality, strong variational inequality, stochastic differential game

AMS subject classifications. 90A70, 93E20, 93E05, 49J40, 35K85

PII. S036301299630495X

1. Introduction.

The central bank has many roles to play in an economy. These vary from country to country depending on federal law and government policy. For example, in 1975, the governor of the Bank of Canada announced policies to deal with inflation which can be summarized as follows (cf. Binhammer [4, pp. 586–587]).

- The Bank of Canada has the responsibility and the means to keep the rate of monetary expansion under control.
- The restoration of price stability requires an average rate of growth of the money supply no higher than the long term average rate of growth of production of goods and services.
- The pursuit of a policy of stable monetary expansion requires that nominal interest rates and the foreign exchange rate be allowed to achieve their levels independently.
- Nominal interest rates can be reduced in the long run by reducing inflation. This is to be achieved by a short term rise in interest rates, producing a decline in the growth rate of the money supply. Over the long term interest rates should settle to lower levels.

As noted above, in Canada, the bank controls the demand for money by influencing interest rates through intervention in the weekly auction of 91-day T-bills and by trading in the secondary market for T-bills. The bank rate is then set at 1/4% above the T-bill rate set at the weekly auction. Reserves are provided to the commercial banks to support this action; hence the monetary base is only controlled passively. The bank followed this policy with the exception of the independence of the exchange rate and interest rates. In fact, the short term swings in the bank rate are mostly due to a policy of defending the Canadian dollar against the U.S. dollar. Nevertheless, this situation is somewhat particular to Canada, and so in this work we shall ignore the influence of the exchange rate on interest rates. The general conclusion to be drawn from the above policy statements is that the bank wants to control inflation

*Received by the editors June 10, 1996; accepted for publication (in revised form) April 14, 1997. This work was supported by Natural Sciences and Engineering Research Council of Canada grant 88051.

http://www.siam.org/journals/sicon/36-3/30495.html

†Istituto di Matematica Finanziaria, Università degli Studi di Bari, via Camillo Rosalba 53, 70124 Bari, Italy (albano@tno.it).

‡Department of Mathematics, University of British Columbia, 1984 Mathematics Road, Vancouver, BC, Canada V6T 1Z2 (uhaus@math.ubc.ca).

by raising nominal interest rates using the bank rate. This policy seems to have been successful during the 1980s and 1990s.

An essential issue at this point is to model the dynamic evolution of the interest and inflation rates over time. Interest rates have received widespread attention over the past decade, usually in the context of the term structure of interest rates; cf. Longstaff and Schwartz [20] and references therein. In general, however, the problem of studying the changes of interest rates due to monetary decisions is quite new in the literature, although several authors have pointed out the relationship between interest rates and inflation. For example, Richard determines the term structure of interest rate in a model whose uncorrelated state variables are the real interest rate and the inflation rate in [24]. The inflation rate appears also in a paper by Cox, Ingersoll, and Ross [8, section 7], but there is a variable without any monetary objective. The effect of monetary policy changes on the relationship between short and long term interest rates is studied by Turnovsky [25], and an equilibrium asset-pricing model in which the real interest rate is negatively correlated with the inflation rate is constructed and estimated by Pennacchi [23]. General observations about correlations among interest rates, inflation, and money supply growth can be found in the work of Mishkin [22].

The recent paper of Fusai [12] analyzes how the term structure of interest rates changes when the central bank tries to stabilize the inflation rate by acting on the money supply. This is an interesting approach, although there may be some mathematical difficulties in the derivation of the model. The problem is set as a stochastic linear regulator, hence a *classical* control problem; the control is given by the money supply and enters the dynamics of the real interest rate and of the inflation rate.

Our model has the following features. The two-dimensional state variable is given by the nominal interest rate $X_{1,s}$ (deduced from the stochastic Fisher law) and the inflation rate $X_{2,s}$; these processes are jointly dependent. This is consistent with some (but not all) models in the literature, e.g., Pennachi [23]. The central bank may modify the dynamics of $X_{1,s}$ by adding a stochastic process $k \in V[0,T]$ of finite variation over the time interval $[0,T]$; the change in k represents the change in the bank rate which is assumed to affect the interest rate directly and additively. Inflation targeting is achieved by minimizing over $V[0,T]$ a certain cost functional. This produces a two-dimensional *singular* stochastic control problem with the control (which, as a function of time, could possibly be singular with respect to the Lebesgue measure) acting only on the x_1-component. We show that the variational formulation of this problem leads to a two-player, zero-sum stochastic differential game with stopping times whose value coincides with the x_1-derivative of the value function of the original problem. The differential game might be interpreted as a game played between the conservative and the expansionist tendencies of the bank. At certain times the conservative tendency asks for low inflation and hence interest rates are increased, whereas at other opportune times the expansionist tendency lowers interest rates in order to stimulate the economy.

Our mathematical model falls in the class of the so-called *bounded variation follower problems with finite horizon*, in the language of the recent singular stochastic control literature. A d-dimensional follower problem for the control of a diffusion process with linear-in-x drift and constant-in-x diffusion coefficient has been sketched by Menaldi and Robin [21]; there compactness methods are used to establish the existence of an optimal control process. More general existence results were established by Haussmann and Suo [15]. On the other hand, the one-dimensional *reflected* follower problem (i.e., a bounded variation follower problem with a reflecting barrier at the origin) for the control of Brownian motion has been extensively studied by El Karoui

and Karatzas [9], Karatzas [16], and Karatzas and Shreve [17], among others. It has been shown that the x-derivative of the value function of the reflected follower problem is the optimal risk of an optimal stopping problem with absorption at the origin. Therefore, our differential game problem seems to be the natural generalization to more than one dimension (but with one-dimensional control) of the stopping problem arising in dimension one. The cost rate of our game depends on the x_2-derivative of the value function of the original problem, i.e., the partial derivative in the direction not controlled, so the game cannot be solved without solving the original problem. It is only in the one-dimensional case where there are no "other" directions, that the derived problem can be stated independently of the original value function.

This paper is organized as follows. In section 2 we formulate the control problem and state the main results. In particular, we find that the value function (derived utility function), $v(x,t)$, satisfies a variational inequality (corresponding to the Bellman equation of dynamic programming), but that an associated problem satisfied by $v_{x_1}(x,t)$ is more relevant than the original variational inequality. Moreover, a stochastic differential game with stopping times is deduced from the associated problem. It is shown that the game has value v_{x_1} and admits a saddle point $(\hat{\theta}_1, \hat{\theta}_2)$, where $\hat{\theta}_1$ is the optimal time to increase interest rates in order to contain the inflation and $\hat{\theta}_2$ is the optimal time to lower interest rates in order to stimulate the economy.

Sections 3, 4, and 5 contain the mathematical analysis of the problem. The nonmathematical reader may skip to section 6. In section 3 we derive some properties of the value function $v(x,t)$. In section 4 we introduce a penalized control problem whose value function $v^\varepsilon(x,t)$ approximates $v(x,t)$ and is regular enough to allow differentiation with respect to x_1 of the corresponding Hamilton–Jacobi–Bellman equation. Then, by taking limits as $\varepsilon \to 0$, in section 5 we obtain a weak variational inequality solved by v_{x_1}. (We point out that the Mignot–Puel method employed in the *monotone* control case to obtain a weak variational inequality (cf. [7, p. 875]) cannot be applied to problems with general bounded variation controls since the penalty term of the differentiated Hamilton–Jacobi–Bellman equation is not nonnegative, and hence it does not correspond to a penalization operator in the sense of Mignot and Puel.) We conclude section 5 by appealing to general existence and uniqueness results established in section 7, to show that v is the unique solution of a pointwise variational inequality. In section 6 we summarize our conclusions regarding the economics problem.

Section 7 stands alone and gives results about variational inequalities. Many of the mathematical techniques used here were inspired by the work of Bensoussan and Lions [2], but in fact our problem falls outside the scope of their work because of an unbounded domain, unbounded coefficients in the state equations, and lack of differentiability of v_{x_2} with respect to time. Using penalization we show that a strong variational inequality has a unique solution, which is related to the unique solution of a weak variational inequality (the one satisfied by $e^{-\lambda(T-t)}v_{x_1}$ in section 5). Finally we show that the solution in fact satisfies a pointwise variational inequality. The appendix contains some technical results used in section 7.

2. Formulation of the problem and results. Pennacchi [23] postulates an economy with a single capital-consumption good and a single technology to transform capital into output. One of the implications of the model is that if $X_{1,s}$, $X_{2,s}$ are the nominal (spot) interest rate and expected rate of inflation, respectively, then they satisfy

$$dX_s = (a + bX_s)ds + \sigma dW_{s-t}, \qquad s \in (t,T]. \tag{2.1}$$

This model does not take into consideration the direct influence of the central bank on the interest rate. It also allows these rates, in particular, the nominal interest rate, to go negative! Nevertheless, as a first step we shall adopt this model except that we add a control term which reflects the actions of the bank. In [14] it is shown that the model, when applied to Canadian data from 1983 through 1988, is not unreasonable; i.e., statistical tests checking for nonnormality of the residuals corresponding to ΔW are not significant. To be fair, we must add that for other time segments of the data this is not the case. If the inflation rate is ignored, i.e., $X_2 = 0$, then the model for the interest rate has been used in derivative pricing; cf. [5].

Let a be a constant vector in $\mathbb{R}^2$ and $e_1 = \binom{1}{0}$, let $b = \binom{b_{11}\, b_{12}}{b_{21}\, b_{22}}$ and σ be constant 2×2 matrices such that $\sigma\sigma^*$ is positive definite. Let $T > 0$ be fixed; then $X_s = \binom{X_{1,s}}{X_{2,s}}$ is the process starting at time $t \in [0, T]$ from $x \in \mathbb{R}^2$ and governed by the stochastic differential equation

$$\begin{cases} dX_s = (a + bX_s)ds + \sigma dW_{s-t} + e_1\, dk_{s-t}, & s \in (t, T], \\ X_t = x + e_1\, k_0 \end{cases} \tag{2.2}$$

on some filtered probability space $(\Omega, \mathcal{F}, \mathcal{F}_s, P)$ with the filtration $\{\mathcal{F}_s, s \in [0, T]\}$ satisfying the usual conditions, where $\{W_s, s \in [0, T]\}$ is a standard two-dimensional Brownian motion and the control $\{k_s, s \in [0, T]\}$ is a real-valued càdlàg (i.e., right-continuous with left limits), $\mathcal{F}_s$-adapted process, almost surely (a.s.) of finite variation. We denote by $V[0, T]$ the set of all such control processes; we refer to (x, t) as "the initial condition" and to T as "the terminal time."

Let $f(x) = \frac{1}{2}[\nu(x_1)^2 + (x_2)^2]$ for $x = (x_1, x_2)$; $\in \mathbb{R}^2$ and $\nu \geq 0$; then the function f is convex. Let $\rho > 0$ be a given discount factor; then the cost corresponding to the control process k is

$$J_{x,t}(k) = E\bigg\{ \int_t^T f(X_s) e^{-\rho(s-t)} ds + \int_t^T e^{-\rho(s-t)} d|k|_{s-t} + |k_0| \bigg\}, \tag{2.3}$$

where $|k|$ is the total variation process of k. The term involving X_1 reflects the desire to keep interest rates low so as to stimulate the economy; the term involving X_2 reflects the desire to keep the inflation rate at zero (we could have chosen any constant), and the terms involving k reflect the bank's reluctance to make large changes in the rate (i.e., it prefers to provide stable interest rates). The coefficient ν is determined by the bank to reflect the weight it wishes to give to fighting inflation compared to stimulating the economy.

The problem is to minimize J and to find the value function

$$v(x, t) = \inf \Big\{ J_{x,t}(k) : k \in V[0, T] \Big\}. \tag{P}$$

In order to state the results we need some notation. Let $p > 1$, $m \in \mathbb{N}$, and let Q denote an open set in $\mathbb{R}^2$; then we set

- $C^{2,1}(Q \times (0, T))$ = the set of all functions u continuous on $Q \times (0, T)$ with continuous partial derivatives $u_{x_i}, u_{x_i x_j}, u_t$, $i, j = 1, 2$;
- $C^{2,1}_{\text{pol}}(Q \times (0, T))$ = the set of all functions $u \in C^{2,1}(Q \times (0, T))$ which satisfy a polynomial growth condition on $Q \times (0, T)$; i.e., for some constants $C > 0$ and $m \in \mathbb{N}$,

$$|u(x, t)| \leq C(1 + |x|^m) \qquad \forall (x, t) \in Q \times (0, T);$$

- $W^{m;p}(Q)$ = the space of all functions g which have weak derivatives $D^\alpha g$ in $L^p(Q)$ for all $|\alpha| \le m$;
- $W^{2,1;p}(Q \times (0,T))$ = the space of all functions u which have weak derivatives $u_t, u_{x_i}, u_{x_i x_j}$, $i,j = 1,2$, in $L^p(Q \times (0,T))$;
- $W^{2,1;p}_{\text{loc}}(\mathbb{R}^2 \times (0,T))$ = the space of all functions u that, for all bounded $Q \subset \mathbb{R}^2$, belong to $W^{2,1;p}(Q \times (0,T))$;
- $\mathcal{L}u(x,t) = \frac{1}{2}\text{trace}[\sigma\sigma^* \mathcal{D}^2 u(x,t)] + (a+bx)\cdot \nabla u(x,t)$, where $\mathcal{D}^2 u(x,t)$ and $\nabla u(x,t)$ are the Hessian matrix and the gradient of $u(x,t)$ with respect to the x-variables, respectively.

The value function can now be characterized by the following theorem.

THEOREM 2.1. *The value function v is the unique solution in $W^{2,1;\infty}_{\text{loc}}(\mathbb{R}^2 \times (0,T))$ of*

$$\begin{cases} v(x,T) = 0 \text{ almost everywhere (a.e.) in } \mathbb{R}^2, \\ \max\Big\{ -v_t + [-\mathcal{L} + \rho]v - f,\ -v_{x_1} - 1,\ v_{x_1} - 1 \Big\} \le 0, \\ (-v_t + [-\mathcal{L} + \rho]v - f)(-v_{x_1} - 1)(v_{x_1} - 1) = 0, \\ \text{a.e. in } \mathbb{R}^2 \times (0,T). \end{cases} \tag{2.4}$$

A proof of this result can be given along the lines of the proof of the next theorem, but as it is similar, we refrain from doing so.

The problem (2.4) is difficult to solve and difficult to interpret in economic terms; an associated problem turns out to have a useful form. Its main actor is the function v_{x_1}.

THEOREM 2.2. *There exists a positive constant C such that*

$$|v_{x_i}(x,t)| \le C(1+|x|), \quad |v_{x_i x_j}(x,t)| \le C, \quad |v_t(x,t)| \le C(1+|x|^2) \tag{2.5}$$

a.s. in $\mathbb{R}^2 \times [0,T]$, $i,j = 1,2$. Assume that $\rho \ge b_{11}$ and set

$$\hat{u} = b_{21} v_{x_2} + \nu x_1.$$

Then the function v_{x_1} is the unique solution in $W^{2,1;6}_{\text{loc}}(\mathbb{R}^2 \times (0,T))$ of the pointwise variational inequality

$$\begin{cases} -1 \le v_{x_1} \le +1 \text{ a.e. in } \mathbb{R}^2 \times (0,T),\ v_{x_1}(x,T) = 0 \text{ a.e. in } \mathbb{R}^2, \\ -v_{x_1 t} + [-\mathcal{L} + (\rho - b_{11})]v_{x_1} = \hat{u} \quad \text{if } -1 < v_{x_1}(x,t) < 1, \\ -v_{x_1 t} + [-\mathcal{L} + (\rho - b_{11})]v_{x_1} \le \hat{u} \quad \text{if } v_{x_1}(x,t) = +1, \\ -v_{x_1 t} + [-\mathcal{L} + (\rho - b_{11})]v_{x_1} \ge \hat{u} \quad \text{if } v_{x_1}(x,t) = -1, \\ \text{a.e. in } \mathbb{R}^2 \times (0,T). \end{cases} \tag{2.6}$$

The proof of this result is technical and is given in the following three sections.

Note that the uniqueness here is for a fixed $\hat{u}$, i.e., for given v_{x_2}. In terms of the unknowns (v_{x_1}, v_{x_2}), this theorem gives existence of solutions, but not uniqueness. The more general uniqueness does not hold without further side conditions; what such conditions are poses an interesting problem. In fact, the above theorem is an intermediate result; what we are after is the interpretation of v_{x_1} as the value of a differential game!

We are now going to show that v_{x_1} is the value of a two-player, zero-sum stochastic differential game. In essence, we have moved from the control problem (P) to the variational inequality satisfied by its value function v, we have differentiated to find a variational inequality for v_{x_1}, and then we have found the corresponding control problem, which happens to be a differential game.

For every $(x,t) \in \mathbb{R}^2 \times [0,T]$ consider the control-free diffusion process starting at time t from x; i.e.,

$$X^0_{s'} = x + \int_t^{s'} (a + bX^0_r)dr + \sigma \int_t^{s'} dW_{r-t} \qquad \forall s' \in [t,T],$$

so if $Y^0_s := X^0_{s+t} = X^0_{s'}$ for $s = s' - t$ then

$$Y^0_s = x + \int_0^s (a + bY^0_r)dr + \sigma \int_0^s dW_r \qquad \forall s \in [0, T-t]. \tag{2.7}$$

For every r, r' with $0 \le r < r' < +\infty$ we set

- $\mathcal{S}_{r,r'}$ = the collection of all stopping times relative to the underlying filtration $\{\mathcal{F}_s\}_{s\in[0,T]}$ and taking values in $[r, r']$.

Let θ_1, θ_2 be stopping times in $\mathcal{S}_{0,T-t}$; if θ_1 and θ_2 play the role of strategies, then we define the *evaluation function* of the game by setting

$$\begin{aligned} G_{x,t}(\theta_1,\theta_2) = E\Big\{ &\int_0^{\theta_1\wedge\theta_2} \hat{u}(Y^0_s, t+s) e^{-(\rho-b_{11})s} ds \\ &- e^{-(\rho-b_{11})\theta_1} \mathbb{I}_{\substack{\theta_1\le\theta_2 \\ \theta_1<T-t}} + e^{-(\rho-b_{11})\theta_2} \mathbb{I}_{\theta_2<\theta_1} \Big\}. \end{aligned} \tag{2.8}$$

Also, we define

- $\hat{\theta}_1 = \inf\{s \in [0,T-t] : v_{x_1}(Y^0_s, t+s) = -1\} \wedge (T-t)$,
- $\hat{\theta}_2 = \inf\{s \in [0,T-t] : v_{x_1}(Y^0_s, t+s) = +1\} \wedge (T-t)$;

then we prove that the differential game has a solution.

THEOREM 2.3. *Assume that $\rho \ge b_{11}$. Then for every initial condition (x,t), the evaluation function $G_{x,t}$ has a saddle point at $(\hat{\theta}_1, \hat{\theta}_2)$ with value $v_{x_1}(x,t)$; i.e.,*

$$\begin{cases} G_{x,t}(\theta_1, \hat{\theta}_2) \le G_{x,t}(\hat{\theta}_1, \hat{\theta}_2) \le G_{x,t}(\hat{\theta}_1, \theta_2) \qquad \forall \theta_1, \theta_2 \in \mathcal{S}_{0,T-t}, \\ v_{x_1}(x,t) = G_{x,t}(\hat{\theta}_1, \hat{\theta}_2), \end{cases} \tag{2.9}$$

and

$$\begin{aligned} v_{x_1}(x,t) &= \inf_{\theta_2\in\mathcal{S}_{0,T-t}} \sup_{\theta_1\in\mathcal{S}_{0,T-t}} G_{x,t}(\theta_1,\theta_2) \\ &= \sup_{\theta_1\in\mathcal{S}_{0,T-t}} \inf_{\theta_2\in\mathcal{S}_{0,T-t}} G_{x,t}(\theta_1,\theta_2). \end{aligned} \tag{2.10}$$

Proof. Clearly (2.10) follows directly from (2.9). Fix (x,t) in $\mathbb{R}^2 \times [0,T]$; for $R > 0$ let τ_R be the first exit time of Y^0 from $B_R = \{x \in \mathbb{R}^2 : |x| < R\}$. Since v_{x_1} is in $W^{2,1;6}_{\mathrm{loc}}(\mathbb{R}^2 \times (0,T))$ (cf. Theorem 2.2), continuous on $\mathbb{R}^2 \times [0,T]$ by Sobolev imbedding, we may apply to $e^{-(\rho-b_{11})s} v_{x_1}(Y^0_s, t+s)$ a generalization of Ito's formula (cf. [2, Chapter 2, Theorem 8.5, p. 185]); hence for any stopping time $\theta \in \mathcal{S}_{0,T-t}$ we obtain

$$\begin{aligned} v_{x_1}(x,t) &= E\Big\{ e^{-(\rho-b_{11})\theta} v_{x_1}(Y^0_\theta, t+\theta) \mathbb{I}_{\theta\le\tau_R} \Big\} \\ &+ E\Big\{ e^{-(\rho-b_{11})\tau_R} v_{x_1}(Y^0_{\tau_R}, t+\tau_R) \mathbb{I}_{\theta>\tau_R} \Big\} \\ &+ E\Big\{ \int_0^{\theta\wedge\tau_R} e^{-(\rho-b_{11})r} \Big(-v_{x_1 t} + [-\mathcal{L} + (\rho - b_{11})] v_{x_1} \Big)(Y^0_r, t+r) dr \Big\}. \end{aligned} \tag{2.11}$$

Since $\tau_R \to +\infty$ a.s. as $R \to +\infty$ and $|v_{x_1}| \le 1$ (cf. (2.6)), we have

$$\begin{cases} \lim\limits_{R\to+\infty} E\Big\{e^{-(\rho-b_{11})\theta} v_{x_1}(Y^0_\theta, t+\theta)\mathbb{I}_{\theta\le\tau_R}\Big\} = E\Big\{e^{-(\rho-b_{11})\theta} v_{x_1}(Y^0_\theta, t+\theta)\Big\}, \\ \lim\limits_{R\to+\infty} E\Big\{e^{-(\rho-b_{11})\tau_R} v_{x_1}(Y^0_{\tau_R}, t+\tau_R)\mathbb{I}_{\theta>\tau_R}\Big\} = 0 \end{cases}$$

by the bounded convergence theorem.

Now, for any stopping time $\theta_2 \in \mathcal{S}_{0,T-t}$, we set

$$\theta = \theta_2 \wedge \hat\theta_1,$$

and we see that (2.6) implies

$$\Big(-v_{x_1 t} + [-\mathcal{L} + (\rho - b_{11})]v_{x_1}\Big)(Y^0_s, t+s) \le \hat u(Y^0_s, t+s) \qquad \forall s \in [0, \hat\theta_1]$$

with equality if $s \le \hat\theta_1 \wedge \hat\theta_2$; therefore, we have

$$\begin{aligned} E\Big\{\int_0^{\theta\wedge\tau_R} e^{-(\rho-b_{11})r}\Big(-v_{x_1 t} + [-\mathcal{L} + (\rho-b_{11})]v_{x_1}\Big)(Y^0_r, t+r)dr\Big\} \\ \le E\Big\{\int_0^{\theta\wedge\tau_R} e^{-(\rho-b_{11})r}\hat u(Y^0_r, t+r)dr\Big\} \end{aligned} \tag{2.12}$$

with equality if $\theta_2 \le \hat\theta_2$. But $|\hat u(Y^0_r, t+r)| \le C(1+|Y^0_r|)$ by $(2.5)_1$, so by applying the dominated convergence theorem we obtain

$$\begin{aligned} \lim_{R\to+\infty} E\Big\{\int_0^{\theta\wedge\tau_R} e^{-(\rho-b_{11})r}\Big(-v_{x_1 t} + [-\mathcal{L} + (\rho-b_{11})]v_{x_1}\Big)(Y^0_r, t+r)dr\Big\} \\ \le E\Big\{\int_0^{\theta} e^{-(\rho-b_{11})r}\hat u(Y^0_r, t+r)dr\Big\} \end{aligned}$$

with equality if $\theta_2 \le \hat\theta_2$. Thus, passing to the limit as $R \to +\infty$ in (2.11) gives

$$\begin{aligned} v_{x_1}(x,t) &\le E\Big\{e^{-(\rho-b_{11})\theta} v_{x_1}(Y^0_\theta, t+\theta)\Big\} \\ &\quad + E\Big\{\int_0^{\theta} e^{-(\rho-b_{11})r}\hat u(Y^0_r, t+r)dr\Big\} \end{aligned} \tag{2.13}$$

with equality if $\theta_2 \le \hat\theta_2$. Finally, $v_{x_1}(Y^0_{T-t}, T) = 0$, $v_{x_1}(Y^0_{\hat\theta_1}, t+\hat\theta_1) = -1$ (this follows from the continuity of v_{x_1} and Y^0), and $v_{x_1} \le 1$ imply

$$\begin{aligned} &E\Big\{e^{-(\rho-b_{11})\theta} v_{x_1}(Y^0_\theta, t+\theta)\Big\} \\ &= E\Big\{-e^{-(\rho-b_{11})\hat\theta_1}\mathbb{I}_{\substack{\theta_2\ge\hat\theta_1 \\ \hat\theta_1<(T-t)}} + e^{-(\rho-b_{11})\theta_2} v_{x_1}(Y^0_{\theta_2}, t+\theta_2)\mathbb{I}_{\theta_2<\hat\theta_1}\Big\} \\ &\le E\Big\{-e^{-(\rho-b_{11})\hat\theta_1}\mathbb{I}_{\substack{\theta_2\ge\hat\theta_1 \\ \hat\theta_1<(T-t)}} + e^{-(\rho-b_{11})\theta_2}\mathbb{I}_{\theta_2<\hat\theta_1}\Big\} \end{aligned} \tag{2.14}$$

with equality if $\theta_2 = \hat\theta_2$. Now from (2.13) and (2.14) it follows that

$$v_{x_1}(x,t) = G_{x,t}(\hat\theta_1, \hat\theta_2) \le G_{x,t}(\hat\theta_1, \theta_2) \qquad \forall \theta_2 \in \mathcal{S}_{0,T-t}.$$

Similarly, by taking $\theta = \theta_1 \wedge \hat{\theta}_2$ for any stopping time $\theta_1 \in \mathcal{S}_{0,T-t}$, we can show that $G_{x,t}(\theta_1, \hat{\theta}_2) \le G_{x,t}(\hat{\theta}_1, \hat{\theta}_2) = v_{x_1}(x,t) \, \forall \theta_1 \in \mathcal{S}_{0,T-t}$. □

The financial interpretation of Theorem 2.3 discloses the complexity of the problem of containing inflation. In fact, (2.10) can be interpreted as a fictitious game between the conservative forces in the central bank, whose goal is to pursue monetary stability, and those forces, either in the bank or outside (e.g., in the government) aiming to stimulate the economy. (The running cost in the game does, however, depend on v_{x_2}, i.e., a derivative of the original value function, so to obtain the control one would not try to solve the game.) The strategies of the conservative forces are given by all stopping times when interest rates are increased in order to reduce the inflation, whereas the strategies of the expansionary forces are all stopping times when interest rates are lowered to favor economic expansion. In this game the role of the "referee" is played by the financial market; according to some economists (cf. Vaciago [26]) the result of the game depends on the efficiency of the financial market. Only if the financial market is "perfect," and hence guarantees long periods of equilibrium, can one assume the complete independence of the central bank from its government, and the stability of prices can then be optimally achieved.

3. Preliminary results. We fix some notation which will be used in the rest of the paper. Let $p > 1$ and let Q denote an open set in $\mathbb{R}^2$ with closure $\bar{Q}$:

- $H^{m+\mu,(m+\mu)/2}(\bar{Q} \times [0,T])$ = the space of all functions u continuous on $\bar{Q} \times [0,T]$ with continuous partial derivatives of the form $D_t^r D_x^j u$ for $2r + j \le m$, whose derivatives of the form $D_t^r D_x^j u$, $2r + j = m$ are Hölder continuous with respect to x (exponent $\mu \in (0,1)$) and whose derivatives of the form $D_t^r D_x^j u$, $0 < m + \mu - 2r - j < 2$, are Hölder continuous with respect to t (exponent $\frac{m+\mu-2r-j}{2} \in (0,1)$) on $\bar{Q} \times [0,T]$ (cf. [19, p. 7], for the norm);
- $H^{m+\mu,(m+\mu)/2}(Q \times (0,T))$ = the set of all functions u such that $u \in H^{m+\mu,(m+\mu)/2}(\bar{\Omega})$ for all Ω such that $\bar{\Omega} \subset Q \times (0,T)$;
- $H^{m+\mu,(m+\mu)/2}_{\text{loc}}(\mathbb{R}^2 \times [0,T])$ is the set of all functions u such that $u \in H^{m+\mu,(m+\mu)/2}(\bar{Q} \times [0,T])$ for all bounded $Q \subset \mathbb{R}^2$;
- $L^p(0,T;Y)$ = the space of all p-integrable functions on $(0,T)$ with values in the Hilbert space Y (notice that

$$W^{2,1;2}(Q \times (0,T)) = \{u \in L^2(0,T; W^{2;2}(Q)), u_t \in L^2(0,T; L^2(Q))\});$$

- $B_N = \{x \in \mathbb{R}^2 : |x| < N\}$;
- $\Omega_N = B_N \times (0,T)$;
- $\pi(x) = (1 + |x|^2)^{-l}$ (any integer $l > 4$);
- $\varphi_i(x) = 2l x_i (1 + |x|^2)^{-1}$ for $x = (x_1, x_2) \in \mathbb{R}^2$, $i = 1, 2$ (notice that $\pi_{x_i}(x) = -\varphi_i(x)\pi(x)$);
- $L^p_\pi(\mathbb{R}^2 \times [0,T]) = \{u : \pi^{1/p} u \in L^p(\mathbb{R}^2 \times [0,T])\}$, i.e., L^p under the measure $\pi(x)dxdt$;
- $L^p_\pi(\mathbb{R}^2) = \{g : \pi^{1/p} g \in L^p(\mathbb{R}^2)\}$;
- $H = L^2_\pi(\mathbb{R}^2)$ with $\|g\|_H^2 = \displaystyle\int_{\mathbb{R}^2} |g(x)|^2 \pi(x) dx$;
- $(g,h) = \int_{\mathbb{R}^2} g(x)h(x)\pi(x)dx$ denotes the inner product in H;
- $V = \{g \in H : g_{x_1}, g_{x_2} \in H\}$ with $\|g\|_V^2 = \|g\|_H^2 + \|g_{x_1}\|_H^2 + \|g_{x_2}\|_H^2$;
- V' = the dual of V (notice that H is identified with its dual and we have $V \subset H \subset V'$);
- $\langle \cdot, \cdot \rangle$ denotes the pairing between V' and V;

- $H_\circ = \{g \in H : |x|g \in H\}$;
- $V_\circ = V \cap H_\circ$;
- $W(0,T) = \{w \in L^2(0,T;V) : w_t \in L^2(0,T;V')\}$ with the Hilbert norm

$$\|w\|_{W(0,T)} = \Big(\int_0^T \|w(t)\|_V^2 \, dt + \int_0^T \|w_t(t)\|_{V'}^2 \, dt \Big)^{\frac{1}{2}}$$

(notice that if $w \in W(0,T)$ then it is continuous in $[0,T] \mapsto H$ and $\frac{d}{dt}\|w(t)\|^2 = 2\langle w(t), w'(t)\rangle$);

- $W_\circ(0,T) = \{w \in W(0,T) : w(t) \in H_\circ \text{ a.e. } t \in [0,T]\}$.

The following simple lemma will often be used.

LEMMA 3.1. *If X_s^0 is the uncontrolled process starting at time t from x, then there exists a positive constant $C > 0$ independent of x and t such that*

$$E\{|X_s^0|^2\} \le C(1 + |x|^2). \tag{3.1}$$

There exists a positive constant C such that for every $k \in V[0,T]$, $x, y \in \mathbb{R}^2$, and $t \in [0,T]$, if X_s^x and X_s^y denote the processes controlled by k and starting at time t from x and y, respectively, then

$$E\{|X_s^y - X_s^x|\} \le C|y - x|, \tag{3.2}$$

$$E\{|X_s^y - X_s^x|^2\} \le C|y - x|^2 \tag{3.3}$$

for every $s \in (t,T]$.

We collect the main properties of $v(x,t)$ in the following theorem. Note that

$$f(x) - f(y) \le C(|x| + |y|)|x - y| \tag{3.4}$$

for all x, y in $\mathbb{R}^2$.

THEOREM 3.2. *There exists a positive constant C such that for every $\lambda \in (0,1)$ and for all x, x' in $\mathbb{R}^2$, $|x'| \le 1$, and t in $[0,T]$, $h > 0$, one has*

$$0 \le v(x,t) \le C(1 + |x|^2); \tag{3.5}$$

$$|v(x,t) - v(x + x', t)| \le C(1 + |x| + |x'|)|x'|; \tag{3.6}$$

$$0 \le v(x + \lambda x', t) + v(x - \lambda x', t) - 2v(x,t) \le C\lambda^2; \tag{3.7}$$

$$v(x, t + h) - v(x,t) \le 0, \qquad h \in (0, T - t]; \tag{3.8}$$

$$v(x, t - h) - v(x,t) \le C(1 + |x|^2)h, \qquad h \in (0,t]. \tag{3.9}$$

Hence $v(x,t)$ is convex in x, nonincreasing in t, and $v \in W^{2,1;p}_{\text{loc}}(\mathbb{R}^2 \times (0,T))$ (any $p \le \infty$) with

$$|v_{x_i}(x,t)| \le C(1 + |x|), \quad |v_{x_i x_j}(x,t)| \le C, \quad |v_t(x,t)| \le C(1 + |x|^2) \tag{3.10}$$

a.s. in $\mathbb{R}^2 \times [0,T]$, $i, j = 1, 2$. In particular, for any $\mu \in (0,1)$, $v \in H^{1+\mu,(1+\mu)/2}_{\text{loc}}$ $(\mathbb{R}^2 \times [0,T])$.

Proof. Our proof borrows from [6] and [7]. Clearly it suffices to consider only those controls which give a cost not greater than $J_{x,t}(0)$, and (3.1) implies

$$J_{x,t}(0) \le C(1+|x|^2) \tag{3.11}$$

and hence also

$$v(x,t) \le C(1+|x|^2) \qquad \forall (x,t) \in \mathbb{R} \times [0,T]. \tag{3.12}$$

Let x, x', t, k be fixed and denote by X_s^x and $X_s^{x+x'}$ the processes controlled by k and starting at time t from x and $x+x'$, respectively. Then (3.4) implies

$$\begin{aligned}
&v(x+x',t) - v(x,t) \\
&\le \sup_{k \in V[0,T]} E\Big\{ \int_t^T \big(f(X_s^{x+x'}) - f(X_s^x)\big) e^{-\rho(s-t)} ds \Big\} \\
&\le C \sup_{k \in V[0,T]} E\Big\{ \int_t^T \big(\big|X_s^{x+x'} - X_s^x\big| + 2\big|X_s^x\big|\big)\big|X_s^{x+x'} - X_s^x\big| e^{-\rho(s-t)} ds \Big\} \\
&\le C \sup_{k \in V[0,T]} \Big(\int_t^T E\big\{\big|X_s^{x+x'} - X_s^x\big|^2\big\} e^{-\rho(s-t)} ds \Big) \\
&\quad + C \sup_{k \in V[0,T]} \Big(\int_t^T E\big\{\big|X_s^{x+x'} - X_s^x\big|^2\big\} e^{-\rho(s-t)} ds \Big)^{\frac{1}{2}} \Big(J_{x,t}(0) \Big)^{\frac{1}{2}}.
\end{aligned}$$

Now we deduce (3.6) and hence $(3.10)_1$ from Lemma 3.1 and (3.11).

We now observe that $J_{x,t}(k)$ is jointly convex in (x,k) since X is affine in k and x, and both the set $V[0,T]$ and the function f are convex; i.e.,

$$J_{\lambda x + (1-\lambda)y, t}(\lambda k + (1-\lambda)k') \le \lambda J_{x,t}(k) + (1-\lambda) J_{y,t}(k')$$

for $\lambda \in [0,1]$. Therefore, the value function $v(x,t)$ is convex in x, and hence the first inequality in (3.7) follows. In order to prove the second one we fix x, x', λ, t, k, and we denote by X_s^x, $X_s^{x+\lambda x'}$, and $X_s^{x-\lambda x'}$ the processes controlled by k and starting at time t from x, $x+\lambda x'$, and $x-\lambda x'$, respectively. Then we have

$$\begin{aligned}
&v(x+\lambda x',t) + v(x-\lambda x',t) - 2v(x,t) \\
&\le \sup_{k \in V[0,T]} E\Big\{ \int_t^T \big(f(X_s^{x+\lambda x'}) + f(X_s^{x-\lambda x'}) - 2f(X_s^x)\big) e^{-\rho(s-t)} ds \Big\}.
\end{aligned}$$

But (2.2) implies

$$X_s^{x+\lambda x'} = X_s^x + \lambda e^{b(s-t)} x',$$

hence from

$$(z_1 + \lambda z_2)^2 + (z_1 - \lambda z_2)^2 - 2{z_1}^2 = 2\lambda^2 {z_2}^2, \qquad z_1, z_2, \lambda \in \mathbb{R},$$

we have

$$\begin{aligned}
&v(x+\lambda x',t) + v(x-\lambda x',t) - 2v(x,t) \\
&\le \lambda^2 (1+\nu) \int_t^T \|e^{b(s-t)}\|^2 e^{-\rho(s-t)} ds
\end{aligned}$$

since $|x'| \le 1$. Now (3.7) and $(3.10)_2$ follow.

Let $t \in (0,T)$ and $h > 0$ such that $t + h \le T$; denote by X_s^t and X_s^{t+h} the processes controlled by k and starting from x at time t and $t+h$, respectively. Then it is easy to check that

$$X_{t+h+s}^{t+h} = X_{t+s}^t \quad \text{a.s.}$$

for all $s \in [0, T-(t+h)]$. Therefore it follows that

$$v(x,t+h) - v(x,t) \le \sup_{k \in V[0,T]} E\Big\{ \int_{T-h}^{T} -\frac{1}{2}[\nu(X_{1,s}^t)^2 + (X_{2,s}^t)^2] e^{-\rho(s-t)} ds \Big\} \le 0;$$

i.e., (3.8) holds.

On the other hand, if $h > 0$ is such that $0 \le t-h$ and if $X_s^{0,t-h}$ is the uncontrolled process (i.e., $k = 0$) starting at time $t-h$ from x, then for every $t-h \le s \le T$ we have

$$v(x,t-h) \le E\Big\{ \int_{t-h}^{s} f(X_r^{0,t-h}) e^{-\rho(r-t+h)} dr + v(X_s^{0,t-h}, s) e^{-\rho(s-t+h)} \Big\},$$

and hence, using (3.1),

$$\begin{aligned} &v(x,t-h) - v(x,t) \\ &\le E\Big\{ \int_{t-h}^{t} f(X_r^{0,t-h}) e^{-\rho(r-t+h)} dr \Big\} + E\Big\{ v(X_t^{0,t-h}, t) e^{-\rho h} - v(x,t) \Big\} \\ &\le \int_{t-h}^{t} C(1+|x|^2) e^{-\rho(r-t+h)} dr + E\Big\{ v(X_t^{0,t-h}, t) e^{-\rho h} - v(x,t) \Big\}. \end{aligned}$$

But for $t \in [0,T]$ fixed we have already shown that $v(\cdot\,,t)$ is a continuous function of polynomial growth whose partial derivatives $v_{x_i}(\cdot\,,t)$ and $v_{x_i x_j}(\cdot\,,t)$ are in $L^p_{\mathrm{loc}}(\mathbb{R}^2)$; that is, $v(\cdot\,,t) \in W^{2;p}_{\mathrm{loc}}(\mathbb{R}^2)$ for any p finite. Hence a generalized Ito's formula holds for $v(\cdot\,,t) e^{-\rho(s-t+h)}$ (cf. [10] or [18]) and we have

$$\begin{aligned} E\Big\{ v(X_t^{0,t-h}, t) e^{-\rho h} - v(x,t) \Big\} &= \int_{t-h}^{t} E\Big\{ \mathcal{L} v(X_r^{0,t-h}, t) \Big\} e^{-\rho(r-t+h)} dr \\ &\le C \int_{t-h}^{t} E\Big\{ 1 + |X_r^{0,t-h}|^2 \Big\} e^{-\rho(r-t+h)} dr \end{aligned}$$

due to the estimates (3.5), $(3.10)_1$, and $(3.10)_2$. Note that in the above, we consider v only as a function of x! Finally, we use (3.1) to obtain

$$E\Big\{ v(X_t^{0,t-h}, t) e^{-\rho h} - v(x,t) \Big\} \le C(1+|x|^2) h,$$

and we conclude that

$$v(x,t-h) - v(x,t) \le C(1+|x|^2) h,$$

and hence $(3.10)_3$ also follows. Finally, the Hölder continuity of v and v_{x_i} follows from an imbedding theorem concerning the space $W^{2,1;p}(\Omega_N)$ with $p > 4$ (cf. [19, Lemma II-3.3, p. 80]). □

4. Penalization. In order to obtain a variational formulation of the problem (P) we introduce a penalized problem as follows. Let β be a $C^\infty(\mathbb{R})$, convex, nondecreasing function such that $\beta(r) = 0$ for $r \le 0$, $\beta(r) > 0$ for $r \in (0,1)$, and $\beta(r) = 2r - 1$ for $r \ge 1$. Then for every $\varepsilon > 0$, the set

$$U^\varepsilon = \Big\{(\eta,\xi) \in \mathbb{R}\times[0,\infty) : |\eta|\theta - \frac{1}{\varepsilon}\beta(\theta(\theta+2)) \le \xi \le \frac{1}{\varepsilon} \text{ for all } \theta \ge 0\Big\} \tag{4.1}$$

is convex and compact. We denote by $V^\varepsilon[0,T]$ the set of all measurable, $\mathcal{F}_s$-adapted processes $(\eta,\xi) : [0,T] \to U^\varepsilon$, and we consider the penalized problem

$$v^\varepsilon(x,t) = \inf\Big\{J^\varepsilon_{x,t}(\eta,\xi) : (\eta,\xi) \in V^\varepsilon[0,T]\Big\}, \tag{P^ε}$$

where

$$J^\varepsilon_{x,t}(\eta,\xi) = E\Big\{\int_t^T \big[f(X_s) + |\eta_{s-t}| + \xi_{s-t}\big]e^{-\rho(s-t)}ds\Big\} \tag{4.2}$$

and X_s is the diffusion determined by the stochastic differential equation

$$\begin{cases} dX_s = (a + bX_s + e_1\,\eta_{s-t})ds + \sigma dW_{s-t}, & s \in (t,T], \\ X_t = x. \end{cases} \tag{4.3}$$

Since X is affine in η and x, the cost $J^\varepsilon_{x,t}(\eta,\xi)$ is simultaneously convex in (η,ξ) and x; i.e.,

$$J^\varepsilon_{\lambda x+(1-\lambda)y,t}(\lambda(\eta,\xi) + (1-\lambda)(\eta',\xi')) = \lambda J^\varepsilon_{x,t}(\eta,\xi) + (1-\lambda)J^\varepsilon_{y,t}(\eta',\xi')$$

for $\lambda \in [0,1]$. Hence the value function $v^\varepsilon(x,t)$ is convex in x.

Estimates analogous to those of Theorem 3.2 hold for v^ε uniformly in ε as stated in the following theorem.

THEOREM 4.1. *There exists a positive constant C such that for every $\varepsilon > 0$, for every $\lambda \in (0,1)$, and for all x, x' in $\mathbb{R}^2$, $|x'| \le 1$, and t in $[0,T]$, $h > 0$, one has*

$$0 \le v^\varepsilon(x,t) \le C(1 + |x|^2); \tag{4.4}$$

$$|v^\varepsilon(x,t) - v^\varepsilon(x+x',t)| \le C(1 + |x| + |x'|)|x'|; \tag{4.5}$$

$$0 \le v^\varepsilon(x+\lambda x',t) + v^\varepsilon(x-\lambda x',t) - 2v^\varepsilon(x,t) \le C\lambda^2; \tag{4.6}$$

$$v^\varepsilon(x,t+h) - v^\varepsilon(x,t) \le 0, \qquad h \in (0, T-t]; \tag{4.7}$$

$$v^\varepsilon(x,t-h) - v^\varepsilon(x,t) \le C(1+|x|^2)h, \qquad h \in (0,t]. \tag{4.8}$$

Hence $v^\varepsilon(x,t)$ is convex in x, nonincreasing in t, and $v^\varepsilon \in W^{2,1;\infty}_{\mathrm{loc}}(\mathbb{R}^2 \times (0,T))$ with

$$|v^\varepsilon_{x_i}(x,t)| \le C(1+|x|), \quad |v^\varepsilon_{x_ix_j}(x,t)| \le C, \quad |v^\varepsilon_t(x,t)| \le C(1+|x|^2) \tag{4.9}$$

a.s. in $\mathbb{R}^2 \times [0,T]$. In particular, $v^\varepsilon, v^\varepsilon_{x_i} \in C(\mathbb{R}^2 \times (0,T))$, $i,j = 1,2$.

Moreover, for each initial condition $(x,t) \in \mathbb{R}^2 \times [0,T]$,

$$\lim_{\varepsilon \to 0+} v^\varepsilon(x,t) = v(x,t). \tag{4.10}$$

Proof. The properties (4.4)–(4.9) can be proved by using arguments similar to those employed in the proof of (3.5)–(3.10).

The pointwise convergence (4.10) is obtained as follows. Since $\xi \geq 0$, then $v^\varepsilon(x,t) \geq v(x,t)$ for every $\varepsilon > 0$ and every $(x,t) \in \mathbb{R}^2 \times [0,T]$. On the other hand, as in [7, Theorem 2.2], one can check that

$$v(x,t) = \inf\Big\{J_{x,t}(k) : k \text{ Lipschitz continuous in } [0,T]\Big\};$$

therefore, for every $\delta > 0$ there exists a Lipschitz continuous control $k_s = \int_0^s \eta_r dr$ in $V[0,T]$ such that

$$J_{x,t}(k) \leq v(x,t) + \frac{\delta}{2}.$$

Then by taking $\xi_s = \frac{\delta}{2}\rho(1 - e^{-\rho(T-t)})^{-1}$ for all $s \in [0,T]$, we can find $\varepsilon_\circ > 0$ such that

$$\begin{cases} (\eta_s, \xi_s) \in U^\varepsilon, \\ J^\varepsilon_{x,t}(\eta,\xi) = J_{x,t}(k) + \dfrac{\delta}{2} \leq v(x,t) + \delta \end{cases}$$

for all $\varepsilon < \varepsilon_\circ$. Thus $v^\varepsilon(x,t) \leq v(x,t) + \delta$ for all $\varepsilon < \varepsilon_\circ$ and (4.10) is proved. □

It can be shown that the Hamilton–Jacobi–Bellman equation for the (P^ε)-problem is

$$-v^\varepsilon_t - \mathcal{L}v^\varepsilon + \rho v^\varepsilon + \frac{1}{\varepsilon}\beta([v^\varepsilon_{x_1}]^2 - 1) = f \tag{4.11}$$

with the boundary condition

$$v^\varepsilon(x,T) = 0 \qquad \forall x \in \mathbb{R}^2. \tag{4.12}$$

PROPOSITION 4.2. *The value function v^ε of the penalized problem (P^ε) is continuous on $\mathbb{R}^2 \times [0,T]$ and is the unique solution in $C^{2,1}_{\text{pol}}(\mathbb{R}^2 \times (0,T))$ of the Hamilton–Jacobi–Bellman equation* (4.11) *with the boundary data* (4.12).

Proof. It suffices to apply Theorem VI-6.2, Theorem VI-6.3, and Corollary VI-4.1 of [10]. □

Remark 4.3. Actually $v^\varepsilon_t, v^\varepsilon_{x_i x_j}$, $i,j = 1,2$, are locally Hölder continuous as shown in the proof of [10, Theorem VI-6.2].

PROPOSITION 4.4. *The value function v^ε is in $H^{4+\mu,(4+\mu)/2}_{\text{loc}}(\mathbb{R}^2 \times [0,T])$ for some $\mu \in (0,1)$.*

Proof. From Proposition 4.2 and Remark 4.3 it follows that

$$v^\varepsilon \in H^{2+\mu,(2+\mu)/2}_{\text{loc}}(\mathbb{R}^2 \times [0,T]);$$

therefore, if we write (4.11) as

$$-v^\varepsilon_t - \mathcal{L}v^\varepsilon + \rho v^\varepsilon = f - \frac{1}{\varepsilon}\beta([v^\varepsilon_{x_1}]^2 - 1),$$

we see that the right-hand side of the equation above is in $H^{1+\mu,(1+\mu)/2}_{\text{loc}}(\mathbb{R}^2 \times [0,T])$. Hence we may apply Theorem III-10 of [11] (with its $p = 1$) to obtain that the derivatives $v^\varepsilon_{x_i}, v^\varepsilon_{x_i x_j}, v^\varepsilon_{x_i x_j x_k}, v^\varepsilon_t, v^\varepsilon_{x_i t}$ are locally Hölder continuous; that is,

$$v^\varepsilon \in H^{3+\mu,(3+\mu)/2}_{\text{loc}}(\mathbb{R}^2 \times [0,T]).$$

So we deduce that the right-hand side above is in $H^{2+\mu,(2+\mu)/2}_{\text{loc}}(\mathbb{R}^2\times[0,T])$, and by an application of Theorem III-11 in [11] (with its $p=2$ and its $q=1$) we conclude that the derivatives $v^\varepsilon_{tt}, v^\varepsilon_{tx_i}, v^\varepsilon_{tx_ix_j}, v^\varepsilon_{x_ix_jx_kx_l}$ are locally Hölder continuous. This completes the proof. ∎

The regularity of v^ε provided by Proposition 4.4 allows us to differentiate (4.11) with respect to x_1; in fact, we have

$$\begin{cases} -u^\varepsilon_t - \mathcal{L}u^\varepsilon + (\rho - b_{11})u^\varepsilon + \dfrac{2}{\varepsilon}\beta'([u^\varepsilon]^2-1)u^\varepsilon u^\varepsilon_{x_1} = b_{21}\tilde{u}^\varepsilon + \nu x_1, \\ u^\varepsilon(x,T) = 0 \qquad \forall x\in\mathbb{R}^2, \end{cases} \tag{4.13}$$

where we have set

$$u^\varepsilon = v^\varepsilon_{x_1} \quad \text{and} \quad \tilde{u}^\varepsilon = v^\varepsilon_{x_2}. \tag{4.14}$$

We will now show that (P^ε) provides an approximation (in some sense) of the original problem (P), and hence we obtain some regularity of the derivatives $v_{x_i}(x,t)$ of the value function.

PROPOSITION 4.5. *There exists a sequence $\{\varepsilon_n\}_{n\in\mathbb{N}}$, $\varepsilon_n\downarrow 0$ as $n\to\infty$, such that if $u^n := u^{\varepsilon_n}$ and $\tilde{u}^n := \tilde{u}^{\varepsilon_n}$, $n\in\mathbb{N}$, then*

$$\begin{cases} \lim\limits_{n\to\infty} u^n = v_{x_1} \text{ weakly in } L^2(0,T;V); \\ \lim\limits_{n\to\infty} \tilde{u}^n = v_{x_2} \text{ weakly in } L^2(0,T;V). \end{cases}$$

Proof. Thanks to the estimates (4.9), there is a constant $C>0$ such that

$$\begin{cases} \|u^\varepsilon(\cdot\,,t)\|_V \le C, \\ \|\tilde{u}^\varepsilon(\cdot\,,t)\|_V \le C \end{cases} \tag{4.15}$$

for $\varepsilon>0, t\in[0,T]$; hence $\{u^\varepsilon\}_{\varepsilon>0}, \{\tilde{u}^\varepsilon\}_{\varepsilon>0}$ lie in a bounded set in $L^\infty(0,T;V)$ which is a subset of $L^2(0,T;V)$. Therefore, there exists a sequence $\{\varepsilon_n\}_{n\in\mathbb{N}}$, $\varepsilon_n\downarrow 0$, as $n\to\infty$, such that $u^n := u^{\varepsilon_n}$ and $\tilde{u}^n := \tilde{u}^{\varepsilon_n}$ converge weakly in $L^2(0,T;V)$. But $v^{\varepsilon_n}\to v$ pointwise by (4.10), and since weak limits are unique, the weak limits of u^n and $\tilde{u}^n$ must be u and $\tilde{u}$, respectively. ∎

LEMMA 4.6. *There exists a positive constant C such that*

$$\|u^\varepsilon_t\|_{L^\infty(0,T;V')} \le C, \tag{4.16}$$

$$\|\tilde{u}^\varepsilon_t\|_{L^\infty(0,T;V')} \le C \tag{4.17}$$

for all $\varepsilon>0$.

Proof. From (4.9) it follows that $|(\mathcal{L}u^\varepsilon(\cdot\,,t),g)| \le C\|g\|_V$ for all $g\in V$, so $\|(\mathcal{L}u^\varepsilon(\cdot\,,t))\|_{V'}\le C$. Moreover, from (4.11) and (4.9), we have

$$\begin{aligned} &\left\| \frac{1}{\varepsilon}\beta([u^\varepsilon(\cdot\,,t)]^2-1)\right\|_H \\ &\qquad = \left\| f + v^\varepsilon_t(\cdot\,,t) + \mathcal{L}v^\varepsilon(\cdot\,,t) - \rho v^\varepsilon(\cdot\,,t)\right\|_H \le C \end{aligned} \tag{4.18}$$

for $t\in[0,T]$ a.e. Also, from (4.15) and $|\beta'(y)|\le 2$, it follows that

$$\|(\beta'([u^\varepsilon]^2-1)u^\varepsilon u^\varepsilon_{x_i})(\cdot\,,t)\|_H \le C,$$

thus $(\beta'([u^\varepsilon]^2-1)u^\varepsilon u^\varepsilon_{x_i})(\cdot\,,t)$ is in H and hence in V' with

$$\begin{aligned}\left\|\frac{2}{\varepsilon}(\beta'([u^\varepsilon]^2-1)u^\varepsilon u^\varepsilon_{x_i})(\cdot\,,t)\right\|_{V'} &= \sup_{\|g\|_V\le 1}\left|\left(\frac{1}{\varepsilon}\frac{\partial}{\partial x_i}\beta([u^\varepsilon(\cdot\,,t)]^2-1),g\right)\right| \\ &= \sup_{\|g\|_V\le 1}\left|-\left(\frac{1}{\varepsilon}\beta([u^\varepsilon(\cdot\,,t)]^2-1),g_{x_i}-\varphi_i g\right)\right| \\ &\le C\left\|\frac{1}{\varepsilon}\beta([u^\varepsilon(\cdot\,,t)]^2-1)\right\|_H;\end{aligned}$$

hence from (4.18)

$$\left\|\frac{2}{\varepsilon}(\beta'([u^\varepsilon]^2-1)u^\varepsilon u^\varepsilon_{x_i})(\cdot\,,t)\right\|_{V'}\le C \qquad \text{a.e. } t\in[0,T]. \tag{4.19}$$

Now (4.16) follows from (4.13).

Similarly, to prove (4.17) we differentiate (4.11) with respect to x_2 and proceed as above. □

PROPOSITION 4.7. *There exists a positive constant $C>0$ such that*

$$\begin{cases}\|v_{x_1t}\|_{L^\infty(0,T;V')}\le C,\\ \|v_{x_2t}\|_{L^\infty(0,T;V')}\le C.\end{cases} \tag{4.20}$$

Proof. Because of (4.16) there exists a sequence u^n_t converging to v_{x_1t} weak-$\star$ in $L^\infty(0,T;V')$ as $n\to\infty$, and this proves $(4.20)_1$. Similarly, from (4.17), $(4.20)_2$ follows. □

Remark 4.8. We point out that from (3.10), (4.9), and $(4.20)_1$, (4.16) follows from $v_{x_1}\in W_\circ(0,T)$ and $u^n\in W_\circ(0,T)$ for all $n\in\mathbb{N}$.

PROPOSITION 4.9. *There exists a sequence $\{\varepsilon_n\}_{n\in\mathbb{N}}$, $\varepsilon_n\downarrow 0$, as $n\to\infty$, such that if $u^n:=u^{\varepsilon_n}$, $n\in\mathbb{N}$, then*

$$\lim_{n\to\infty}u^n=v_{x_1}\quad \textit{pointwise}$$

and

$$\lim_{n\to\infty}u^n=v_{x_1}\ \textit{in}\ L^2_\pi(\mathbb{R}^2\times[0,T]).$$

Proof. For $N<\infty$, (4.9) implies that $u^n(\cdot\,,t)$ lies in a bounded set in $W^{1;\infty}(B_N)$, hence in a compact set in $H^\mu(B_N)$ for any $\mu\in(0,1)$ (by the compact imbedding of Sobolev spaces; cf. [13, Theorem 7.26]). Hence for each t, there exists a subsequence converging uniformly, and by the earlier pointwise convergence of v^n, this limit is v_{x_1}; i.e., for ϕ differentiable with support in B_N,

$$\begin{aligned}\lim_m\int_{B_N}u^{n_m}(x,t)\phi(x)dx &= -\lim_m\int_{B_N}v^{n_m}(x,t)\phi_{x_1}(x)dx\\ &= -\int_{B_N}v(x,t)\phi_{x_1}(x)dx\\ &= \int_{B_N}v_{x_1}(x,t)\phi(x)dx.\end{aligned}$$

Since the limit is unique then the whole sequence must converge to it. As N and t are arbitrary, then $u^n(x,t)\to v_{x_1}(x,t)$ pointwise. The second result now follows from (4.9). □

5. Variational formulations. We would like to obtain a variational inequality for v_{x_1} by taking limits as $\varepsilon \downarrow 0$ in (4.13), but the mixed derivatives $v_{x_i t}$ are not regular enough to allow for a strong variational formulation. We begin by showing that operator $e^{-\lambda(T-t)}v_{x_1}$ solves (uniquely, as will be seen later) a *weak* variational inequality corresponding to the coercive operator $A + \lambda$ (as defined below). General results, developed in section 7, then give a pointwise variational inequality satisfied by v_{x_1}.

We set

- $\hat{u}(x,t) = b_{21}v_{x_2}(x,t) + \nu x_1$;
- $\mathcal{K} = \{g \in V : |g(x)| \le 1 \text{ a.e.}\}$;
- $W^{\mathcal{K}}(0,T) = \{w \in W(0,T) : w(t) \in \mathcal{K} \text{ a.e. } t \in [0,T]\}$;
- $W^{\mathcal{K}}_{\lambda}(0,T) = \left\{\bar{w} \in W(0,T) : \bar{w}(t) = w(t)e^{-\lambda(T-t)} \text{ with } w \in W^{\mathcal{K}}(0,T)\right\}$;
- $\alpha = \frac{1}{2}\sigma\sigma^*$.

Note that $W^{\mathcal{K}}(0,T) \subset W_\circ(0,T)$. There exists $\eta > 0$ such that

$$\sum_{i,j=1}^{2} \alpha_{ij}\xi_i\xi_j \ge 2\eta \sum_{i=1}^{2} \xi_i^2 \qquad \forall \xi_1, \xi_2 \in \mathbb{R}, \tag{5.1}$$

since, by assumption, $\sigma\sigma^*$ is positive definite.

We define the operator $A : V \times V_\circ \to \mathbb{R}$ by setting

$$\begin{aligned} A(g,h) = &\sum_{i,j=1}^{2} \alpha_{ij}(g_{x_i}, h_{x_j} - \varphi_j\, h) \\ &- \sum_{i=1}^{2}([a+bx]_i\, g_{x_i}, h) + (\rho - b_{11})(g,h), \end{aligned} \tag{5.2}$$

where the subscript "i" refers to the ith component (notice that if $g(x)$ is smooth, then $A(g,h) = \langle -\mathcal{L}g + (\rho - b_{11})g, h\rangle$ for every $h \in V_\circ$). Obviously, $h \in V_\circ$ is required to make sense of the second term on the right-hand side above; however, that term can still be defined for $h = g \in V$ by an integration by parts, since

$$\begin{aligned} -([a+bx]_i\, g_{x_i}, g) &= (g, [a+bx]_i\, g_{x_i} + \{b_{ii} - [a+bx]_i\, \varphi_i\}g) \\ &= \frac{1}{2}(\{b_{ii} - [a+bx]_i\, \varphi_i\}g, g) \end{aligned}$$

and $\{b_{ii} - [a+bx]_i\, \varphi_i\}$ is bounded. So $A(g,g)$ is well defined in V and given by

$$\begin{aligned} A(g,g) = &\sum_{i,j=1}^{2} \alpha_{ij}(g_{x_i}, g_{x_j} - \varphi_j\, g) \\ &+ \frac{1}{2}\sum_{i=1}^{2}(\{b_{ii} - [a+bx]_i\, \varphi_i\}g, g) + (\rho - b_{11})(g,g). \end{aligned} \tag{5.3}$$

We begin with the following lemma.

LEMMA 5.1. *The function v_{x_1} is in $W^{\mathcal{K}}(0,T)$.*

Proof. Remark 4.8 provides us with $v_{x_1} \in W_\circ(0,T)$; we need only show that

$$v_{x_1}(\cdot\,, t) \in \mathcal{K} \qquad \text{a.e. } t \in [0,T].$$

In fact, (4.18) implies

$$\|\beta([u^n(\cdot\,,t)]^2-1)\|_H \le C\varepsilon_n \qquad \forall n\in\mathbb{N},\ \text{a.e. } t\in[0,T], \tag{5.4}$$

with u^n as in the previous section; also, β is continuous and $u^n(x,t)\to v_{x_1}(x,t)$ for almost every $t\in[0,T]$, hence

$$\lim_{n\to\infty}\beta([u^n(x,t)]^2-1)=\beta([v_{x_1}(x,t)]^2-1) \qquad \text{a.e. } t\in[0,T].$$

Finally, (4.9) implies

$$0\le\beta([u^n(x,t)]^2-1)\le 2([u^n(x,t)]^2-1)\le C(1+|x|^2);$$

therefore, we can pass to the limit in (5.4) to obtain $\|\beta([v_{x_1}(\cdot\,,t)]^2-1)\|_H^2\le 0$, and hence $v_{x_1}(\cdot\,,t)\in\mathcal{K}$ for $t\in[0,T]$ a.e. □

THEOREM 5.2. *The function $\bar v = v_{x_1}e^{-\lambda(T-t)}$ solves the weak variational inequality*

$$\begin{aligned} -\int_0^T\langle\bar w_t,\bar w-\bar v\rangle dt+\frac{1}{2}\|\bar w(T)\|_H^2+\int_0^T[A+\lambda](\bar v,\bar w-\bar v)dt \\ \ge\int_0^T e^{-\lambda(T-t)}(\hat u,\bar w-\bar v)dt \qquad \forall\bar w\in W_\lambda^{\mathcal{K}}(0,T) \end{aligned} \tag{5.5}$$

with $\bar v\in W_\lambda^{\mathcal{K}}(0,T)$.

Proof. The proof is suggested by one in [27]. Let u^n be as in the previous section, let $\bar u^n = e^{-\lambda(T-t)}u^n$, and let $\hat u^n(x,t)=b_{21}\tilde u^n(x,t)+\nu x_1$. Let $\bar w = e^{-\lambda(T-t)}w\in W_\lambda^{\mathcal{K}}(0,T)$; since integration by parts holds in $W(0,T)$, then we obtain

$$\begin{aligned} \int_0^T(\bar u_t^n,\bar w-\bar u^n)dt=\frac{1}{2}\|\bar w(0)-\bar u^n(0)\|_H^2-\frac{1}{2}\|\bar w(T)\|_H^2 \\ +\int_0^T\langle\bar w_t,\bar w-\bar u^n\rangle dt. \end{aligned} \tag{5.6}$$

Since $\bar u^n\in W_\circ(0,T)$ by Remark 4.8, we may multiply (4.13) by $e^{-\lambda(T-t)}(\bar w-\bar u^n)$ and use (5.6) to obtain

$$\begin{aligned} -\int_0^T\langle\bar w_t,\bar w-\bar u^n\rangle dt-\frac{1}{2}\|\bar w(0)-\bar u^n(0)\|_H^2+\frac{1}{2}\|\bar w(T)\|_H^2 \\ +\int_0^T[A+\lambda](\bar u^n,\bar w-\bar u^n)dt \\ +\frac{2}{\varepsilon_n}\int_0^T e^{\lambda(T-t)}(\beta'([u^n]^2-1)\bar u^n\bar u^n_{x_1},\bar w-\bar u^n)dt \\ =\int_0^T e^{-\lambda(T-t)}(\hat u^n,\bar w-\bar u^n)dt \end{aligned} \tag{5.7}$$

for all $\bar w\in W_\lambda^{\mathcal{K}}(0,T)$. Now we proceed to take limits in (5.7) as $n\to\infty$. Since $u^n\to v_{x_1}$ weakly in $L^2(0,T;V)$ as $n\to\infty$ by Proposition 4.5, then

$$\int_0^T\langle\bar w_t,\bar w-\bar u^n\rangle dt\to\int_0^T\langle\bar w_t,\bar w-\bar v\rangle dt$$

and

$$\int_0^T [A+\lambda](\bar{u}^n,\bar{w})dt \to \int_0^T [A+\lambda](\bar{v},\bar{w})dt$$

as $n\to\infty$. Also, since $\hat{u}^n \to \hat{u}$ weakly in $L^2(0,T;H)$ as $n\to\infty$ (again by Proposition 4.5) and $u^n \to v_{x_1}$ in $L^2_\pi(\mathbb{R}^2\times[0,T])$ as $n\to\infty$ (by Proposition 4.9), then as $n\to\infty$

$$\int_0^T (e^{-\lambda(T-t)}\hat{u}^n, \bar{w}-\bar{u}^n)dt \to \int_0^T (e^{-\lambda(T-t)}\hat{u}, \bar{w}-\bar{v})dt$$

and

$$\int_0^T [A-A_\circ](\bar{u}^n,\bar{u}^n)dt \to \int_0^T [A-A_\circ](\bar{v},\bar{v})dt,$$

where $A_\circ(g,g)=\sum_{i,j}\alpha_{ij}(g_{x_i},g_{x_j})+\lambda\|g\|_H^2$. Finally, since $\Big(A_\circ(g,g)\Big)^{1/2}$ defines a norm on V and $\bar{u}^n\to\bar{v}$ weakly, we have

$$\liminf_{n\to\infty} A_\circ(\bar{u}^n,\bar{u}^n) \ge A_\circ(\bar{v},\bar{v}),$$

hence

$$\liminf_{n\to\infty} A(\bar{u}^n,\bar{u}^n) \ge A(\bar{v},\bar{v}).$$

To prove (5.5) it remains only to show that the penalty term in (5.7) is nonpositive; we examine two cases.

Case 1. On $\{(x,t) : \bar{w}^2(x,t) \ge (\bar{u}^n(x,t))^2\}$ we have $(u^n(x,t))^2 \le 1$ since $w\in W^{\mathcal{K}}(0,T)$, hence

$$\beta'([u^n(x,t)]^2-1)=0.$$

Case 2. On $\{(x,t) : \bar{w}^2(x,t) < (\bar{u}^n(x,t))^2\}$ it follows that

$$\begin{cases} 2\bar{u}^n(x,t)\Big(\bar{w}(x,t)-\bar{u}^n(x,t)\Big) \le (\bar{u}^n(x,t))^2+\bar{w}^2(x,t)-2(\bar{u}^n(x,t))^2<0,\\ \beta'([u^n(x,t)]^2-1)\ge 0,\\ \bar{u}^n_{x_1}=e^{-\lambda(T-t)}v^{\varepsilon_n}_{x_1x_1}\ge 0, \end{cases}$$

since v^{ε_n} is convex by Theorem 4.1. Hence we conclude that

$$\frac{2}{\varepsilon_n}\int_0^T (\beta'([u^n]^2-1)\bar{u}^n\bar{u}^n_{x_1},\bar{w}-\bar{u}^n)dt\le 0,$$

and thus (5.5) is proved. □

The general results in section 7, i.e., Theorems 7.9 and 7.11, now imply that (5.5) has a unique solution which is related to the solution of a pointwise variational inequality; cf. Remark 7.1. In fact, we obtain the following theorem.

THEOREM 5.3. *Assume that $\rho\ge b_{11}$. Then the function v_{x_1} is the unique solution of the pointwise variational inequality*

$$\text{(5.8)}\qquad \begin{cases} v_{x_1}\in W^{2,1;6}_{\mathrm{loc}}(\mathbb{R}^2\times(0,T)),\\ -1\le v_{x_1}\le +1 \text{ a.e. in } \mathbb{R}^2\times(0,T),\ v_{x_1}(x,T)=0 \text{ a.e. in } \mathbb{R}^2,\\ -v_{x_1t}+[-\mathcal{L}+(\rho-b_{11})]v_{x_1}=\hat{u} \quad \text{if } -1<v_{x_1}(x,t)<+1,\\ -v_{x_1t}+[-\mathcal{L}+(\rho-b_{11})]v_{x_1}\le\hat{u} \quad \text{if } v_{x_1}(x,t)=+1,\\ -v_{x_1t}+[-\mathcal{L}+(\rho-b_{11})]v_{x_1}\ge\hat{u} \quad \text{if } v_{x_1}(x,t)=-1,\\ \qquad\qquad \text{a.e. in } \mathbb{R}^2\times(0,T). \end{cases}$$

6. Conclusion. The central bank singular control problem analyzed in this paper gives rise to a free boundary problem; in fact, (5.8) or (2.6) is the variational formulation of a *two-obstacle* problem determining two moving boundaries, ∂_{-1} and ∂_{+1}, which at each time t split the $\mathbb{R}^2$-plane into three regions:

- the *inflation region* $\mathbf{I_t} = \{v_{x_1}(\cdot\,, t) = -1\}$, where interest rates are so low as to produce significant inflation;
- the *inflation-output tradeoff region* $\mathbf{T_t} = \{-1 < v_{x_1}(\cdot\,, t) < +1\}$, where a sort of equilibrium is found between inflation and output growth;
- the *deflation region* $\mathbf{D_t} = \{v_{x_1}(\cdot\,, t) = +1\}$, where interest rates are too high and prevent economic expansion.

In the inflation region $\mathbf{I_t}$ it is optimal for the conservative forces of the Bank to intervene by increasing interest rates; in the inflation-output tradeoff region $\mathbf{T_t}$ the conservative and the expansionist forces of the bank find an equilibrium, and hence it is optimal for both to do nothing; finally, in the deflation region $\mathbf{D_t}$ it is optimal for the expansionist forces to intervene by lowering interest rates.

We point out that Theorem 2.1, which gives a variational problem for the original value function v, might be obtained from (4.11) by taking limits (in some sense) as $\varepsilon \to 0$; that is,

$$\begin{cases} v \in W^{2,1;\infty}_{\mathrm{loc}}(\mathbb{R}^2 \times (0,T)), \quad v(x,T) = 0 \ \text{ a.e. in } \mathbb{R}^2, \\ \max\Big\{ -v_t + [-\mathcal{L} + \rho]v - f,\ -v_{x_1} - 1,\ v_{x_1} - 1 \Big\} \le 0, \\ (-v_t + [-\mathcal{L} + \rho]v - f)(-v_{x_1} - 1)(v_{x_1} - 1) = 0, \\ \qquad\qquad \text{a.e. in } \mathbb{R}^2 \times (0,T). \end{cases} \tag{6.1}$$

Then (6.1) or (2.4) gives us some insight into the optimal control $\hat{k}$ of the original problem (P); $\hat{k}$ must be singular with respect to the Lebesgue measure as a function of time (as in [6]). In fact, by applying the dynamic programming equation we see that $\hat{k}$ should be inactive when the state is in the region $\mathbf{T_t}$, whereas it should make the state jump to a convenient point on the boundary $\partial_{-1} \cup \partial_{+1}$ when the state is in either the inflation or the deflation regions. At the boundary, the optimal control $\hat{k}$ should be active just enough to keep the process inside $\mathbf{T_t}$. Hence the resulting optimal process is a diffusion reflected at the boundary, possibly with an initial jump if the state starts in $\mathbf{I_t}$ or $\mathbf{D_t}$.

Usually the construction of $\hat{k}$ is not an easy task as it requires some regularity of the boundaries ∂_{-1} and ∂_{+1} of the *two-obstacle* problem (5.8); work in this direction is in progress.

Regarding (5.8) or (2.6) from an economics point of view, they correspond to a differential game with stopping times where the players are the conservative and the expansionist forces of the central bank whose primary target is the control of inflation. In order to interpret the evaluation function (2.8) of the game we take $b_{21} = -1$ and $\nu = 0$ for simplicity (notice that it is reasonable to take $b_{21} < 0$ since inflation decreases if interest rates increase). Suppose that the diffusion starts at time t from some point x inside $\mathbf{T_t}$; then the bank will act in such a way as to minimize $J_{x,t}(k)$. However, the conservative forces will not object to any drastic monetary policies aiming to contain inflation, whereas the expansionist forces will be unwilling to accept policies that prevent economic expansion. It is easy to see that, under the present conditions, from (5.8) it follows that

$$\begin{cases} v_{x_2}(\cdot\,, s) \ge 0 \ \text{ in } \mathbf{I_s}, \\ v_{x_2}(\cdot\,, s) \le 0 \ \text{ in } \mathbf{D_s} \end{cases}$$

for $s \in (t,T]$. But $v(\cdot\,,s)$ is convex, hence $v_{x_i}(\cdot\,,s)$ is nondecreasing in the x_i direction; this implies that $\mathbf{I_s}$ must be on the left and above $\mathbf{D_s}$ for all $s \in [0,T]$, and

$$\begin{cases} v_{x_2}(\cdot\,,s) \geq 0, & v_{x_1}(\cdot\,,s) = -1 \quad \text{in } \mathbf{I_s}, \\ v_{x_2}(\cdot\,,s) \leq 0, & v_{x_1}(\cdot\,,s) = +1 \quad \text{in } \mathbf{D_s}. \end{cases}$$

Therefore, the first player (who represents the conservative tendency of the bank) will try to keep the state out of the region $\mathbf{I_s}$, since there the value function v increases with the inflation x_2; on the other hand, the second player will not allow the state to diffuse into the region $\mathbf{D_s}$ where v increases with the interest rate x_1. Thus, knowing that the second player will select the strategic time $\hat{\theta}_2$ in order to avoid high interest rates, the first player will aim to minimize

$$E\Bigg\{ \int_0^{\theta_1 \wedge \hat{\theta}_2} v_{x_2}(Y_s^0, t+s) e^{-(\rho - b_{11})s} ds - e^{-(\rho-b_{11})\theta_1} \mathbb{I}_{\substack{\theta_1 \leq \hat{\theta}_2 \\ \theta_1 < T-t}} + e^{-(\rho-b_{11})\hat{\theta}_2} \mathbb{I}_{\hat{\theta}_2 < \theta_1 \wedge (T-t)} \Bigg\},$$

and this is exactly what Theorem 2.3 states.

There are various directions in which to continue this work. In the first instance, what can we say about the (free) boundary of $\mathbf{T_t}$? What about numerical approximations to $\mathbf{T_t}$? Now we need to estimate the parameters of the model. On a more basic level, we can ask: how realistic is the model (2.2)? Probably, the dynamics we use do not reflect reality very well—certainly other models have been used in the literature. For example, Cox, Ingersoll, and Ross [8] use a diffusion coefficient, volatility, which is not constant but is proportional to the square root of the interest/inflation rate. Also, as mentioned before, for the Canadian economy the exchange rate with the U.S. dollar is certainly important in the decision-making process of the Bank of Canada, and this should be reflected in the cost functional J.

7. Variational inequalities. We begin by establishing existence and uniqueness results for the following weak and strong variational inequalities. The weak inequality is

$$\begin{aligned} -\int_0^T \langle \bar{w}_t, \bar{w} - \bar{v} \rangle dt + \frac{1}{2}\|\bar{w}(T)\|_H^2 + \int_0^T [A + \lambda](\bar{v}, \bar{w} - \bar{v}) dt \\ \geq \int_0^T e^{-\lambda(T-t)} (\hat{u}, \bar{w} - \bar{v}) dt \qquad \forall \bar{w} \in W_\lambda^{\mathcal{K}}(0,T), \end{aligned} \tag{7.1}$$

and the strong one is

$$\begin{cases} -(\hat{v}_t(t), g - \hat{v}(t)) + A(\hat{v}(t), g - \hat{v}(t)) \\ \qquad \geq (\hat{u}(t), g - \hat{v}(t)) \qquad \text{a.e. } t \in [0,T] \quad \forall g \in \mathcal{K}, \\ \hat{v}(T) = 0. \end{cases} \tag{7.2}$$

DEFINITION 7.1. *A solution of the* strong *variational inequality* (7.2) *is a function* $\hat{v} \in L^2(0,T;V)$ *with* $\hat{v}_t \in L^2(0,T;H)$ *and* $\hat{v}(t) \in \mathcal{K}$ *for a.e.* $t \in [0,T]$, *which satisfies* (7.2).

Remark 7.1. The main result is (cf. Theorems 7.9 and 7.11) that both equations have unique solutions; moreover, the solution of (7.2) solves a pointwise variational inequality. The assumptions on $a, b, \sigma, \rho, \nu, \pi, \phi$ are as before, and we assume that

$\hat{u} \in L^2(0,t;V) \cap H^{\mu,\mu/2}_{\text{loc}}(\mathbb{R}^2 \times [0,T])$ for some $\mu \in (0,1)$, $\hat{u}_t \in L^\infty(0,t;V')$, and $|\hat{u}(x,t)| \le K(1+|x|)$ for some constant K. Observe that our previously defined $\hat{u} = b_{21}v_{x_2} + \nu x_1$ satisfies these hypotheses (cf. Theorem 3.2 and Propositions 4.5 and 4.7), hence Theorem 5.2 implies that the unique solution of (7.1) is $e^{-\lambda(T-t)}v_{x_1}$ and the unique solution of (7.2) and (5.8) is v_{x_1}.

DEFINITION 7.2. *The operator A is said to be weakly coercive in V if there exists $\eta > 0$ and $\lambda \ge 0$ such that*

$$A(g,g) \ge \eta \|g\|_V^2 - \lambda \|g\|_H^2 \qquad \forall g \in V. \tag{7.3}$$

It is easy to see that A as defined in (5.2) is weakly coercive in V. This property gives uniqueness.

PROPOSITION 7.2. *The weak variational inequality* (7.1) *has at most one solution* $\bar{v} \in W^{\mathcal{K}}_\lambda(0,T)$.

Proof. We use an idea due to Bensoussan and Lions [2]. Let $\bar{u} = e^{-\lambda(T-t)}u$ and $\bar{z} = e^{-\lambda(T-t)}z$ be two solutions of (7.1), set $\tilde{w} = e^{-\lambda(T-t)}w$ where

$$w = \frac{1}{2}(u+z) \in \mathcal{K},$$

and let w_q be the solution of

$$\begin{cases} -q\dot{w}_q + w_q = w, \\ w_q(T) = 0, \end{cases}$$

where "$\cdot$" denotes the derivative with respect to time. Solving for w_q we notice that $w_q(t) \in \mathcal{K}$ since $\mathcal{K}$ is convex and $0 \in \mathcal{K}$. Hence $w_q \in W^{\mathcal{K}}(0,T)$ and

$$\lim_{q \downarrow 0} w_q(t) = \lim_{q \downarrow 0} \int_0^{(T-t)/q} w(t+qs)e^{-s}ds = w(t)$$

since $W(0,T) \subset C([0,T];H)$. Thus $\bar{w}_q := e^{-\lambda(T-t)}w_q \in W^{\mathcal{K}}_\lambda(0,T)$.

We take $\bar{w} = \bar{w}_q$ in (7.1) for $\bar{u}$ and $\bar{z}$ and add to obtain

$$\begin{aligned} -2\int_0^T \left\langle \frac{\bar{w}_q - \tilde{w}}{q} + \lambda \bar{w}_q, \bar{w}_q - \tilde{w} \right\rangle dt + \int_0^T [A+\lambda](\bar{u}, \bar{w}_q - \bar{u})dt \\ + \int_0^T [A+\lambda](\bar{z}, \bar{w}_q - \bar{z})dt \\ \ge 2\int_0^T e^{-\lambda(T-t)}(\hat{u}, \bar{w}_q - \tilde{w})dt, \end{aligned}$$

hence

$$\begin{aligned} &\int_0^T \Big\{ [A+\lambda](\bar{u}, \bar{w}_q - \bar{u}) + [A+\lambda](\bar{z}, \bar{w}_q - \bar{z}) \Big\} dt \\ &\ge 2\int_0^T e^{-\lambda(T-t)}(\hat{u}, \bar{w}_q - \tilde{w})dt + 2\lambda \int_0^T (\bar{w}_q, \bar{w}_q - \tilde{w})dt. \end{aligned}$$

As $q \downarrow 0$, the right-hand side converges to zero and the left to

$$\int_0^T \Big\{ [A+\lambda](\bar{u}, \tilde{w} - \bar{u}) + [A+\lambda](\bar{z}, \tilde{w} - \bar{z}) \Big\} dt = -\frac{1}{2}\int_0^T [A+\lambda](\bar{u} - \bar{z}, \bar{u} - \bar{z})dt.$$

Therefore, (7.3) implies

$$\frac{1}{2}\int_0^T \eta\|\bar{u}-\bar{z}\|_V^2 dt \le 0,$$

hence $\|\bar{u}-\bar{z}\|_V^2 = 0$ for all $t\in[0,T]$; i.e., $\bar{u}=\bar{z}$ in $W(0,T)$. ∎

LEMMA 7.3. *If $\hat{v}$ solves the strong variational inequality* (7.2) *then $e^{-\lambda(T-t)}\hat{v}$ solves the weak variational inequality* (7.1). *Hence* (7.2) *admits at most one solution.*

Proof. The last result follows by Proposition 7.2. Let $\hat{v}$ be a solution of (7.2) and $\bar{v}=e^{-\lambda(T-t)}\hat{v}$. Then $\bar{v}$ is a solution of

$$\begin{cases} -(\bar{v}_t(t),\bar{g}-\bar{v}(t))+[A+\lambda](\bar{v}(t),\bar{g}-\bar{v}(t)) \\ \qquad \ge (\hat{u}(t)e^{-\lambda(T-t)},\bar{g}-\bar{v}(t)) \qquad \text{a.e. } t\in[0,T] \quad \forall \bar{g}=e^{-\lambda(T-t)}\mathcal{K}, \\ \bar{u}(T)=0. \end{cases}$$

If $\bar{w}\in W_\lambda^{\mathcal{K}}(0,T)$, then $\bar{w}(t)\in\mathcal{K}$ for almost every $t\in[0,T]$ and hence

$$\begin{aligned} -(\bar{v}_t(t),\bar{w}(t)-\bar{v}(t))+[A+\lambda](\bar{v}(t),\bar{w}(t)-\bar{v}(t)) \\ \ge (\hat{u}(t),\bar{w}(t)-\bar{v}(t)) \qquad \text{a.e. } t\in[0,T]. \end{aligned}$$

By integrating over $[0,T]$ and after an integration by parts on the first term, it follows that

$$\begin{aligned} -\int_0^T \langle \bar{w}_t(t),\bar{w}(t)-\bar{v}(t)\rangle dt+\frac{1}{2}\|\bar{w}(T)\|_H^2-\frac{1}{2}\|\bar{w}(0)-\bar{v}(0)\|_H^2 \\ +\int_0^T [A+\lambda](\bar{v}(t),\bar{w}(t)-\bar{v}(t))dt \\ \ge \int_0^T (\hat{u}(t),\bar{w}(t)-\bar{v}(t))dt. \end{aligned}$$

But $\|\bar{w}(0)-\bar{v}(0)\|_H^2\ge 0$ may be dropped to obtain (7.1). ∎

To establish existence of solutions of (7.2) we start by penalizing (7.2). Let $\zeta:\mathbb{R}\to\mathbb{R}$ be defined by $\zeta(r)=(r-1)^+-(r+1)^-$ for $r\in\mathbb{R}$; then ζ is monotone nondecreasing, Lipschitz, and piecewise linear. Also,

$$|r|\ge|\zeta(r)| \qquad \forall r\in\mathbb{R},$$

and hence $r\zeta(r)\ge\zeta^2(r)\ge 0$ for all $r\in\mathbb{R}$, and

$$\zeta'(r)=\mathbb{I}_{\{|r|>1\}} \qquad \forall r\in\mathbb{R}.$$

Then, for every $\varepsilon>0$, we look for a solution $w^\varepsilon\in W_\circ(0,T)$ of the penalized problem

$$\begin{cases} -(w_t^\varepsilon(t),g)+A(w^\varepsilon(t),g)+\dfrac{1}{\varepsilon}(\zeta(w^\varepsilon(t)),g)=(\hat{u}(t),g) \\ \qquad\qquad \text{a.e. } t\in[0,T] \quad \forall g\in V_\circ, \\ w^\varepsilon(T)=0. \end{cases} \tag{7.4}$$

PROPOSITION 7.4. *For every $\varepsilon>0$ there exists at most one solution w^ε in $W_\circ(0,T)$ of the penalized problem* (7.4).

Proof. This follows from the monotonicity of ζ and is similar to the proof in [2, Theorem 2.3, p. 239]. ∎

We now turn to the proof of existence of a solution of (7.4). Due to the lack of a priori bounds of the type (4.9), we take care of the linear-in-x term in A by regularization, and by Galerkin's method we prove (cf. Proposition 8.3) the following proposition.

PROPOSITION 7.5. *There exists a solution $w^\varepsilon \in W(0,T)$ of the penalized variational problem* (7.4). *Moreover, the estimates*

$$\begin{cases} \|w^\varepsilon\|^2_{L^\infty(0,T;H)} + \|w^\varepsilon\|^2_{L^2(0,T;V)} \le C, \\ \|w^\varepsilon_t\|^2_{L^\infty(0,T;H)} + \|w^\varepsilon_t\|^2_{L^2(0,T;V)} \le C, \\ \dfrac{1}{\sqrt{\varepsilon}}\|\zeta(w^\varepsilon)\|_{L^2(0,T;H)} \le C \end{cases} \tag{7.5}$$

hold uniformly in $\varepsilon > 0$.

In order to show that the function w^ε is in $W_\circ(0,T)$ (i.e., $w^\varepsilon(\cdot\,,t) \in V_\circ$ for almost all $t \in [0,T]$), we obtain a bound for w^ε from the pointwise version of (7.4); i.e.,

$$\begin{cases} -\tilde w^\varepsilon_t(x,t) + [-\mathcal{L} + (\rho - b_{11})]\tilde w^\varepsilon(x,t) + \dfrac{1}{\varepsilon}\zeta(\tilde w^\varepsilon(x,t)) = \hat u(x,t) \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{a.e. } (x,t) \in \mathbb{R}^2 \times [0,T], \\ \tilde w^\varepsilon(x,T) = 0. \end{cases} \tag{7.6}$$

We will show that (7.6) has a solution $\tilde w^\varepsilon \in H^{2+\mu,(2+\mu)/2}_{\rm loc}$; since any solution of (7.6) is also a solution of (7.4), and this has a unique solution, then $\tilde w^\varepsilon = w^\varepsilon$. However, the pointwise formulation (7.6) will only allow an estimate uniform in ε on the function itself, so we still need the estimates (7.5) and hence the lengthy Galerkin approach for convergence proofs as $\varepsilon \to 0$.

THEOREM 7.6. *For every $\varepsilon > 0$ the function w^ε is a solution of the penalized problem* (7.6) *with $w^\varepsilon \in H^{2+\mu,(2+\mu)/2}_{\rm loc}$ and*

$$|w^\varepsilon(x,t)| \le K(1+|x|) \qquad \text{a.e. in } \mathbb{R}^2 \times [0,T], \tag{7.7}$$

K independent of $\varepsilon > 0$.

Proof. It suffices to show that (7.6) admits a solution $\tilde w^\varepsilon$ enjoying not only the properties above but also $\tilde w^\varepsilon \in W(0,T)$ (see (7.4)); then $\tilde w^\varepsilon = w^\varepsilon$ by uniqueness (cf. Proposition 7.4). Our proof is a modification of one found in [2, Theorem 5.1, p. 464].

We start by taking care of the unbounded coefficients in (7.6), so for every $N \in \mathbb{N}$ and $r \in \mathbb{R}$ we define $\chi^N(r) = N\mathbb{I}_{\{r \ge N\}} - N\mathbb{I}_{\{r \le -N\}}$ and $\chi^N(x) = (\chi^N(x_1), \chi^N(x_2))$ for $x = (x_1, x_2) \in \mathbb{R}^2$, so $|\chi^N(x)|^2 \le 2N^2$. We denote by $\mathcal{L}^N$ the operator $\mathcal{L}$ with $(a + bx)$ replaced by $(a + b\chi^N(x))$. Then, since $\hat u(x,t) \le C(1+|x|)$, we take care of the unbounded right-hand side of (7.6) by the change of variable $\tilde z(x,t) = e^{-\beta t}(1 + |x|^2)^{-1/2} z(x,t)$ with $\beta < 0$ to be chosen later. So we have

$$-\tilde z_t + [-\tilde{\mathcal{L}}^N + (\rho - b_{11})]\tilde z = e^{-\beta t}(1+|x|^2)^{-1/2}\{-z_t + [-\mathcal{L}^N + (\rho - b_{11})]z\}$$

with $\tilde{\mathcal{L}}^N \tilde z := \mathcal{L}^N \tilde z - 2\sum_{i,j}\alpha_{ij}x_i(1+|x|^2)^{-1}\tilde z_{x_j} + E(x)\tilde z$, the bounded function $E(x)$ being given by

$$E(x) := -\mathrm{trace}[\alpha](1+|x|^2)^{-1} + x^*\alpha x(1+|x|^2)^{-2} - (a + b\chi^N(x)) \cdot x(1+|x|^2)^{-1} - \beta.$$

Let $C_b(\Omega)$ denote the space of bounded, continuous functions on Ω with sup norm. So with

$$\begin{cases} \tilde u(x,t) := \hat u(x,t)e^{-\beta t}(1+|x|^2)^{-1/2} \in C_b(\mathbb{R}^2 \times [0,T]), \\ \tilde f^\varepsilon := \tilde u(x,t) + \dfrac{1}{\varepsilon}\dfrac{[2e^{\beta t}(1+|x|^2)^{1/2}\tilde\varphi - \zeta(e^{\beta t}(1+|x|^2)^{1/2}\tilde\varphi)]}{e^{\beta t}(1+|x|^2)^{1/2}} \end{cases}$$

for every $\tilde{\varphi} \in C_b(\mathbb{R}^2 \times [0,T])$, we seek a solution of

$$\begin{cases} -\tilde{z}_t^N(x,t) + [-\tilde{\mathcal{L}}^N + (\rho - b_{11})]\tilde{z}^N(x,t) + \dfrac{2}{\varepsilon}\tilde{z}^N = \tilde{f}^\varepsilon \\ \qquad\qquad\qquad\qquad \text{a.e. } (x,t) \in \mathbb{R}^2 \times [0,T], \\ \tilde{z}(x,T) = 0. \end{cases} \tag{7.8}$$

But $\tilde{f}^\varepsilon$ is in $L^2(0,T;H) \cap L^p(0,T;L^p_\pi(\mathbb{R}^2))$, hence there exists a unique solution $\tilde{z}^N$ of (7.8) in $L^2(0,T;V)$; moreover,

$$\tilde{z}^N \in L^p(0,T;L^p_\pi(\mathbb{R}^2)) \cap L^\infty(0,T;V) \cap W^{2,1;p}_{\text{loc}}(\mathbb{R}^2 \times [0,T]),$$

and for every bounded open set $G \subset \mathbb{R}^2 \times [0,T]$ and every open subset $G' \subset\subset G$:

$$\|\tilde{z}^N\|_{W^{2,1;p}(G')} \le C(\|\tilde{z}^N\|_{L^p(G)} + \|\tilde{f}^\varepsilon\|_{L^p(G)}), \tag{7.9}$$

where the constant C is independent of N and depends only on G, G', and the bounds of the coefficients of $[-\mathcal{L} + (\rho - b_{11})] + \frac{2}{\varepsilon}$ on G (cf. [2, Theorem 6.6, p. 134, Theorem 6.7, p. 139, Remark 6.13, p. 141, and Theorem 6.4, p. 131]). Hence, by Sobolev imbedding, $\tilde{z}^N$ is continuous.

On the other hand, the Feynman–Kac representation (cf. [3, Theorem 8.1, p. 281]) of the solution $\tilde{z}^N$ gives

$$\tilde{z}^N(x,t) = E\left\{ \int_t^T \tilde{f}^\varepsilon(y^N(s),s) e^{-\int_t^s (\rho - b_{11} + E(y^N(r)) + 2/\varepsilon)dr} ds \right\},$$

where $y^N(s)$ is the diffusion starting at time t from x and governed by the stochastic differential equation

$$dy^N = \left[a + b\chi^N(y^N) + 2\frac{\alpha y^N}{1 + |y^N|^2} \right] ds + \sigma dW_{s-t}.$$

But for β sufficiently negative (independent of N and ε) we can find $\delta > 0$ such that

$$\rho - b_{11} + E(x) \ge \delta > 0 \qquad \forall x \in \mathbb{R}^2.$$

Hence

$$\sup_{x,t} |\tilde{z}^N(x,t)| \le \|\tilde{f}^\varepsilon\|_{C_b(\mathbb{R}^2 \times [0,T])} \int_t^T e^{-(\delta + 2/\varepsilon)(s-t)} ds < \infty;$$

i.e., $\tilde{z}^N \in C_b(\mathbb{R}^2 \times [0,T])$.

Now, if $S(\tilde{\varphi}) := \tilde{z}^N$ for $\tilde{\varphi}_1, \tilde{\varphi}_2 \in C_b(\mathbb{R}^2 \times [0,T])$, then we have

$$\begin{aligned} |S(\tilde{\varphi}_1) - S(\tilde{\varphi}_2)| &\le \|\tilde{\varphi}_1 - \tilde{\varphi}_2\|_{C_b(\mathbb{R}^2 \times [0,T])} \left\{ \int_t^T \frac{2}{\varepsilon} e^{-(\delta + 2/\varepsilon)(s-t)} ds \right\} \\ &\le \frac{2}{\varepsilon} \, \frac{1}{\delta + 2/\varepsilon} \, \|\tilde{\varphi}_1 - \tilde{\varphi}_2\|_{C_b(\mathbb{R}^2 \times [0,T])} \\ &= C_{\varepsilon,\delta} \, \|\tilde{\varphi}_1 - \tilde{\varphi}_2\|_{C_b(\mathbb{R}^2 \times [0,T])} \end{aligned}$$

with $C_{\varepsilon,\delta} < 1$ independent of N. Hence $S : C_b(\mathbb{R}^2 \times [0,T]) \to C_b(\mathbb{R}^2 \times [0,T])$ is a contraction, and we denote by $\tilde{w}^{\varepsilon,N}$ its unique fixed point.

Let $\tilde{z}_0 = 0$ and $\tilde{z}_n := S(\tilde{z}_{n-1})$. Then

$$\|\tilde{z}_n\|_{C_b(\mathbb{R}^2\times[0,T])} = \left\| \sum_{i=1}^{n} (\tilde{z}_i - \tilde{z}_{i-1}) \right\|_{C_b(\mathbb{R}^2\times[0,T])} \le \frac{1}{1 - C_{\varepsilon,\delta}} \|\tilde{z}_1\|_{C_b(\mathbb{R}^2\times[0,T])}.$$

But

$$\begin{aligned}\|\tilde{z}_1\|_{C_b(\mathbb{R}^2\times[0,T])} &= \left\| E\left\{ \int_t^T \tilde{u}(y^N(s), s) e^{-\int_t^s (\rho - b_{11} + E(y^N(r)) + 2/\varepsilon) dr} ds \right\} \right\|_{C_b} \\ &\le \|\tilde{u}\|_{C_b(\mathbb{R}^2\times[0,t])} \int_t^T e^{-(\delta + 2/\varepsilon)(s-t)} ds,\end{aligned}$$

which implies

$$\|\tilde{z}_n\|_{C_b(\mathbb{R}^2\times[0,T])} \le \frac{1}{\delta} \|\tilde{u}\|_{C_b(\mathbb{R}^2\times[0,T])} \qquad \forall n \in \mathbb{N},$$

and so also

$$\|\tilde{w}^{\varepsilon,N}\|_{C_b(\mathbb{R}^2\times[0,T])} \le \frac{1}{\delta} \|\tilde{u}\|_{C_b(\mathbb{R}^2\times[0,T])}. \tag{7.10}$$

If $w^{\varepsilon,N}(x,t) := e^{\beta t}(1 + |x|^2)^{1/2}\ \tilde{w}^{\varepsilon,N}(x,t)$, then $w^{\varepsilon,N}$ is a continuous function verifying the polynomial growth condition

$$|w^{\varepsilon,N}(x,t)| \le K(1 + |x|) \qquad \forall (x,t) \in \mathbb{R}^2 \times [0,T]$$

with the constant K independent of ε and N. Moreover, $w^{\varepsilon,N}$ satisfies

$$\begin{cases} -w_t^{\varepsilon,N}(x,t) + [-\mathcal{L}^N + (\rho - b_{11})] w^{\varepsilon,N}(x,t) = \hat{u}(x,t) - \dfrac{1}{\varepsilon} \zeta(w^{\varepsilon,N}(x,t)) \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{a.e. } (x,t) \in \mathbb{R}^2 \times [0,T], \\ w^{\varepsilon,N}(x,T) = 0, \end{cases}$$

and

$$\left| \left(\hat{u} - \frac{1}{\varepsilon} \zeta(w^{\varepsilon,N}) \right)(x,t) \right| \le C \left(1 + \frac{1}{\varepsilon} \right) (1 + |x|).$$

Therefore, (7.9) implies (cf. [10, $(E.9)$, p. 207])

$$\begin{aligned}\|w^{\varepsilon,N}\|_{H^{1+\mu,(1+\mu)/2}(G')} &\le C_1 \|w^{\varepsilon,N}\|_{W^{2,1;p}(G')} \\ &\le C \left(\|w^{\varepsilon,N}\|_{L^p(G)} + \left\| \hat{u} - \frac{1}{\varepsilon} \zeta(w^{\varepsilon,N}) \right\|_{L^p(G)} \right) \le K_\varepsilon\end{aligned}$$

for $p > 4, \mu = 1 - \frac{4}{p}, C_1 = C_1(G', p), K_\varepsilon = K_\varepsilon(G, G')$, and G, G' bounded open subsets of $\mathbb{R}^2 \times [0,T]$ with $G' \subset\subset G$. Hence $w^{\varepsilon,N}$ is locally Hölder continuous in $\mathbb{R}^2 \times [0,T]$.

Now let $Q_0 \subset\subset Q$ be bounded open sets in $\mathbb{R}^2$ and let $\psi(x)$ be a smooth function with compact support in Q such that $\psi(x) = 1$ on Q_0. Then the function $w^{\varepsilon,N}(x,t)\psi(x)$ satisfies

$$\begin{cases} -\Big(w^{\varepsilon,N}(x,t)\psi(x) \Big)_t + [-\mathcal{L}^N + (\rho - b_{11})] \Big(w^{\varepsilon,N}(x,t)\psi(x) \Big) \\ \qquad = \hat{u}(x,t)\psi(x) - \dfrac{1}{\varepsilon} \zeta(w^{\varepsilon,N}(x,t))\psi(x) \\ \qquad + g\Big(x, \psi_{x_i}(x) w^{\varepsilon,N}(x,t), \psi_{x_i}(x) w_{x_j}^{\varepsilon,N}(x,t), \psi_{x_i x_j}(x) w^{\varepsilon,N}(x,t) \Big) \\ \qquad\qquad\qquad\qquad \text{a.e. } (x,t) \in Q \times [0,T], \\ w^{\varepsilon,N}(x,T)\psi(x) = 0, \end{cases}$$

with the right-hand side Hölder continuous. Hence (cf. [19, Theorem 5.1, p. 320]) $w^{\varepsilon,N}\psi \in H^{2+\mu,(2+\mu)/2}(\mathbb{R}^2 \times [0,T])$, and so also

$$w^{\varepsilon,N} \in H^{2+\mu,(2+\mu)/2}(\bar{Q_0} \times [0,T]).$$

Then, by [19, Theorem 10.1, p. 351], for all G, G' bounded open subsets of $\mathbb{R}^2 \times [0,T]$ with $G' \subset\subset G$ we also have

$$\|w^{\varepsilon,N}\|_{H^{2+\mu,(2+\mu)/2}(G')} \le K\Big(\Big\|\hat{u} - \frac{1}{\varepsilon}\zeta(w^{\varepsilon,N})\Big\|_{H^{\mu,\mu/2}(G)} + \|w^{\varepsilon,N}\|_{H^{\mu,\mu/2}(G)}\Big) \le K_\varepsilon$$

for all $N \in \mathbb{N}$ (since $\hat{u}$, $w^{\varepsilon,N}$ are locally Hölder continuous, ζ is Lipschitz continuous) and the Hölder norm above is independent of N. Thus, $\{w^{\varepsilon,N}\}_{N\in\mathbb{N}}$ lies in a compact subset of $H^{2+\mu',(2+\mu')/2}(G')$ with $\mu' < \mu$, and hence $w^{\varepsilon,N}$ converges on G' along a subsequence; i.e., there exists $\tilde{w}^\varepsilon$ such that $w^{\varepsilon,N} \to \tilde{w}^\varepsilon$ in $H^{2+\mu',(2+\mu')/2}(G')$ as $N \to \infty$. As $G' \uparrow \mathbb{R}^2 \times [0,T]$ a standard diagonalization argument yields $w^{\varepsilon,N} \to \tilde{w}^\varepsilon$ locally uniformly in $\mathbb{R}^2 \times [0,T]$ with all the x_i-, x_ix_j-, t-derivatives. It follows that $\tilde{w}^\varepsilon$ satisfies (7.6) and the growth condition (7.7).

It remains only to show that $\tilde{w}^\varepsilon \in W(0,T)$. Let $\{\psi_n\}_{n\in\mathbb{N}}$ be a sequence of nonnegative $C^2(\mathbb{R}^2)$-functions such that $\psi_n(x) = 1$ if $|x| < n$ and $\psi_n(x) = 0$ if $|x| > n+2$, with $|(\psi_n)_{x_i}| \le 1$ and $|(\psi_n)_{x_ix_j}| \le 1$. Multiplying (7.6) by $\tilde{w}^\varepsilon(x,t)\psi_n(x)$ and integrating over $\mathbb{R}^2$ yields (after an integration by parts)

$$-\frac{d}{dt}\|\tilde{w}^\varepsilon\|^2_{H_n} + A_n(\tilde{w}^\varepsilon, \tilde{w}^\varepsilon) + \int_{\mathbb{R}^2}\Big(\frac{1}{\varepsilon}\zeta(\tilde{w}^\varepsilon) - \hat{u}\Big)\tilde{w}^\varepsilon\psi_n dx + \int_{\mathbb{R}^2}\sum_{i,j}\alpha_{ij}\tilde{w}^\varepsilon_{x_i}\tilde{w}^\varepsilon\pi\psi_n dx = 0$$

for every $t \in [0,T]$, where H_n and A_n are similar to H and A but with weight $\pi\psi_n$ rather than π. By integrating by parts in the second integral above it is easy to see that it is bounded by a constant independent of n and also of ε (thanks to the uniform growth condition (7.7)); then, by the weak coercivity of A_n, the growth condition (7.7), and the usual energy inequalities, we obtain for some K_ε independent of n

$$\int_0^T \int_{\mathbb{R}^2} (\tilde{w}^\varepsilon_{x_i})^2\pi\psi_n dx \le K_\varepsilon.$$

So, by taking limits as $n \to \infty$, the monotone convergence theorem implies

$$\int_0^T \|\tilde{w}^\varepsilon\|^2_V dt \le K_\varepsilon.$$

Finally we show $\tilde{w}^\varepsilon_t \in L^2(0,T;V')$. In fact, we multiply (7.6) by $u \in L^2(0,T;V_\circ)$ and we integrate over $[0,T]$ (notice that there is no loss of generality since $V_\circ$ is dense in V); then

$$\begin{aligned}\Big\|\int_0^T (\tilde{w}^\varepsilon_t, u)dt\Big\| &\le C\Big(1 + \frac{1}{\varepsilon}\Big)\|\tilde{w}^\varepsilon\|_{L^2(0,T;V)}\|u\|_{L^2(0,T;V)} \\ &\quad + C\|\hat{u}\|_{L^2(0,T;V)}\|u\|_{L^2(0,T;V)} \\ &\le K_1(\varepsilon)\|u\|_{L^2(0,T;V)}.\end{aligned}$$

Therefore, $\tilde{w}^\varepsilon \in W(0,T)$ follows, and so also $\tilde{w}^\varepsilon \in W_\circ(0,T)$ because of the growth condition (7.7). □

Now Propositions 7.4 and 7.5 and Theorem 7.6 imply the following theorem.

THEOREM 7.7. *For every $\varepsilon > 0$ there exists a unique solution w^ε in $W_\circ(0,T)$ of the penalized problem* (7.4).

Finally, we show the existence of a solution to the strong variational problem (7.2) by taking limits as $\varepsilon \downarrow 0$.

LEMMA 7.8. *There exists $\varepsilon_o > 0$ and a positive constant $C = C(\lambda, \hat{u})$ such that for every $\varepsilon < \varepsilon_o$*

$$\frac{1}{\varepsilon}\|\zeta(w^\varepsilon(t))\|_{L^6_\pi(\mathbb{R}^2\times[0,T])} \le C.$$

This is just Lemma 8.4 (in the Appendix) with $p = 6$.

THEOREM 7.9. *The strong variational inequality* (7.2) *has a unique solution $\hat{v}$, and $e^{-\lambda(T-t)}\hat{v}$ is the unique solution of* (7.1). *Moreover, $w^\varepsilon \to \hat{v}$ in $L^2_\pi(\mathbb{R}^2 \times (0,T))$ and weakly in $L^2(0,T;V)$.*

Proof. The bounds (7.5) imply that there exists a function $\hat{v} \in W(0,T)$ with $\hat{v}_t \in L^\infty(0,T;H)$ such that, along a subsequence,

$$\lim_{\varepsilon\downarrow 0} w^\varepsilon = \hat{v} \quad \text{ weakly in } L^2(0,T;V).$$

On the other hand, w^ε solves (7.6), i.e.,

$$-w^\varepsilon_t + [-\mathcal{L} + (\rho - b_{11})]w^\varepsilon = \hat{u}(x,t) - \frac{1}{\varepsilon}\zeta(w^\varepsilon),$$

and (8.7) implies that the right-hand side above is bounded in $L^6_\pi(\mathbb{R}^2 \times [0,T])$, uniformly in $\varepsilon < \varepsilon_0$. Hence if Q is a bounded open set in $\mathbb{R}^2$ we have

$$\left\| \hat{u} - \frac{1}{\varepsilon}\zeta(w^\varepsilon) \right\|_{L^6(Q\times[0,T])} \le C;$$

then, if G, G' are bounded open sets in $\mathbb{R}^2 \times [0,T]$ with $G' \subset\subset G$, it follows from [19, (10.12), p. 355] that

$$\|w^\varepsilon\|_{W^{2,1;6}(G')} \le C\Big(\left\| \hat{u} - \frac{1}{\varepsilon}\zeta(w^\varepsilon) \right\|_{L^6(G)} + \|w^\varepsilon\|_{L^6(G)} \Big) \le C$$

with $C = C(G)$, because of the uniform growth condition (7.7). So also (cf. [10, $(E.9)$, p. 207])

$$\|w^\varepsilon\|_{H^{1+\mu,(1+\mu)/2}(G')} \le C_1 \|w^\varepsilon\|_{W^{2,1;6}(G')} \le C_2$$

for $\mu = \frac{1}{3}, C_1 = C_1(G'), C_2 = C_2(G)$; in particular,

$$\|w^\varepsilon\|_{H^{\mu,\mu/2}(G')} \le C_2,$$

hence w^ε lies in a compact subset of $H^{\mu',\mu'/2}(G')$ with $\mu' < \mu$. This implies that (along a subsequence) w^ε converges to u uniformly in G'; thus, w^ε converges to u locally uniformly in $\mathbb{R}^2 \times [0,T]$ and hence

$$\lim_{\varepsilon\downarrow 0} w^\varepsilon = \hat{v} \quad \text{ in } L^2_\pi(\mathbb{R}^2 \times (0,T)).$$

It remains to show that $\hat{v}$ solves (7.2). Let $w \in W^{\mathcal{K}}(0,T)$; then $w(t) - w^\varepsilon(t) \in V_\circ$ a.e. $t \in [0,T]$ because of Theorem 7.7. Since the simple functions are dense in $L^2(0,T;V)$, then we may set $g = w(t) - w^\varepsilon(t)$ in (7.4) and integrate over $(t,t']$ to obtain

$$\begin{aligned} &-\int_t^{t'} (w_t(s), w(s) - w^\varepsilon(s))ds \\ &\quad + \frac{1}{2}\Big(\|w(t') - w^\varepsilon(t')\|_H^2 - \|w(t) - w^\varepsilon(t)\|_H^2\Big) \\ &\quad + \int_t^{t'} \Big[A(w^\varepsilon(s), w(s)) - A(w^\varepsilon(s), w^\varepsilon(s))\Big]ds \\ &\geq \int_t^{t'} (\hat{u}(s), w(s) - w^\varepsilon(s))ds \end{aligned} \tag{7.11}$$

after taking into account that

$$(\zeta(w^\varepsilon(t)), w(t) - w^\varepsilon(t)) = -(\zeta(w(t)) - \zeta(w^\varepsilon(t)), w(t) - w^\varepsilon(t)) \leq 0$$

because ζ is monotone nondecreasing and $\zeta(w(t)) = 0$ a.e. $t \in [0,T]$. Now, using arguments similar to those employed for the weak variational formulation (cf. Theorem 5.2), we may pass to the limit in (7.11) to obtain

$$\begin{aligned} -\int_t^{t'} (\hat{v}_t(s), w(s) - \hat{v}(s))ds + \int_t^{t'} \Big[A(\hat{v}(s), w(s)) - A(\hat{v}(s), \hat{v}(s))\Big]ds \\ \geq \int_t^{t'} (\hat{u}(s), w(s) - \hat{v}(s))ds. \end{aligned}$$

Since t and t' are arbitrary, we deduce that

$$-(\hat{v}_t(t), g - \hat{v}(t)) + A(\hat{v}(t), g) - A(\hat{v}(t), \hat{v}(t)) \geq (\hat{u}(t), g - \hat{v}(t)) \tag{7.12}$$

for all $g \in \mathcal{K}$ and almost every $t \in [0,T]$.

Also, $(7.5)_3$ implies (again along a subsequence)

$$\lim_{\varepsilon \downarrow 0} \|\zeta(w^\varepsilon)\|_{L^2(0,T;H)} = 0,$$

and, since ζ is Lipschitz, then

$$\lim_{\varepsilon \downarrow 0} \|\zeta(w^\varepsilon) - \zeta(\hat{v})\|_{L^2(0,T;H)} = 0;$$

i.e., $\zeta(\hat{v}(x,t)) = 0$ a.e., so $\hat{v} \in W^{\mathcal{K}}(0,T)$. It follows that $A(\hat{v},\hat{v})$ can be defined by (5.2), hence $A(\hat{v}(t), g) - A(\hat{v}(t), \hat{v}(t)) = A(\hat{v}(t), \hat{v}(t) - g)$ a.e., and hence $\hat{v}$ is a solution of the strong variational inequality (7.2).

Uniqueness and the remainder of the results follow from Lemma 7.3. □

The fact that the function $\hat{v}$ is the unique solution of the strong variational inequality (7.2) says nothing about the derivatives of order two, $(\hat{v})_{x_i x_j}$; however, the approximating functions w^ε and their estimates (7.5) will allow us to obtain more regularity of $\hat{v}$ and hence a variational inequality in the a.e. sense.

LEMMA 7.10. *Assume* $\rho \geq b_{11}$. *Then*

$$\hat{v} \in W^{2,1;6}_{\text{loc}}(\mathbb{R}^2 \times (0,T)). \tag{7.13}$$

Proof. From (8.7) it follows that for some constant C and for all $\varepsilon < \varepsilon_0$

$$\begin{aligned}&\| - w_t^\varepsilon + [-\mathcal{L} + (\rho - b_{11})]w^\varepsilon \|_{L^6_\pi(\mathbb{R}^2\times[0,T])}\\ &= \left\| -\frac{1}{\varepsilon}\zeta(w^\varepsilon) + \hat{u}(t) \right\|_{L^6_\pi(\mathbb{R}^2\times[0,T])} \le C,\end{aligned}$$

and after taking limits as $\varepsilon \downarrow 0$ (because of (7.5))

$$\| - \hat{v}_t + [-\mathcal{L} + (\rho - b_{11})]\hat{v} \|_{L^6_\pi(\mathbb{R}^2\times[0,T])} \le C.$$

Thus $-\hat{v}_t + [-\mathcal{L} + (\rho - b_{11})]\hat{v} := \tilde{u} \in L^6_\pi(\mathbb{R}^2 \times [0,T])$.

Observe that if ψ is a function in $C^\infty(\mathbb{R}^2)$ with compact support in B_N (for some $N \in \mathbb{N}$), then $w^\varepsilon\psi$ also satisfies (cf. (7.4))

$$\begin{cases} -(Z_t(t), g) + A(Z(t), g) \\ \qquad = (\psi[\hat{u}(t) - \frac{1}{\varepsilon}\zeta(w^\varepsilon)], g) - (w^\varepsilon \mathcal{L}\psi + 2(\nabla w^\varepsilon)^* \alpha \nabla\psi, g) \\ \qquad\qquad\qquad\qquad \text{a.e. } t \in [0,T] \quad \forall g \in V_\circ, \\ Z(T) = 0, \end{cases} \tag{7.14}$$

and taking the limit in (7.14) shows that $z := \psi\hat{v}$ satisfies

$$\begin{cases} -(Z_t(t), g) + A(Z(t), g) \\ \qquad = (\psi\tilde{u}(t), g) - (\hat{v}\mathcal{L}\psi + 2(\nabla\hat{v})^* \alpha \nabla\psi, g) \\ \qquad\qquad\qquad\qquad \text{a.e. } t \in [0,T] \quad \forall g \in V_\circ, \\ Z(T) = 0, \end{cases} \tag{7.15}$$

and this equation has a unique solution in $W(0,T)$. Moreover, the unique solution in $W^{2,1;6}(B_N \times (0,T))$ of

$$\begin{cases} -Z_t + [-\mathcal{L} + (\rho - b_{11})]Z = \psi\tilde{u} - \hat{v}\mathcal{L}\psi - 2(\nabla\hat{v})^*\alpha\nabla\psi, \\ Z(T) = 0, \\ Z|_{\partial B_N \times (0,T)} = 0, \end{cases}$$

satisfies (7.15), and hence is z. We conclude that $\hat{v} \in W^{2,1;6}_{\text{loc}}(\mathbb{R}^2 \times (0,T))$. ∎

THEOREM 7.11. *Assume that $\rho \ge b_{11}$. Then the function $\hat{v}$ is the unique solution of the pointwise variational inequality*

$$\begin{cases} \hat{v} \in W^{2,1;6}_{\text{loc}}(\mathbb{R}^2 \times (0,T)), \\ -1 \le \hat{v} \le +1 \ \textit{a.e. in} \ \mathbb{R}^2 \times (0,T), \ \hat{v}(x,T) = 0 \ \textit{a.e. in} \ \mathbb{R}^2, \\ -\hat{v}_t + [-\mathcal{L} + (\rho - b_{11})]\hat{v} = \hat{u} \quad \textit{if} \quad -1 < \hat{v}(x,t) < +1, \\ -\hat{v}_t + [-\mathcal{L} + (\rho - b_{11})]\hat{v} \le \hat{u} \quad \textit{if} \quad \hat{v}(x,t) = +1, \\ -\hat{v}_t + [-\mathcal{L} + (\rho - b_{11})]\hat{v} \ge \hat{u} \quad \textit{if} \quad \hat{v}(x,t) = -1, \\ \qquad\qquad \textit{a.e. in} \ \mathbb{R}^2 \times (0,T). \end{cases} \tag{7.16}$$

Proof. We know that $\hat{v}$ is the unique solution of (7.2) by Theorem 7.9; due to the regularity of $\hat{v}$ we may write (7.2) as

$$\begin{cases} \hat{v}(T) = 0, \\ (-\hat{v}_t(t) + [-\mathcal{L} + (\rho - b_{11})]\hat{v}(t) - \hat{u}(t), g - \hat{v}(t)) \ge 0 \\ \qquad\qquad\qquad\qquad \text{a.e. } t \in [0,T] \quad \forall g \in \mathcal{K}, \end{cases}$$

and by taking $g = \hat{v}(t) + \psi$ with $\psi \in C^\infty(\mathbb{R}^2)$ with compact support, we obtain (7.16). ∎

8. Appendix. The assumptions are as in section 7. We aim first to prove Proposition 7.5. Due to the lack of a priori bounds of the type (4.9), we take care of the linear-in-x term in A by regularization, so we set

- $A_1(g,h) = (x^*b^*\nabla g, x^*b^*\nabla h)$ for $g,h \in V$;
- $\mathcal{V} = \{g \in V : A_1(g,g) < +\infty\}$ with the Hilbert norm

$$\|g\|_{\mathcal{V}}^2 = \|g\|_V^2 + A_1(g,g)$$

(notice that $\mathcal{V} \subset H \subset \mathcal{V}'$ with continuous dense injection);

- $\mathcal{W}(0,T) = \{w \in L^2(0,T;\mathcal{V}) : w_t \in L^2(0,T;\mathcal{V}')\}$.

Clearly, for $\gamma > 0$ the continuous bilinear form $A + \gamma A_1 : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ verifies

$$A(g,g) + \gamma A_1(g,g) \geq \eta\|g\|_V^2 - \lambda\|g\|_H^2 + \gamma\|x^*b^*\nabla g\|_H^2$$

since A is weakly coercive in V.

Since $\mathcal{V}$ is dense in V, in the limit as $\gamma \downarrow 0$ we will obtain w^ε from the solution of the following γ-approximating problem.

THEOREM 8.1. *For every $\gamma > 0$ there exists a unique solution $w^{\varepsilon,\gamma}$ in $\mathcal{W}(0,T)$ of*

$$\begin{cases} -(w_t(t),g) + A(w(t),g) + \gamma A_1(w(t),g) + \dfrac{1}{\varepsilon}(\zeta(w(t)),g) \\ \qquad\qquad = (\hat{u}(t),g) \qquad \text{a.e. } t \in [0,T] \quad \forall g \in \mathcal{V}, \\ w(T) = 0. \end{cases} \tag{8.1}$$

Proof. We use Galerkin's method. Let $\{\psi_i\}_{i\in\mathbb{N}}$ be a complete orthonormal basis of the separable space H such that $\{\psi_i\}_{i\in\mathbb{N}} \subset \mathcal{V}$ and set

$$w^m(x,t) = \sum_{i=1}^m G_i^m(t)\psi_i(x), \qquad m \in \mathbb{N},$$

where the scalar functions $G_i^m(t)$ are such that

$$\begin{cases} -(w_t^m(t),\psi_j) + A(w^m(t),\psi_j) + \gamma A_1(w^m(t),\psi_j) \\ \qquad + \dfrac{1}{\varepsilon}(\zeta(w^m(t)),\psi_j) = (\hat{u}(t),\psi_j), \qquad j = 1,\dots,m, \\ w^m(T) = 0. \end{cases} \tag{8.2}$$

Now $\zeta^2(r) \leq r\zeta(r)$ and standard energy estimates show that there exists a constant $C > 0$ such that for all $m \in \mathbb{N}$

$$\begin{aligned} \|w^m\|_{L^\infty(0,T;H)}^2 + \|w^m\|_{L^2(0,T;V)}^2 + \frac{1}{\varepsilon}\|\zeta(w^m)\|_{L^2(0,T;H)}^2 \\ + \gamma\int_0^T A_1(w^m(t),w^m(t))dt \leq C; \end{aligned} \tag{8.3}$$

see [2, p. 125] or [1, p. 80] for the idea.

We recall that $\hat{u}_t \in L^\infty(0,T;V')$ and $\psi_j \in \mathcal{V} \subset V$, so $\ddot{G}_i^m$ exists and we have $w_{tt}^m(x,t) = \sum_{i=1}^m \ddot{G}_i^m(t)\psi_i(x)$; thus, from (8.2) it follows that

$$\begin{cases} -(w_{tt}^m(t),\psi_j) + A(w_t^m(t),\psi_j) + \gamma A_1(w_t^m(t),\psi_j) + \dfrac{1}{\varepsilon}(\zeta'(w^m(t))w_t^m(t),\psi_j) \\ \qquad\qquad = \langle \hat{u}_t(t),\psi_j\rangle, \quad j = 1,\dots,m, \\ -(w_t^m(T),\psi_j) = (\hat{u}(T),\psi_j), \qquad\qquad j = 1,\dots,m. \end{cases}$$

Again, standard estimates show that there exists a constant $C > 0$ such that for all $m \in \mathbb{N}$

$$\|w_t^m\|^2_{L^\infty(0,T;H)} + \|w_t^m\|^2_{L^2(0,T;V)} + \frac{1}{\varepsilon}\int_0^T \left\|\frac{d}{dt}\zeta(w^m(t))\right\|^2_H dt + \gamma\int_0^T A_1(w_t^m(t), w_t^m(t))dt \le C. \tag{8.4}$$

It follows from (8.3) and (8.4) that we can extract a subsequence such that

$$\begin{cases} w^m \to w^{\varepsilon,\gamma} \text{ weakly in } L^2(0,T;V) \text{ and weak-}\star \text{ in } L^\infty(0,T;H); \\ w_t^m \to w_t^{\varepsilon,\gamma} \text{ weakly in } L^2(0,T;V) \text{ and weak-}\star \text{ in } L^\infty(0,T;H); \\ \sqrt{\gamma}x^*b^*\nabla w^m \to \sqrt{\gamma}x^*b^*\nabla w^{\varepsilon,\gamma} \text{ weakly in } L^2(0,T;H); \\ \sqrt{\gamma}x^*b^*\nabla w_t^m \to \sqrt{\gamma}x^*b^*\nabla w_t^{\varepsilon,\gamma} \text{ weakly in } L^2(0,T;H); \\ \dfrac{1}{\sqrt{\varepsilon}}\zeta(w^m) \to \dfrac{1}{\sqrt{\varepsilon}}\chi \text{ weakly in } L^2(0,T;H), \end{cases}$$

for some $w^{\varepsilon,\gamma}$ in $\mathcal{W}(0,T)$ and χ in $L^2(0,T;H)$.

To identify the function χ we digress to the following lemma.

LEMMA 8.2. *Let $\{v^m\}$ be a sequence converging weakly in $L^2(0,T;V)$ to v such that*

$$\begin{cases} \|v^m\|^2_{L^2(0,T;V)} \le C, \\ \|v_t^m\|^2_{L^\infty(0,T;H)} \le C. \end{cases} \tag{8.5}$$

Then there exists a subsequence $\{v^{m_k}\}$ converging a.e. to v on $\mathbb{R}^2 \times (0,T)$.

Proof. Fix $N \in \mathbb{N}$ and recall that $\Omega_N = B_N \times (0,T)$ satisfies a uniform interior cone condition. On Ω_N we can drop the weight function π since $\|g\|_{L^2_\pi(B_N)} \le \|g\|_{L^2(B_N)} \le (1+N^2)^l \|g\|_{L^2_\pi(B_N)}$, so from (8.5) it follows that $\|v^m\|_{W^{1,1;2}(\Omega_N)} \le C$; that is, the sequence $\{v^m\}_{m\in\mathbb{N}}$ lies in a bounded set in $W^{1,1;2}(\Omega_N)$ (cf. [13, p. 158 and Problem 7.14]), hence in a compact set in $L^2(\Omega_N)$. Thus a subsequence converges strongly, and the limit must be v, so the whole sequence converges to v in $L^2(\Omega_N)$. Then a subsequence converges to v a.e. on Ω_N, and hence, since N is arbitrary, on $\mathbb{R}^2 \times (0,T)$. ∎

We return now to the proof of the theorem. It follows from the lemma that, at least for a subsequence, again denoted $\{w^m\}$, $w^m \to w^{\varepsilon,\gamma}$ a.e. The bound $|r - \zeta(r)| \le 1$ for all $r \in \mathbb{R}$ allows us to apply the bounded convergence theorem and get

$$\lim_{m\to\infty}(w^m - \zeta(w^m)) = w^{\varepsilon,\gamma} - \zeta(w^{\varepsilon,\gamma}) \quad \text{in } L^2_\pi(\mathbb{R}^2 \times (0,T))$$

since ζ is continuous. But it also converges weakly to $w^{\varepsilon,\gamma} - \chi$; therefore, $\chi = \zeta(w^{\varepsilon,\gamma})$.

Now we go back to (8.2), multiply it by any $\mu(t) \in L^2(0,T)$, and integrate over $[0,T]$ and

$$\int_0^T -(w_t^{\varepsilon,\gamma}(t), \psi_j)\mu(t)dt + \int_0^T [A + \gamma A_1](w^{\varepsilon,\gamma}(t), \psi_j)\mu(t)dt + \int_0^T \frac{1}{\varepsilon}(\zeta(w^{\varepsilon,\gamma}(t)), \psi_j)\mu(t)dt = \int_0^T (\hat{u}(t), \psi_j)\mu(t)dt$$

for every $\mu(t) \in L^2(0,T)$ and $j = 1, 2, \ldots$. It follows that $w^{\varepsilon,\gamma}$ is a solution of (8.1). Moreover, (8.3) and (8.4) imply that for some $C < \infty$ and all $\varepsilon,\ \gamma > 0$,

$$\begin{cases} \|w^{\varepsilon,\gamma}\|^2_{L^\infty(0,T;H)} + \|w^{\varepsilon,\gamma}\|^2_{L^2(0,T;V)} \le C, \\ \|w_t^{\varepsilon,\gamma}\|^2_{L^\infty(0,T;H)} + \|w_t^{\varepsilon,\gamma}\|^2_{L^2(0,T;V)} \le C, \\ \sqrt{\gamma}\|x^* b^* \nabla w^{\varepsilon,\gamma}\|_{L^2(0,T;H)} + \dfrac{1}{\sqrt{\varepsilon}}\|\zeta(w^{\varepsilon,\gamma})\|_{L^2(0,T;H)} \le C. \end{cases} \tag{8.6}$$

Uniqueness follows as in Proposition 7.4. □

We now take limits as $\gamma \downarrow 0$ and we have the following proposition.

PROPOSITION 8.3. *There exists a solution $w^\varepsilon \in W(0,T)$ of the penalized variational problem* (7.4) *and $w^{\varepsilon,\gamma}(x,t) \to w^\varepsilon(x,t)$ a.e. Moreover, w^ε satisfies* (8.6).

Proof. From (8.6) it follows that

$$\int_0^T |A_1(w^{\varepsilon,\gamma}(t), w(t))|dt \le \frac{C}{\sqrt{\gamma}}\|x^* b^* \nabla w\|_{L^2(0,T;H)}$$

for every $w \in \mathcal{V}$; then, after multiplying (8.1) by $\mu(t) \in L^2(0,T)$ and integrating over $[0,T]$, we may take limits as $\gamma \downarrow 0$ (along a subsequence if necessary) to get

$$\begin{aligned} &\int_0^T -(w_t^\varepsilon(t), g)\mu(t)dt + \int_0^T A(w^\varepsilon(t), g)\mu(t)dt \\ &+ \int_0^T \frac{1}{\varepsilon}(\zeta(w^\varepsilon(t)), g)\mu(t)dt = \int_0^T (\hat{u}(t), g)\mu(t)dt \end{aligned}$$

for some $w^\varepsilon \in W(0,T)$ whose existence is guaranteed by the estimates (8.6). In fact, just as above, $\frac{1}{\varepsilon}\zeta(w^{\varepsilon,\gamma})$ converges weakly in $L^2(0,T;H)$ to some function which is then identified as $\frac{1}{\varepsilon}\zeta(w^\varepsilon)$, and $w^{\varepsilon,\gamma}(x,t) \to w^\varepsilon(x,t)$ a.e. thanks to Lemma 8.2. Since $\mu(t)$ is arbitrary in $L^2(0,T)$, we obtain (7.4) for every $g \in \mathcal{V}$, and hence for all $g \in V_\circ$ since $\mathcal{V}$ is dense in $V_\circ$. □

The estimates (7.5) follow from (8.6).

We must also prove Lemma 7.8; it is a special case of the following result.

LEMMA 8.4. *Assume that $2 \le p < 2l - 2$, p even. Then there exists $\varepsilon_o > 0$ and a positive constant $C = C(\lambda, \hat{u})$ such that for every $\varepsilon < \varepsilon_o$*

$$\frac{1}{\varepsilon}\|\zeta(w^\varepsilon(t))\|_{L^p_\pi(\mathbb{R}^2 \times [0,T])} \le C. \tag{8.7}$$

Proof. Recall that $|\zeta(r)| \le |r|$ for all $r \in \mathbb{R}$, hence

$$|\zeta(w^\varepsilon(x,t))| \le |w^\varepsilon(x,t)| \le K(1 + |x|)$$

by (7.7). The relationship between l and p insures that x, hence $\zeta(w^\varepsilon(t))$, is in $L^p_\pi(\mathbb{R}^2)$, and allows us to multiply (7.6) by $(\zeta(w^\varepsilon(t)))^{p-1}\pi$ (recall that $w^\varepsilon_{x_i x_j}$ is well defined because of Theorem 7.6). Also notice that $r\zeta(r) = (\zeta(r))^2 + |\zeta(r)|$. Then, after an integration over $\mathbb{R}^2$ we obtain

$$\begin{aligned} &-\frac{1}{p}\frac{d}{dt}\|\zeta(w^\varepsilon(t))\|^p_{L^p_\pi(\mathbb{R}^2)} \\ &\quad +(p-1)\sum_{i,j=1}^2 \alpha_{ij}([\zeta(w^\varepsilon(t))]^{p/2-1}\zeta(w^\varepsilon(t))_{x_i}, [\zeta(w^\varepsilon(t))]^{p/2-1}\zeta(w^\varepsilon(t))_{x_j}) \end{aligned}$$

$$
\begin{aligned}
&-\sum_{i,j=1}^{2}\alpha_{ij}([\zeta(w^\varepsilon(t))]^{p/2-1}(\zeta(w^\varepsilon(t)))_{x_i},[\zeta(w^\varepsilon(t))]^{p/2}\varphi_j)\\
&-\sum_{i=1}^{2}([a+bx]_i[\zeta(w^\varepsilon(t))]^{p/2-1}(\zeta(w^\varepsilon(t)))_{x_i},[\zeta(w^\varepsilon(t))]^{p/2})\\
&+(\rho-b_{11})([\zeta(w^\varepsilon(t))]^2+|\zeta(w^\varepsilon(t))|,[\zeta(w^\varepsilon(t))]^{p-2})\\
&+\frac{1}{\varepsilon}\|\zeta(w^\varepsilon(t))\|^p_{L^p_\pi(\mathbb{R}^2)}\\
&=(\hat u(t),[\zeta(w^\varepsilon(t))]^{p-1}).
\end{aligned}
$$

Integration by parts can be used to remove all terms involving partial derivatives except the second term in the above equation, but its integrand is nonnegative, so the term is well defined. Hence

$$
\begin{aligned}
&-\frac{1}{p}\frac{d}{dt}\|\zeta(w^\varepsilon(t))\|^p_{L^p_\pi(\mathbb{R}^2)}\\
&\quad+\frac{2}{p}A([\zeta(w^\varepsilon(t))]^{p/2},[\zeta(w^\varepsilon(t))]^{p/2})\\
&\quad+\Big(1-\frac{2}{p}\Big)\frac{2}{p}\int_{\mathbb{R}^2}\nabla\Big([\zeta(w^\varepsilon(t))]^{p/2}\Big)^*\alpha\nabla\Big([\zeta(w^\varepsilon(t))]^{p/2}\Big)\pi dx\\
&\quad+\Big(1-\frac{2}{p}\Big)(\rho-b_{11})\|\zeta(w^\varepsilon(t))\|^p_{L^p_\pi(\mathbb{R}^2)}\\
&\quad+(\rho-b_{11})\|\zeta(w^\varepsilon(t))\|^{p-1}_{L^{p-1}_\pi(\mathbb{R}^2)}+\frac{1}{\varepsilon}\|\zeta(w^\varepsilon(t))\|^p_{L^p_\pi(\mathbb{R}^2)}\\
&=(\hat u(t),[\zeta(w^\varepsilon(t))]^{p-1}).
\end{aligned}
$$

The weak coercivity of A, integration over $(t,T]$, and Young's inequality give

$$
\begin{aligned}
\frac{1}{\varepsilon}\delta\int_t^T\|\zeta(w^\varepsilon(s))\|^p_{L^p_\pi(\mathbb{R}^2)}ds&\le\int_t^T\|\hat u(s)\|_{L^p_\pi(\mathbb{R}^2)}\|\zeta(w^\varepsilon(s))\|^{p-1}_{L^p_\pi(\mathbb{R}^2)}ds\\
&\le\frac{1}{p}\gamma^p\int_t^T\|\hat u(s)\|^p_{L^p_\pi(\mathbb{R}^2)}ds\\
&\quad+\frac{1}{2\varepsilon}\delta\int_t^T\|\zeta(w^\varepsilon(s))\|^p_{L^p_\pi(\mathbb{R}^2)}ds,
\end{aligned}
$$

where $\delta=1-\frac{2}{p}\varepsilon\lambda$ and $\gamma=[\frac{2\varepsilon}{p-2\varepsilon\lambda}(p-1)]^{(p-1)/p}$. ($\lambda$ comes from the weak coercivity of A.) So

$$
\Big(\frac{1}{\varepsilon}\Big)^p\int_t^T\|\zeta(w^\varepsilon(s))\|^p_{L^p_\pi(\mathbb{R}^2)}ds\le\frac{1}{p}\Big(\frac{p-1}{p}\Big)^{p-1}\Big(\frac{2}{\delta}\Big)^p\int_t^T\|\hat u(s)\|^p_{L^p_\pi(\mathbb{R}^2)}ds;
$$

i.e., for some $\varepsilon_0>0$ and all $\varepsilon<\varepsilon_0$ we have

$$
\frac{1}{\varepsilon}\|\zeta(w^\varepsilon(t))\|_{L^p_\pi(\mathbb{R}^2\times[0,T])}\le C(\lambda,p,\hat u),
$$

and hence (8.7) follows. □

REFERENCES

[1] A. Bensoussan, *Stochastic Control by Functional Analysis Methods*, North-Holland, Amsterdam, 1982.

[2] A. Bensoussan and J. L. Lions, *Applications of Variational Inequalities in Stochastic Control*, North-Holland, Amsterdam, 1982.

[3] A. Bensoussan and J. L. Lions, *Contrôle Impulsionnel et Inéquations Quasi Variationnelles*, Dunod, Paris, 1982.

[4] H. N. Binhammer, *Money, Banking and the Canadian Financial System*, 4th ed., Methuen, Toronto, Canada, 1982.

[5] K. Chan, A. Karolyi, F. Longstaff, and A. Sanders, *An empirical comparison of alternative models of the short-term interest rates*, J. Finance, 47 (1992), pp. 1209–1227.

[6] M. B. Chiarolla and U. G. Haussmann, *The optimal control of the cheap monotone follower*, Stochastics Stochastics Rep., 49 (1994), pp. 99–128.

[7] P. L. Chow, J. L. Menaldi, and M. Robin, *Additive control of stochastic linear systems with finite horizon*, SIAM J. Control Optim., 23 (1985), pp. 858–899.

[8] J. C. Cox, J. E. Ingersoll, and S. A. Ross, *A theory of the term structure of interest rates*, Econometrica, 53 (1985), pp. 385–407.

[9] N. El Karoui and I. Karatzas, *Probabilistic aspects of finite-fuel, reflected follower problems*, Acta Appl. Math., 11 (1988), pp. 223–258.

[10] W. H. Fleming and R. W. Rishel, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.

[11] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.

[12] G. Fusai, *Monetary Policy and Term Structure of Interest Rates*, preprint 1995.

[13] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, New York, 1983.

[14] B. Han, U. G. Haussmann, *Interest Rates and Inflation: Parameter Identification and Control in a Model using Canadian Data*, preprint 1998.

[15] U. G. Haussmann and W. Suo, *Singular optimal stochastic controls* I, II, SIAM J. Control Optim., 33 (1995), pp. 916–936, 937–959.

[16] I. Karatzas, *A class of singular stochastic control problems*, Adv. in Appl. Probals., 15 (1983), pp. 225–254.

[17] I. Karatzas and S. S. Shreve, *Connections between optimal stopping and singular stochastic control* II. *Reflected follower problems*, SIAM J. Control Optim., 23 (1985), pp. 433–451.

[18] N. V. Krylov, *Controlled Diffusion Processes*, Springer-Verlag, New York, 1980.

[19] O. A. Ladyženskaja, V. A. Solonnikov, and N. N. Ural'ceva, *Linear and Quasilinear Equations of Parabolic Type*, AMS, Providence, RI, 1968.

[20] F. A. Longstaff and E. S. Schwartz, *Interest rate volatility and the term structure: A two-factor general equilibrium model*, J. Finance, 47 (1992), pp. 1259–1282.

[21] J. L. Menaldi and M. Robin, *On optimal correction problems with partial information*, Stochastic Anal. Appl., 3 (1985), pp. 63–92.

[22] F. S. Mishkin, *Money, Interest Rates and Inflation*, Edward Elgar, Aldershot, UK, 1993.

[23] G. G. Pennacchi, *Identifying the dynamics of real interest rates and inflation: Evidence using survey data*, Rev. Financial Studies, 4 (1991), pp. 53–86.

[24] S. F. Richard, *An arbitrage model of the term structure of interest rates*, J. Financial Economics, 6 (1978), pp. 33–57.

[25] S. J. Turnovsky, *The term structure of interest rates and the effects of macroeconomic policy*, J. Money, Credit and Banking, 21 (1989), pp. 321–347.

[26] G. Vaciago, *Banca Centrale tra governo e mercato*, XXXII Riunione Scientifica Annuale della Società Italiana degli Economisti: "Il Ruolo della Banca Centrale nella Politica Economica," il Mulino, Bologna, 1992.

[27] S. A. Williams, P. L. Chow, and J. L. Menaldi, *Regularity of the free boundary in singular stochastic control*, J. Differential Equations, 111 (1994), pp. 175–201.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 4, pp. 1133–1146, July 1998

001

CAUSAL INPUT/OUTPUT REPRESENTATION OF 2D SYSTEMS IN THE BEHAVIORAL APPROACH*

SANDRO ZAMPIERI†

Abstract. In this paper the concept of causal input/output representation of two-dimensional (2D) systems in the behavioral approach is introduced. These representations provide an interesting connection between classical 2D systems theory and 2D systems theory in the behavioral approach. Some characterizations of such representations are presented. Finally, a technique that allows us to obtain causal input/output representations is proposed.

Key words. 2D systems, behavioral approach, input/output representation, proper rational matrices, causality

AMS subject classifications. 93A30, 93B25

PII. S0363012995292573

1. Introduction.

The theory of dynamical systems in the behavioral approach has attracted much attention in the last few years. The main characteristic of this approach is that the system variables are not a priori divided into inputs and outputs, and moreover, no causality structure is imposed on the dynamics. This division in inputs and outputs is something that can be obtained a posteriori, analyzing the system equations. This feature is useful above all when there is an unclear distinction between causes and effects, and this happens frequently in the study of multidimensional systems, for instance, systems operating on space-time signals. Moreover, it has been shown recently that in modeling and identification procedures it is in some cases more reasonable to avoid distinctions between inputs and outputs.

On the other hand, if we want to apply the classical control and filtering strategies in this setup, the extraction of the input/output structure becomes an essential step. In the one-dimensional (1D) case this problem has been completely solved by Willems in [10]. In that paper Willems introduces a notion of input/output representation that takes into account a causality relation between inputs and outputs with respect to a time direction. Such representations are obtained exploiting the properties of row proper polynomial matrices. In the two-dimensional (2D) case this has been done only partially in [6, 12, 5, 13], since the input/output representations proposed there do not obey any causality assumption.

In this paper we will propose input/output representations satisfying weakly causal relations between inputs and outputs [1]. In the 1D case [10] these input/output representations are called nonanticipating. In the 2D case we prefer to call them causal input/output representations, which seems to be a more suitable terminology for multidimensional systems. We will analyze the properties of such representations and introduce the concept of impulse response that, as we will show, is strictly connected with the controllable part of the system. Moreover, we will propose a method for verifying if an input/output representation is causal, and finally we will present an algorithm that allows us to obtain causal input/output representations.

*Received by the editors October 2, 1995; accepted for publication (in revised form) May 6, 1997; published electronically May 15, 1998.

http://www.siam.org/journals/sicon/36-4/29257.html

†Dipartimento di Elettronica ed Informatica, Università di Padova, Via Gradenigo 6/a, 35131 Padova, Italy (zampi@paola.dei.unipd.it).

During the review process, the possibility of extending the results presented in this paper to the n-dimensional (nD) case has been pointed out by one of the referees.

2. 2D systems theory in the input/output and in the behavioral approach. This section will be devoted to the introduction of some preliminaries on 2D systems theory both in the behavioral approach and in the classical input/output approach. As far as the behavioral approach is concerned, we refer to [6, 8, 12], while for the classical input/output approach, we refer to [2].

In the behavioral approach, a dynamical system is described by a triple

$$\Sigma = (T, W, \mathcal{B}),$$

where T is the time set, W is the signal alphabet and $\mathcal{B}$ is a subset of the set of all signals W^T that is called the *behavior* of the system. In this paper we are interested in the so-called discrete 2D linear shift-invariant complete systems. Discrete 2D systems are dynamical systems whose time set is $\mathbb{Z}^2$ and signal alphabet is $\mathbb{R}^q$ ($\mathbb{R}$ can be replaced by any field). A discrete 2D system is linear shift-invariant and complete if, moreover, the behavior $\mathcal{B}$ is a linear subspace of the vector space $(\mathbb{R}^q)^{\mathbb{Z}^2}$ which is invariant with respect to the backward shifts σ_1, σ_2 in the two directions of $\mathbb{Z}^2$ and satisfies the following requirement:

$$w \in \mathcal{B} \quad \Leftrightarrow \quad w_{|I} \in \mathcal{B}_{|I} \quad \text{for all finite } I \subset \mathbb{Z}^2.$$

Notice that, as shown in [6, 8], the behavior of a linear shift-invariant system is complete if and only if it is a closed set with respect to the pointwise convergence topology in $(\mathbb{R}^q)^{\mathbb{Z}^2}$.

It can be shown that a discrete 2D system is linear shift-invariant complete if and only if its behavior can be specified by a difference equation. More precisely, given any polynomial matrix $R \in \mathbb{R}[z_1, z_2, z_1^{-1}, z_2^{-1}]^{l \times q}$ (the symbol $\mathbb{R}[z_1, z_2, z_1^{-1}, z_2^{-1}]$ denotes the ring of Laurent polynomials that are polynomials where the indeterminates may have negative exponent), then R can be written as follows:

$$\sum_{(i,j) \in S} R_{ij} z_1^i z_2^j,$$

where $R_{ij} \in \mathbb{R}^{l \times q}$ and S is a finite subset of $\mathbb{Z}^2$. Then we can associate with R an operator from $(\mathbb{R}^q)^{\mathbb{Z}^2}$ to $(\mathbb{R}^l)^{\mathbb{Z}^2}$, denoted as $R(\sigma_1, \sigma_2)$, operating as follows:

$$(R(\sigma_1, \sigma_2)w)(h, k) := \sum_{(i,j) \in S} R_{ij} w(h + i, k + j), \quad (h, k) \in \mathbb{Z}^2.$$

These operators will be called difference operators. A 2D system $\Sigma = (\mathbb{Z}^2, \mathbb{R}^q, \mathcal{B})$ is a linear shift-invariant and complete system if and only if there exists a positive integer l and a polynomial matrix $R \in \mathbb{R}[z_1, z_2, z_1^{-1}, z_2^{-1}]^{l \times q}$ such that

$$\mathcal{B} = \ker R(\sigma_1, \sigma_2).$$

Since the behaviors of these systems are characterized by a difference equation, they are also called autoregressive (AR) 2D systems.

For our purpose it is important to note that different difference operators may produce the same behavior. More precisely, given two polynomial matrices R_1, R_2, we have $\ker R_1(\sigma_1, \sigma_2) \subseteq \ker R_2(\sigma_1, \sigma_2)$ if and only if there exists a polynomial matrix

X such that $R_2 = XR_1$ and so $\ker R_1(\sigma_1, \sigma_2) = \ker R_2(\sigma_1, \sigma_2)$ if and only if there exist polynomial matrices X_1, X_2 such that $R_2 = X_1 R_1$ and $R_1 = X_2 R_2$.

An important class of 2D AR systems is given by the controllable 2D AR systems. A 2D system $\Sigma = (\mathbb{Z}^2, \mathbb{R}^q, \mathcal{B})$ is said to be controllable if there exists $g \geq 0$ such that, given any pair of trajectories $w_1, w_2 \in \mathcal{B}$ and given any pair of subsets $T_1, T_2 \subseteq \mathbb{Z}^2$ such that the Euclidean distance between T_1 and T_2 is greater than g, there exists $w \in \mathcal{B}$ such that $w_{|T_1} = w_{1|T_1}$ and $w_{|T_2} = w_{2|T_2}$. Controllability has been extensively studied in [6, 9]. If $\Sigma = (\mathbb{Z}^2, \mathbb{R}^q, \mathcal{B})$ is a 2D AR system, then it has been shown in [3] that there exists the greatest controllable AR subsystem of Σ, which is called the controllable subsystem of Σ and is denoted by $\Sigma_c = (\mathbb{Z}^2, \mathbb{R}^q, \mathcal{B}_c)$. Moreover, if $\mathcal{B} = \ker R(\sigma_1, \sigma_2)$, where $R \in \mathbb{R}[z_1, z_2, z_1^{-1}, z_2^{-1}]^{l \times q}$, then there exists a factor left prime matrix $R' \in \mathbb{R}[z_1, z_2, z_1^{-1}, z_2^{-1}]^{l' \times q}$ (that is, a full row rank polynomial matrix whose square left factors must be unimodular) and a full column rank matrix $F \in \mathbb{R}[z_1, z_2, z_1^{-1}, z_2^{-1}]^{l \times l'}$ such that [4]

$$R = FR'. \tag{1}$$

It can be proved that [3]

$$\mathcal{B}_c = \ker R'(\sigma_1, \sigma_2).$$

Notice that when we refer to the rank of a polynomial matrix we mean its rank as a matrix with entries in the field of fractions of the domain $\mathbb{R}[z_1, z_2, z_1^{-1}, z_2^{-1}]$.

Let's finally define the concept of free components of a 2D AR system $\Sigma = (\mathbb{Z}^2, \mathbb{R}^q, \mathcal{B})$. We say that components $i_1, \ldots, i_m$ of the system Σ are free if for all $v_1, \ldots, v_m \in \mathbb{R}^{\mathbb{Z}^2}$, there exists $w = (w_1, \ldots, w_q)^T \in \mathcal{B}$ such that $w_{i_1} = v_1, \ldots, w_{i_m} = v_m$.

We pass now to the introduction of 2D systems theory in the classical input/output approach. Consider the spaces

$$\mathcal{U} := \{u \in (\mathbb{R}^m)^{\mathbb{Z}^2} : \text{supp } (\sigma_1^{t_1} \sigma_2^{t_2} u) \subseteq \mathbb{N}^2 \text{ for some } (t_1, t_2) \in \mathbb{Z}^2\},$$
$$\mathcal{Y} := \{y \in (\mathbb{R}^p)^{\mathbb{Z}^2} : \text{supp } (\sigma_1^{t_1} \sigma_2^{t_2} y) \subseteq \mathbb{N}^2 \text{ for some } (t_1, t_2) \in \mathbb{Z}^2\},$$

where $\mathbb{N}$ is the set of nonnegative integers and where supp $(\cdot)$ means the support of 2D sequences. The trajectories in $\mathcal{U}$ are called inputs, while the trajectories in $\mathcal{Y}$ are called outputs. A 2D system in the input/output approach is essentially given by an operator

$$\Psi : \mathcal{U} \longrightarrow \mathcal{Y}.$$

This operator is assumed to satisfy the following properties.

1. Ψ is a linear operator.
2. Ψ is shift-invariant; i.e., for all $u \in \mathcal{U}$ and for all $(t_1, t_2) \in \mathbb{Z}^2$ we have

$$\Psi(\sigma_1^{t_1} \sigma_2^{t_2} u) = \sigma_1^{t_1} \sigma_2^{t_2} \Psi(u).$$

3. Ψ is quarter-plane causal; i.e., if $u_1, u_2 \in \mathcal{U}$ are such that $u_1(h_1, h_2) = u_2(h_1, h_2)$ for all $h_1 \leq t_1, h_2 \leq t_2$ and if $y_1 = \Psi(u_1), y_2 = \Psi(u_2)$, then $y_1(t_1, t_2) = y_2(t_1, t_2)$.

Using the linearity and the shift-invariance it can be shown that the causality property has the following equivalent expression. If $u \in \mathcal{U}$ is such that supp $(u) \subseteq \mathbb{N}^2$, then supp $(\Psi(u)) \subseteq \mathbb{N}^2$.

The causality requirement can be weakened by substituting a general cone $C \subseteq \mathbb{Z}^2$ instead of $\mathbb{N}^2$ in the definitions of $\mathcal{U}$ and $\mathcal{Y}$ and in the causality requirement. 2D systems that are causal with respect to cones are called weakly causal. These systems are extensively treated in [1]. This generalization is essential for our objectives.

Consider the inputs $\delta^{(i)}$ defined for each $i = 1, \dots, m$ as

$$\delta^{(i)}(t) := \begin{cases} e_i & \text{if } t = (0,0), \\ 0 & \text{if } t \neq (0,0), \end{cases} \tag{2}$$

where $e_1, \dots, e_m$ denotes the canonical base in $\mathbb{R}^m$, and let $y^{(i)} := \Psi(\delta^{(i)})$. Then the impulse response associated with the 2D system is the matrix-valued 2D sequence defined as $Y := [y^{(1)}\ y^{(2)} \cdots y^{(m)}] \in (\mathbb{R}^{p\times m})^{\mathbb{Z}^2}$, which is supported in $\mathbb{N}^2$. As shown in [2], the operator Ψ is completely determined by the impulse response. Actually, suppose that u is any input and let $y \in (\mathbb{R}^p)^{\mathbb{Z}^2}$ be defined as follows:

$$y(t_1, t_2) := \sum_{(k_1,k_2)\in\mathbb{Z}^2} Y(t_1 - k_1, t_2 - k_2)u(k_1, k_2); \tag{3}$$

then it can be shown that $y = \Psi(u)$. Note that the sum in the previous formula is finite, and moreover, since Y and u have support in $\mathbb{N}^2$, y also has support in $\mathbb{N}^2$. The operation described in (3) is called convolution.

The theory of 2D systems in the classical input/output approach considers at this point another restriction on the operator Ψ or equivalently on its impulse response Y. Actually, the formal power series

$$\bar{Y} := \sum_{(t_1,t_2)\in\mathbb{Z}^2} Y(t_1, t_2) z_1^{-t_1} z_2^{-t_2} \tag{4}$$

associated with the impulse response Y is assumed to be rational, which means that there exists a nonsingular square polynomial matrix $P \in \mathbb{R}[z_1, z_2]^{p\times p}$ such that $P\bar{Y} = Q$ is a polynomial matrix in $\mathbb{R}[z_1, z_2]^{p\times m}$. The power series $\bar{Y}$ is called a transfer matrix of the 2D system and the representation

$$\bar{Y} = P^{-1}Q$$

is called matrix fraction description (MFD) of the transfer matrix $\bar{Y}$. Notice that the matrices P and Q could be taken left coprime, and in this case the MFD is called left coprime [4]. As shown in [2], rationality ensures that the 2D system is realizable through a state space model.

3. Passing from input/output systems to behavioral systems. Suppose now that we have a 2D input/output system; i.e., we have an input/output operator

$$\Psi : \mathcal{U} \longrightarrow \mathcal{Y}.$$

Then the set of trajectories

$$\left\{ \begin{pmatrix} \Psi(u) \\ u \end{pmatrix} \in (\mathbb{R}^{p+m})^{\mathbb{Z}^2} : u \in \mathcal{U} \right\} \tag{5}$$

is clearly a linear and shift-invariant behavior, but it is not complete or equivalently closed with respect to the pointwise convergence topology. If we denote by $\mathcal{B}(\Psi)$ the

closure of such a set, then $\Sigma(\Psi) = (\mathbb{Z}^2, \mathbb{R}^{p+q}, \mathcal{B}(\Psi))$ is a linear shift-invariant and complete 2D system in the behavioral approach. This is the linear shift-invariant and complete 2D system with the smallest possible behavior containing all the trajectories generated by the input/output map. This is therefore the most natural way to associate with a 2D input/output system a 2D behavioral system. Not every linear shift-invariant and complete 2D system can be obtained in this way. The following proposition shows that the 2D systems that come from 2D input/output systems are always controllable.

PROPOSITION 1. *Let $\Psi : \mathcal{U} \longrightarrow \mathcal{Y}$ be an input/output operator associated with a linear shift-invariant and quarter-plane causal 2D system. Then $\Sigma(\Psi)$ is controllable.*

The proof of this proposition is a direct consequence of the following two lemmas.

LEMMA 1. *Let $P \in \mathbb{R}[z_1, z_2, z_1^{-1}, z_2^{-1}]^{l \times p}$ be a full column rank matrix. Then $y \in \ker\, P(\sigma_1, \sigma_2)$ and $y \in \mathcal{Y}$ if and only if $y = 0$.*

Proof. One direction is trivial. Suppose conversely that $y \in \ker\, P(\sigma_1, \sigma_2) \cap \mathcal{Y}$. Since P is full column rank, there exists a polynomial matrix X such that $XP = fI$, where I is the identity matrix and f is a nonzero polynomial. Then each component of y is in the kernel of $f(\sigma_1, \sigma_2)$. Since the support of such components is in $\mathbb{N}^2$, a simple computation (see also [3]) shows that $y = 0$. □

LEMMA 2. *Let $\Sigma = (\mathbb{Z}^2, \mathbb{R}^q, \mathcal{B})$ be an AR system and $\Sigma_c = (\mathbb{Z}^2, \mathbb{R}^q, \mathcal{B}_c)$ be its controllable subsystem. If $w \in \mathcal{B}$ and if w has support in $\mathbb{N}^2$, then $w \in \mathcal{B}_c$*

Proof. Suppose that $\mathcal{B} = \ker\, R(\sigma_1, \sigma_2)$ and that $R = FR'$, where R' is factor left prime and F is full column rank. Then $\mathcal{B}_c = \ker\, R'(\sigma_1, \sigma_2)$. Suppose that $w \in \mathcal{B}$ and that w has support in $\mathbb{N}^2$. Then $v := R'(\sigma_1, \sigma_2)w \in \ker\, F(\sigma_1, \sigma_2)$. Since F is full column rank and since the support of v is contained in some translation of $\mathbb{N}^2$, then by Lemma 1 $v = 0$ and so $w \in \ker\, R'(\sigma_1, \sigma_2) = \mathcal{B}_c$. □

Remark. Lemma 2 cannot be extended directly to the nD case, since it is based on the factorization (1), which is always possible in the 2D case (see [4]) but is not always possible in the general nD case, when $n > 2$ (see [11]). This is the reason why the extension to the nD case, pointed out by one of the referees, is not straightforward with the techniques used in this paper.

The previous proposition holds also for nonrational 2D input/output systems. If the system is rational we are able to obtain an AR representation of the 2D behavioral system associated with the input/output system directly from a left coprime matrix fraction description of the transfer matrix of the system.

PROPOSITION 2. *Let $\Psi : \mathcal{U} \longrightarrow \mathcal{Y}$ be an input/output operator associated with a linear shift-invariant and quarter-plane causal rational 2D system. Let $\bar{Y} = P^{-1}Q$ be a left coprime MFD of the transfer matrix. Then $\Sigma(\Psi)$ admits the following AR representation:*

$$\mathcal{B}(\Psi) = \ker\, [P(\sigma_1, \sigma_2) \ -Q(\sigma_1, \sigma_2)].$$

Proof. We have to show that

$$\mathcal{B}(\Psi) = \ker\, [P(\sigma_1, \sigma_2) \ -Q(\sigma_1, \sigma_2)].$$

We start showing $\subseteq$. First notice that $P\bar{Y} = Q$ if and only if

$$Q_{k_1,k_2} = \sum_{h_1,h_2} P_{h_1,h_2} Y(-k_1 + h_1, -k_2 + h_2). \tag{6}$$

Suppose that $u \in \mathcal{U}$ and that $y = \Psi(u)$. Then we have

$$\begin{aligned}(P(\sigma_1,\sigma_2)y)(t_1,t_2) &= \sum_{h_1,h_2} P_{h_1,h_2} \sum_{k_1,k_2} Y(t_1+h_1-k_1, t_2+h_2-k_2)u(k_1,k_2)\\ &= \sum_{k_1,k_2} \left(\sum_{h_1,h_2} P_{h_1,h_2} Y(t_1+h_1-k_1, t_2+h_2-k_2) \right) u(k_1,k_2)\\ &= \sum_{k_1,k_2} Q_{k_1-t_1,k_2-t_2} u(k_1,k_2) = (Q(\sigma_1,\sigma_2)u)(t_1,t_2).\end{aligned}$$

We can argue that

$$\left\{ \begin{pmatrix} \Psi(u) \\ u \end{pmatrix} \in (\mathbb{R}^{p+m})^{\mathbb{Z}^2} : u \in \mathcal{U} \right\} \subseteq \ker\, [P(\sigma_1,\sigma_2) \ -Q(\sigma_1,\sigma_2)],$$

and since the behavior $\ker\, [P(\sigma_1,\sigma_2) \ -Q(\sigma_1,\sigma_2)]$ is closed, we are done.

Suppose on the other hand that $\mathcal{B}_f$ is the set of finite supported trajectories in $\ker\, [P(\sigma_1,\sigma_2) \ -Q(\sigma_1,\sigma_2)]$. By controllability [9, Theorem 1], we have that

$$\overline{\mathcal{B}_f} = \ker\, [P(\sigma_1,\sigma_2) \ -Q(\sigma_1,\sigma_2)],$$

where $\overline{\cdot}$ means closure. To prove the assertion we have to show that

$$\mathcal{B}_f \subseteq \left\{ \begin{pmatrix} \Psi(u) \\ u \end{pmatrix} \in (\mathbb{R}^{p+m})^{\mathbb{Z}^2} : u \in \mathcal{U} \right\}.$$

Suppose that u_f, y_f are finite supported trajectories such that

$$P(\sigma_1,\sigma_2)y_f = Q(\sigma_1,\sigma_2)u_f.$$

By the first part of the proof we can argue that

$$P(\sigma_1,\sigma_2)\Psi(u_f) = Q(\sigma_1,\sigma_2)u_f,$$

and so applying Lemma 1 we argue that $y_f = \Psi(u_f)$. □

4. Causal input/output representations. In the last section we have proposed a way to pass from a 2D input/output system to a 2D behavioral system. The following three sections are concerned with the inverse problem of passing from a 2D behavioral system to a 2D input/output system. This problem will be solved by resorting to the concept of causal input/output representation. The concept of input/output representation (without causality) has been introduced and studied in the 2D case in [6, 12, 13]. The construction of this representation is rather simple, since it is connected with the extraction of the free components of a system.

The notion of causal input/output system has been given in [10] only in the 1D case. In the 2D case this notion is a little more involved since it allows much more freedom in the choice of the causality cone. This is specified simply by a pair of elements $d_1, d_2 \in \mathbb{Z}^2$ that generates $\mathbb{Z}^2$ as a group and is defined as

$$C := \{\alpha d_1 + \beta d_2 \in \mathbb{Z}^2 : \alpha, \beta \in \mathbb{N}\}.$$

This definition of cone corresponds to the definition of "causality cone" introduced in [1].

DEFINITION 1. *Let* $P \in \mathbb{R}[z_1, z_2, z_1^{-1}, z_2^{-1}]^{l\times p}$, $Q \in \mathbb{R}[z_1, z_2, z_1^{-1}, z_2^{-1}]^{l\times m}$. *The difference equation*

$$P(\sigma_1, \sigma_2)y = Q(\sigma_1, \sigma_2)u, \tag{7}$$

where $y \in (\mathbb{R}^p)^{\mathbb{Z}^2}$ *and* $u \in (\mathbb{R}^m)^{\mathbb{Z}^2}$, *is an input/output representation of the* 2*D AR system* $\Sigma = (\mathbb{Z}^2, \mathbb{R}^q, \mathcal{B})$ *if*

1. *The difference equation* (7) *determines* $\mathcal{B}$ *up to a permutation of its components: it means that* $p + m = q$ *and that there exists a permutation matrix* $S \in \{0, 1\}^{q\times q}$ *such that*

$$S\begin{bmatrix} y \\ u \end{bmatrix} \in \mathcal{B} \quad \Leftrightarrow \quad P(\sigma_1, \sigma_2)y = Q(\sigma_1, \sigma_2)u.$$

2. u *is free: for all* $u \in (\mathbb{R}^m)^{\mathbb{Z}^2}$ *there exists* $y \in (\mathbb{R}^p)^{\mathbb{Z}^2}$ *such that* $P(\sigma_1, \sigma_2)y = Q(\sigma_1, \sigma_2)u$.

3. *No other components in* y *are free: the* 2*D AR system* $(\mathbb{Z}^2, \mathbb{R}^p, \ker P(\sigma_1, \sigma_2))$ *has no free components.*

The difference equation (7) *is said to be a causal input/output representation with respect to the cone* C *if, moreover,*

4. *the action of* u *on* y *is causal with respect to the cone* C*: if* u *is any signal in* $(\mathbb{R}^m)^{\mathbb{Z}^2}$ *with support in* C, *then there exists a signal* $y \in (\mathbb{R}^p)^{\mathbb{Z}^2}$ *with support in* C *such that*

$$P(\sigma_1, \sigma_2)y = Q(\sigma_1, \sigma_2)u.$$

Notice that it is always possible to perform a change of coordinates and a corresponding change of variables in $\mathbb{R}[z_1, z_2, z_1^{-1}, z_2^{-1}]$ in such a way that the cone C coincides with $\mathbb{N}^2$.

Define $(\mathbb{R}^m)^{(C)}$ and $(\mathbb{R}^p)^{(C)}$ to be the set of signals in $(\mathbb{R}^m)^{\mathbb{Z}^2}$ or in $(\mathbb{R}^p)^{\mathbb{Z}^2}$ with support in C. By this notation we can say that an input/output representation

$$P(\sigma_1, \sigma_2)y = Q(\sigma_1, \sigma_2)u \tag{8}$$

is causal if for any $u \in (\mathbb{R}^m)^{(C)}$ there exists $y \in (\mathbb{R}^p)^{(C)}$ such that (8) holds. As a direct consequence of Lemma 1 we have that such a y is also unique. Therefore, it can be defined by an operator

$$\Psi : (\mathbb{R}^m)^{(C)} \to (\mathbb{R}^p)^{(C)}$$

associating with $u \in (\mathbb{R}^m)^{(C)}$ the unique element $y = \Psi(u) \in (\mathbb{R}^m)^{(C)}$ such that (8) holds. This operator is called *input/output operator* associated with the causal input/output representation. It can be extended directly to the set of trajectories whose support is contained in a suitable shift of C, providing in this way a 2D input/output system. Therefore, there exists a direct and natural way to obtain a 2D input/output system from a causal input/output representation.

Notice that the existence of the impulse response is a necessary and sufficient condition for the difference equation (7) to be a causal input/output representation. More precisely the difference equation (7) is a causal input/output representation if and only if for each $i = 1, \ldots, m$ there exists $y^{(i)} \in (\mathbb{R}^p)^{(C)}$ such that $P(\sigma_1, \sigma_2)y^{(i)} = Q(\sigma_1, \sigma_2)\delta^{(i)}$, where $\delta^{(i)}$ are the trajectories defined in (2).

The first useful fact concerning these representations is provided by the following proposition that shows that, in finding causal input/output representations, only the controllable subsystem is really relevant. Moreover, this proposition shows that the 2D input/output system that we obtain from a causal input/output representation is always rational and that a left coprime MFD of its transfer matrix can be obtained directly from the AR representation of the controllable subsystem. This part of the following proposition can be considered the counterpart of Proposition 2.

PROPOSITION 3. *Suppose that* $\Sigma = (\mathbb{Z}^2, \mathbb{R}^q, \mathcal{B})$ *is a 2D AR system and that* $\Sigma_c = (\mathbb{Z}^2, \mathbb{R}^q, \mathcal{B}_c)$ *is its controllable subsystem. Let, moreover,* P, Q, P_c, Q_c *be polynomial matrices of suitable dimensions such that* $[P \ -Q] = F[P_c \ -Q_c]$, *where* F *is full column rank and* $[P_c \ Q_c]$ *is factor left prime. Then the difference equation*

$$P(\sigma_1, \sigma_2)y = Q(\sigma_1, \sigma_2)u$$

is a causal input/output representation of Σ *with respect to the cone* C *if and only if*

$$P_c(\sigma_1, \sigma_2)y = Q_c(\sigma_1, \sigma_2)u$$

is a causal input/output representation of Σ_c *with respect to the cone* C. *Moreover, the input/output operators associated with the previous causal input/output representations coincide, and the associated transfer matrix is*

$$\bar{Y} = P_c^{-1}Q_c.$$

Proof. The fact that if $P_c(\sigma_1, \sigma_2)y = Q_c(\sigma_1, \sigma_2)u$ is a causal input/output representation of Σ_c, then $P(\sigma_1, \sigma_2)y = Q(\sigma_1, \sigma_2)u$ is a causal input/output representation of Σ is trivial. Suppose conversely that $P(\sigma_1, \sigma_2)y = Q(\sigma_1, \sigma_2)u$ is a causal input/output representation of Σ and let u be any signal in $(\mathbb{R}^m)^{(C)}$. Then there exists $y \in (\mathbb{R}^p)^{(C)}$ such that $P(\sigma_1, \sigma_2)y = Q(\sigma_1, \sigma_2)u$. Then

$$S \begin{bmatrix} y \\ u \end{bmatrix} \in \mathcal{B},$$

where S is the permutation matrix introduced in Definition 1 and, moreover, the support of such a trajectory is included in C. Then by Lemma 2 we have that

$$S \begin{bmatrix} y \\ u \end{bmatrix} \in \mathcal{B}_c,$$

and consequently $P_c(\sigma_1, \sigma_2)y = Q_c(\sigma_1, \sigma_2)u$. This also shows that the previous causal input/output representations have the same impulse response and so the same transfer matrix. Finally, by (6), $P_c(\sigma_1, \sigma_2)y^{(i)} = Q_c(\sigma_1, \sigma_2)\delta^{(i)}$, for all $i = 1, \ldots, m$, implies that $P_c\bar{Y} = Q_c$. □

Remark. Since, as shown in the previous proposition, the formal power series associated with the impulse response Y is rational, then it is possible to construct a state space representation of the input/output operator Ψ. Notice that such a state space representation provides a way to obtain an input/state/output representation of the 2D AR system $\Sigma = (\mathbb{Z}^2, \mathbb{R}^q, \mathcal{B})$ associated with the causal input/output representation, similar to what Willems has done in the 1D case. Here the state variable is Markov with respect to a suitable modification of the south-west Markov property introduced in [6, 7].

Two different 2D AR systems may be described by the same input/output operator Ψ and so by the same impulse response. A direct consequence of Propositions 2 and 3 is that the input/output operator Ψ is completely determined by the controllable part of the 2D AR system and vice versa the controllable part uniquely determines Ψ.

5. Characterization of causal input/output representations. In this section we will present a method that allows us to check whether a difference equation like

$$P(\sigma_1, \sigma_2)y = Q(\sigma_1, \sigma_2)u \tag{9}$$

provides a causal input/output representation with respect to a given cone C. First we need to introduce the concept of proper rational matrix with respect to a cone. A rational matrix $H \in \mathbb{R}(z_1, z_2, z_1^{-1}, z_2^{-1})^{p\times m}$ is said to be proper with respect to a cone C if there exists $P = \sum P_{ij} z_1^i z_2^j \in \mathbb{R}[z_1, z_2, z_1^{-1}, z_2^{-1}]^{p\times p}$ and $Q \in \mathbb{R}[z_1, z_2, z_1^{-1}, z_2^{-1}]^{p\times m}$, both supported in $-C$ such that

$$H = P^{-1}Q$$

and such that P_{00} is invertible. In order to prove the next proposition we need the following lemma.

LEMMA 3. *Let $\bar{Y} \in \mathbb{R}[[z_1, z_2]]^{p\times m}$ be a formal power series and let $P = \sum P_{ij} z_1^i z_2^j \in \mathbb{R}[z_1, z_2]^{p\times p}$ and $Q \in \mathbb{R}[z_1, z_2]^{p\times m}$ be factor left coprime polynomial matrices such that P is nonsingular and*

$$P\bar{Y} = Q.$$

Then P_{00} is invertible.

Proof. First consider the scalar case $p = m = 1$. Let $P, Q \in \mathbb{R}[z_1, z_2]$ be coprime polynomials such that $P\bar{Y} = Q$. There exists $M, N \in \mathbb{R}[z_1, z_2]$ such that $F := PM + QN \in \mathbb{R}[z_1]$. Then $P(M + N\bar{Y}) = F$. Let

$$\hat{Y} := M + N\bar{Y} = \sum_{i=l}^{\infty} \hat{Y}_i(z_2) z_1^i, \quad \hat{Y}_l(z_2) \neq 0$$

and

$$P = \sum_{i=h}^{k} P_i(z_2) z_1^i, \quad P_h(z_2) \neq 0.$$

Then $P_h\hat{Y}_l \in \mathbb{R} \setminus \{0\}$ and so P_h has nonzero constant term. With the same reasoning, exchanging z_1 with z_2, we argue that the constant term of P must be nonzero.

Consider now the general case. It is clear that since $\bar{Y}$ is rational, then for each $i = 1, \ldots, p$ and $j = 1, \ldots, m$ there exist coprime $d_{ij}, n_{ij} \in \mathbb{R}[z_1, z_2]$ such that $d_{ij}\bar{Y}_{ij} = n_{ij}$. Therefore we have that $d_{ij}(0, 0) \neq 0$. Let

$$D := \left(\prod_{i=1}^{p} \prod_{j=1}^{m} d_{ij} \right) I,$$

where I is the $p \times p$ identity matrix. Then $N := D\bar{Y} \in \mathbb{R}[z_1, z_2]^{p\times m}$ and the constant term in D is invertible. If $P \in \mathbb{R}[z_1, z_2]^{p\times p}$ and $Q \in \mathbb{R}[z_1, z_2]^{p\times m}$ are factor left

coprime polynomial matrices such that P is nonsingular and $P\bar{Y} = Q$, then we have $P^{-1}Q = D^{-1}N$. Then by Lemma 5.3 in [4] there exists a polynomial matrix $X \in \mathbb{R}[z_1, z_2]^{p\times p}$ such that $D = XP$ and this implies that the constant term of P is also invertible. □

THEOREM 1. *Let $P \in \mathbb{R}[z_1, z_2, z_1^{-1}, z_2^{-1}]^{l\times p}$, $Q \in \mathbb{R}[z_1, z_2, z_1^{-1}, z_2^{-1}]^{l\times m}$. The difference equation*

$$P(\sigma_1, \sigma_2)y = Q(\sigma_1, \sigma_2)u \tag{10}$$

is an input/output representation if and only if $rank P = rank[P\ Q] = p$. The difference equation (10) *is a causal input/output representation with respect to the cone C if and only if, in addition, the rational matrix H such that $Q = PH$ is proper with respect to C.*

Proof. The proof of the first part of the theorem can be found in [12]. Consider the second part. Without loss of generality we can suppose that $C = \mathbb{N}^2$. If the rational matrix H such that $Q = PH$ is proper with respect to C, then there exist polynomial matrices $\bar{P} = \sum \bar{P}_{ij} z_1^{-i} z_2^{-j}, \bar{Q} = \sum \bar{Q}_{ij} z_1^{-i} z_2^{-j}$ with support in $-\mathbb{N}^2$ such that $\bar{P}$, $\bar{Q}$ are factor left coprime, $\bar{P}_{00} = I$, and $\bar{P}^{-1}\bar{Q} = H$. Therefore, $Q = P\bar{P}^{-1}\bar{Q}$, and so by Lemma 5.3 in [4] there exists a polynomial matrix $X \in \mathbb{R}[z_1, z_2, z_1^{-1}, z_2^{-1}]^{l\times p}$ such that $[P\ Q] = X[\bar{P}\ \bar{Q}]$. Suppose now that $u \in (\mathbb{R}^m)^{(C)}$ and let $y \in (\mathbb{R}^p)^{(C)}$ be defined for all $(h, k) \in \mathbb{N}^2$ recursively as follows:

$$y(h,k) := -\sum_{(i,j)\neq(0,0)} \bar{P}_{ij} y(h-i, k-j) + \sum \bar{Q}_{ij} u(h-i, k-j). \tag{11}$$

Since the equation (11) holds for all $(h, k) \in \mathbb{Z}^2$, we have that $\bar{P}(\sigma_1, \sigma_2)y = \bar{Q}(\sigma_1, \sigma_2)u$, and consequently $P(\sigma_1, \sigma_2)y = Q(\sigma_1, \sigma_2)u$.

Suppose conversely that $P(\sigma_1, \sigma_2)y = Q(\sigma_1, \sigma_2)u$ is a causal input/output representation and let $Y \in (\mathbb{R}^{p\times m})^{(C)}$ be its impulse response. Define

$$\bar{Y} := \sum_{i,j\in\mathbb{N}} Y(t_1, t_2) z_1^{-t_1} z_2^{-t_2}$$

as the formal power series in $\mathbb{R}[[z_1^{-1}, z_2^{-1}]]^{p\times m}$ associated with Y. Then we have that

$$P\bar{Y} = Q.$$

Let $\bar{P} \in \mathbb{R}[z_1^{-1}, z_2^{-1}]^{p\times p}$ and $\bar{Q} \in \mathbb{R}[z_1^{-1}, z_2^{-1}]^{p\times m}$ be coprime polynomial matrices such that

$$\bar{P}\bar{Y} = \bar{Q}.$$

By Lemma 3 the constant term of $\bar{P}$ is invertible. Notice now that $P(\sigma_1, \sigma_2)y = Q(\sigma_1, \sigma_2)u$ and $\bar{P}(\sigma_1, \sigma_2)y = \bar{Q}(\sigma_1, \sigma_2)u$ have the same impulse response. If $\mathcal{B}$ is the behavior associated with $P(\sigma_1, \sigma_2)y = Q(\sigma_1, \sigma_2)u$ and $\bar{\mathcal{B}}$ is the behavior associated with $\bar{P}(\sigma_1, \sigma_2)y = \bar{Q}(\sigma_1, \sigma_2)u$, then, since $\bar{\mathcal{B}}$ is controllable, $\bar{\mathcal{B}} = \mathcal{B}_c \subseteq \mathcal{B}$. As mentioned in the preliminaries on the behavioral approach, this implies that there exists a polynomial matrix X such that $P = \bar{P}X$ and $Q = \bar{Q}X$ and so $Q = P\bar{P}^{-1}\bar{Q}$. □

Given a cone C and an input/output representation

$$P(\sigma_1, \sigma_2)y = Q(\sigma_1, \sigma_2)u, \tag{12}$$

the previous theorem suggests an algorithmic method that allows us to verify if (12) is causal with respect to the cone C. It consists of two steps.

1. Perform a change of coordinates in $\mathbb{Z}^2$ and the corresponding change of variables in P, Q in such a way as to transform C in $\mathbb{N}^2$. Let P', Q' be the matrices P, Q after this change of variables.

2. Find $\bar{P} \in \mathbb{R}[z_1^{-1}, z_2^{-1}]^{p\times p}$ and $\bar{Q} \in \mathbb{R}[z_1^{-1}, z_2^{-1}]^{p\times m}$ that are left coprime as matrices with entries in $\in \mathbb{R}[z_1^{-1}, z_2^{-1}]$ and such that

$$P'\bar{P}^{-1}\bar{Q} = Q'.$$

Then (12) is causal with respect to C if and only if the constant term of $\bar{P}$ is invertible.

Notice that the pair of matrices $\bar{P}, \bar{Q}$ through formula (11) provides a computational way to represent the input/output operator associated with the causal input/output representation (12).

6. Existence and construction of causal input/output representations. In the 1D case it can be shown that, given an AR system Σ and a given cone (that in this case can be either $\mathbb{N}$ or $-\mathbb{N}$), it is always possible to find a causal input/output representation of Σ with respect to the cone. In this section we will show that also in the 2D case this is possible with some restrictions. The fact that this representation cannot be found in general for any given cone is shown in the following example.

Example 1. Let $\Sigma = (\mathbb{Z}^2, \mathbb{R}^2, \mathcal{B})$ be a 2D AR system with $\mathcal{B} = \ker R(\sigma_1, \sigma_2)$ and $R = [z_1 - z_2 \mid -1]$. Therefore a trajectory $(w_1, w_2)^T$ is in $\mathcal{B}$ if and only if it satisfies the difference equation

$$w_1(i+1, j) - w_1(i, j+1) - w_2(i, j) = 0. \tag{13}$$

Consider the cone $C = \mathbb{N}^2$. We want to show that there exists no causal input/output representations of Σ with respect to the cone C. This can be verified by applying Theorem 1. Actually, consider the input/output representation

$$P(\sigma_1, \sigma_2)w_1 = Q(\sigma_1, \sigma_2)w_2, \tag{14}$$

where $P = z_1 - z_2$ and $Q = 1$, then $Q = P\bar{P}^{-1}\bar{Q}$, where $\bar{P} = z_1^{-1}z_2^{-1}$ and $\bar{Q} = z_2^{-1} - z_1^{-1}$ are coprime in $\mathbb{R}[z_1^{-1}, z_2^{-1}]$. Observe finally that the constant term of $\bar{P}$ is zero. Similarly, it can be shown that (14) is not causal with respect to $\mathbb{N}^2$ even if we suppose that w_1 is the input and w_2 the output.

The same fact can be shown directly. Suppose we consider the input/output representation (14), where w_1 is considered as the input and w_2 as the output. Suppose that w_1 is 1 in $(0, 0)$ and 0 elsewhere and evaluate the difference equation for $i = -1, j = 0$. Then we obtain that $w_2(-1, 0) = 1$, and so this representation cannot be causal with respect to the cone C.

If conversely we suppose that w_1 is the output and w_2 is the input and if we suppose that w_2 is 1 in $(0, 0)$ and 0 elsewhere, then evaluating the difference equation (13) for $(i, j) = (0, 0)$, we obtain that $w_1(1, 0) - w_1(0, 1) = 1$. Moreover, evaluating the difference equation (13) for $(i, j) = (-1, 1)$ and $(i, j) = (1, -1)$, we see that $w_1(2, -1) = w_1(1, 0)$ and that $w_1(-1, 2) = w_1(0, 1)$ and so $w_1(2, -1) - w_1(-1, 2) = 1$. It follows that $w_1(2, -1), w_1(-1, 2)$ cannot both be equal to zero, and this implies that this input/output representation also cannot be causal with respect to the cone C.

The following theorem shows what can be done.

THEOREM 2. *Let $\Sigma = (\mathbb{Z}^2, \mathbb{R}^q, \mathcal{B})$ be a 2D AR system with $\mathcal{B} = \ker R(\sigma_1, \sigma_2)$. Moreover, let C be any cone in $\mathbb{Z}^2$. Then there exists a cone C' containing C such that there exists a causal input/output representation of Σ with respect to C'.*

In order to prove the previous theorem we need some notations and a technical result. Consider the total ordering $<_T$ in $\mathbb{Z}^2$ defined in the following way:

$$(n_1, n_2) <_T (m_1, m_2) \quad \Leftrightarrow \quad \begin{array}{l} n_1 < m_1 \text{ or} \\ n_1 = m_1 \text{ and } n_2 < m_2. \end{array}$$

Let $f \in \mathbb{R}[z_1, z_2, z_1^{-1}, z_2^{-1}]$. Then f can be expressed as follows:

$$f = \sum f_{ij} z_1^i z_2^j, \qquad f_{ij} \in \mathbb{R}.$$

The finite set of $(i, j) \in \mathbb{Z}^2$ such that $f_{ij} \neq 0$ is said to be in the support of f, while the greatest element in the support of f with respect to the total ordering $<_T$ is called the degree of f and is denoted by $\deg f$.

LEMMA 4. *Let $R \in \mathbb{R}[z_1, z_2, z_1^{-1}, z_2^{-1}]^{l \times q}$ be a rank p polynomial matrix. Then there exists a full row rank polynomial matrix $U \in \mathbb{R}[z_1, z_2, z_1^{-1}, z_2^{-1}]^{p \times l}$ and a permutation matrix $S \in \{0,1\}^{q \times q}$ such that*

$$URS = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} & \cdots & r_{1q} \\ 0 & r_{22} & \cdots & r_{2p} & \cdots & r_{2q} \\ \vdots & \vdots & \ddots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & r_{pp} & \cdots & r_{pq} \end{bmatrix} \tag{15}$$

with $r_{ii} \neq 0$ and $\deg r_{ii} \geq \deg r_{ij}$ for all $j \geq i$.

Proof. First we suppose that R is full row rank. We will show the assertion of the lemma by induction on p. If $p = 1$, then the assertion is clearly true. Suppose now that the assertion is true for $p = k-1$ and let $R \in \mathbb{R}[z_1, z_2, z_1^{-1}, z_2^{-1}]^{k \times q}$ be a full row rank polynomial matrix. By the postmultiplication by a suitable permutation matrix $\bar{S} \in \{0,1\}^{q \times q}$ we are able to obtain that the polynomial in position $(1,1)$ in $R\bar{S}$ has the greatest degree among all the polynomials in the first row of $R\bar{S}$. Let $\bar{V} \in \mathbb{R}[z_1, z_2, z_1^{-1}, z_2^{-1}]^{k-1 \times k}$ be a full row rank polynomial matrix such that the first column of $\bar{V}R\bar{S}$ is zero and let

$$\bar{U} := \begin{bmatrix} 1 & 0 & \cdots & 0 \\ & & \bar{V} & \end{bmatrix} \in \mathbb{R}[z_1, z_2, z_1^{-1}, z_2^{-1}]^{k \times k}.$$

Consequently

$$\bar{U}R\bar{S} = \begin{bmatrix} r'_{11} & r'_1 \\ 0 & R' \end{bmatrix},$$

where $r'_1 \in \mathbb{R}[z_1, z_2, z_1^{-1}, z_2^{-1}]^{1 \times q-1}$ and $R' \in \mathbb{R}[z_1, z_2, z_1^{-1}, z_2^{-1}]^{k-1 \times q-1}$. We want to show now that R' is full row rank. Suppose that $v \in \mathbb{R}[z_1, z_2, z_1^{-1}, z_2^{-1}]^{k-1}$ is such that $v^T R' = 0$. Then

$$v^T \bar{V} R\bar{S} = v^T \begin{bmatrix} 0 & R' \end{bmatrix} = 0,$$

and so, since $\bar{S}, R$, and $\bar{V}$ are full row rank, we have $v = 0$. The fact that R' is full row rank implies that $\bar{U}$ is full row rank. By induction, there exists a nonsingular polynomial matrix $U' \in \mathbb{R}[z_1, z_2, z_1^{-1}, z_2^{-1}]^{k-1 \times k-1}$ and a permutation matrix $S' \in \{0,1\}^{q-1 \times q-1}$ such that

$$U'R'S' = \begin{bmatrix} r'_{11} & r'_{12} & \cdots & r'_{1\ k-1} & \cdots & r'_{1\ q-1} \\ 0 & r'_{22} & \cdots & r'_{2\ k-1} & \cdots & r'_{2\ q-1} \\ \vdots & \vdots & \ddots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & r'_{k-1\ k-1} & \cdots & r'_{k-1\ q-1} \end{bmatrix}$$

with $r'_{ii} \neq 0$ and $\deg r'_{ii} \geq \deg r'_{ij}$ for all $j \geq i$. Finally, defining

$$U := \begin{bmatrix} 1 & 0 \\ 0 & U' \end{bmatrix} \bar{U}, \quad S := \bar{S} \begin{bmatrix} 1 & 0 \\ 0 & S' \end{bmatrix},$$

a simple computation shows that URS satisfies the thesis.

We now consider the general case. It is not restrictive to assume that the first p rows of R constitute a full row rank polynomial matrix R'. Then there exist a nonsingular polynomial matrix $U' \in \mathbb{R}[z_1, z_2, z_1^{-1}, z_2^{-1}]^{p \times p}$ and a permutation matrix $S \in \{0,1\}^{q \times q}$ such that $U'R'S$ is in the form (15). Put $U := [\, U' \quad 0\,] \in \mathbb{R}[z_1, z_2, z_1^{-1}, z_2^{-1}]^{p \times l}$. Then URS is in the form (15). □

Proof of the theorem. As usual, it is not restrictive to suppose that C coincides with $\mathbb{N}^2$. Suppose that R has rank p. By Lemma 4 there exists a full row rank polynomial matrix $U \in \mathbb{R}[z_1, z_2, z_1^{-1}, z_2^{-1}]^{l \times p}$ and a permutation matrix $S \in \{0,1\}^{q \times q}$ such that (15) holds with $r_{ii} \neq 0$ and $\deg r_{ii} \geq \deg r_{ij}$ for all $j \geq i$. It is not restrictive to assume that $\deg r_{11} = \deg r_{22} = \cdots = \deg r_{pp} = (0,0)$ (this can be obtained by premultiplying URS by a diagonal matrix having suitable monomials on the diagonal). Then there exists a cone C' containing $\mathbb{N}^2$ such that the supports of each polynomial in URS is included in $-C'$. Partition

$$RS = [\, P \quad -Q\,]$$

with $P \in \mathbb{R}[z_1, z_2, z_1^{-1}, z_2^{-1}]^{l \times p}$ and $Q \in \mathbb{R}[z_1, z_2, z_1^{-1}, z_2^{-1}]^{l \times m}$, where $m = q - p$. We want to show that

$$P(\sigma_1, \sigma_1)y = Q(\sigma_1, \sigma_1)u \tag{16}$$

is a causal input/output representation with respect to the cone C'. The difference equation (16) is an input/output representation. Indeed, rank $[P \ -Q] = \text{rank } R = p$, while on one hand rank $P \leq p$, and on the other hand rank $P \geq \text{rank } UP = p$, since UP is a square $p \times p$ upper triangular matrix. Now perform a change of coordinates in $\mathbb{Z}^2$ and the corresponding change of variables in such a way that C' is transformed into $\mathbb{N}^2$. In this way $URS \in \mathbb{R}[z_1^{-1}, z_2^{-1}]^{p \times q}$. Partition

$$URS = [\, P' \quad -Q'\,]$$

with $P' \in \mathbb{R}[z_1^{-1}, z_2^{-1}]^{p \times p}$ and $Q' \in \mathbb{R}[z_1^{-1}, z_2^{-1}]^{p \times m}$. By construction the constant term of P' is an invertible matrix. Let H be the rational matrix such that $Q = PH$. Then $Q' = P'H$ and so $H = P'^{-1}Q'$ is proper with respect to $\mathbb{N}^2$. By Theorem 1 this implies that the input/output representation (16) is causal with respect to C'.

Example 2. Let $\Sigma = (\mathbb{Z}^2, \mathbb{R}^2, \mathcal{B})$ be a 2D AR system with $\mathcal{B} = \ker R(\sigma_1, \sigma_2)$ and

$$R = \begin{bmatrix} z_1 - z_2^2 & 0 & 2z_1z_2 - 1 \\ 1 & z_1 - z_2 & 1 \end{bmatrix}.$$

We want to find a causal input/output representation of Σ with respect to a cone C containing $\mathbb{N}^2$. We apply the algorithm shown in Lemma 4. Since the polynomial with greatest degree in the first row is the third one, postmultiply R by the permutation matrix

$$S = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

which performs the exchange of the first and third columns of R. Notice that if $\bar{V} := [\,-1 \quad 2z_1z_2 - 1\,]$, then

$$\begin{bmatrix} 1 & 0 \\ & \bar{V} \end{bmatrix} RS = \begin{bmatrix} 2z_1z_2 - 1 & 0 & z_1 - z_2^2 \\ 0 & 2z_1^2z_2 - 2z_1z_2^2 - z_1 + z_2 & -z_1^2 + z_2 + 2z_1z_2 - 1 \end{bmatrix}.$$

Therefore, premultiplying by $\operatorname{diag}\{z_1^{-1}z_2^{-1}, z_1^{-2}z_2^{-1}\}$, we obtain a polynomial matrix $U \in \mathbb{R}[z_1, z_2, z_1^{-1}, z_2^{-1}]^{2\times 2}$ such that

$$URS = \begin{bmatrix} 2 - z_1^{-1}z_2^{-1} & 0 & z_2^{-1} - z_1^{-1}z_2 \\ 0 & 2 - 2z_1^{-1}z_2 - z_1^{-1}z_2^{-1} + z_1^{-2} & -z_2^{-1} + z_1^{-2} + 2z_1^{-1} - z_1^{-2}z_2^{-1} \end{bmatrix}.$$

Let $C := \{\alpha(1,-1) + \beta(0,1) \in \mathbb{Z}^2 : \alpha, \beta \in \mathbb{N}\}$. Then $C \supseteq \mathbb{N}^2$, and moreover, all the entries of URS have support in $-C$. Therefore, if we let

$$P := \begin{bmatrix} 2z_1z_2 - 1 & 0 \\ 1 & z_1 - z_2 \end{bmatrix}, \qquad Q := \begin{bmatrix} z_1 - z_2^2 \\ 1 \end{bmatrix},$$

then $P(\sigma_1, \sigma_1)y = Q(\sigma_1, \sigma_1)u$ is a causal input/output representation of Σ with respect to the cone C.

Remark. The construction of causal input/output representations suggested by Theorem 2 is in some way partial. We would achieve a more satisfactory and complete solution to the problem of constructing causal input/output representations if, given an input/output representation

$$P(\sigma_1, \sigma_1)y = Q(\sigma_1, \sigma_1)u, \tag{17}$$

we were able to parametrize the set of all the cones C such that (17) is causal with respect to C. At the moment, the solution to this problem is unknown to us and its investigation is the object of our future research.

REFERENCES

[1] R. Eising, *State-space realization and inversion of* 2*-D systems*, IEEE Trans. Circuits Systems, CAS-27 (1980), pp. 612–619.

[2] E. Fornasini and G. Marchesini, *State-space realization theory of two-dimensional filters*, IEEE Trans. Automat. Control, AC-21 (1976), pp. 484–492.

[3] E. Fornasini, P. Rocha, and S. Zampieri, *State space realization of* 2*D finite dimensional behaviours*, SIAM J. Control Optim., 31 (1993), pp. 1502–1517.

[4] S. Kung, B. Lévy, M. Morf, and T. Kailath, *New results in* 2*D systems theory. Part* 1, Proc. IEEE, 65 (1977), pp. 861–872.

[5] U. Oberst, *Multidimensional constant linear systems*, Acta Appl. Math., 20 (1990), pp. 1–175.

[6] P. Rocha, *Structure and Representation of* 2*-D Systems*, Ph.D. thesis, University of Groningen, the Netherlands, 1990.

[7] P. Rocha and J. Willems, *State for* 2*D systems*, Linear Algebra Appl., 122/123/124 (1989), pp. 1003–1038.

[8] P. Rocha and J. Willems, *Canonical computational forms for AR* 2*-D systems*, Multidimens. Systems Signal Process., 2 (1990), pp. 251–278.

[9] P. Rocha and J. Willems, *Controllability of* 2*-D systems*, IEEE Trans. Automat. Control, AC-36 (1991), pp. 413–423.

[10] J. Willems, *Models for dynamics*, Dynamics Reported, 2 (1988), pp. 171–269.

[11] D. Youla and G. Gnavi, *Notes on n-dimensional system theory*, IEEE Trans. Circuits Systems, 26 (1979), pp. 105–111.

[12] S. Zampieri, *A solution of the Cauchy problem for multidimensional discrete linear shift-invariant systems*, Linear Algebra Appl., 202 (1994), pp. 143–162.

[13] E. Zerz and U. Oberst, *The canonical Cauchy problem for linear systems of partial difference equations with constant coefficients over the complete* r*-dimensional integral lattice* $\mathbb{Z}^r$, Acta Appl. Math., 31 (1993), pp. 249–273.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 4, pp. 1147–1170, July 1998

002

RISK-SENSITIVE PRODUCTION PLANNING OF A STOCHASTIC MANUFACTURING SYSTEM*

W. H. FLEMING† AND Q. ZHANG‡

Abstract. This paper is concerned with long-run average risk-sensitive control of production planning in a manufacturing system with machines that are subject to breakdown and repair. By using a logarithmic transformation, it is shown that the associated Hamilton–Jacobi–Bellman equation has a viscosity solution. The risk-sensitive control problem has a dynamic stochastic game interpretation. Finally, a limiting problem is obtained when the rates of machine breakdown and repair go to infinity.

Key words. risk-sensitive control, production planning, logarithmic transformation, irreducible Markov chain

AMS subject classifications. 93E20, 93B35, 90B30

PII. S036301299631034X

1. Introduction.

In this paper we consider a manufacturing system which consists of machines that are subject to breakdown and repair. The objective of the problem is to choose a production planning to minimize a risk-sensitive cost criterion over the infinite horizon. In risk-sensitive control theory, typically an exponential-of-integral cost criterion is considered. Such cost functions heavily penalize state trajectories and controls which give large values to the exponent. The risk-sensitive approach has been applied to the so-called disturbance attenuation problem; see, for example, Whittle [17], Fleming and McEneaney [7], Basar and Bernhard [1] and Barron and Jensen [2], Glover and Doyle [12], and references therein. In Fleming and McEneaney [7], risk-sensitive control problems of controlled diffusions are considered. By using the associated dynamic programming equations, they show that as the system noise goes to zero, the value function of the risk-sensitive control problem converges to the value function of a differential game problem.

In this paper, we consider the risk-sensitive control of manufacturing systems with stochastic production capacity. The machine capacity process will be assumed to be an irreducible finite state (jump) Markov chain, with generator Q/ε, where $\varepsilon > 0$ is a parameter related to the frequency of machine breakdown and repair relative to an underlying production time scale. For simplicity, we consider a one-part-type manufacturing system with constant demand. The control is the production rate, which is subject to a random machine capacity constraint, as in Sethi and Zhang [15]. For fixed $\varepsilon > 0$ the goal is to find a control policy which minimizes the long-term growth rate of an expected exponential-of-integral criterion, as in formula (2.2). The minimum growth rate λ^ε and an associated cost potential function w^ε satisfy dynamic programming equations (3.1), which form a system of first-order nonlinear partial differential equations. The function w^ε is a viscosity sense solution.

*Received by the editors October 9, 1996; accepted for publication (in revised form) May 13, 1997; published electronically May 15, 1998.

http://www.siam.org/journals/sicon/36-4/31034.html

†Division of Applied Mathematics, Brown University, Providence, RI 02912 (whf@cfm.brown.edu). The research of this author was partially supported by AFOSR grant F49620-92-J-0081, ARO grant DAAL03-92-G-0115, and NSF grant DMS-9300048.

‡Department of Mathematics, University of Georgia, Athens, GA 30602 (qingz@math.uga.edu). The research of this author was partially supported by ONR grant N00014-96-1-0263 and the University of Georgia Faculty Research Grant.

Theorem 3.3 states that a solution exists. It is proved using the so-called vanishing discount approach often used for analysis of average cost minimization problems. This is done by using a logarithmic transformation introduced by Bensoussan and Nagai [3] to obtain an equivalent problem that is easier to deal with. In section 4, we discuss the asymptotic property of the problem as the rate of fluctuation of the production capacity process goes to infinity ($\varepsilon \to 0$). We show that the risk-sensitive control problem can be approximated by a limiting problem in which the stochastic capacity process can be averaged out and replaced by its average. This procedure is analogous to passing in the disturbance attenuation problem from the risk-sensitive model with small noise intensity to the deterministic robust control limit.

In our model, we assume a positive deterioration rate a for items in storage (formula (2.1)). This corresponds to a stability condition typically imposed for disturbance attenuation problems on an infinite time horizon (see Fleming and McEneaney [7]), and this assumption is essential in the proof of technical estimates in Lemma 3.2. Nevertheless, it would be interesting to weaken the assumption that $a > 0$.

2. Problem formulation. Let us consider a one-part-type and parallel-machine manufacturing system with stochastic production capacity and constant demand for its production over time. For $t \geq 0$, let $x(t)$, $u(t)$, and z denote the surplus level (the state variable), the production rate (the control variable), and the constant demand rate, respectively. We assume $x(t) \in \mathbb{R} = (-\infty, \infty)$, $u(t) \in \mathbb{R}^+ = [0, \infty)$, $t \geq 0$, and z a positive constant. They satisfy the following differential equation:

$$\dot{x}(t) = -ax(t) + u(t) - z, \qquad x(0) = x, \tag{2.1}$$

where $a > 0$ is a constant, representing the deterioration rate (or spoilage rate) of the finished product.

Let $(\Omega, \mathcal{F}, P)$ denote a probability space. Let $\alpha^\varepsilon(t) \in \mathcal{M} = \{0, 1, 2, \ldots, m\}$, $t \geq 0$, denote a Markov process generated by Q/ε, where $\varepsilon > 0$ is a small parameter and $Q = (q_{ij})$, $i, j \in \mathcal{M}$, is an $(m+1) \times (m+1)$ matrix such that $q_{ij} \geq 0$ for $i \neq j$ and $q_{ii} = -\sum_{i \neq j} q_{ij}$. We let $\alpha^\varepsilon(t)$ represent the maximum production capacity of the system at time t. The representation for $\mathcal{M}$ usually stands for the case of m identical machines, each with a unit capacity and having two states: up and down.

The production constraints are given by the inequalities:

$$0 \leq u(t) \leq \alpha^\varepsilon(t), \qquad t \geq 0.$$

Definition 2.1. *A production control process $u(\cdot) = \{u(t),\, t \geq 0\}$ is admissible if* (i) *$u(t)$ is $\sigma\{\alpha^\varepsilon(s), 0 \leq s \leq t)$ progressively measurable and* (ii) *$0 \leq u(t) \leq \alpha^\varepsilon(t)$ for all $t \geq 0$. Let $\mathcal{A}^\varepsilon$ denote the class of admissible controls.*

Let $L(x, u)$ denote a cost function of the surplus and the production. The objective of the problem is to choose $u(\cdot) \in \mathcal{A}^\varepsilon$ to minimize

$$J^\varepsilon(u(\cdot)) = \limsup_{T \to \infty} \frac{\varepsilon}{T} \log E \exp\left(\frac{1}{\varepsilon} \int_0^T L(x(t), u(t))dt\right), \tag{2.2}$$

where $x(\cdot)$ is the surplus process corresponding to the production process $u(\cdot)$. Let $\lambda^\varepsilon = \inf_{u(\cdot) \in \mathcal{A}^\varepsilon} J^\varepsilon(u(\cdot))$.

A motivation for choosing such an exponential cost criterion is that such criteria are sensitive to large values of the exponent which occur with small probability, for example, rare sequences of unusually many machine failures resulting in shortages ($x(t) < 0$).

Remark 2.2. In Zhang [18] a discounted cost criterion

$$J^\varepsilon(u(\cdot)) = \sqrt{\varepsilon}\log E\exp\left(\frac{1}{\sqrt{\varepsilon}}\int_0^\infty e^{-\rho t}L(x(t),u(t))dt\right)$$

is considered. The scale parameter in the cost is $\sqrt{\varepsilon}$ instead of ε as in (2.2). This is because the convergence involving a discounted cost is mainly affected by the convergence rate of $\alpha^\varepsilon(\cdot)$ to its equilibrium distribution, which is of order $\sqrt{\varepsilon}$.

Remark 2.3. The positive spoilage rate a appears in certain crucial estimates (see Lemma 3.2 (ii)). It also implies a uniform bound for $x(t)$. Note that the control $u(\cdot)$ is bounded between 0 and m. This implies that a solution $x(\cdot)$ to (2.1) must satisfy

$$|x(t)| \le |x|e^{-at} + (m+z)\int_0^t e^{-a(t-s)}ds \le |x|e^{-at} + \frac{m+z}{a}. \tag{2.3}$$

For a multidimensional problem, in order to obtain such a bound for $x(t)$, one may replace $a > 0$ by a matrix A with eigenvalues having positive real parts.

We assume that the cost function $L(x,u)$ and the production capacity process $\alpha^\varepsilon(\cdot)$ satisfy the following.

(A1) $L(x,u) \ge 0$ is continuous, bounded, and uniformly Lipschitz in x.

Remark 2.4. In a manufacturing system the running cost function $L(x,u)$ is usually chosen to be of the form $L(x,u) = h(x) + c(u)$ with piecewise linear $h(x)$ and $c(u)$. Note that piecewise linear functions are not bounded as required in (A1). However, this is not important, in view of the uniform bounds on $u(t)$ and on $x(t)$ for initial state $x = x(0)$ in any bounded set.

(A2) Q is *irreducible* in the following sense: the equations

$$\nu Q = 0 \quad \text{and} \quad \sum_{i=0}^m \nu_i = 1$$

have a unique solution $\nu = (\nu_0, \nu_1, \ldots, \nu_m)$ with $\nu_k > 0$, $k = 0, 1, \ldots, m$. The vector ν is called the *equilibrium distribution* of the Markov chain $\alpha^\varepsilon(\cdot)$.

Remark 2.5. One may also consider the model in which the demand rate $z = z(t)$ is a finite state Markov chain. In this case, one needs to consider various rates of fluctuation of $z(t)$ in comparison with that of $\alpha^\varepsilon(t)$. We refer the reader to Sethi and Zhang [15, Chap. 11] for related discussions in connection with production and marketing.

In the next section, we discuss the dynamics of the system and the associated Hamilton–Jacobi–Bellman (HJB) equations.

3. HJB equations. Formally, we can write the associated HJB equation as follows:

$$\frac{\lambda^\varepsilon}{\varepsilon} = \inf_{0\le u\le\alpha}\left\{(-ax+u-z)\frac{w_x^\varepsilon(x,\alpha)}{\varepsilon} + \exp\left(-\frac{w^\varepsilon(x,\alpha)}{\varepsilon}\right)\frac{Q}{\varepsilon}\exp\left(\frac{w^\varepsilon(x,\cdot)}{\varepsilon}\right)(\alpha) + \frac{L(x,u)}{\varepsilon}\right\},$$

where $w^\varepsilon(x,\alpha)$ is the potential function, $w_x^\varepsilon(x,\alpha)$ denotes the partial derivative of $w^\varepsilon(x,\alpha)$ with respect to x, and $Qf(\cdot)(i) := \sum_{j\ne i} q_{ij}(f(j) - f(i))$ for a function f on

$\mathcal{M}$. By multiplying ε on both sides of this equation, we have

$$\lambda^\varepsilon = \inf_{0\le u\le\alpha}\Big\{(-ax+u-z)w_x^\varepsilon(x,\alpha) + \exp\Big(-\frac{w^\varepsilon(x,\alpha)}{\varepsilon}\Big)Q\exp\Big(\frac{w^\varepsilon(x,\cdot)}{\varepsilon}\Big)(\alpha)+L(x,u)\Big\}. \tag{3.1}$$

As in almost all long-run average cost problems, an immediate question is if the equation (3.1) has a solution in some sense. In this paper, we will show that (3.1) indeed has a solution in the viscosity sense. We use a vanishing discount approach. Let $\rho>0$ denote a discount factor and let $w_\rho^\varepsilon(x,\alpha)$ denote the corresponding value function. Then, the associated HJB equation has the form

$$\rho w_\rho^\varepsilon(x,\alpha) = \inf_{0\le u\le\alpha}\Big\{(-ax+u-z)(w_\rho^\varepsilon)_x(x,\alpha) +\exp\Big(-\frac{w_\rho^\varepsilon(x,\alpha)}{\varepsilon}\Big)\frac{Q}{\varepsilon}\exp\Big(\frac{w_\rho^\varepsilon(x,\cdot)}{\varepsilon}\Big)(\alpha)+L(x,u)\Big\}. \tag{3.2}$$

Let

$$\psi_\rho^\varepsilon(x,\alpha)=\exp\Big(\frac{w_\rho^\varepsilon(x,\alpha)}{\varepsilon}\Big).$$

Then, (3.2) becomes

$$\rho\psi_\rho^\varepsilon(x,\alpha)\log\psi_\rho^\varepsilon(x,\alpha) = \inf_{0\le u\le\alpha}\Big\{(-ax+u-z)(\psi_\rho^\varepsilon)_x(x,\alpha) +\frac{L(x,u)\psi_\rho^\varepsilon(x,\alpha)}{\varepsilon}+\frac{Q}{\varepsilon}\psi_\rho^\varepsilon(x,\cdot)(\alpha)\Big\}. \tag{3.3}$$

We would like to get rid of the term $\psi_\rho^\varepsilon(x,\alpha)\log\psi_\rho^\varepsilon(x,\alpha)$. One way of doing so is to use the transform device introduced by Bensoussan and Nagai [3] based on the following expression:

$$-r\log r=\inf_y\{yr+e^{-(y+1)}\}\ \text{ for any } r>0, \tag{3.4}$$

where the minimum is obtained at $y+1=-\log r$. Letting $r=\psi_\rho^\varepsilon(x,\alpha)$, we have

$$\psi_\rho^\varepsilon(x,\alpha)\log\psi_\rho^\varepsilon(x,\alpha)=-\inf_y\{y\psi_\rho^\varepsilon(x,\alpha)+e^{-(y+1)}\}.$$

In view of this and (3.3), the discounted HJB equation (3.2) has the form

$$0=\inf_{0\le u\le\alpha,y}\Big\{(-ax+u-z)(\psi_\rho^\varepsilon)_x(x,\alpha) +\frac{L(x,u)\psi_\rho^\varepsilon(x,\alpha)}{\varepsilon}+\frac{Q}{\varepsilon}\psi_\rho^\varepsilon(x,\cdot)(\alpha)+\rho y\psi_\rho^\varepsilon(x,\alpha)+\rho e^{-(y+1)}\Big\}.$$

By adding $\rho\psi_\rho^\varepsilon(x,\alpha)$ to both sides of this equation and changing $(y+1)$ to y, we obtain

$$\rho\psi_\rho^\varepsilon(x,\alpha) = \inf_{0\le u\le\alpha,y}\Big\{(-ax+u-z)(\psi_\rho^\varepsilon)_x(x,\alpha) +\frac{L(x,u)\psi_\rho^\varepsilon(x,\alpha)}{\varepsilon}+\frac{Q}{\varepsilon}\psi_\rho^\varepsilon(x,\cdot)(\alpha)+\rho y\psi_\rho^\varepsilon(x,\alpha)+\rho e^{-y}\Big\}. \tag{3.5}$$

Remark 3.1. Note that the HJB equation (3.5) consists of a set of equations coupled by a discrete variable α. A viscosity solution of such HJB equations has been considered by Soner [16] and Fleming, Sethi, and Soner [8]; see also Fleming and Soner [10] for a more general setting and Sethi and Zhang [15] for several equivalent definitions.

We consider the following control problem so that the value function is a viscosity solution to this equation:

$$\begin{aligned}
&\text{Minimize}\\
&\quad J_\rho^\varepsilon(x,\alpha,u(\cdot),y(\cdot))\\
&\qquad = E\int_0^\infty e^{-\rho t}\exp\left[\int_0^t\left(\frac{L(x(s),u(s))}{\varepsilon}+\rho y(s)\right)ds\right](\rho e^{-y(t)})dt\\
&\text{subject to}\\
&\quad \dot{x}(t) = -ax(t)+u(t)-z,\ x(0)=x \text{ and}\\
&\quad \text{both } u(\cdot) \text{ and } y(\cdot) \text{ are } \alpha^\varepsilon(\cdot) \text{ adapted such that}\\
&\quad 0\le u(t)\le \alpha^\varepsilon(t) \text{ and } -M_0\le y(t)\le 0,\ t\ge 0\\
&\quad \text{for any constant } M_0 \ge \sup_{x,u}|L(x,u)|/(\rho\varepsilon).
\end{aligned} \tag{3.6}$$

With a little abuse of notation, let $\psi_\rho^\varepsilon(x,\alpha)$ denote the value function of this control problem. We next show that such $\psi_\rho^\varepsilon(x,\alpha)$ is a viscosity solution to (3.5) with some a priori estimates.

LEMMA 3.2. (i) *For all x and α,*

$$1\le \psi_\rho^\varepsilon(x,\alpha)\le \exp\left(\frac{C_1}{\rho\varepsilon}\right),$$

where $C_1 = ||L(x,\alpha)|| := \sup_{x,u}|L(x,u)|$.

(ii) *For all x, $\tilde{x}$, and α,*

$$\exp\left(-\frac{C_2|x-\tilde{x}|}{\varepsilon}\right)\le \frac{\psi_\rho^\varepsilon(x,\alpha)}{\psi_\rho^\varepsilon(\tilde{x},\alpha)}\le \exp\left(\frac{C_2|x-\tilde{x}|}{\varepsilon}\right),$$

where $C_2 = ||L_x(x,u)||/a$.

(iii) *For each $r>0$ there is a constant $C_3>0$ independent of $\rho\le 1$ and ε such that, for all α, $\tilde{\alpha}$, and $|x|\le r$,*

$$e^{-C_3}\le \frac{\psi_\rho^\varepsilon(x,\alpha)}{\psi_\rho^\varepsilon(x,\tilde{\alpha})}\le e^{C_3}.$$

(iv) *$\psi_\rho^\varepsilon(x,\alpha)$ is a viscosity solution to* (3.5).

Proof. We begin with (i). We first show $\psi_\rho^\varepsilon(x,\alpha)\ge 1$. In view of the nonnegativity of $L(x,u)$, it suffices to show for all deterministic Borel measurable $-M_0\le y(t)\le 0$, $t\ge 0$,

$$\int_0^\infty \rho\exp\left(\int_0^t \rho(y(s)-1)ds - y(t)\right)dt \ge 1. \tag{3.7}$$

Let

$$\gamma = \inf_{y(\cdot)}\int_0^\infty \rho\exp\left(\int_0^t \rho(y(s)-1)ds - y(t)\right)dt.$$

First, it is easy to see that $\gamma \geq 0$. By taking $y(t) = 0$, $t \geq 0$, we obtain

$$\int_0^\infty \rho \exp\left(\int_0^t \rho(y(s) - 1)ds - y(t)\right) dt = \int_0^\infty \rho \exp\left(-\rho t\right) dt = 1.$$

Thus, by definition, $\gamma \leq 1$.

In view of the control problem defined in (3.6), we may consider γ as the value function of a problem with no state and control costs, i.e., with $L(x, u)$ replaced by 0. Then, following the standard dynamic programming approach as in Sethi and Zhang [15], we can show that the constant γ is the unique solution to the following HJB equation:

$$\rho\gamma = \inf_{-M_0 \leq y \leq 0} \{\rho y\gamma + \rho e^{-y}\}.$$

The only solution to this equation is $\gamma = 1$. Thus, the inequality (3.7) follows.

Let $y(t) = -M$ for all $t \geq 0$ and let $C_1 = ||L(x, u)||$. Then, for all admissible $u(\cdot)$,

$$\begin{aligned}\psi_\rho^\varepsilon(x, \alpha) &\leq E \int_0^\infty \rho \exp\left[\int_0^t \left(\frac{C_1}{\varepsilon} - \rho(M + 1)\right) ds + M\right] dt \\ &= \frac{\varepsilon\rho e^M}{|C_1 - \rho\varepsilon(M + 1)|}.\end{aligned}$$

Let $M = C_1/(\rho\varepsilon)$. Then, $\varepsilon\rho e^M/|C_1 - \rho\varepsilon(M + 1)| = e^M = e^{C_1/(\rho\varepsilon)}$. This proves (i).

We now prove (ii). Let $(u(\cdot), y(\cdot))$ denote a pair of admissible controls and let $x(t)$ and $\tilde{x}(t)$ denote the corresponding trajectories with initial values x and $\tilde{x}$, respectively. Then,

$$x(t) - \tilde{x}(t) = (x - \tilde{x})e^{-at} \quad \text{for all } t \geq 0.$$

In view of this and the Lipschitz property of $L(x, u)$, we have

$$L(x(t), u(t)) \leq L(\tilde{x}(t), u(t)) + C_0|x - \tilde{x}|e^{-at},$$

where $C_0 = ||L_x(x, u)||$. For notational simplification, let

$$\eta(t, y(\cdot)) = \rho \exp\left(\int_0^t \rho(y(s) - 1)ds - y(t)\right).$$

Then, we have

$$\begin{aligned}&E \int_0^\infty \left[\exp\left(\int_0^t \frac{L(x(s), u(s))}{\varepsilon} ds\right)\right] \eta(t, y(\cdot))dt \\ &\quad \leq E \int_0^\infty \left[\exp\left(\int_0^t \frac{L(\tilde{x}(s), u(s))}{\varepsilon} ds\right) \exp\left(\int_0^t \frac{C_0|x - \tilde{x}|e^{-as}}{\varepsilon} ds\right)\right] \eta(t, y(\cdot))dt \\ &\quad = \exp\left(\int_0^\infty \frac{C_0|x - \tilde{x}|e^{-as}}{\varepsilon} ds\right) E \int_0^\infty \left[\exp\left(\int_0^t \frac{L(\tilde{x}(s), u(s))}{\varepsilon} ds\right)\right] \eta(t, y(\cdot))dt \\ &\quad = \exp\left(\frac{C_0|x - \tilde{x}|}{a\varepsilon}\right) E \int_0^\infty \left[\exp\left(\int_0^t \frac{L(\tilde{x}(s), u(s))}{\varepsilon} ds\right)\right] \eta(t, y(\cdot))dt.\end{aligned}$$

Hence,

$$\psi_\rho^\varepsilon(x, \alpha) \leq \left[\exp\left(\frac{C_0|x - \tilde{x}|}{a\varepsilon} ds\right)\right] \psi_\rho^\varepsilon(\tilde{x}, \alpha).$$

Similarly, we can show the other inequality in (ii).

We now show (iii). Let $\alpha^\varepsilon(0)=\alpha$ and τ denote the first time $\alpha^\varepsilon(\cdot)$ jumps to $\tilde\alpha$. Let

$$G(s)=\frac{L(x(s),u(s))}{\varepsilon}+\rho(y(s)-1)\quad\text{and}\quad h(t)=\rho e^{-y(t)}.$$

Then, the dynamic programming principle with the random stopping time τ (see the Appendix for a sketch of the proof) gives

$$\psi^\varepsilon_\rho(x,\alpha)=\inf_{u(\cdot),y(\cdot)}E\Bigg\{\int_0^\tau\left(\exp\int_0^t G(s)ds\right)h(t)dt + \left(\exp\int_0^\tau G(s)ds\right)\psi^\varepsilon_\rho(x(\tau),\tilde\alpha)\Bigg\}.$$

Using $L\ge 0$ and $h>0$, we have

$$\psi^\varepsilon_\rho(x,\alpha)\ge E\left\{\left(\exp\int_0^\tau\rho(y(s)-1)ds\right)\psi^\varepsilon_\rho(x(\tau),\tilde\alpha)\right\}.$$

For x in any bounded interval, by (ii) and $|x(\tau)-x|\le K\tau(1+|x|)\le K\tau(1+r)$,

$$\frac{\psi^\varepsilon_\rho(x(\tau),\tilde\alpha)}{\psi^\varepsilon_\rho(x,\tilde\alpha)}\ge\exp\left(-\frac{C_2K\tau(1+|x|)}{\varepsilon}\right)\ge\exp\left(-\frac{C_2K\tau(1+r)}{\varepsilon}\right).$$

We can assume $y(t)\ge -||L||/\rho\varepsilon$, which implies

$$\exp\int_0^t\rho(y(s)-1)ds\ge\exp\left(-\left[\frac{||L||}{\varepsilon}+\rho\right]\tau\right).$$

Take $B>||L||+\varepsilon\rho+C_2K(1+r)$. Change of time scale $t=\varepsilon t'$ sends $Q/\varepsilon\to Q$, which implies

$$Ee^{-B\tau/\varepsilon}\ge e^{-C_3}\quad\text{for some } C_3>0.$$

Therefore,

$$\psi^\varepsilon_\rho(x,\alpha)\ge\psi^\varepsilon_\rho(x,\tilde\alpha)e^{-C_3}.$$

Exchange α and $\tilde\alpha$ to get the opposite inequality.

Finally it can be shown as in Sethi and Zhang [15] that $\psi^\varepsilon_\rho(x,\alpha)$ is a viscosity solution to (3.5) under the constraint $-M_0\le y\le 0$; i.e.,

$$\begin{aligned}\rho\psi^\varepsilon_\rho(x,\alpha)=\inf_{0\le u\le\alpha,-M_0\le y\le 0}\Bigg\{&(-ax+u-z)(\psi^\varepsilon_\rho)_x(x,\alpha)\\&+\frac{L(x,u)\psi^\varepsilon_\rho(x,\alpha)}{\varepsilon}+\frac{Q}{\varepsilon}\psi^\varepsilon_\rho(x,\cdot)(\alpha)+\rho y\psi^\varepsilon_\rho(x,\alpha)+\rho e^{-y}\Bigg\}.\end{aligned}\tag{3.8}$$

Since $\psi^\varepsilon_\rho(x,\alpha)\ge 1$ and the minimum in (3.8) is obtained at $y=-\log\psi^\varepsilon_\rho(x,\alpha)\le 0$, $\psi^\varepsilon_\rho(x,\alpha)$ is also a viscosity solution to (3.5). The proof of the lemma is complete. □

THEOREM 3.3. *The HJB equation* (3.1) *has a viscosity solution* $(\lambda^\varepsilon,w^\varepsilon(x,\alpha))$.

Proof. In this proof, ε is fixed. In view of the logarithmic transformation

$$\psi_\rho^\varepsilon(x,\alpha) = \exp\left(\frac{w_\rho^\varepsilon(x,\alpha)}{\varepsilon}\right),$$

we have $w_\rho^\varepsilon(x,\alpha) = \varepsilon \log \psi_\rho^\varepsilon(x,\alpha)$. It follows from Lemma 3.2 that

(i) $0 \le \rho w_\rho^\varepsilon(x,i) \le C_1$, uniformly in ρ;

(ii) $|w_\rho^\varepsilon(x,\alpha) - w_\rho^\varepsilon(\tilde{x},\alpha)| \le C_2|\tilde{x}-x|$, uniformly in ρ;

(iii) for each $r>0$, $|x|\le r$, $\alpha,\tilde{\alpha}\in\mathcal{M}$, $|w_\rho^\varepsilon(x,\alpha) - |w_\rho^\varepsilon(x,\tilde{\alpha})| \le \varepsilon \log C_3$, uniformly in ρ.

Then, in view of these and the Arzela–Ascoli theorem, it is easy to see that, for each (x,α), there exist a sequence $\rho_n \to 0$ such that $\rho_n w_{\rho_n}^\varepsilon(0,0) \to \lambda^\varepsilon$ and

$$\begin{aligned} w_{\rho_n}^\varepsilon(x,\alpha) - w_{\rho_n}^\varepsilon(0,0) &= (w_{\rho_n}^\varepsilon(x,\alpha) - w_{\rho_n}^\varepsilon(0,\alpha)) \\ &\quad +(w_{\rho_n}^\varepsilon(0,\alpha) - w_{\rho_n}^\varepsilon(0,0)) \to w^\varepsilon(x,\alpha) \end{aligned}$$

on any compact subset of $\mathbb{R}\times\mathcal{M}$. Therefore,

$$\begin{aligned} \rho_n w_{\rho_n}^\varepsilon(x,\alpha) &= \rho_n(w_{\rho_n}^\varepsilon(x,\alpha) - w_{\rho_n}^\varepsilon(0,\alpha)) \\ &\quad +\rho_n(w_{\rho_n}^\varepsilon(0,\alpha) - \rho_n w_{\rho_n}^\varepsilon(0,0)) + \rho_n w_{\rho_n}^\varepsilon(0,0) \to \lambda^\varepsilon. \end{aligned}$$

Finally, it can be shown, as in Fleming and Soner [10], that the limit $(\lambda^\varepsilon, w^\varepsilon(x,\alpha))$ is a viscosity solution to the HJB equation (3.1). □

COROLLARY 3.4. *The pair $(\lambda^\varepsilon, w^\varepsilon(x,\alpha))$ obtained in Theorem 3.3 satisfies the following conditions.*

For some constant C independent of $\varepsilon>0$,

(i) $0 \le \lambda^\varepsilon \le C_1$ *and*

(ii) $|w^\varepsilon(x,\alpha) - w^\varepsilon(\tilde{x},\alpha)| \le C_2|x-\tilde{x}|$.

Proof. It is easy to check from the proof of Theorem 3.3 that (i) holds and

$$|w_{\rho_n}^\varepsilon(x,\alpha) - w_{\rho_n}^\varepsilon(\tilde{x},\alpha)| \le C_2|x-\tilde{x}|,$$

$$|w^\varepsilon(x,\alpha) - w^\varepsilon(\tilde{x},\alpha)| = \lim_{\rho_n\to 0} |w_{\rho_n}^\varepsilon(x,\alpha) - w_{\rho_n}^\varepsilon(\tilde{x},\alpha)| \le C_2|x-\tilde{x}|. \qquad \square$$

THEOREM 3.5. *Let $(\lambda^\varepsilon, w^\varepsilon(x,\alpha))$ be a viscosity solution to the HJB equation in (3.1). Assume $w^\varepsilon(x,\alpha)$ to be Lipschitz continuous in x. Then*

$$\lambda^\varepsilon = \inf_{u(\cdot)\in\mathcal{A}^\varepsilon} J^\varepsilon(u(\cdot)),$$

where $J^\varepsilon(u(\cdot))$ is defined in (2.2).

Proof. We divide the proof into two steps.

Step 1. $\lambda^\varepsilon \le \inf_{u(\cdot)\in\mathcal{A}^\varepsilon} J^\varepsilon(u(\cdot))$.

Let $\psi^\varepsilon(x,\alpha) = \exp(w^\varepsilon(x,\alpha)/\varepsilon)$. Then, the HJB equation (3.1) becomes

$$\frac{\lambda^\varepsilon \psi^\varepsilon(x,\alpha)}{\varepsilon} = \inf_{0\le u\le\alpha}\left\{(-ax+u-z)\psi_x^\varepsilon(x,\alpha) + \frac{L(x,u)}{\varepsilon}\psi^\varepsilon(x,\alpha) + \frac{Q}{\varepsilon}\psi^\varepsilon(x,\cdot)(\alpha)\right\}.$$

It is equivalent to

$$0 = \inf_{0\le u\le\alpha}\left\{(-ax+u-z)\psi_x^\varepsilon(x,\alpha) + \frac{L(x,u)-\lambda^\varepsilon}{\varepsilon}\psi^\varepsilon(x,\alpha) + \frac{Q}{\varepsilon}\psi^\varepsilon(x,\cdot)(\alpha)\right\}.$$

It is easy to see that $\psi^\varepsilon(x,\alpha)$ is a viscosity solution to the following time-dependent equation for $\phi(T,x,\alpha)$:

$$\begin{cases} \dfrac{\partial\phi}{\partial T} = \inf\limits_{0\le u\le\alpha}\left\{(-ax+u-z)\phi_x + \dfrac{L-\lambda^\varepsilon}{\varepsilon}\phi + \dfrac{Q}{\varepsilon}\phi\right\}, \\ \phi(0,x,\alpha) = \psi^\varepsilon(x,\alpha). \end{cases} \tag{3.9}$$

As can be shown as in Sethi and Zhang [15, Appendix G], this HJB equation has a unique viscosity solution. Moreover, if we define

$$\phi^\varepsilon(T,x,\alpha) = \inf_{u(\cdot)\in\mathcal{A}^\varepsilon} E\left(\psi^\varepsilon(x(T),\alpha^\varepsilon(T))\exp\int_0^T \frac{L(x(t),u(t))-\lambda^\varepsilon}{\varepsilon}dt\right),$$

then, using the dynamic programming principle (see Appendix), it can be shown that $\phi^\varepsilon(T,x,\alpha)$ is also a viscosity solution to (3.9). Thus, $\phi^\varepsilon(T,x,\alpha) = \psi^\varepsilon(x,\alpha)$ for all $T\ge 0$. Namely,

$$\psi^\varepsilon(x,\alpha) = \inf_{u(\cdot)\in\mathcal{A}^\varepsilon} E\left(\psi^\varepsilon(x(T),\alpha^\varepsilon(T))\exp\int_0^T \frac{L(x(t),u(t))-\lambda^\varepsilon}{\varepsilon}dt\right). \tag{3.10}$$

It follows that for all $u(\cdot)\in\mathcal{A}^\varepsilon$,

$$\begin{aligned} \psi^\varepsilon(x,\alpha) &\le E\left(\psi^\varepsilon(x(T),\alpha^\varepsilon(T))\exp\int_0^T \frac{L(x(t),u(t))-\lambda^\varepsilon}{\varepsilon}dt\right) \\ &= E\left(\psi^\varepsilon(x(T),\alpha^\varepsilon(T))\exp\int_0^T \frac{L(x(t),u(t))}{\varepsilon}dt\right)\exp\left(-\frac{\lambda^\varepsilon T}{\varepsilon}\right). \end{aligned}$$

Taking the logarithm of both sides, we have

$$\log\psi^\varepsilon(x,\alpha) \le \log E\left(\psi^\varepsilon(x(T),\alpha^\varepsilon(T))\exp\int_0^T \frac{L(x(t),u(t))}{\varepsilon}dt\right) - \frac{\lambda^\varepsilon T}{\varepsilon}. \tag{3.11}$$

Recall the Lipschitz property of $w^\varepsilon(x,\alpha)$ in x. It follows for all x and $\tilde{x}$ that

$$\frac{\psi^\varepsilon(\tilde{x},\alpha)}{\psi^\varepsilon(x,\alpha)} = \exp\left(\frac{w^\varepsilon(\tilde{x},\alpha)-w^\varepsilon(x,\alpha)}{\varepsilon}\right) \le \exp\left(\frac{C_2|\tilde{x}-x|}{\varepsilon}\right).$$

Replacing $\tilde{x}$ by $x(T)$ and α by $\alpha^\varepsilon(T)$, respectively, we obtain

$$\psi^\varepsilon(X(T),\alpha^\varepsilon(T)) \le \psi^\varepsilon(x,\alpha^\varepsilon(T))\exp\left(\frac{C_2|x(T)-x|}{\varepsilon}\right).$$

Note also that $|x(T)-x|\le K(1+|x|)$ for some constant K (see (2.3)) and

$$\psi^\varepsilon(x,\alpha^\varepsilon(T)) \le M(x) := \max_{j\in\mathcal{M}}\psi^\varepsilon(x,j).$$

We have

$$\psi^\varepsilon(X(T),\alpha^\varepsilon(T)) \le M(x)\exp\left(\frac{C_2K(1+|x|)}{\varepsilon}\right). \tag{3.12}$$

Combining this inequality with (3.11), we obtain

$$\log \psi^\varepsilon(x,\alpha) \le \log M(x) + \frac{C_2K(1+|x|)}{\varepsilon} + \log E\exp\int_0^T \frac{L(x(t),u(t))}{\varepsilon}dt - \frac{\lambda^\varepsilon T}{\varepsilon}.$$

Dividing both sides by T and letting $T\to\infty$ yields

$$\lambda^\varepsilon \le J^\varepsilon(u(\cdot)) \quad \text{for all } u(\cdot)\in\mathcal{A}^\varepsilon.$$

Thus, $\lambda^\varepsilon \le \inf_{u(\cdot)\in\mathcal{A}^\varepsilon} J^\varepsilon(u(\cdot))$ for all $u(\cdot)\in\mathcal{A}^\varepsilon$.

Step 2. $\lambda^\varepsilon \ge \inf_{u(\cdot)\in\mathcal{A}^\varepsilon} J^\varepsilon(u(\cdot))$.

Let

$$V^\varepsilon(T,x,\alpha) = \inf_{u(\cdot)\in\mathcal{A}^\varepsilon} \log E\exp\left(\int_0^T \frac{L(x(t),u(t))}{\varepsilon}dt\right).$$

We first show that

$$\frac{\lambda^\varepsilon}{\varepsilon} = \lim_{T\to\infty}\frac{1}{T}V^\varepsilon(T,x,\alpha), \tag{3.13}$$

uniformly for x in any compact set.

In fact, as in (3.12), we can show that there exist positive constants K_1 and K_2 such that for all $x=x(0)$ and $T>0$,

$$\exp\left(-\frac{K_1(1+|x|)}{\varepsilon}\right) \le \psi^\varepsilon(x(T),\alpha^\varepsilon(T)) \le \exp\left(\frac{K_2(1+|x|)}{\varepsilon}\right).$$

In view of this and (3.10), we have

$$\begin{aligned}
&\exp\left(-\frac{K_1(1+|x|)}{\varepsilon}\right)\inf_{u(\cdot)\in\mathcal{A}^\varepsilon} E\exp\left(\int_0^T \frac{L(x(t),u(t))}{\varepsilon}dt\right)\\
&\le \psi^\varepsilon(x,\alpha)\exp\left(\frac{\lambda^\varepsilon T}{\varepsilon}\right)\\
&\le \exp\left(\frac{K_2(1+|x|)}{\varepsilon}\right)\inf_{u(\cdot)\in\mathcal{A}^\varepsilon} E\exp\left(\int_0^T \frac{L(x(t),u(t))}{\varepsilon}dt\right).
\end{aligned}$$

Taking the logarithm on both sides and noting that $\inf_{u(\cdot)\in\mathcal{A}^\varepsilon}\log(\cdots) = \log\inf_{u(\cdot)\in\mathcal{A}^\varepsilon}(\cdots)$, we obtain

$$-\frac{K_1(1+|x|)}{\varepsilon} + V^\varepsilon(T,x,\alpha) \le \log\psi^\varepsilon(x,\alpha) + \frac{\lambda^\varepsilon T}{\varepsilon} \le \frac{K_2(1+|x|)}{\varepsilon} + V^\varepsilon(T,x,\alpha).$$

Dividing both sides by T and sending $T\to\infty$, we arrive at (3.13).

In view of (2.3), for any fixed $r>0$, there exists $r_1>0$ such that $|x(t)|\le r_1$ for all $t\ge 0$, $\alpha\in\mathcal{M}$, and $|x|\le r$. Therefore, for each $\delta>0$ there exists T_0 such that

$$\left|\frac{\lambda^\varepsilon}{\varepsilon} - \frac{1}{T_0}V^\varepsilon(T_0,x,\alpha)\right| \le \delta$$

for all $\alpha\in\mathcal{M}$ and $|x|\le r_1$. Hence,

$$V^\varepsilon(T_0,x,\alpha) \le \frac{\lambda^\varepsilon T_0}{\varepsilon} + T_0\delta \tag{3.14}$$

for all $\alpha \in \mathcal{M}$ and $|x| \le r_1$.

On $[0, T_0)$, choose an admissible $u^{(1)}(t)$ such that

$$E \exp\left(\int_0^{T_0} \frac{L(x(t), u^{(1)}(t))}{\varepsilon} dt\right) \le \exp\left(V^\varepsilon(T_0, x, \alpha) + \delta T_0\right) \le \exp\left(\frac{\lambda^\varepsilon T_0}{\varepsilon} + 2\delta T_0\right).$$

Let $\mathcal{G}_{T_0} = \sigma\{(x(t), \alpha^\varepsilon(t)) : \ t \le T_0\}$. On $[T_0, 2T_0)$, if we choose $u^{(2)}(t)$ to be $\sigma\{\alpha^\varepsilon(s) : \ T_0 \le s \le t\}$ measurable, then

$$E\left\{\exp\left(\int_{T_0}^{2T_0} \frac{L(x(t), u^{(2)}(t))}{\varepsilon} dt\right) \Big| \mathcal{G}_{T_0}\right\}$$

is a function of $(T_0, x(T_0), \alpha^\varepsilon(T_0))$. More precisely, if we let

$$\Phi(T_0, x, \alpha, u(\cdot)) = E\left\{\exp\left(\int_{T_0}^{2T_0} \frac{L(x(t), u(t))}{\varepsilon} dt\right) \Big| x(T_0) = x, \alpha^\varepsilon(T_0) = \alpha\right\},$$

then

$$\Phi(T_0, x(T_0), \alpha^\varepsilon(T_0), u^{(2)}(\cdot)) = E\left\{\exp\left(\int_{T_0}^{2T_0} \frac{L(x(t), u^{(2)}(t))}{\varepsilon} dt\right) \Big| \mathcal{G}_{T_0}\right\}.$$

Moreover, by changing the variable $t \to (t - T_0)$, we have

$$V^\varepsilon(T_0, x, \alpha) = \inf_{u(\cdot) \in \mathcal{A}^\varepsilon} \log \Phi(T_0, x, \alpha, u(\cdot)).$$

Similarly, as in the proof of Lemma 3.2 (ii), we can show for some constant C,

$$|V^\varepsilon(T, \tilde{x}, \alpha) - V^\varepsilon(T, x, \alpha)| \le \frac{CT|\tilde{x} - x|}{\varepsilon}$$

for all T, $\tilde{x}$, x, and $\alpha \in \mathcal{M}$.

Let B_1, B_2,..., B_l be a partition of $\{x : \ |x| \le r_1\}$. For any given $\delta > 0$, if the diameter of the B_j's is small enough, then for all $\tilde{x}$ and x in B_j, and $u(\cdot) \in \mathcal{A}^\varepsilon$,

$$|V^\varepsilon(T_0, \tilde{x}, \alpha) - V^\varepsilon(T_0, x, \alpha)| \le \delta T_0 \quad \text{and} \quad \frac{\Phi(T_0, \tilde{x}, \alpha, u(\cdot))}{\Phi(T_0, x, \alpha, u(\cdot))} \le e^{\delta T_0}.$$

For $j = 1, 2, \ldots, l$, pick out $x_j \in B_j$. For each (j, α), choose $u^{(2)}_{j,\alpha}(t)$ on $[T_0, 2T_0)$ such that

$$\Phi(T_0, x_j, \alpha) \le \exp(V^\varepsilon(T_0, x_j, \alpha) + \delta T_0) \le \exp\left(\frac{\lambda^\varepsilon T_0}{\varepsilon} + 2\delta T_0\right).$$

On $[0, 2T_0)$, define

$$u(t) = \begin{cases} u^{(1)}(t) & \text{if } 0 \le t < T_0, \\ \displaystyle\sum_{j,\alpha} I_{\{(x(T_0), \alpha^\varepsilon(T_0)) \in B_j \times \{\alpha\}\}} u^{(2)}_{j,\alpha}(t) & \text{if } T_0 \le t < 2T_0, \end{cases} \tag{3.15}$$

where I_F is the indicator function of a set F. It follows that

$$\begin{aligned} & E\left\{\exp\left(\int_{T_0}^{2T_0}\frac{L(x(t),u(t))}{\varepsilon}dt\right)\Big|\mathcal{G}_{T_0}\right\} \\ &\quad = \sum_{j,\alpha} I_{\{x(T_0)\in B_j\}} I_{\{\alpha^\varepsilon(T_0)=\alpha\}} \Phi(T_0, x(T_0), \alpha, u^{(2)}_{j,\alpha}(t)) \\ &\quad \le \sum_{j,\alpha} I_{\{x(T_0)\in B_j\}} I_{\{\alpha^\varepsilon(T_0)=\alpha\}} \Phi(T_0, x_j, \alpha, u^{(2)}_{j,\alpha}(t)) e^{\delta T_0} \\ &\quad \le \sum_{j,\alpha} I_{\{x(T_0)\in B_j\}} I_{\{\alpha^\varepsilon(T_0)=\alpha\}} \exp\left(\frac{\lambda^\varepsilon T_0}{\varepsilon} + 3\delta T_0\right). \end{aligned}$$

Note that

$$\begin{aligned} & E\exp\left(\int_0^{2T_0}\frac{L(x(t),u(t))}{\varepsilon}dt\right) \\ &\quad = E\left\{\exp\left(\int_0^{T_0}\frac{L(x(t),u(t))}{\varepsilon}dt\right) E\left[\exp\left(\int_{T_0}^{2T_0}\frac{L(x(t),u(t))}{\varepsilon}dt\right)\Big|\mathcal{G}_{T_0}\right]\right\}. \end{aligned}$$

It follows that

$$E\exp\left(\int_0^{2T_0}\frac{L(x(t),u(t))}{\varepsilon}dt\right) \le \exp\left(\frac{\lambda^\varepsilon(2T_0)}{\varepsilon} + 5\delta T_0\right).$$

Continuing this procedure on $[(N-1)T_0, NT_0)$ for $N = 3, \ldots$, we can construct an admissible control $u(t)$ as in (3.15) such that

$$E\exp\left(\int_0^{NT_0}\frac{L(x(t),u(t))}{\varepsilon}dt\right) \le \exp\left(\frac{\lambda^\varepsilon NT_0}{\varepsilon} + \delta(3N-1)T_0\right).$$

Hence,

$$\frac{1}{NT_0}\log E\exp\left(\int_0^{NT_0}\frac{L(x(t),u(t))}{\varepsilon}dt\right) \le \frac{\lambda^\varepsilon}{\varepsilon} + \frac{\delta(3N-1)T_0}{NT_0} \to \frac{\lambda^\varepsilon}{\varepsilon} + 3\delta.$$

Since δ is arbitrary, $\lambda^\varepsilon \ge \inf_{u(\cdot)\in\mathcal{A}^\varepsilon} J^\varepsilon(u(\cdot))$ follows. □

This theorem implies that λ^ε in $(\lambda^\varepsilon, w^\varepsilon(x,\alpha))$ as a viscosity solution is unique.

We next give a verification theorem. In order to incorporate nondifferentiability of the value function, we consider the superdifferential of the function. Let $D^+f(x)$ denote the superdifferential of a function $f(x)$, i.e.,

$$D^+f(x) = \left\{r\in\mathbb{R} : \limsup_{h\to 0}\frac{f(x+h)-f(x)-hr}{|h|} \le 0\right\}.$$

THEOREM 3.6. *Let $(\lambda^\varepsilon, w^\varepsilon(x,\alpha))$ be a viscosity solution to the HJB equation in (3.1). Assume that $w^\varepsilon(x,\alpha)$ is Lipschitz continuous in x. Let $\psi^\varepsilon(x,\alpha) = \exp(w^\varepsilon(x,\alpha)/\varepsilon)$. Suppose that there are $u^*(\cdot)$, $x^*(\cdot)$, and $r^*(t)$ such that*

$$\dot{x}^*(t) = -ax^*(t) + u^*(t) - z, \qquad x^*(0) = x,$$

$r^*(t) \in D^+\psi_x^\varepsilon(x^*(t), \alpha^\varepsilon(t))$ *satisfying*

$$\begin{aligned}\frac{\lambda^\varepsilon}{\varepsilon}\psi^\varepsilon(x^*(t),\alpha^\varepsilon(t)) &= (-ax^*(t)+u^*(t)-z)r^*(t)\\ &\quad+\frac{L(x^*(t),u^*(t))}{\varepsilon}\psi^\varepsilon(x^*(t),\alpha^\varepsilon(t))+\frac{Q}{\varepsilon}\psi^\varepsilon(x^*(t),\cdot)(\alpha^\varepsilon(t))\end{aligned} \tag{3.16}$$

almost everywhere (a.e.) in t *and with probability* 1 *(w.p.*1*). Then,* $\lambda^\varepsilon = J^\varepsilon(u^*(\cdot))$.

Proof. First, note that the HJB equation in (3.1) is equivalent to

$$\begin{aligned}\frac{\lambda^\varepsilon}{\varepsilon}\psi^\varepsilon(x,\alpha) = \inf_{0\le u\le\alpha}\Bigg\{&(-ax+u-z)\psi_x^\varepsilon(x,\alpha)\\ &+\frac{L(x,\alpha)}{\varepsilon}\psi^\varepsilon(x,\alpha)+\frac{Q}{\varepsilon}\psi^\varepsilon(x,\cdot)(\alpha)\Bigg\}.\end{aligned} \tag{3.17}$$

The Lipschitz property of $\psi^\varepsilon(x,\alpha)$ implies that $\psi^\varepsilon(x(t),\alpha)$ is Lipschitz in t. For each $t \ge 0$ such that $(d/dt)\psi^\varepsilon(x(t),\alpha)$ exists and

$$\int_t^{t+h}(-ax(s)+u(s)-z)ds = h(-ax(t)+u(t)-z)+o(h),$$

we have

$$\begin{aligned}\frac{d\psi^\varepsilon(x(t),\alpha)}{dt} &= \lim_{h\to0^+}\frac{1}{h}\left(\psi^\varepsilon(x(t+h),\alpha)-\psi^\varepsilon(x(t),\alpha)\right)\\ &= \lim_{h\to0^+}\frac{1}{h}\left(\psi^\varepsilon\left(x(t)+\int_t^{t+h}(-ax(s)+u(s)-z)ds,\alpha\right)-\psi^\varepsilon(x(t),\alpha)\right)\\ &= \lim_{h\to0^+}\frac{1}{h}\left(\psi^\varepsilon(x(t)+h(-ax(t)+u(t)-z)+o(h),\alpha)-\psi^\varepsilon(x(t),\alpha)\right)\\ &= \lim_{h\to0^+}\frac{1}{h}\left(\psi^\varepsilon(x(t)+h(-ax(t)+u(t)-z),\alpha)-\psi^\varepsilon(x(t),\alpha)\right)\\ &\le (-ax(t)+u(t)-z)r\end{aligned} \tag{3.18}$$

for $r \in D^+\psi^\varepsilon(x(t),\alpha)$; see Zhou [19, Lemma 2.1]. In view of (3.18) and the proof of the Feynman–Kac formula (see Fleming and Soner [10]), we can show, for any $T \ge 0$,

$$\begin{aligned}&E\left[\psi^\varepsilon(x(T),\alpha^\varepsilon(T))\exp\int_0^T\left(\frac{L(x(t),u(t))-\lambda^\varepsilon}{\varepsilon}\right)dt\right]-\psi^\varepsilon(x,\alpha)\\ &\quad= E\int_0^T\frac{d}{dt}\left(\psi^\varepsilon(x(t),\alpha^\varepsilon(t))\exp\int_0^t\left(\frac{L(x(s),u(s))-\lambda^\varepsilon}{\varepsilon}\right)ds\right)dt\\ &\quad\le E\int_0^T\exp\int_0^t\left(\frac{L(x^*(s),u^*(s))-\lambda^\varepsilon}{\varepsilon}\right)ds\\ &\qquad\cdot\Bigg[\left(\frac{L(x^*(t),u^*(t))-\lambda^\varepsilon}{\varepsilon}\right)\psi^\varepsilon(x^*(t),\alpha^\varepsilon(t))\\ &\qquad+(-ax^*(t)+u^*(t)-z)r^*(t)+\frac{Q}{\varepsilon}\psi^\varepsilon(x^*(t),\cdot)(\alpha^\varepsilon(t))\Bigg]dt=0.\end{aligned} \tag{3.19}$$

Note that for any given initial value x, the corresponding trajectory $x(t)$ is bounded. Thus, for each x and $\varepsilon > 0$, there exist positive constants M_1 and M_2 such that

$$0 < M_1 \le \psi^\varepsilon(x(T),\alpha^\varepsilon(T)) \le M_2 \text{ for all } T \ge 0.$$

Hence, it follows from (3.17) and (3.19) that

$$M_1 E\left[\exp\int_0^T\left(\frac{L(x^*(t),u^*(t))}{\varepsilon}\right)dt\right]\exp\left(-\frac{\lambda^\varepsilon}{\varepsilon}\right)\le\psi^\varepsilon(x,\alpha).$$

Taking the logarithm on both sides and dividing by T leads to

$$\frac{\log M_1}{T}+\frac{1}{T}\log E\exp\int_0^T\left(\frac{L(x^*(t),u^*(t))}{\varepsilon}\right)dt-\frac{\lambda^\varepsilon}{\varepsilon}\le\frac{\psi^\varepsilon(x,\alpha)}{T}.$$

Sending $T\to\infty$ yields

$$\lambda^\varepsilon\ge\limsup_{T\to\infty}\frac{\varepsilon}{T}\log E\exp\int_0^T\left(\frac{L(x(t),u(t))}{\varepsilon}\right)dt.$$

Hence, in view of Theorem 3.5, $\lambda^\varepsilon=J^\varepsilon(u^*(\cdot))$. □

4. Limiting problem. In this section, we analyze the asymptotic properties of the HJB equation (3.3.1) as $\varepsilon\to0$. First of all, note that this HJB equation is similar to that for an ordinary long-run average cost problem except for the term involving the exponential functions. In order to get rid of such a term, we make use of the logarithmic transformation in Fleming and Soner [10, p. 275].

Let $\mathcal{V}=\{v=(v(0),\dots,v(m))\in\mathbb{R}^{m+1}:\ v(i)>0,i=0,1,\dots,m\}$. Define

$$Q^v=(q_{ij}^v)\text{ such that }q_{ij}^v=q_{ij}\frac{v(j)}{v(i)}\text{ for }i\ne j\text{ and }q_{ii}^v=-\sum_{j\ne i}q_{ij}^v.$$

Then, in view of the logarithmic transformation, we have, for each $i\in\mathcal{M}$,

$$\begin{aligned}&\exp\left(-\frac{w^\varepsilon(x,\alpha)}{\varepsilon}\right)Q\exp\left(\frac{w^\varepsilon(x,\cdot)}{\varepsilon}\right)(i)\\&\quad=\sup_{v\in\mathcal{V}}\left\{\frac{Q^v}{\varepsilon}w^\varepsilon(x,\cdot)(i)+\frac{Qv(\cdot)(i)}{v(i)}-Q^v(\log v(\cdot))(i)\right\}.\end{aligned}$$

The supermum is obtained at $v(i)=\exp(-w^\varepsilon(x,i)/\varepsilon)$.

The logarithmic transformation suggests that the HJB equation is equivalent to an Isaacs equation of a two-player zero-sum dynamic stochastic game. The Isaacs equation is given as follows:

$$\lambda^\varepsilon=\inf_{0\le u\le\alpha}\sup_{v\in\mathcal{V}}\left\{(-ax+u-z)w_x^\varepsilon(x,\alpha)+\widetilde{L}(x,u,v,\alpha)+\frac{Q^v}{\varepsilon}w^\varepsilon(x,\cdot)(\alpha)\right\},\tag{4.1}$$

where

$$\widetilde{L}(x,u,v,i)=L(x,u)+\frac{Qv(\cdot)(i)}{v(i)}-Q^v(\log v(\cdot))(i)\tag{4.2}$$

for $i\in\mathcal{M}$.

Remark 4.1. Note that if $v=(1,\dots,1)$, then $\widetilde{L}(x,u,v,i)=L(x,u)$ and $Q^v=Q$.

Remark 4.2. In the results to follow, we will not give a precise description of the stochastic dynamic game with Isaacs equation (4.1) since this interpretation will not be used in proving our results about the deterministic limit $\varepsilon\to0$. In the game,

$u(t)$ and $v(t)$ represent minimizing and maximizing controls, based on information available at time t. Note that the maximizing control v produces a change in transition rates, from q_{ij} to q_{ij}^v. This imprecise idea can be made precise using Elliott–Kalton-type strategies (Fleming and Souganidis [11]). Since the order in (4.1) is $\inf(\sup(\cdots))$ rather than $\sup(\inf(\cdots))$, λ^ε turns out to be the upper game value for the game payoff

$$\limsup_{T\to\infty} \frac{1}{T} E \int_0^T \widetilde{L}(x(t), u(t), v(t), \alpha^\varepsilon(t))dt.$$

We consider the limit of the problem as $\varepsilon \to 0$. In order to define a limiting problem, we first define control sets for the limiting problem. Let

$$\Gamma_u = \{U = (u^0, \ldots, u^m);\ 0 \le u^i \le i,\ i = 0, \ldots, m\}$$

and

$$\Gamma_v = \{V = (v^0, \ldots, v^m);\ v^i = (v^i(0), \ldots, v^i(m)) \in \mathcal{V},\ i = 0, \ldots, m\}.$$

For each $V \in \Gamma_v$, let $\overline{Q}^V := (q_{ij}^V)$ such that

$$q_{ij}^{v^i} = q_{ij}^V = \frac{q_{ij} v^i(j)}{v^i(i)} \quad \text{for } i \ne j \quad \text{and} \quad q_{ii}^V = -\sum_{j\ne i} q_{ij}^V,$$

and let $\nu^V = (\nu_0^V, \ldots, \nu_m^V)$ denote the equilibrium distribution of $\overline{Q}^V$. The next lemma says $\overline{Q}^V$ is irreducible. Therefore, there exists a unique positive ν^V for each $V \in \Gamma_v$. Moreover, ν^V depends continuously on V.

LEMMA 4.3. *For each $V \in \Gamma_v$, $\overline{Q}^V$ is irreducible.*

Proof. We divide the proof into three steps.

Step 1. $\operatorname{rank}(\overline{Q}^V) = m$.

First, it is easy to see that the irreducibility of Q implies $q_{kk}^V < 0$ for $k = 0, 1, \ldots, m$. We multiply the first row of $\overline{Q}^V$ by $-q_{k0}^V/q_{00}^V$ and add to the kth row, $k = 1, \ldots, m$, to make the first component of that row 0. Let $Q^{V,1} = (q_{ij}^{V,1})$ denote the resulting matrix. Then, $Q^{V,1}$ must satisfy

$$\begin{aligned}
&q_{0j}^{V,1} = q_{0j}^V, \quad j = 0, 1, \ldots, m,\\
&q_{k0}^{V,1} = 0, \quad k = 1, \ldots, m,\\
&q_{kk}^{V,1} \le 0, \quad k = 1, \ldots, m, \text{ and}\\
&\sum_{j=0}^m q_{kj}^{V,1} = 0 \quad \text{for } k = 0, 1, \ldots, m.
\end{aligned}$$

We now show that $q_{kk}^{V,1} < 0$ for $k = 1, \ldots, m$. For $k = 1$, if $q_{11}^{V,1} \not< 0$, then it must be equal to 0, which implies

$$(q_{12}^V, \ldots, q_{1m}^V) - \left(\frac{q_{10}^V}{q_{00}^V}\right)(q_{02}^V, \ldots, q_{0m}^V) = 0. \tag{4.3}$$

Recall that $q_{11}^V \ne 0$. One must have $q_{10}^V > 0$ since otherwise $q_{10}^V = 0$ implies $q_{11}^V = q_{11}^{V,1} = 0$, which contradicts the fact that $q_{kk}^V < 0$ for $k = 0, 1, \ldots, m$. Thus, $-q_{10}^V/q_{00}^V >$

0. This together with the nonnegativity of q_{ij}^V, $i \neq j$, imply that both of the vectors in (4.3) must be equal to 0, i.e.,

$$(q_{12}^V, \dots, q_{1m}^V) = 0 \quad \text{and} \quad (q_{02}^V, \dots, q_{0m}^V) = 0.$$

These equations contradict the irreducibility of Q since a state in $\{2, 3, \dots, m\}$ is not accessible from a state in $\{0, 1\}$. Therefore, one must have $q_{11}^{V,1} < 0$. Similarly, we can show $q_{kk}^{V,1} < 0$ for $k = 2, \dots, m$.

We repeat this procedure in a similar way by multiplying the second row of $Q^{V,1}$ by $-q_{k1}^{V,1}/q_{11}^{V,1}$, $k = 2, \dots, m$, and add to the kth row. Let $Q^{V,2)} = (q_{ij}^{V,2})$ denote the resulting matrix. Then one has

$$\begin{aligned}
&q_{ij}^{V,2} = q_{ij}^{V,1}, \quad i = 0, 1, \quad j = 0, 1, \dots, m,\\
&q_{ij}^{V,2} = 0, \quad i = 2, \dots, m, \quad j = 0, 1\\
&q_{kk}^{V,2} \leq 0, \quad k = 2, \dots, m, \text{ and}\\
&\sum_{j=0}^{m} q_{kj}^{V,2} = 0 \quad \text{for } k \in \mathcal{M}.
\end{aligned}$$

Similarly, we can show $q_{kk}^{V,2} < 0$ for $k = 2, \dots, m$.

We continue this procedure and transform $Q \to Q^{V,1} \to \cdots \to Q^{V,m-1}$ with $Q^{V,m-1} = (q_{ij}^{V,m-1})$ such that

$$\begin{aligned}
&q_{ij}^{V,m-1} = 0, \quad i > j\\
&q_{kk}^{V,m-1} < 0, \quad k = 0, 1, \dots, m-1,\\
&\sum_{j=0}^{m} q_{kj}^{V,m-1} = 0 \quad \text{for } k \in \mathcal{M} \text{ and}\\
&q_{mm}^{V,m-1} = 0.
\end{aligned}$$

Notice that the prescribed transformations do not change the rank of the original matrix. Thus,

$$\operatorname{rank}(\overline{Q}^V) = \operatorname{rank}(Q^{V,1}) = \cdots = \operatorname{rank}(Q^{V,m-1}) = m.$$

Step 2. $\overline{Q}^V$ is weakly irreducible.

Consider an $(m+1)$ row vector $b = (b_0, \dots, b_m)$ such that

$$b\overline{Q}^V = 0 \quad \text{and} \quad b_0 + \cdots + b_m = 1.$$

It follows as in Sethi and Zhang [15, Lemma C.1] that $(b' \vdots \cdots \vdots b') = \lim_{t\to\infty} \exp(\overline{Q}^V t)$, where A' denotes the transpose of a matrix A. Since $\exp(\overline{Q}^V t)$ represents the transition probabilities, the limit b must be nonnegative. Thus, $b := (\nu_0^V, \dots, \nu_m^V)$ is an equilibrium distribution of $\overline{Q}^V$. Note that $\ker(\overline{Q}^V)' = \operatorname{span}\{(\nu_0^V, \dots, \nu_m^V)\}$ since $\operatorname{rank}(\overline{Q}^V)' = \operatorname{rank}(\overline{Q}^V) = m$. Then, $c = (\nu_0^V, \dots, \nu_m^V)$ is the unique nonnegative solution to $b\overline{Q}^V = 0$ and $b_0 + \cdots + b_m = 1$. Hence $\overline{Q}^V$ is weakly irreducible.

Step 3. $\overline{Q}^V$ is irreducible; i.e., $(\nu_0^V, \dots, \nu_m^V) > 0$.

If not, then, without loss of generality, we may assume $\nu_0^V > 0, \dots, \nu_{k_0}^V > 0$ and $\nu_{k_0+1}^V = 0, \dots, \nu_m^V = 0$ for some $k_0 = 0, 1, \dots, m$. Note that $(\nu_0^V, \dots, \nu_m^V)\overline{Q}^V = 0$ implies that $q_{ij}^V = 0$ and thus $q_{ij} = 0$ for $i = 0, \dots, k_0$ and $j = k_0 + 1, \dots, m$. This in turn implies that Q is not irreducible since the process $\alpha^\varepsilon(\cdot)$ cannot jump from a state in $\{0, 1, \dots, k_0\}$ to a state in $\{k_0 + 1, \dots, m\}$. The contradiction yields the irreducibility of $\overline{Q}^V$. □

THEOREM 4.4. *Let $\varepsilon_n \to 0$ be a sequence such that $\lambda^{\varepsilon_n} \to \lambda^0$ and $w^{\varepsilon_n}(x, \alpha) \to w^0(x, \alpha)$. Then,*

(i) *$w^0(x, \alpha)$ is independent of α, i.e., $w^0(x, \alpha) = w^0(x)$;*

(ii) *$w^0(x)$ is Lipschitz; and*

(iii) *$(\lambda^0, w^0(x))$ is a viscosity solution to the following Isaacs equation:*

$$\lambda^0 = \inf_{U \in \Gamma_u} \sup_{V \in \Gamma_v} \left\{ \left(-ax + \sum_{i=0}^m \nu_i^V u^i - z \right) w_x^0(x) + \sum_{i=0}^m \nu_i^V L(x, u^i) + \left(\sum_{i=0}^m \nu_i^V \frac{Qv^i(\cdot)(i)}{v^i(i)} - \sum_{i=0}^m \nu_i^V \overline{Q}^V (\log v^i(\cdot))(i) \right) \right\}. \tag{4.4}$$

Proof. Note that Lemma 3.2 (iii) implies that

$$|w_\rho^\varepsilon(x, \alpha) - w_\rho^\varepsilon(x, \tilde{\alpha})| \le \varepsilon \log C_3$$

for x in any finite interval. Thus, the limit of $w_\rho^\varepsilon(x, \alpha)$ must be independent of α, i.e.,

$$w^0(x, 0) = \cdots = w^0(x, m) =: w^0(x).$$

The Lipschitz property of $w^0(x)$ follows from the Lipschitz property of $w^\varepsilon(x, \alpha)$.

Finally, note that

$$\nu^V \overline{Q}^V = (\nu_0^V, \dots, \nu_m^V)\overline{Q}^V = 0.$$

It follows that

$$\sum_{i=0}^m \nu_i^V Q^{v^i} w^\varepsilon(x, \cdot)(i) = 0. \tag{4.5}$$

The remaining proof of (iii) is standard and can be carried out as in Fleming, Sethi, and Soner [8]. □

Remark 4.5. We would like to point out that the last term in (4.4) is nonnegative. This can be seen as follows: note that for each $v \in \mathcal{V}$ and $i \in \mathcal{M}$, we have

$$\begin{aligned} &\frac{Qv(\cdot)(i)}{v(i)} - Q^v(\log v(\cdot))(i) \\ &\quad = \sum_{j \ne i} q_{ij} \left(\frac{v(j)}{v(i)} - 1 \right) - \sum_{j \ne i} q_{ij} \frac{v(j)}{v(i)} \log \frac{v(j)}{v(i)} \\ &\quad = \sum_{j \ne i} q_{ij} \left(\frac{v(j)}{v(i)} - 1 - \frac{v(j)}{v(i)} \log \frac{v(j)}{v(i)} \right) \le 0, \end{aligned}$$

because the function $(x - 1 - x \log x)$ is nonnegative on $(0, \infty)$. It follows that

$$\sum_{i=0}^m \nu_i^V \frac{Qv^i(\cdot)(i)}{v^i(i)} - \sum_{i=0}^m \nu_i^V Q^{v^i} (\log v^i(\cdot))(i) \le 0.$$

Let

$$\widehat{L}(x,U,V)=\sum_{i=0}^{m}\nu_i^V L(x,u^i)+\sum_{i=0}^{m}\nu_i^V\frac{Qv^i(\cdot)(i)}{v^i(i)}-\sum_{i=0}^{m}\nu_i^V\overline{Q}^V(\log v^i(\cdot))(i).$$

Note that $\widehat{L}(x,U,V)\le||L||$, where $||\cdot||$ is the sup norm. Moreover, since $L\ge 0$, $\widehat{L}(x,U,1)\ge 0$ where $V=1$ means $v^i(j)=1$ for all i,j. Then, the equation in (4.4) is an Isaacs equation associated with a two-player, zero-sum dynamic game with objective

$$J^0(U(\cdot),V(\cdot))=\limsup_{T\to\infty}\frac{1}{T}\int_0^T\widehat{L}(x(t),U(t),V(t))dt$$

subject to

$$\dot{x}(t)=-ax(t)+\sum_{i=0}^{m}\nu_i^{V(t)}u^i(t)-z,\qquad x(0)=x,$$

where $U(\cdot)$ and $V(\cdot)$ are Borel measurable functions and $U(t)\in\Gamma_u$ and $V(t)\in\Gamma_v$ for $t\ge 0$.

Remark 4.6. Let $I(\mu)$ be the Donsker–Varadhan function, defined for any probability vector $\mu=(\mu_0,\dots,\mu_m)>0$, i.e., $\mu_i>0$ and $\sum_{i=0}^m\mu_i=1$. Then (see the Appendix)

$$-I(\mu)=\sup_{V,\nu^V=\mu}\left\{\sum_{i=0}^{m}\nu_i^V\frac{Qv^i(\cdot)(i)}{v^i(i)}-\sum_{i=0}^{m}\nu_i^V\overline{Q}^V(\log v^i(\cdot))(i)\right\}.\tag{4.6}$$

Thus (4.4) is equivalent to

$$\lambda^0=\inf_{U\in\Gamma_u}\sup_{\mu}\left\{\Big(-ax+\langle\mu,u\rangle-z\Big)w_x^0(x)+\langle\mu,L(x,u)\rangle-I(\mu)\right\}.$$

Similarly, the dynamics of $x(t)$ can be written

$$\dot{x}(t)=-ax(t)+\langle\mu(t),u(t)\rangle-z.$$

Let

$$\begin{aligned}H(x,p)=\inf_{U\in\Gamma_u}\sup_{V\in\Gamma_v}\Bigg\{&\left(-ax+\sum_{i=0}^{m}\nu_i^V u^i-z\right)p+\sum_{i=0}^{m}\nu_i^V L(x,u^i)\\&+\left(\sum_{i=0}^{m}\nu_i^V\frac{Qv^i(\cdot)(i)}{v^i(i)}-\sum_{i=0}^{m}\nu_i^V\overline{Q}^V(\log v^i(\cdot))(i)\right)\Bigg\}.\end{aligned}$$

Then,

$$\begin{aligned}|H(\tilde{x},p)-H(x,p)|&\le(a|p|+||L_x||)|\tilde{x}-x|,\\|H(x,\tilde{p})-H(x,p)|&\le(a|x|+m+z)|\tilde{p}-p|.\end{aligned}$$

These conditions imply the uniqueness of viscosity solution to the following finite time problem:

$$\begin{cases} \dfrac{\partial \Psi}{\partial T} = H(x, \Psi_x) - \lambda^0, \quad T > 0, \\ \Psi(0, x) = w^0(x). \end{cases}$$

Uniqueness is in the class of continuous viscosity solution $\Psi(x, T)$ such that $\Psi(\cdot, T)$ satisfies a uniform Lipschitz condition on every finite time interval $0 \le T \le T_1$; see Crandall and Lions [4] and Ishii [13]. A more general uniqueness theorem in which $\Psi(\cdot, T)$ satisfies a uniform local Lipschitz condition is given in McEneaney [14].

The method of Evans and Souganidis [6] shows that

$$\text{upper value} \left\{ \int_0^T \left(\widehat{L}(x(t), U(t), V(t)) - \lambda^0 \right) dt + w^0(x(T)) \right\}$$

is such a viscosity solution and $w^0(x)$ is also a viscosity solution. So $w^0(x) = \Psi(T, x)$. Namely,

$$w^0(x) = \text{ upper value} \left\{ \int_0^T \left(\widehat{L}(x(t), U(t), V(t)) - \lambda^0 \right) dt + w^0(x(T)) \right\}.$$

In Evans and Souganidis [6] the control spaces for both players are assumed compact, and Γ_v is not compact. This requires minor changes in the arguments in [6] using the special form of the game dynamics and payoff function $\widehat{L} - \lambda_0$; see the Appendix for more details. Using the above equality, one can show as in Fleming and McEneaney [7] that

$$\lambda^0 = \inf_{U(\cdot)} \sup_{V(\cdot)} J^0(U(\cdot), V(\cdot)),$$

which implies the uniqueness of λ^0.

Finally, we would like to comment on how to use the solution to the limiting problem to obtain a control for the original problem. Typically an explicit solution is not available to either of the problems. A numerical scheme has to be used to obtain an approximate solution. The advantage of the limiting problem is its dimensionality, which is much smaller than that of the original problem if the number of states in $\mathcal{M}$ is large.

Let $(U^*(x), V^*(x))$ denote a solution to the upper value problem. Suggested by the ideas of hierarchical control (see Sethi and Zhang [15]), it is expected that the control

$$u(x, \alpha) = \sum_{j=0}^{m} I_{\{\alpha = j\}} u^{i*}(x)$$

is nearly optimal for the original problem.

5. Concluding remarks. This paper deals with the risk-sensitive control with a long-run average cost arising in a failure-prone manufacturing system. Typically the problem with a long-run average cost requires the stability of the system. In this paper, a model with product deterioration is considered. Such a deterioration con-

dition is used to guarantee the desired stability without undue technical difficulties. It would be interesting to study the stability without such a deterioration condition. One possible direction for attacking the problem is to use a "diminishing deterioration" approach by sending the deterioration rate $a \to 0$. In order to obtain the desired convergence of the potential function $w^\varepsilon(x,\alpha)$ as $a \to 0$, it is necessary to have the uniform equicontinuity property that typically is guaranteed by the Lipschitz condition uniform with respect to $a > 0$. A major difficulty, however, is the absence of such a uniform Lipschitz property. This can be seen from (ii) in Lemma 3.2 in which the Lipschitz constant depends on a.

In this paper, a single machine, single product model is considered. It would also be interesting to generalize the results to more general manufacturing systems such as flowshops and jobshops; see Sethi and Zhang [15].

6. Appendix. In section 3 we used the following dynamic programming principle. For brevity, let us write $U(t) = (u(t), y(t))$. Let $G(s)$ and $h(t)$ be as in the proof of Lemma 3.2 (iii). Then for every stopping time τ

$$\text{(DP)} \qquad \psi^\varepsilon_\rho(x,\alpha) = \inf_{U(\cdot)} E\biggl\{\int_0^\tau \biggl(\exp\int_0^t G(s)ds\biggr) h(t)dt + \biggl(\exp\int_0^\tau G(s)ds\biggr)\psi^\varepsilon_\rho(x(\tau),\alpha(\tau))\biggr\}.$$

In the proof of Lemma 3.2 (iii), τ is the first t such that $\alpha^\varepsilon(t) = \tilde\alpha$. To prove that ψ^ε_ρ is a viscosity solution of (3.8), property (DP) is needed for any nonrandom τ. While results of this kind are considered well known, the authors did not find a convenient reference which applies to the class of stochastic control problems considered in this paper. For completeness we sketch a proof of (DP). In the proof of Theorem 3.5, a dynamic programming principle is used, for which an entirely similar proof can be given. Indeed, the argument is slightly simpler since only nonrandom stopping times need to be considered.

Sketch of proof of (DP). It suffices to consider the "canonical" sample space $(\Omega, \{\mathcal{F}_t\}, P)$ with $\Omega = D([0,\infty);\mathcal{M})$ the space of possible $\alpha(\cdot)$ paths, $\mathcal{F}_t = \sigma\{\alpha(s) : 0 \le s \le t\}$ and $P = P^\varepsilon_\alpha$ the probability distribution of a Markov chain $\alpha^\varepsilon(\cdot)$ with generator Q/ε and initial state $\alpha(0) = \alpha$. We wish to establish (DP) for any $\mathcal{F}_t$ -stopping time τ. By using the Lipschitz property in Lemma 3.2 (ii) which does not depend on (DP), it suffices to consider τ with finitely many values $0 < t_1 < t_2 < \cdots < t_n$ because one may approximate τ by a step function $\sum t_k I_{\{t_k \le \tau < t_{k+1}\}}$. Let $\Gamma_k = \{\tau = t_k\}$ and on Γ_k we identify $U(t)$ for $t \ge t_k$ with $U_k(t - t_k, \alpha_{1k}(\cdot), \alpha_{2k}(\cdot))$, where

$$\begin{aligned} \alpha_{1k}(t) &= \alpha(t), \quad 0 \le t \le t_k, \\ \alpha_{2k}(t) &= \alpha(t_k + t), \quad t \ge 0. \end{aligned}$$

If $U(\cdot)$ is admissible (progressively measurable and satisfying the control constraints on $u(t)$ and $y(t)$), then for each fixed $\alpha_{1k}(\cdot)$, $U_k(\cdot)$ is also admissible. Moreover, if $\mathcal{F}_\infty$ is the least σ-algebra containing all $\mathcal{F}_t$, then for any bounded $\mathcal{F}_\infty$- measurable Φ,

$$E(I_{\Gamma_k}\Phi) = \int_{\Gamma_k}\int \Phi(\alpha_{1k},\alpha_{2k}) P^\varepsilon_{\alpha_{1k}(t_k)}(d\alpha_{2k}) P^\varepsilon_\alpha(d\alpha_{1k}).$$

A routine calculation then gives

$$\begin{aligned} E\int_\tau^\infty &\left(\exp\int_0^t G(s)ds\right)h(t)dt \\ &= \sum_{k=1}^n E\left[I_{\Gamma_k}\left(\exp\int_0^{t_k} G(s)ds\right)J_\rho^\varepsilon(x(t_k),\alpha(t_k),U_k(\cdot))\right] \\ &\geq \sum_{k=1}^n E\left[I_{\Gamma_k}\left(\exp\int_0^{t_k} G(s)ds\right)\psi_\rho^\varepsilon(x(t_k),\alpha(t_k))\right] \\ &= E\left(\exp\int_0^{\tau} G(s)ds\right)\psi_\rho^\varepsilon(x(\tau),\alpha(\tau)). \end{aligned}$$

Since $\psi_\rho^\varepsilon(x,\alpha)$ is the inf of $J_\rho^\varepsilon(x,\alpha,U(\cdot))$ taken over all admissible $U(\cdot)$, this implies that

$$\psi_\rho^\varepsilon(x,\alpha) \geq \text{ right side of (DP).}$$

It remains to outline a proof that

$$\psi_\rho^\varepsilon(x,\alpha) \leq \text{ right side of (DP).}$$

Given an initial $x(0)=x$, formula (2.3) implies that $|x(t)|\leq r_1$ for some r_1. As in the proof of Theorem 3.5, given $\delta>0$, partition $\{|x|\leq r_1\}$ into intervals $B_1,B_2,\ldots,B_l$ of length $<\delta$ and choose $x_j\in B_j$ for $j=1,\ldots,l$. Given $\eta>0$, choose admissible $U_{ij}(\cdot)$ such that

$$J_\rho^\varepsilon(x_j,i,U_{ij}(\cdot)) < \psi_\rho^\varepsilon(x_j,i)+\eta.$$

Given admissible $U(\cdot)$, we define $\widetilde{U}(\cdot)$ by

$$\widetilde{U}(t)=U(t) \quad \text{for } 0\leq t<\tau,$$

and for $\tau=t_k$, $\alpha(t_k)=i$, $x(t_k)\in B_j$,

$$\widetilde{U}(t)=U_{ij}(t-t_k), \quad t\geq t_k.$$

Then $\widetilde{U}(\cdot)$ is admissible, and a routine calculation using Lemma 3.2 (ii) gives

$$\psi_\rho^\varepsilon(x,\alpha)\leq J_\rho^\varepsilon(x,\alpha,\widetilde{U}(\cdot)) \leq \text{ right side of (DP) } + F(\delta,\eta),$$

where $F(\delta,\eta)\to 0$ as $\delta,\eta\to 0$.

Remark on upper values and viscosity solutions. In section 4 we used a slight modification of a result of Evans and Souganidis [6]. Let us sketch the changes in [6] needed to account for the fact that the maximizing players' control space Γ_v is not compact. The game dynamics are

$$\dot{x}(t)=\hat{f}(x(t),U(t),V(t)),$$

where

$$\hat{f}(x,U,V)=-ax+\sum_{i=0}^m \nu_i^V u^i - z.$$

Let

$$\widetilde{\Psi}(x,T) = \text{ upper value } \left\{ \int_0^T \widehat{L}(x(t),U(t),V(t))dt + w^0(x(T)) \right\},$$

where the upper value is in the Elliott–Kalton sense. Equivalently,

$$\Psi(x,T) = \widetilde{\Psi}(x,T) - \lambda^0 T$$

is the upper value considered in section 5.

The assertion is that $\widetilde{\Psi}$ is a viscosity solution to

$$\frac{\partial \widetilde{\Psi}}{\partial T} = H(x, \widetilde{\Psi}_x)$$

and that $\widetilde{\Psi}$ is continuous with $\widetilde{\Psi}(\cdot,T)$ satisfying a uniform Lipschitz condition on any finite interval $0 \le T \le T_1$. The first step is to prove the dynamic programming principle (see [6, Theorem 3.1]): for $0 < \tau < T$,

$$\text{(DP)} \qquad \widetilde{\Psi}(x,T) = \text{ upper value } \left\{ \int_0^\tau \widehat{L}(x(t),U(t),V(t))dt + \widetilde{\Psi}(x(\tau),\tau) \right\}.$$

That argument is unchanged. Next, the facts that $\hat{f}_x = -a$ and $\widehat{L}_x = L_x$ is bounded imply a uniform Lipschitz condition for $\widetilde{\Psi}(\cdot,T)$ on any finite time interval. As noted in section 5, $\widehat{L} \le ||L||$ and

$$\sup_V \widehat{L}(x,U,V) \ge \widehat{L}(x,U,1) \ge 0.$$

Moreover, by the form of $\hat{f}$ and compactness of the minimizer's control set Γ_u, for every R, there exists K_R such that $|x| \le R$ implies

$$|x(\tau) - x| \le K_R \tau.$$

By subtracting $\widetilde{\Psi}(x,\tau)$ from both sides of (DP) we then obtain a uniform local Lipschitz condition for $\widetilde{\Psi}(x,\cdot)$.

Finally, to show that $\widetilde{\Psi}$ is a viscosity solution we proceed as in [6, section 4]. Minor changes in the proof of [6, Lemma 4.3] are needed, since Γ_v is not compact. For this we use the inequality for compact maximizer's control space. (The proof of [6, Lemma 4.3(a)] does use compactness of the minimizer's control space, which holds in our case.)

Proof of (4.6). Recall that the Donsker–Varadhan function is defined as $I(\mu) = \sup_{\beta \in \mathcal{V}}[-\langle \mu, \beta^{-1} Q\beta\rangle]$; see Fleming, Sheu, and Soner [9]. For each $V \in \Gamma_v$, let

$$K^V(i) = \overline{Q}^V(\log v^i(\cdot))(i) - \frac{1}{v^i(i)} Q v^i(\cdot)(i).$$

Then (4.6) can be written as $I(\mu) = \inf_{\nu^V = \mu} \langle \mu, K^V \rangle$. We first show that

$$I(\mu) \ge \inf_{\nu^V = \mu} \langle \mu, K^V \rangle. \tag{6.1}$$

It is elementary to show that there exists $\beta^* \in \mathcal{V}$ such that $I(\mu) = \langle \mu, (\beta^*)^{-1} Q \beta^* \rangle$. Then, in view of Lemma 3.2 in [9], we have

$$\mu = \nu^{V^*}, \quad \text{where } v^{i*}(j) = \beta^*(i)/\beta^*(j).$$

It follows that $\langle \mu, K^{V^*} \rangle = -\langle \mu, (\beta^*)^{-1} Q \beta^* \rangle$, because $\langle \mu, \overline{Q}^{V^*}(\phi) \rangle = \langle \nu^{V^*}, \overline{Q}^{V^*} \phi \rangle = 0$. This implies (6.1).

To show the opposite inequality, note that the logarithmic transformation

$$e^{-\phi(i)} Q(e^{\phi(\cdot)})(i) = \sup_{V \in \Gamma_v} [\overline{Q}^V \phi(\cdot)(i) - K^V(i)]$$

for all ϕ. Let $\phi = \beta^*$. Then for each V such that $\nu^V = \mu$, we have

$$\frac{1}{\beta^*} Q \beta^* \geq \overline{Q}^V \phi - K^V.$$

Hence, for $\nu^V = \mu$, we obtain

$$\left\langle \mu, \frac{1}{\beta^*} Q \beta^* \right\rangle \geq \langle \mu, \overline{Q}^V \phi \rangle - \langle \mu, K^V \rangle = -\langle \mu, K^V \rangle.$$

This implies that $I(\mu) \leq \inf_{\nu^V = \mu} \langle \mu, K^V \rangle$. The proof is complete.

Acknowledgments. We would like to thank S.-J. Sheu for pointing out Remark 4.6 and providing its proof. We also thank the referees for comments and suggestions that led to improvement of the paper.

REFERENCES

[1] T. Basar and P. Bernhard, *H^∞-Optimal Control and Related Minimax Design Problems*, Birkhauser, Boston, 1991.

[2] E. N. Barron and R. Jensen, *Total risk aversion, stochastic optimal control, and differential games*, Appl. Math. Optim., 19 (1989), pp. 313–327.

[3] A. Bensoussan and H. Nagai, *An ergodic control problem arising from the principal eigenfunction of an elliptic operator*, J. Math. Soc. Japan, 43 (1991), pp. 49–65.

[4] M. G. Crandall and P.-L. Lions, *Remarks on the existence and uniqueness of unbounded viscosity solutions of Hamilton-Jacobi equations*, Illinois J. Math., 31 (1987) pp. 665–688.

[5] M. H. A. Davis, *Markov Models and Optimization*, Chapman & Hall, New York, 1993.

[6] L. C. Evans and P. E. Souganidis, *Differential games and representation formulas for solutions of Hamilton-Jacobi-Isaacs equations*, Indiana Univ. Math. J., 33 (1984), pp. 773–797.

[7] W. H. Fleming and W. M. McEneaney, *Risk sensitive control on an infinite horizon*, SIAM J. Control Optim., 33 (1995), pp. 1881–1921.

[8] W. H. Fleming, S. P. Sethi, and H. M. Soner, *An optimal stochastic production planning problem with random fluctuating demand*, SIAM J. Control Optim., 25 (1987), pp. 1494–1502.

[9] W. H. Fleming, S.-J. Sheu, and H. M. Soner, *A remark on the large deviations of an ergodic Markov process*, Stochastics, 22 (1987), pp. 187–199.

[10] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1992.

[11] W. H. Fleming and P. E. Souganidis, *On the existence of value functions of two-player, zero-sum stochastic different games*, Indiana Univ. Math. J., 38 (1989), pp. 293–314.

[12] K. Glover and J. C. Doyle, *State space formulae for all stabilizing controllers that satisfy an H^∞ norm bound and relations to risk sensitivity*, Systems Control Lett., 11 (1988), pp. 167–172.

[13] H. Ishii, *Uniqueness of unbounded viscosity solution of Hamilton-Jacobi equations*, Indiana Univ. Math. J., 33 (1984), pp. 721–748.

[14] W. M. McEneaney, *Uniqueness of viscosity solutions of nonstationary HJB equations with some a priori conditions (with applications)*, SIAM J. Control Optim., 33 (1995), pp. 1560–1576.
[15] S. P. Sethi and Q. Zhang, *Hierarchical Decision Making in Stochastic Manufacturing Systems*, Birkhauser, Boston, 1994.
[16] H. M. Soner, *Optimal control with state space constraints* II, SIAM J. Control Optim., 24 (1986), pp. 1110–1122.
[17] P. Whittle, *Risk-Sensitive Optimal Control*, Wiley, New York, 1990.
[18] Q. Zhang, *Risk sensitive production planning of stochastic manufacturing systems: A singular perturbation approach*, SIAM J. Control Optim., 33 (1995), pp. 498–527.
[19] X. Y. Zhou, *Verification theorem within the framework of viscosity solutions*, J. Math. Anal. Appl., 177 (1993), pp. 208–225.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 4, pp. 1171–1189, July 1998

003

STUDY OF AN OPTIMAL CONTROL PROBLEM FOR DIFFUSIVE NONLINEAR ELLIPTIC EQUATIONS OF LOGISTIC TYPE*

A. CAÑADA†, J. L. GÁMEZ†, AND J. A. MONTERO†

Abstract. An optimal control problem for a nonlinear elliptic equation of logistic type is considered. Under certain assumptions, the existence of at least an optimal control is shown and an optimality system is derived. Then this system is used for proving the uniqueness of and a constructive approximation to the optimal control.

Key words. optimal control, logistic elliptic equations, existence, uniqueness, approximation

AMS subject classifications. 49J20, 49K20, 49M05, 92D25

PII. S0363012995293323

1. Introduction. The aim of this paper is to study an optimal control problem for a nonlinear elliptic equation of the Volterra–Lotka type. More precisely, we consider the equation

$$\begin{aligned} -\Delta u(x) &= u(x)[a(x) - f(x) - b(x)u(x)],\ x \in \Omega, \\ u(x) &= 0,\ x \in \partial\Omega, \end{aligned} \tag{1.1}$$

where Ω is a bounded and regular domain in $\mathbb{R}^n$.

The previous equation arises from population dynamics, the study of the evolution of some biological species, and it models the steady-state solutions of the corresponding nonlinear evolution problem (see [12]). Here function u is the species concentration, a represents its intrinsic growth rate, b is the crowding effect, and f plays the role of control. The Laplacian operator Δ shows the diffusive character of the species u in the domain Ω, and the boundary condition in (1.1) may be interpreted as the condition that the species may not stay on $\partial\Omega$.

Under certain assumptions (see Hypothesis H below), equation (1.1) will have, for each given function f, a unique maximal nonnegative solution denoted by u_f, and we will be interested in maximizing the payoff functional

$$J(f) = \int_\Omega (\lambda u_f f - f^2), \tag{1.2}$$

which represents the difference between economic revenue and cost. The real constant λ, which will be taken to be strictly positive, describes the quotient between the price of the species and the cost of the control, whose role is to influence the growth rate of the species just to get better quality with the purpose of obtaining a greater benefit from the harvest. In the second section, some preliminary results are summarized, including the existence of an optimal control, i.e., a function f such that the profitability of the harvest is maximized. In the following, some necessary conditions for a control to be an optimal control are obtained; in particular, the optimality system

*Received by the editors October 18, 1995; accepted for publication (in revised form) June 16, 1997; published electronically May 15, 1998. This research was partially supported by Dirección General de Enseñanza Superior, Ministry of Education and Science (Spain) grant PB95-1190 and by EEC contract (Human Capital and Mobility program) ERBCHRXCT 940494.

http://www.siam.org/journals/sicon/36-4/29332.html

†Departamento de Análisis Matemático, Universidad de Granada, 18071 Granada, Spain (acanada@goliat.ugr.es, jlgamez@goliat.ugr.es, jmontero@goliat.ugr.es).

is deduced. With the help of some estimations for the solutions of such a system, we prove, in section 4, the uniqueness of the optimal control in two cases: when the parameter λ of the functional (1.2) is small and when the function b in (1.1) is a constant sufficiently large. The same ideas may be useful for studying the case where b is not necessarily constant; in this situation, it is possible to impose a restriction on the quotient between the supremum and the infimum (which, again, must be sufficiently large) of the function b to assure uniqueness (see [5]). In the last section, we give, for λ sufficiently small, a constructive scheme which provides a sequence of functions converging to some special solutions of the optimality system; this will be useful for approximating the optimal control. Of course, similar procedures are valid in other cases such as those considered in the fourth section. Related problems have been considered in [8, 9, 13]. In fact, our work was motivated by [9], where the authors study a problem like (1.1) but with Neumann boundary conditions and the controls are restricted to members of the set

$$C_\delta = \{g \in L^\infty(\Omega) : 0 \le g(x) \le \delta \text{ a.e. in } \Omega\},$$

where $0 < \delta < \inf_{x\in\Omega} a(x)$. The main novelty of our results is that we study the case of Dirichlet boundary conditions, which seems to be different in many aspects from the Neumann case (see [3], [7]). Also, our control space $L^\infty_+(\Omega)$, the set of functions of $L^\infty(\Omega)$ that are nonnegative a.e. in Ω, is different from that of [9] (see the final Remark 1), and we prove not only existence but also uniqueness of the optimal control, which is very important for its possible approximation as is seen in this paper.

2. Preliminary results. In this section we present some previous results and notation which will be useful below. The details may be seen in [1], [3], [6], [7], [10], and [11]. If $e \in L^\infty(\Omega)$, we denote $\underline{e} = \operatorname{ess\,inf}_{x\in\Omega} e(x)$, $\overline{e} = \operatorname{ess\,sup}_{x\in\Omega} e(x)$.

Let us consider equation (1.1). From now on, we assume the following hypothesis.

Hypothesis H. $a, b \in L^\infty(\Omega)$, $\underline{b} > 0$, $f \in L^\infty_+(\Omega) = \{g \in L^\infty(\Omega) : g(x) \ge 0 \text{ a.e. in } \Omega\}$.

For a function $q \in L^\infty(\Omega)$, we define $\sigma_1(q)$ to be the principal eigenvalue of the eigenvalue problem

$$\begin{gathered} -\Delta u(x) + q(x)u(x) = \sigma u(x),\ x \in \Omega, \\ u(x) = 0,\ x \in \partial\Omega. \end{gathered}$$

This principal eigenvalue can be expressed variationally as

$$\sigma_1(q) = \inf_{u\in H^1_0(\Omega)\setminus\{0\}} \frac{\displaystyle\int_\Omega |\nabla u|^2 + \int_\Omega q|u|^2}{\displaystyle\int_\Omega |u|^2}, \tag{2.3}$$

where $H^1_0(\Omega)$ is the usual Sobolev space. It is known that the algebraic multiplicity of $\sigma_1(q)$ is equal to one, and it is possible to choose an associated eigenfunction $\phi_1(q)$ (where the previous infimum is attained, becoming a minimum) such that $\phi_1(q) \in C^{1,\alpha}(\overline{\Omega})$ $\forall \alpha \in (0,1)$, $\phi_1(q) > 0$ in Ω, $\|\phi_1(q)\|_{L^\infty} = 1$.

As a direct consequence of the previous variational characterization, we obtain the following properties:

(i) $\sigma_1(q)$ is strictly increasing with respect to the weight function q; i.e., if $q_1 \le q_2$, we have $\sigma_1(q_1) \le \sigma_1(q_2)$ and this last inequality is strict if, moreover, $q_1(x) < q_2(x)$ on a subset of positive measure.

(ii) $\forall M \in \mathbb{R}$, $\sigma_1(q+M) = \sigma_1(q) + M$.

(iii) $\sigma_1(q)$ is continuous with respect to $q \in L^\infty(\Omega)$.

By using techniques of sub- and supersolutions, one may prove [1] that (1.1) has a (weak) nontrivial and nonnegative solution u_f iff $\sigma_1(-a+f) < 0$. In this case, the solution u_f is the unique nontrivial and nonnegative solution of (1.1), and it verifies the estimates

$$\frac{-\sigma_1(-a+f)}{\overline{b}}\ \phi_1(-a+f)(x) \le u_f(x) \le \frac{\overline{a}-\underline{f}}{\underline{b}}\ \forall x \in \Omega. \tag{2.4}$$

It is known [11] that every solution obtained by this method is a priori stable when considered as an equilibrium solution of an associated time dependent evolution equation. Also, the sub- and supersolutions provide an estimate of the extent of stability. Moreover, any bounded subsolution w of (1.1) must satisfy $w \le u_f$. If $\sigma_1(-a+f) < 0$, then $u_f(x) > 0\ \forall x \in \Omega$, and, therefore, taking into account the previous property (i), we have

$$\sigma_1(-a+f+2bu_f) > \sigma_1(-a+f+bu_f) = 0. \tag{2.5}$$

Note that Lyapunov's indirect argument analyzes stability by considering $u = u_f + \beta v \exp^{\mu t}$ in the time dependent problem, leading to the self-adjoint eigenvalue problem

$$\mu v - \Delta v + [-a+f+2bu_f]v = 0.$$

So, (2.5) provides the condition for linearized stability. Now we may extend the definition of u_f. For this, for each $f \in L^\infty_+(\Omega)$ (and in the same way, for each $f \in L^\infty(\Omega)$), we will denote by u_f the maximal nonnegative solution of equation (1.1). Then $u_f \equiv 0$ iff $\sigma_1(-a+f) \ge 0$ and u_f is strictly positive in Ω iff $\sigma_1(-a+f) < 0$. The following monotonicity property of u_f is easily proved, with respect to f, and is fundamental in many assumptions contained in this paper: if $f, g \in L^\infty(\Omega)$, and $f \le g$, then, $u_f \ge u_g$.

Let λ be a real positive constant and define the functional $J : L^\infty_+(\Omega) \to \mathbb{R}$, given by the expression

$$J(g) = \int_\Omega (\lambda u_g g - g^2).$$

The existence of an optimal control in $L^\infty_+(\Omega)$, i.e., $f \in L^\infty_+(\Omega)$, satisfying

$$J(f) = \sup_{g \in L^\infty_+(\Omega)} J(g)$$

has been done by the authors in [3]. More precisely, we have the following result.

THEOREM 2.1. *Consider problem* (1.1) *under Hypothesis* H. *Then the optimal control problem has a solution in the space* $L^\infty_+(\Omega)$.

The basic idea for the proof of this theorem is that the possible optimal controls must be bounded; in fact, if $f \in L^\infty_+(\Omega)$ and $g = \min\{f, \frac{\overline{a}\lambda}{\underline{b}}\}$, then $J(g) \ge J(f)$. To see this, we consider two cases:

(a) $u_f \equiv 0$. Then $J(f) = -\int_\Omega f^2 \le -\int_\Omega g^2 \le J(g)$.

(b) $u_f > 0$ in Ω. Then by the monotonicity property of u_f, with respect to f, we have $u_g \ge u_f > 0$ in Ω. By careful discussion, one may prove

$$\lambda u_g g - g^2 \ge \lambda u_f f - f^2 \quad \text{a.e. in } \Omega,$$

which implies $J(g) \ge J(f)$.

In particular, if $f \in L^\infty_+(\Omega)$ is an optimal control and $\overline{a} > 0$, then we may assume that

$$f \leq \frac{\overline{a}\lambda}{\underline{b}} \text{ a.e. in } \Omega. \tag{2.6}$$

Now from (2.4) and (2.6) it is easy to prove that if $\{f_n\}$ is any maximizing sequence in $L^\infty_+(\Omega)$ for J, then there exists $f \in L^\infty_+(\Omega)$ such that, for a convenient subsequence,

$$\begin{aligned} f_n &\rightharpoonup f \text{ weakly in } L^2(\Omega), \\ u_{f_n} &\to u_f \text{ strongly in } H^1_0(\Omega). \end{aligned}$$

Consequently, $J(f) = \sup_{g \in L^\infty_+(\Omega)} J(g)$.

Theorem 2.1 assures the existence of optimal control. Moreover, we may give conditions to guarantee the positivity of the benefit, i.e., to conclude that the quantity $\sup_{g \in L^\infty_+(\Omega)} J(g)$ is strictly positive. To do so, note that if this is true, then there exists $g \in L^\infty_+(\Omega)$ such that $J(g) > 0$. So, $u_g > 0$ and $\sigma_1(-a) \leq \sigma_1(-a+g) < 0$. Reciprocally, if $\sigma_1(-a) < 0$ is assumed, then observe that for any $f \in L^\infty_+(\Omega)$ we may write

$$\lambda u_f f - f^2 = \frac{\lambda^2 u_f^2}{4} + \left(f - \frac{\lambda u_f}{2} \right)^2.$$

Also, the condition $\sigma_1(-a) < 0$ implies the existence of a positive f such that $f = \frac{\lambda u_f}{2}$. Consequently, $J(f) > 0$ (see [3] for the details, where some additional estimates for the profit are also given). Therefore,

$$\sup_{g \in L^\infty_+(\Omega)} J(g) > 0 \;\Leftrightarrow\; \sigma_1(-a) < 0,$$

which justifies the hypothesis $\sigma_1(-a) < 0$ that we will assume in many of the next results. Now we present some results about elliptic operators of the Schrödinger type $-\Delta u + q(x)u$, $u \in H^1(\Omega)$.

LEMMA 2.2. *Assume*

$$q \in L^\infty(\Omega),\ \sigma_1(q) > 0.$$

Then the following results hold.

(i) *For each $f \in L^2(\Omega)$ the linear problem*

$$\begin{aligned} -\Delta u + q(x)u &= f \text{ in } \Omega, \\ u &= 0 \text{ on } \partial\Omega, \end{aligned} \tag{2.7}$$

has a unique solution $u \in H^1_0(\Omega)$. Moreover, if $f \in L^\infty(\Omega)$, $u \in C^{1,\alpha}(\overline{\Omega})$ $\forall\, \alpha \in (0,1)$.

(ii) *Let $u_1, u_2 \in H^1(\Omega)$ satisfy $u_2 \leq u_1$ on $\partial\Omega$ and $-\Delta u_2 + q u_2 \leq -\Delta u_1 + q u_1$ in the weak sense; i.e., $\forall\ \phi \in H^1_0(\Omega)$, $\phi \geq 0$,*

$$\int_\Omega \nabla u_2 \nabla \phi + \int_\Omega q u_2 \phi \leq \int_\Omega \nabla u_1 \nabla \phi + \int_\Omega q u_1 \phi.$$

Then $u_2 \leq u_1$ in Ω.

(iii) *Consider* $p \in L^\infty(\Omega)$, $p \geq q$ *in* Ω *and* $f \in L^2(\Omega)$, $f \geq 0$ *in* Ω. *Denote by* $\omega(p), \omega(q)$ *the respective solutions of problems*

$$-\Delta u + p(x)u = f \text{ in } \Omega,$$
$$u = 0 \text{ on } \partial\Omega,$$

and (2.7). *Then*

$$\omega(p) \leq \omega(q) \text{ in } \Omega.$$

Proof. The first part may be proved by applying the Lax–Milgram theorem to the bilinear form $L : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ defined by

$$L(u, v) = \int_\Omega \nabla u \nabla v + \int_\Omega quv$$

and the functional $\varphi : H_0^1(\Omega) \to \mathbb{R}$ defined by

$$\varphi(v) = \int_\Omega fv.$$

In fact, from the variational characterization of $\sigma_1(q)$ (see (2.3)) it follows that

$$\int_\Omega |\nabla u|^2 + \int_\Omega qu^2 \geq c \int_\Omega |\nabla u|^2 \ \forall\ u \in H_0^1(\Omega),$$

where

$$c = \frac{\sigma_1(q)}{\sigma_1(q) + \|q\|_\infty}. \tag{2.8}$$

Parts (ii) and (iii) are a direct consequence of the maximum principle (see [10] for more details). □

Remark 1. It is important to observe that if the function q in (2.7) belongs to a subset A of $L^\infty(\Omega)$ such that there exist two positive constants M and μ satisfying

$$\|q\|_\infty \leq M,\ \sigma_1(q) \geq \mu \ \forall\ q \in A,$$

then the constant c in (2.8) may be chosen independent of $q \in A$. In fact, we can take

$$c = \frac{\mu}{\mu + M}.$$

To do so, it is sufficient to see that

$$\frac{\mu}{\mu + M} \leq \frac{\sigma_1(q)}{\sigma_1(q) + \|q\|_\infty} \ \forall\ q \in A.$$

3. The optimality system. In this section we obtain some necessary conditions for an element $f \in L_+^\infty(\Omega)$ to be an optimal control. This will be carried out by deducing the optimality system satisfied for some pair (u_f, P_f) associated with f. Among other things, it will allow us to prove, in the next section, the uniqueness of the optimal control when the data of the problem fulfill some restrictions. We begin by proving a property about the directional differentiability of u_f with respect to f.

LEMMA 3.1. *Let $f \in L^\infty(\Omega)$ such that*

$$\sigma_1(-a+f) < 0. \tag{3.9}$$

Then

$$\frac{u_{f+\beta g} - u_f}{\beta} \to \xi$$

in $H_0^1(\Omega)$ as $\beta \to 0$ for any $g \in L^\infty(\Omega)$. Further, ξ is the unique solution of the linear problem

$$\begin{aligned} -\Delta\xi + [-a+f+2bu_f]\xi &= -gu_f \quad \text{in } \Omega, \\ \xi &= 0 \text{ on } \partial\Omega. \end{aligned} \tag{3.10}$$

In particular, this is true for each $f \in L^\infty_+(\Omega)$ such that $J(f) > 0$.

Proof. If $\beta \neq 0$ and

$$\xi_\beta = \frac{u_{f+\beta g} - u_f}{\beta},$$

then it satisfies

$$\begin{aligned} -\Delta\xi_\beta + [-a+f+b(u_{f+\beta g}+u_f)]\xi_\beta &= -gu_{f+\beta g} \quad \text{in } \Omega, \\ \xi_\beta &= 0 \text{ on } \partial\Omega. \end{aligned} \tag{3.11}$$

Since $\sigma_1(q)$ is increasing with respect to $q \in L^\infty(\Omega)$ and u_f is decreasing with respect to f, for each $\epsilon \in \mathbb{R}^+$ we have

$$\sigma_1(-a+f+b(u_{f+\beta g}+u_f)) \geq \sigma_1(-a+f+b(u_{f+\epsilon\|g\|_\infty}+u_f)) \equiv \mu,$$

provided that $|\beta| \leq \epsilon$, where $\|\cdot\|_\infty$ is the usual norm in $L^\infty(\Omega)$. Also, it is possible to take ϵ such that $\mu > 0$. To see this, note that if ϵ is sufficiently small the continuity of $\sigma_1(q)$ with respect to $q \in L^\infty(\Omega)$ and (3.9) imply

$$\sigma_1(-a+f+\epsilon\|g\|_\infty) < 0.$$

Consequently, $bu_{f+\epsilon\|g\|_\infty}$ is strictly positive in Ω and therefore the strict monotonicity of $\sigma_1(q)$ with respect to $q \in L^\infty(\Omega)$ and the monotonicity of $f \to u_f$ imply

$$\sigma_1(-a+f+b(u_{f+\epsilon\|g\|_\infty}+u_f)) > \sigma_1(-a+f+bu_f) = 0.$$

Taking into account the remark after Lemma 2.2 and (3.11), there is a constant c, independent of $\beta \in (-\epsilon,\epsilon)$, such that

$$\begin{aligned} &c\|\xi_\beta\|^2_{H_0^1(\Omega)} \\ &\leq \int_\Omega \{|\nabla\xi_\beta|^2 + [-a+f+b(u_{f+\beta g}+u_f)]\xi_\beta^2\} \\ &= \int_\Omega -gu_{f+\beta g}\xi_\beta \leq K\|\xi_\beta\|_{H_0^1(\Omega)} \end{aligned}$$

for some positive constant K (see (2.4)). Thus there exists some constant d, independent of $\beta \in (-\epsilon,\epsilon)$, such that

$$\|\xi_\beta\|_{H_0^1(\Omega)} \leq d. \tag{3.12}$$

This implies $u_{f+\beta g} \to u_f$ in $H_0^1(\Omega)$ as $\beta \to 0$. Taking into account that $bu_f > 0$ implies $\sigma_1(-a+f+2bu_f) > \sigma_1(-a+f+bu_f) = 0$ we obtain, from the uniqueness of solutions to (3.10), that $\xi_\beta \rightharpoonup \xi$ in $H_0^1(\Omega)$. Now rewriting (3.11) in the form

$$\begin{gathered} -\Delta\xi_\beta + [-a+f+2bu_f]\xi_\beta = -gu_f - \beta[g+b\xi_\beta]\xi_\beta \ \text{ in } \ \Omega, \\ \xi_\beta = 0 \ \text{ on } \ \partial\Omega \end{gathered}$$

one actually has strong convergence $\xi_\beta \to \xi$ in $H_0^1(\Omega)$. □

The previous result allows us to obtain a new step toward the derivation of the optimality system.

LEMMA 3.2. *Assume*

$$\sigma_1(-a) < 0. \tag{3.13}$$

If $f \in L_+^\infty(\Omega)$ is any optimal control, then

$$f = \frac{\lambda}{2}\, u_f(1-P_f)^+ \ \ \textit{a.e. in } \Omega, \tag{3.14}$$

where P_f is the unique solution of the linear problem

$$\begin{gathered} -\Delta P_f + (-a+f+2bu_f)P_f = f \ \ \textit{in } \Omega, \\ P_f = 0 \ \ \textit{on } \partial\Omega. \end{gathered} \tag{3.15}$$

Proof. Let $f \in L_+^\infty(\Omega)$ be an optimal control and $g \in L^\infty(\Omega)$, so that $f + \beta g \in L_+^\infty(\Omega)$ as $\beta \to 0^+$. Then

$$J(f+\beta g) - J(f) \le 0.$$

Dividing by β, we obtain

$$\int_\Omega \left[\lambda\, \frac{u_{f+\beta g} - u_f}{\beta}\, (f+\beta g) + \lambda u_f g - 2gf - \beta g^2\right] \le 0.$$

Letting $\beta \to 0^+$ and using Lemma 3.1, we have

$$\int_\Omega (\lambda\xi f + \lambda u_f g - 2gf) \le 0, \tag{3.16}$$

where ξ is defined by (3.10). Now multiplying equation (3.10) by P_f, multiplying equation (3.15) by ξ, and integrating and subtracting both expressions, we obtain

$$\int_\Omega f\xi + \int_\Omega g u_f P_f = 0. \tag{3.17}$$

Combining (3.16) and (3.17), we deduce in particular

$$\int_\Omega g[\lambda u_f(1-P_f) - 2f] \le 0 \ \forall\ g \in L_+^\infty(\Omega).$$

Therefore,

$$f \ge \frac{\lambda}{2} u_f(1-P_f) \ \text{ a.e. in } \Omega. \tag{3.18}$$

On the other hand, observe that if we take $g = -f$, then $f + \beta g \in L^\infty_+(\Omega)\ \forall\ \beta \in (0,1)$. Consequently, in the same way as before, we obtain

$$\int_{\Omega\cap\{f>0\}} f[\lambda u_f(1-P_f) - 2f] = \int_\Omega f[\lambda u_f(1-P_f) - 2f] \geq 0.$$

So, from (3.18) we must have

$$f = \frac{\lambda}{2} u_f(1-P_f) \ \text{ a.e. in } \ \Omega \cap \{f > 0\}. \tag{3.19}$$

From (3.18) and (3.19) we conclude (3.14). □

Our next purpose would be to prove that if $f \in L^\infty_+(\Omega)$ is an optimal control, then $P_f \leq 1$ a.e. in Ω. To do so, the basic tool may be the assertions (ii) and (iii) of Lemma 2.2; in fact, $P_f \geq 0$ by (ii). Moreover, if one wants to find an upper bound for P_f, the logical way must be to establish an upper bound for f and a lower bound for the function $-a + f + 2bu_f$. The existence of an upper bound for any optimal control is shown in (2.6). The lower bound for the mentioned function is a direct consequence of the continuity of the mapping u_f with respect to f, as we see in the next result.

LEMMA 3.3. *Assume* (3.13) *and choose* $\epsilon \in \mathbb{R}^+$ *such that*

$$\epsilon < \frac{\sigma_1(-a + 2bu_0)}{2\overline{b}}. \tag{3.20}$$

Then there exists a positive constant Λ_0 *depending on* $a, b,$ *and* Ω *such that if*

$$\lambda \leq \Lambda_0, \tag{3.21}$$

the function P_f *defined in* (3.15) *for any optimal control* f *satisfies the inequality*

$$0 \leq P_f \leq \lambda\ Q \text{ a.e. in } \Omega, \tag{3.22}$$

where Q *is the unique solution of the problem*

$$\begin{gathered} -\Delta Q + (-a + 2b(u_0 - \epsilon))Q = \frac{\overline{a}}{\underline{b}} \ \text{ in } \Omega, \\ Q = 0 \ \text{ on } \partial\Omega. \end{gathered} \tag{3.23}$$

Proof. First, note that from (3.13) u_0 is strictly positive in Ω and, therefore, $\sigma_1(-a + bu_0) = 0$. So $\sigma_1(-a + 2bu_0) > 0$. Now take ϵ satisfying (3.20). Then since the mapping $L^\infty_+(\Omega) \to C^1(\overline{\Omega})$, $f \to u_f$, is continuous [2], there is a positive constant Λ_0 such that

$$g \in L^\infty_+(\Omega),\ g \leq \Lambda_0 \frac{\overline{a}}{\underline{b}} \Rightarrow u_g \geq u_0 - \ \epsilon \ \text{ a.e. in } \ \Omega. \tag{3.24}$$

Also, from (3.20) and Lemma 2.2(i), problem (3.23) has a unique solution Q, and the function λQ satisfies the equation

$$-\Delta(\lambda Q) + (-a + 2b(u_0 - \epsilon))\lambda Q = \lambda \frac{\overline{a}}{\underline{b}}. \tag{3.25}$$

Last, if λ satisfies (3.21) and f is an optimal control, from (2.6), (3.24), and statements (ii) and (iii) of Lemma 2.2, we conclude (3.22). □

COROLLARY 3.4. *Let us suppose* (3.13) *and*

$$\lambda \leq \min\left\{\Lambda_0, \frac{1}{\|Q\|_\infty}\right\} \equiv \Lambda_1. \tag{3.26}$$

Then the function P_f *defined in* (3.15) *for any optimal control* f *satisfies the inequality*

$$0 \leq P_f \leq 1 \quad a.e.\ in\ \Omega.$$

Now we can state the main result of this section.

THEOREM 3.5. *Assume* (3.13) *and* (3.26). *Then any optimal control* $f \in L^\infty_+(\Omega)$ *may be expressed in the form*

$$f = \frac{\lambda}{2} u_f (1 - P_f), \tag{3.27}$$

where the pair $(u_f, P_f) \equiv (u, p)$ *satisfies*

$$0 \leq p \leq 1,\ u > 0 \quad a.e.\ in\ \Omega \tag{3.28}$$

and the optimality system

$$\begin{aligned} -\Delta u &= u\left(a - \left[b + \frac{\lambda}{2}(1-p)\right]u\right)\ in\ \Omega, \\ -\Delta p + p(-a + 2bu) &= \frac{\lambda}{2} u (1-p)^2\ in\ \Omega, \\ u = p &= 0\ on\ \partial\Omega. \end{aligned} \tag{3.29}$$

Proof. The proof is trivial using the previous results. □

The expression (3.27) for the optimal controls may be used for deducing some of their qualitative and quantitative properties. For instance, under the hypotheses of Theorem 3.5, all the optimal controls in $L^\infty_+(\Omega)$ have a suitable regularity: they must belong to the space $C^{1,\alpha}(\overline{\Omega})$ for any $\alpha \in (0,1)$. Also, if $\lambda < \Lambda_1$, then any optimal control f is such that $f > 0$ a.e. in Ω. Moreover, the uniqueness and approximation of the optimal control will be a consequence of (3.27); they will be deduced in the next sections.

4. Uniqueness of the optimal control. In this section we prove the uniqueness of the optimal control when the parameter λ of the functional (1.2) is small. To do so, we take into account that in Theorem 3.5 any optimal control is expressed in the form (3.27) with the pair (u, p) satisfying (3.28) and the optimality system (3.29). As a consequence, the uniqueness of the optimal control will be obtained by proving that for λ sufficiently small the optimality system (3.29) has a unique solution (u, p) verifying (3.28). First it is necessary to give some estimations for the solutions of system (3.29).

LEMMA 4.1. *Assume* (3.13) *and* (3.26) (ϵ *is chosen as in* (3.20)). *Let* (v, q) *be any solution of system* (3.29) *satisfying* (3.28). *Then*

$$u_0(x) - \epsilon \leq w(x) \leq v(x) \leq \frac{\overline{a}}{\underline{b}}, \quad 0 \leq q(x) \leq \lambda Q(x) \quad a.e.\ \ in\ \ \Omega, \tag{4.30}$$

where w *is the maximal nonnegative solution of*

$$\begin{aligned} -\Delta w &= w\left[a - \frac{\lambda}{2} w - bw\right]\ \ in\ \Omega, \\ w &= 0\ \ on\ \partial\Omega, \end{aligned} \tag{4.31}$$

and Q *is the unique solution of* (3.23).

Proof. As $\sigma_1(-a) < 0$, the function w is in fact strictly positive a.e. in Ω. So $w = u_{\frac{\lambda}{2}w}$, and since $\frac{\lambda}{2}w \leq \Lambda_0 \frac{\overline{a}}{\underline{b}}$ (see (2.4)), from (3.24) we obtain $w(x) \geq u_0(x) - \epsilon$ a.e. in Ω. On the other hand, it may be directly verified that w is a subsolution of the first equation of the optimality system (3.29); therefore, the maximality property of v (remember that $v = u_{\frac{\lambda}{2}(1-p)v}$) proves $w(x) \leq v(x)$ a.e. in Ω. The last inequality for v, i.e., $v(x) \leq \frac{\overline{a}}{\underline{b}}$, is trivial, taking into account systems (3.29) and (2.4). In relation to the function q, note that it satisfies the problem

$$\begin{aligned} -\Delta q + q(-a + 2bv) &= \frac{\lambda}{2} v(1-q)^2 \quad \text{in } \Omega, \\ q &= 0 \quad \text{on } \partial\Omega. \end{aligned}$$

Moreover, $-a + 2bv \geq -a + 2b(u_0 - \epsilon)$ and $\frac{\lambda}{2}v(1-q)^2 \leq \lambda \frac{\overline{a}}{\underline{b}}$. Consequently, the inequality $q \leq \lambda Q$ is deduced from (3.25) and statements (ii) and (iii) of Lemma 2.2. □

THEOREM 4.2. *Let us suppose* (3.13), *and take* ϵ *as in* (3.20). *Define* $\delta_0 \equiv \sigma_1(-a + 2b(u_0 - \epsilon))$, *which from* (3.20) *is strictly positive. Then if*

$$\lambda \leq \Lambda_2 \equiv \min \left\{ \Lambda_1, \frac{4\delta_0}{4\overline{b}\|Q\|_\infty + 1 + \frac{\overline{a}^2}{\underline{b}^2}} \right\}, \tag{4.32}$$

there can be only one solution (u, p) *of* (3.29) *satisfying* (3.28).

Proof. Let (u, p) and (v, q) be two solutions of (3.29) verifying (3.28). Then

$$\begin{aligned} 0 = &-\Delta(u-v) - a(u-v) + b(u-v)(u+v) \\ &+ \frac{\lambda}{2}(1-p)(u+v)(u-v) - \frac{\lambda}{2}v^2(p-q) \end{aligned} \tag{4.33}$$

and

$$\begin{aligned} 0 = &-\Delta(p-q) - a(p-q) + 2bu(p-q) \\ &+ 2bq(u-v) - \frac{\lambda}{2}(u-v) + \lambda u(p-q) \\ &+ \lambda q(u-v) - \frac{\lambda}{2}u(p-q)(p+q) - \frac{\lambda}{2}q^2(u-v). \end{aligned} \tag{4.34}$$

Multiplying (4.33) by $(u-v)$, multiplying (4.34) by $(p-q)$, integrating on Ω, and adding the two expressions, we obtain

$$\begin{aligned} 0 = &\int_\Omega \left[|\nabla(u-v)|^2 - a(u-v)^2 + b(u+v)(u-v)^2 \right. \\ &\left. + \frac{\lambda}{2}(1-p)(u+v)(u-v)^2 - \frac{\lambda}{2}v^2(u-v)(p-q) \right] \\ &+ \int_\Omega \left[|\nabla(p-q)|^2 - a(p-q)^2 + 2bu(p-q)^2 \right. \\ &+ (u-v)(p-q)\left(2bq - \frac{\lambda}{2} + \lambda q - \frac{\lambda}{2}q^2 \right) \\ &\left. + \frac{1}{2}\lambda u(2 - (p+q))(p-q)^2 \right]. \end{aligned}$$

Now observe that the terms

$$\frac{\lambda}{2}(1-p)(u+v)(u-v)^2$$

and

$$\frac{1}{2}\lambda u(2-(p+q))(p-q)^2$$

in the previous equality are both nonnegative. Also, from Lemma 4.1, functions u and v are greater than or equal to the function $u_0-\epsilon$. Moreover, the variational characterization of δ_0 (see 2.3) implies

$$\int_\Omega |\nabla r|^2 + \int_\Omega (-a+2b(u_0-\epsilon))r^2 \geq \delta_0 \int_\Omega r^2 \quad \forall\, r \in H_0^1(\Omega). \tag{4.35}$$

Therefore,

$$\begin{aligned} 0 \geq \int_\Omega \Bigg[& \delta_0(u-v)^2 + \delta_0(p-q)^2 \\ & + (u-v)(p-q)\left(2bq - \frac{\lambda}{2} + \lambda q - \frac{\lambda}{2}(q^2+v^2)\right)\Bigg] \end{aligned} \tag{4.36}$$

with strict inequality if $p(x) \neq q(x)$ in any subset of Ω with positive measure.

Also, by using the estimations (4.30) and hypothesis (4.32), we have

$$\begin{aligned} & \left|2bq - \frac{\lambda}{2} + \lambda q - \frac{\lambda}{2}(q^2+v^2)\right| \\ & = \left|2bq - \frac{\lambda}{2}(1-q)^2 - \frac{\lambda}{2}v^2\right| \\ & \leq \lambda\left[2\bar{b}\|Q\|_\infty + \frac{1}{2}\left(1+\frac{\overline{a}^2}{\underline{b}^2}\right)\right] \leq 2\delta_0. \end{aligned}$$

So, the integral that appears in (4.36) is also greater than or equal to zero. This implies $p(x)=q(x)$ a.e. in Ω and consequently $u(x)=v(x)$ a.e. in Ω. □

The main consequence of the previous result is the uniqueness of the optimal control for the considered control problem. Before discussing this result, it is convenient to say something more about notation. Our control problem lies in maximizing the functional (1.2) where, for a given function f, u_f means the maximal nonnegative solution to problem (1.1) (this requires Hypothesis H, which was established at the beginning of section 2 and assumed throughout the paper). So, this control problem is completely defined by the domain Ω, the functions a and b, and the parameter λ. Therefore, it is clear to denote it by $P_{\Omega,a,b,\lambda}$. The next corollary shows that when Ω, a, and b are conveniently fixed, the problem $P_{\Omega,a,b,\lambda}$ has a unique solution for λ sufficiently small.

COROLLARY 4.3. *Let us consider the problem $P_{\Omega,a,b,\lambda}$. Assume that the domain Ω and the functions a and b are fixed satisfying Hypothesis H and* (3.13). *Then if*

$$\lambda \leq \Lambda_2,$$

where Λ_2 is defined in (4.32), *the problem $P_{\Omega,a,b,\lambda}$ has a unique optimal control.*

Remark 2. It is important to notice that the positive constant Λ_2 defined in (4.32) depends only on the domain Ω and the functions a and b; so it would be more correct to denote it as $\Lambda_2(\Omega, a, b)$ instead of Λ_2. This observation allows us to prove the uniqueness of the optimal control in situations different from that considered in the previous corollary. For instance, this may be done if we fix Ω, the function a, and the parameter λ and consider b as a constant function to be chosen in a proper manner such as is indicated in the next corollary.

COROLLARY 4.4. *Let us consider the problem $P_{\Omega,a,b,\lambda}$ with b as a positive constant function. Assume that the constant λ, the domain Ω, and the function a are fixed satisfying Hypothesis H and* (3.13). *Choose $\Lambda_2(\Omega, a, 1)$ as in Corollary* 4.3. *Then if*

$$b \geq \frac{\lambda}{\Lambda_2(\Omega, a, 1)},$$

the problem $P_{\Omega,a,b,\lambda}$ has a unique optimal control.

Proof. It is sufficient to observe that

$$b u_{\Omega,a,b,f} = u_{\Omega,a,1,f}, \tag{4.37}$$

where $u_{\Omega,a,b,f}$ means the maximal nonnegative solution of (1.1). As a consequence, we have that for each $f \in L^\infty_+(\Omega)$,

$$J_{\Omega,a,b,\lambda}(f) \equiv \int_\Omega (\lambda u_{\Omega,a,b,f} f - f^2) = J_{\Omega,a,1,\frac{\lambda}{b}}\ (f).$$

Therefore, Corollary 4.3 may be applied to obtain the desired conclusion. □

Remark 3. If b is a positive constant function, the previous identity shows that when the quantity $\frac{\lambda}{b}$ is sufficiently small, the optimal control is unique. This may be carried out if b is fixed and λ is sufficiently small or if λ is fixed and b is sufficiently large. In addition to the treated cases, it is possible to study the other case where Ω, the function a, and the parameter λ are fixed and b is not necessarily a constant function. It does not seem possible to study this last case, proving a relation similar to (4.37). However, an analogous treatment to that considered in the previous results can be done, obtaining a result about the uniqueness of the optimal control, which involves, first, a restriction on the quantity $\frac{\overline{b}}{\underline{b}}$ (basically, this quantity must belong to the interval $[1, 2)$), and, second, $\underline{b}$ must be sufficiently large (see [5]).

Remark 4. We have proved that when λ is small enough or b is a large enough constant, the optimality system (3.29) admits a unique solution satisfying (3.28). We may ask about the essential implication of these two conditions on the functional J. In fact, if the domain Ω and the functions a and b are fixed, satisfying Hypothesis H and (3.13), there exists a different argument to prove that when λ is sufficiently small the optimal control is unique. This argument uses the regularity of the functional J on suitable subsets of $L^\infty(\Omega)$ (which contain the optimal controls), and in addition the monotonicity of its Fréchet derivative (see final Remark 2).

5. Approximation of the optimal control. The main purpose of this section is to give, for λ sufficiently small, a constructive scheme which provides a sequence of functions converging to the unique solution of the optimality system (3.29) satisfying (3.28). Due to the relation (3.27), this may be useful for approximating the optimal control f. For clarity of the exposition, it is convenient to adopt a more general

framework than (3.29). Consider the elliptic system

$$\begin{array}{ll} -\Delta u(x) = B(x,u(x),p(x)), & x \in \Omega, \\ -\Delta p(x) = C(x,u(x),p(x)) + D(x,u(x),p(x)), & x \in \Omega, \\ u(x) = p(x) = 0, & x \in \partial\Omega, \end{array} \tag{5.38}$$

where Ω is a bounded and regular domain in $\mathbb{R}^n$, and the nonlinearities $B, C, D : \overline{\Omega} \times \mathbb{R}^2 \to \mathbb{R}$ satisfy (regularity condition):

Hypothesis H1. B, C and D are continuous with respect to $(u,p) \in \mathbb{R}^2$ for fixed $x \in \overline{\Omega}$. Moreover, $\forall u, p \in L^\infty(\Omega)$, the functions $B(\cdot, u(\cdot), p(\cdot)), C(\cdot, u(\cdot), p(\cdot))$, and $D(\cdot, u(\cdot), p(\cdot))$ belong to $L^\infty(\Omega)$.

DEFINITION 5.1. *Let $\underline{u}, \overline{u}, \underline{p}, \overline{p} \in H^1(\Omega) \cap L^\infty(\Omega)$. Such functions are said to be a* system of upper-lower-solutions *for system* (5.38) *if they verify:* (a)

$$\begin{cases} \underline{u}(x) \le \overline{u}(x), \quad \underline{p}(x) \le \overline{p}(x) & \text{a.e. in } \Omega, \\ \underline{u} \le 0 \le \overline{u}, \qquad \underline{p} \le 0 \le \overline{p} & \text{in } \partial\Omega. \end{cases}$$

(A function $v \in H^1(\Omega)$ is said to be less than or equal to $w \in H^1(\Omega)$ on $\partial\Omega$ when $(v-w)^+ = \max\{0, v-w\} \in H_0^1(\Omega)$.) (b) $\forall \phi \in H_0^1(\Omega)$, $\phi \ge 0$,

$$\int_\Omega \nabla \overline{u} \cdot \nabla \phi \ge \int_\Omega B(x, \overline{u}, \overline{p})\phi,$$

$$\int_\Omega \nabla \underline{u} \cdot \nabla \phi \le \int_\Omega B(x, \underline{u}, \underline{p})\phi,$$

$$\int_\Omega \nabla \overline{p} \cdot \nabla \phi \ge \int_\Omega C(x, \underline{u}, \overline{p})\phi + \int_\Omega D(x, \overline{u}, \overline{p})\phi,$$

$$\int_\Omega \nabla \underline{p} \cdot \nabla \phi \le \int_\Omega C(x, \overline{u}, \underline{p})\phi + \int_\Omega D(x, \underline{u}, \underline{p})\phi.$$

Also, we will assume (monotonicity condition) the following hypothesis.

Hypothesis H2. $\exists\, M > 0$ such that the function $B(x,u,p) + Mu$ is increasing in u and the functions $C(x,u,p) + \frac{M}{2}p$, $D(x,u,p) + \frac{M}{2}p$ are increasing in p for $(x,u,p) \in \overline{\Omega} \times [\text{inf ess } \underline{u}, \text{sup ess } \overline{u}] \times [\text{inf ess } \underline{p}, \text{sup ess } \overline{p}]$. Moreover, functions B, C, and D satisfy the following monotonicity properties with respect to the other variables: $B(x,u,p)$ is increasing in p, the function $C(x,u,p)$ is decreasing in u, and the function $D(x,u,p)$ is increasing in u for $(x,u,p) \in \overline{\Omega} \times [\text{inf ess } \underline{u}, \text{sup ess } \overline{u}] \times [\text{inf ess } \underline{p}, \text{sup ess } \overline{p}]$.

Observe that if $C \equiv 0$, we have a system of cooperative type, whereas if $D \equiv 0$, we have a system of predator–prey type. For our purpose, the most interesting case is when C and D are both nonidentically zero as happens with system (3.29). For this kind of system, we show the next result whose proof may be carried out following the ideas contained in [4] and [7, Chapter V].

THEOREM 5.2. *Consider system* (5.38) *under the Hypotheses* H1 and H2. *Let us suppose that $\exists\, \underline{u}, \overline{u}, \underline{p}, \overline{p} \in H^1(\Omega) \cap L^\infty(\Omega)$, a system of upper–lower solutions for* (5.38).

Define by induction the sequences $\{u_n\},\{u^n\},\{p_n\},\{p^n\}$, *as*

$$u_1=\underline{u},\ u^1=\overline{u},\ p_1=\underline{p},\ p^1=\overline{p},$$

$$\begin{gathered}-\Delta u_n+Mu_n=B(x,u_{n-1},p_{n-1})+Mu_{n-1}\ in\ \Omega,\\ u_n=0\ on\ \partial\Omega,\end{gathered}$$

$$\begin{gathered}-\Delta u^n+Mu^n=B(x,u^{n-1},p^{n-1})+Mu^{n-1}\ in\ \Omega,\\ u^n=0\ on\ \partial\Omega,\end{gathered}$$

$$\begin{gathered}-\Delta p_n+Mp_n=C(x,u^{n-1},p_{n-1})+\frac{M}{2}p_{n-1}+D(x,u_{n-1},p_{n-1})+\frac{M}{2}p_{n-1}\ in\ \Omega,\\ p_n=0\ on\ \partial\Omega,\end{gathered}$$

$$\begin{gathered}-\Delta p^n+Mp^n=C(x,u_{n-1},p^{n-1})+\frac{M}{2}p^{n-1}+D(x,u^{n-1},p^{n-1})+\frac{M}{2}p^{n-1}\ in\ \Omega,\\ p^n=0\ on\ \partial\Omega.\end{gathered}$$

Then the sequence of functions defined above satisfy the order relation

$$\begin{gathered}u_1\le u_2\le\cdots\le u_n\le u^n\le u^{n-1}\le\cdots\le u^1,\\ p_1\le p_2\le\cdots\le p_n\le p^n\le p^{n-1}\le\cdots\le p^1\end{gathered}$$

for all $x\in\Omega$ *and*

$$u_n\nearrow u_*,\ u^n\searrow u^*,\ p_n\nearrow p_*,\ p^n\searrow p^*$$

(pointwise), where u_*,u^*,p_*,p^* *belong to* $C^{1,\alpha}(\Omega)$ *for any* $\alpha\in(0,1)$ *and satisfy the system*

$$\begin{gathered}-\Delta u_*=B(x,u_*,p_*)\ in\ \Omega,\\ -\Delta u^*=B(x,u^*,p^*)\ in\ \Omega,\\ -\Delta p_*=C(x,u^*,p_*)+D(x,u_*,p_*)\ in\ \Omega,\\ -\Delta p^*=C(x,u_*,p^*)+D(x,u^*,p^*)\ in\ \Omega,\\ u_*=u^*=p_*=p^*=0\ on\ \partial\Omega.\end{gathered}\tag{5.39}$$

Moreover, any solution (u,p) *of* (5.38) *with the property*

$$\underline{u}\le u\le\overline{u},\ \underline{p}\le p\le\overline{p},\tag{5.40}$$

must satisfy, for each $n\in\mathbb{N}$,

$$u_n\le u\le u^n,\ p_n\le p\le p^n,$$

and consequently

$$u_*\le u\le u^*,\ p_*\le p\le p^*.$$

Remark 5. Observe that (u^*,u_*,p^*,p_*) is also a solution of (5.39). So, if we are able to prove that there is only one solution (u,v,p,q) of (5.39) satisfying

$$\begin{gathered}\underline{u}\le u\le\overline{u},\ \underline{u}\le v\le\overline{u},\\ \underline{p}\le p\le\overline{p},\ \underline{p}\le q\le\overline{p},\end{gathered}\tag{5.41}$$

then we must have $u_* = u^*$, $p_* = p^*$, and therefore any solution of (5.38) with the property (5.40) must be (u_*, p_*). This is what happens in our optimal control problem as we are going to see next.

First, we assume all the assumptions of Theorem 4.2. Second, with the objective of applying Theorem 5.2 to system (3.29), we choose

$$\begin{aligned} B(x,u,p) &= u\left(a - \left[b + \frac{\lambda}{2}(1-p)\right]u\right), \\ C(x,u,p) &= p(a-2bu), \\ D(x,u,p) &= \frac{\lambda}{2}u(1-p)^2. \end{aligned}$$

Then Hypothesis H clearly implies Hypothesis H1. Also, a valid system of upper–lower solutions as in Definition 5.1 is

$$\underline{u} = w,\ \overline{u} = \frac{\overline{a}}{\underline{b}},\ \underline{p} = 0,\ \overline{p} = \lambda Q,$$

where w and Q are defined in (4.31) and (3.23), respectively.

On the other hand, Hypothesis H2 holds also for a convenient M, because the functions B, C, and D are of class C^1 with respect to (u,p). Consequently, Theorem 5.2 may be applied to system (3.29), obtaining

$$\begin{aligned} -\Delta u_* &= u_*\left(a - \left[b + \frac{\lambda}{2}\right]u_* + \frac{\lambda}{2}p_*u_*\right) \quad \text{in } \Omega, \\ -\Delta u^* &= u^*\left(a - \left[b + \frac{\lambda}{2}\right]u^* + \frac{\lambda}{2}p^*u^*\right) \quad \text{in } \Omega, \\ -\Delta p_* &= (a-2bu^*)p_* + \frac{\lambda}{2}u_*(1-p_*)^2 \quad \text{in } \Omega, \\ -\Delta p^* &= (a-2bu_*)p^* + \frac{\lambda}{2}u^*(1-p^*)^2 \quad \text{in } \Omega, \\ u_* &= u^* = p_* = p^* = 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{5.42}$$

It happens that for λ sufficiently small, the previous system has a unique solution (u,v,p,q) satisfying (5.41), so that the conclusions that were exposed in Remark 5 may be applied. In fact, we have the following theorem.

THEOREM 5.3. *As in Theorem* 4.2, *let us suppose* (3.13) *and take* ϵ *as in* (3.20). *Define* $\delta_0 \equiv \sigma_1(-a + 2b(u_0 - \epsilon))$. *Then if*

$$\lambda \le \Lambda_3 \equiv \min\left\{\Lambda_2, \frac{\delta_0}{2\overline{b}\|Q\|_\infty}, \frac{2\delta_0\underline{b}^2}{\underline{b}^2 + \overline{a}^2}\right\},$$

there can be only one solution (u,v,p,q) *of* (5.42) *satisfying* (5.41).

Proof. We use ideas similar to those of Theorem 4.2. To do so, let (u,v,p,q), (U,V,P,Q) be two solutions of (5.42) verifying (5.41). Then

$$-\Delta(u-U) - a(u-U) + \left(b + \frac{\lambda}{2}\right)(u^2 - U^2) - \frac{\lambda}{2}(pu^2 - PU^2) = 0$$

or, equivalently,

$$\begin{aligned} -\Delta(u-U) - a(u-U) + \left(b + \frac{\lambda}{2}\right)(u+U)(u-U) \\ -\frac{\lambda}{2}u^2(p-P) - \frac{\lambda}{2}P(u^2 - U^2) = 0. \end{aligned}$$

Multiplying the previous expression by $u-U$ and integrating on Ω, we have

$$\begin{aligned}0=\int_\Omega\Big[&|\nabla(u-U)|^2-a(u-U)^2+b(u+U)(u-U)^2\\&+\frac{\lambda}{2}(u+U)(u-U)^2-\frac{\lambda}{2}P(u+U)(u-U)^2\\&-\frac{\lambda}{2}u^2(p-P)(u-U)\Big].\end{aligned}$$

Analogously, we prove

$$\begin{aligned}0=\int_\Omega\Big[&|\nabla(v-V)|^2-a(v-V)^2+b(v+V)(v-V)^2\\&+\frac{\lambda}{2}(v+V)(v-V)^2-\frac{\lambda}{2}Q(v+V)(v-V)^2\\&-\frac{\lambda}{2}v^2(q-Q)(v-V)\Big].\end{aligned}$$

In the same way, we have

$$\begin{aligned}&-\Delta(p-P)-a(p-P)+2b(vp-VP)\\&-\frac{\lambda}{2}(u-U)+\lambda(up-UP)-\frac{\lambda}{2}(up^2-UP^2)=0\end{aligned}\tag{5.43}$$

or, equivalently,

$$\begin{aligned}&-\Delta(p-P)-a(p-P)+2bv(p-P)\\&+2bP(v-V)-\frac{\lambda}{2}(u-U)+\lambda u(p-P)\\&+\lambda P(u-U)-\frac{\lambda}{2}u(p+P)(p-P)-\frac{\lambda}{2}P^2(u-U)=0.\end{aligned}$$

Multiplying by $(p-P)$ and integrating on Ω, we have the result

$$\begin{aligned}0=\int_\Omega\Big[&|\nabla(p-P)|^2-a(p-P)^2+2bv(p-P)^2\\&+2bP(v-V)(p-P)-\frac{\lambda}{2}(u-U)(p-P)\\&+\lambda u(p-P)^2+\lambda P(u-U)(p-P)\\&-\frac{\lambda}{2}u(p+P)(p-P)^2-\frac{\lambda}{2}P^2(u-U)(p-P)\Big].\end{aligned}$$

Analogously,

$$\begin{aligned}0=\int_\Omega\Big[&|\nabla(q-Q)|^2-a(q-Q)^2+2bu(q-Q)^2\\&+2bQ(u-U)(q-Q)-\frac{\lambda}{2}(v-V)(q-Q)\\&+\lambda v(q-Q)^2+\lambda Q(v-V)(q-Q)\\&-\frac{\lambda}{2}v(q+Q)(q-Q)^2-\frac{\lambda}{2}Q^2(v-V)(q-Q)\Big].\end{aligned}$$

Now using relation (4.35) we deduce

$$
\begin{aligned}
0 \geq \int_\Omega \Big\{ & \delta_0[(u-U)^2 + (v-V)^2 + (p-P)^2 + (q-Q)^2] \\
& -\frac{\lambda}{2}[u^2 + (1-P)^2](p-P)(u-U) \\
& -\frac{\lambda}{2}[v^2 + (1-Q)^2](q-Q)(v-V) \\
& + 2bP(v-V)(p-P) + 2bQ(u-U)(q-Q) \Big\}.
\end{aligned} \tag{5.44}
$$

Moreover, from the hypotheses of Theorem 5.2, we obtain

$$
\left|\frac{\lambda}{2}[u^2 + (1-P)^2]\right| \leq \delta_0, \ \left|\frac{\lambda}{2}[v^2 + (1-Q)^2]\right| \leq \delta_0,
$$
$$
|2bP| \leq \delta_0, \ |2bQ| \leq \delta_0,
$$

so that the integral which appears in (5.44) is also nonnegative; it is strictly positive if $p(x) \neq P(x)$ or $q(x) \neq Q(x)$ on any subset of Ω with positive measure. Therefore, $p(x) = P(x)$ and $q(x) = Q(x)$ a.e. in Ω and, consequently, $u = U$ and $v = V$. □

The same ideas may be developed for the case considered in Corollary 4.4.

6. Final remarks. (1) It is possible to use different control spaces from that of $L^\infty_+(\Omega)$. For instance, such as is done in [9] for the case of Neumann boundary conditions, we may consider equation (1.1) and the benefit-cost functional (1.2) defined on the space of admissible controls $C_\rho = \{g \in L^\infty(\Omega) : 0 \leq g \leq \rho \text{ a.e. in } \Omega\}$, $\rho > 0$.

Trivially, Theorem 2.1 remains true, taking C_ρ instead of $L^\infty_+(\Omega)$. Also, it may be proved that

$$
\sup_{g \in C_\rho} J(g) > 0 \ \Leftrightarrow \ \sigma_1(-a) < 0.
$$

To do so, observe that if the profit is positive, as in section 2, $\sigma_1(-a) < 0$. Reciprocally, if $\sigma_1(-a) < 0$, then u_0 is strictly positive in Ω. Using the fact that the mapping $C_\rho \to C^1(\overline{\Omega})$, $f \to u_f$, is continuous [2], there exists some positive number ϵ, $\epsilon < \rho$, such that $u_\epsilon > 0$ in Ω and

$$
J(\epsilon) = \epsilon \int_\Omega (\lambda u_\epsilon - \epsilon|\Omega|) > 0.
$$

Last, from (2.6) we may prove the same results as in Theorems 4.2 and 5.3 concerning the uniqueness and approximation of the optimal control, respectively, if, in addition, we assume $0 < \lambda \frac{\overline{a}}{\underline{b}} < \rho$.

(2) In this remark we sketch some ideas which show the regularity of the functional J and the monotonicity of its Fréchet derivative. This may be a different proof of the fact that when the parameter λ is sufficiently small the optimal control is unique.

Let us begin with the following proposition.

PROPOSITION 6.1. *Let $E \subset L^\infty(\Omega)$ be the open subset defined as*

$$
E = \{f \in L^\infty(\Omega)/\sigma_1(-a+f) < 0\}. \tag{6.45}
$$

Then $J : E \to \mathbb{R}$, $f \to J(f)$, is Fréchet continuously differentiable and

$$J'(f)(g) = \int_\Omega (\lambda \xi_{f,g} f + \lambda u_f g - 2fg) \ \forall\ f \in E \ \forall\ g \in L^\infty(\Omega), \tag{6.46}$$

where $\xi_{f,g}$ is the function given in (3.10).

Proof. If $f \in E$ and $g \in L^\infty(\Omega)$, then, as in Lemma 3.2,

$$\frac{J(f+\beta g) - J(f)}{\beta} \ \to \int_\Omega (\lambda \xi_{f,g} f + \lambda u_f g - 2fg) \quad \text{as} \quad \beta \to 0.$$

If P_f is the function defined in (3.15), then

$$\int_\Omega f \xi_{f,g} + \int_\Omega g u_f P_f = 0,$$

so that

$$\frac{J(f+\beta g) - J(f)}{\beta} \ \to \int_\Omega (-\lambda g u_f P_f + \lambda u_f g - 2fg) \quad \text{as} \quad \beta \to 0.$$

Because of the regularity of the functions u_f and P_f, the mapping $L^\infty(\Omega) \to \mathbb{R}$, $g \to \int_\Omega (-\lambda g u_f P_f + \lambda u_f g - 2fg)$, is linear and continuous. This proves that J is Gateaux differentiable at any $f \in E$ and that

$$J'_G(f)(g) = \int_\Omega (-\lambda u_f P_f g + \lambda u_f g - 2fg) \ \forall\ g \in L^\infty(\Omega). \tag{6.47}$$

Moreover, if $f_n \to f \in E$, then

$$\begin{aligned} &\sup_{g \in B_{L^\infty(\Omega)}(0;1)} |J'_G(f_n)(g) - J'_G(f)(g)| \\ &\leq \sup_{g \in B_{L^\infty(\Omega)}(0;1)} \left| \int_\Omega \left[-\lambda (u_{f_n} P_{f_n} - u_f P_f) + \lambda (u_{f_n} - u_f) - 2(f_n - f) \right] g \right|, \end{aligned}$$

where $B_{L^\infty(\Omega)}(0;1)$ is the closed ball in $L^\infty(\Omega)$ of center 0 and radius 1. Since $f_n \to f$ in $L^\infty(\Omega)$, we have that $u_{f_n} \to u_f$ in $C^1(\overline{\Omega})$ [2] and, therefore, $P_{f_n} \to P_f$, in $C^1(\overline{\Omega})$. Consequently,

$$\sup_{g \in B_{L^\infty(\Omega)}(0;1)} |J'_G(f_n)(g) - J'_G(f)(g)| \to 0,$$

and J is Fréchet continuously differentiable in E. The next step to get the monotonicity of J' requires the parameter λ to be sufficiently small. If this is the case, then, under the hypotheses of the previous proposition,

$$\left[0, \lambda \frac{\overline{a}}{\underline{b}}\right]_{L^\infty(\Omega)} \subset E,$$

and

$$\begin{aligned} &(J'(f) - J'(g))(f-g) \\ &= \int_\Omega [(\lambda u_f (1 - P_f) - 2f)(f-g) - (\lambda u_g (1 - P_g) - 2g)(f-g)] \\ &\forall\ f, g \in \left[0, \lambda \frac{\overline{a}}{\underline{b}}\right]_{L^\infty(\Omega)}. \end{aligned} \tag{6.48}$$

By using similar arguments to those of Lemma 3.1, it is possible to prove that if λ is sufficiently small, the mappings $u, P : \left[0, \lambda\frac{\overline{a}}{\underline{b}}\right] \to L^\infty(\Omega)$ are Lipschitz continuous (to see this, it is sufficient to consider, instead of ξ_β, the function $\xi_h = \frac{u_{f+h}-u_f}{\|h\|_\infty}$, with $f \in \left[0, \lambda\frac{\overline{a}}{\underline{b}}\right]$ and h belonging to a convenient bounded subset of $L^\infty(\Omega)$, and then repeat the reasoning given there). Therefore, the mapping $\left[0, \lambda\frac{\overline{a}}{\underline{b}}\right]_{L^\infty(\Omega)} \to L^\infty(\Omega),\ f \to u_f(1 - P_f)$ is Lipschitz continuous with Lipschitz constant L. Then

$$(J'(f) - J'(g))(f - g) \le \int_\Omega \left[\lambda L(f-g)^2 - 2(f-g)^2\right], \tag{6.49}$$

which proves that if λ is sufficiently small, the optimal control is unique (see (2.6)).

(3) Finally, it is interesting to point out that, a priori, the payoff functional $J(f)$ could seem to be well defined for every $f \in L^2(\Omega)$. As far as we know, it is not possible to define properly the principal eigenvalue $\sigma_1(q)$ for every $q \in L^2(\Omega)$. Accordingly, we cannot define either the solution u_f or the functional $J(f)$ if $f \in L^2(\Omega)$ is not bounded. This is the reason we chose a control space contained in $L^\infty(\Omega)$. □

Acknowledgments. The authors wish to thank the referees for several important remarks and suggestions on the original manuscript.

REFERENCES

[1] H. Berestycki and P.L. Lions, *Some applications of the method of super and subsolutions*, Lecture Notes Math. 782, Springer-Verlag, New York, 1980, pp. 16–42.
[2] J. Blat and K.J. Brown, *Bifurcation of steady-state solutions in predator-prey and competition systems*, Proc. Roy. Soc. Edinburgh, 97 (1984), pp. 21–34.
[3] A. Cañada, J.L. Gámez, and J.A. Montero, *An optimal control problem for a nonlinear elliptic equation arising from population dynamics*, in Proceedings of the Second European Conference on Elliptic and Parabolic Problems, Pont-a-Mousson, France, 1994, Longman, Pitman Research Notes in Math. 326, C. Bandle, J. Bemelmans, M. Chipot, J. Saint Jean Paulin, and I. Shafrir, eds., Longman Scientific and Technical, Harlow, U.K., 1995, pp. 35–40.
[4] A. Cañada, P. Drábek, and J.L. Gámez, *Existence of positive solutions for some problems with nonlinear diffusion*, Trans. Amer. Math. Soc., 349 (1997), pp. 4231–4249.
[5] J.L. Gámez and J.A. Montero, *Uniqueness of the optimal control for a Lotka-Volterra control problem with a large crowding effect*, European Series in Applied and Industrial Mathematics: Control Optimization and Calculus of Variations, 2 (1997), pp. 1–12; also available online from http//www.emath.fr/COCV.
[6] P. Hess, *Periodic-Parabolic Boundary Value Problems and Positivity*, Longman Group U.K. Limited, London, UK, 1991.
[7] A. Leung *Systems of Nonlinear Partial Differential Equations*, Kluwer Academic Publishers, The Netherlands, 1989.
[8] A. Leung and S. Stojanovic, *Direct methods for some distributed games*, Differential Integral Equations, 3 (1990), pp. 1113–1125.
[9] A. Leung and S. Stojanovic, *Optimal control for elliptic Volterra-Lotka equations*, J. Math. Anal. Appl., 173 (1993), pp. 603–619.
[10] L. Li and R. Logan, *Positive solutions to general elliptic competition models*, Differential Integral Equations, 4 (1991), pp. 817–834.
[11] D.H. Sattinger, *Monotone methods in nonlinear elliptic and parabolic boundary value problems*, Indiana Univ. Math. J., 21 (1972), pp. 979–1000.
[12] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, 1983.
[13] S. Stojanovic, *Optimal damping control and nonlinear elliptic systems*, SIAM J. Control Optim., 29 (1991), pp. 594–608.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 4, pp. 1190–1206, July 1998

004

IDENTIFYING A SPATIAL BODY FORCE IN LINEAR ELASTODYNAMICS VIA TRACTION MEASUREMENTS*

MAURIZIO GRASSELLI† AND MASAHIRO YAMAMOTO‡

Abstract. An elastic and compressible material occupies a bounded domain $\Omega \subset \mathbf{R}^n$, $n \geq 2$, for any time $t \in [0,T]$, $T > 0$ being given. This material is subject to a body force $\mathbf{F} : [0,T] \times \Omega \to \mathbf{R}^n$ of the form $\mathbf{F}(t,x) := \varphi(t)\mathbf{f}(x)$, where $\mathbf{f} : \Omega \to \mathbf{R}^n$ is supposed to be unknown. The evolution of the displacement vector field is described by a Cauchy–Dirichlet problem for the linear elastodynamics system. We study the inverse problem of identifying $\mathbf{f}$ by measuring the traction $\mathbf{g}$ exerted on a portion Γ of the boundary $\partial\Omega$ over the time interval $[0,T]$. Using exact controllability methods, we show uniqueness and continuous dependence results. Also, we prove a representation formula for $\mathbf{f}$ in terms of $\mathbf{g}$.

Key words. inverse problems, exact controllability, linear elastodynamics

AMS subject classifications. 35R30, 35B37, 73C02

PII. S0363012996300288

1. Introduction.

Consider an elastic and compressible material which occupies, for any time $t \in [0,T]$, $T > 0$, a bounded domain Ω in $\mathbf{R}^3$, for instance. Set $Q_T := (0,T) \times \Omega$ and suppose that the material is subject to a body force field $\mathbf{F} : Q_T \to \mathbf{R}^3$ of the form

$$\mathbf{F}(t,x) := \varphi(t)\mathbf{f}(x), \qquad (t,x) \in Q_T, \tag{1.1}$$

where $\varphi : (0,T) \to \mathbf{R}$ and $\mathbf{f} := (f^1, f^2, f^3) : \Omega \to \mathbf{R}^3$.

Denote by $\mathbf{u} := (u^1, u^2, u^3) : Q_T \to \mathbf{R}^3$ the displacement with respect to the unstressed state. Then, in the linear setting, the evolution of $\mathbf{u}$ is described by the motion equation (see, e.g., [MH, Chap. 6, section 6.2])

$$\varrho \mathbf{u}_{tt} = \nabla \cdot \sigma(\mathbf{u}) + \mathbf{F} \qquad \text{in } Q_T, \tag{1.2}$$

where $\varrho : \overline{\Omega} \to (0, +\infty)$ denotes the medium density, $\nabla\cdot$ stands for the spatial divergence operator, and $\sigma(\mathbf{u}) := [\sigma_{ij}(\mathbf{u})]$, $i,j = 1,2,3$, represents the stress tensor defined by the constitutive law

$$\sigma(\mathbf{u}) = C : \varepsilon(\mathbf{u}). \tag{1.3}$$

Here $C := [C_{ijlk}]$ is the elasticity tensor, $\varepsilon(\mathbf{u}) := [\varepsilon_{lk}(\mathbf{u})]$ is the linear strain, where $\varepsilon_{lk}(\mathbf{u}) = \frac{1}{2}(u^l_{x_k} + u^k_{x_l})$, and the colon indicates the standard product between tensors of order $2m$ and tensors of order m for any $m \in \mathbf{N}$.

Let us associate with (1.2) a set of Cauchy and Dirichlet data, i.e.,

$$\mathbf{u}(0) = \mathbf{u}_0, \qquad \mathbf{u}_t(0) = \mathbf{u}_1 \qquad \text{in } \Omega, \tag{1.4}$$

$$\mathbf{u} = \mathbf{h} \qquad \text{on } (0,T) \times \partial\Omega, \tag{1.5}$$

*Received by the editors March 8, 1996; accepted for publication (in revised form) May 27, 1997; published electronically May 15, 1998.
http://www.siam.org/journals/sicon/36-4/30028.html

†Dipartimento di Matematica, Politecnico di Milano, Via E. Bonardi 9, 20133 Milano, Italy (maugra@mate.polimi.it).

‡Department of Mathematical Sciences, University of Tokyo, 3-8-1 Komaba Meguro, Tokyo 153, Japan (myama@ms.u-tokyo.ac.jp).

where $\mathbf{u}_0, \mathbf{u}_1 : \Omega \to \mathbf{R}^3$ and $\mathbf{h} : (0,T) \times \partial\Omega \to \mathbf{R}^3$ are prescribed functions. Consequently, $\mathbf{u}$ is uniquely determined by equation (1.2) and conditions (1.4)–(1.5), provided that the data are smooth enough (see, e.g., [MH, Chap. 6, section 6.2]). Nevertheless, here we assume that the spatial body force $\mathbf{f}$ is a priori unknown. We are thus led to consider the inverse problem of determining $\mathbf{f}$ (see also [BK]). Our main aim consists in showing that $\mathbf{f}$ can be uniquely identified by a boundary traction exerted on a portion Γ of the boundary $\partial\Omega$ over the time interval $[0,T]$, provided that both $T > 0$ and the two-dimensional Lebesgue measure of Γ, say *meas* Γ, are large enough. To be more precise, we suppose

$$\sigma(\mathbf{u}) : \nu = \mathbf{g} \qquad \text{on } (0,T) \times \Gamma, \tag{1.6}$$

where $\nu := (\nu^1, \nu^2, \nu^3) : \partial\Omega \to \mathbf{R}^3$ stands for the outward normal vector field to $\partial\Omega$, $\mathbf{g} := (g^1, g^2, g^3) : (0,T) \times \Gamma \to \mathbf{R}^3$ is the measured traction, and $\Gamma \subset \partial\Omega$ is prescribed. Then, our inverse problem can be formulated thus.

Problem (P_1). Find $\mathbf{f}$ from $\mathbf{g}$, provided that $\mathbf{u}$ fulfills (1.2) and (1.4)–(1.6).

Concerning (P_1), we are able to prove that, for T and *meas* Γ large enough,

(i) the mapping $\mathcal{G} : \mathbf{f} \mapsto \mathbf{g}$ has a Lipschitz continuous inverse;

(ii) $\mathbf{f}$ can be *reconstructed* by means of the eigenfunctions associated with the linear operator $-\varrho^{-1}\nabla \cdot \sigma(\cdot)$ which satisfy homogeneous Dirichlet boundary conditions;

(iii) the range of the adjoint operator $\mathcal{G}^*$ can be partially characterized.

All these results are obtained by extending the technique devised in [Y] to deal with the wave equation. This approach bears upon exact controllability methods presented in [L]. The plan goes as follows. In section 2, some basic preliminary results are introduced. Section 3 is devoted to stating our theorems rigorously. They are then proved in sections 4, 5, and 6. Finally, a proof of a basic technical lemma is given in section 7.

2. Some preliminary results. Here and below, $\Omega \subset \mathbf{R}^n$, $n \geq 2$, is an open, bounded, and connected subset with a boundary $\partial\Omega$ of class $\mathcal{C}^2$ (see Remark 2.1 below, however). Just for the sake of simplicity, we deal with an isotropic material. Therefore, relationship (1.3) becomes

$$\sigma(\mathbf{u}) = \lambda \operatorname{Tr} \varepsilon(\mathbf{u})\delta + 2\mu\varepsilon(\mathbf{u}), \tag{2.1}$$

where $\operatorname{Tr}\varepsilon$ denotes the trace of the linear strain tensor, while δ indicates the Kronecker tensor.

Concerning ϱ, λ, and μ, we assume

$$\varrho,\ \lambda,\ \mu \in C^1(\overline{\Omega}), \tag{2.2}$$

$$\varrho \geq \varrho_0,\ \mu \geq \mu_0,\ \lambda + \mu \geq \alpha_0 \quad \text{in } \overline{\Omega}, \tag{2.3}$$

where ϱ_0, μ_0, and α_0 are positive constants.

Also, we take (cf. (1.5))

$$\mathbf{h} \equiv \mathbf{0} \tag{2.4}$$

and set $H := (L^2(\Omega))^n$, $V := (H_0^1(\Omega))^n$, $V' := (H^{-1}(\Omega))^n$. From now on, $\langle \cdot, \cdot \rangle$ stands for the duality pairing between V' and V, while $(\cdot,\cdot)_X$, X being a real Hilbert space, indicates the *natural* inner product in X. If, in particular, $X = \mathbf{R}^n$, then we set $(\cdot,\cdot) := (\cdot,\cdot)_{\mathbf{R}^n}$. In addition, the symbol $\mathcal{D}(\Theta)$, $\Theta \subseteq \mathbf{R}^n$ being an open set, denotes the space of C^∞ functions having compact support in Θ and taking values in $\mathbf{R}^n$.

We first state a couple of results about the well-posedness of the Cauchy–Dirichlet problem for the elastodynamic system. Their proofs can be derived arguing, e.g., as in [L, Chap. I, sections 3 and 4] (see also [LLT, section 4]).

Let

$$\mathbf{F}\in L^1(0,T;H), \tag{2.5}$$

$$\mathbf{u}_0 \in V, \tag{2.6}$$

$$\mathbf{u}_1 \in H, \tag{2.7}$$

and consider the following problem.

Problem (P_2). Find $\mathbf{u}$ such that

$$\mathbf{u}\in C^1([0,T];H)\cap C^0([0,T];V), \tag{2.8}$$

$$\mathbf{u}_\nu := (\nabla\mathbf{u},\nu) \in L^2(0,T;L^2(\partial\Omega;\mathbf{R}^n)), \tag{2.9}$$

$$\varrho\mathbf{u}_{tt} = \nabla\cdot\sigma(\mathbf{u}) + \mathbf{F} \qquad \text{in } V', \text{ a.e. in } (0,T), \tag{2.10}$$

$$\mathbf{u}(0) = \mathbf{u}_0, \qquad \mathbf{u}_t(0) = \mathbf{u}_1 \qquad \text{a.e. in } \Omega. \tag{2.11}$$

We have the following theorem.

THEOREM 2.1. *Let the assumptions* (2.1)–(2.7) *hold. Then Problem* (P_2) *admits a unique solution. Moreover, there exists a positive constant* Λ_0 *such that*

$$\begin{aligned}&\|\mathbf{u}\|_{C^1([0,T];H)} + \|\mathbf{u}\|_{C^0([0,T];V)} + \|\mathbf{u}_\nu\|_{L^2(0,T;L^2(\partial\Omega;\mathbf{R}^n))}\\ &\le \Lambda_0\left\{\|\mathbf{F}\|_{L^1(0,T;H)} + \|\mathbf{u}_0\|_V + \|\mathbf{u}_1\|_H\right\}.\end{aligned} \tag{2.12}$$

Here Λ_0 *only depends on* Ω, T, ϱ, λ, *and* μ.

Strengthening the assumptions (2.5)–(2.7), i.e., assuming

$$\mathbf{F}\in W^{1,1}(0,T;H), \tag{2.13}$$

$$\mathbf{u}_0 \in V,\ \nabla\cdot\sigma(\mathbf{u}_0)\in H, \tag{2.14}$$

$$\mathbf{u}_1 \in V, \tag{2.15}$$

one gets from Theorem 2.1 the following regularity result (cf. also [Y, Appendix]).

THEOREM 2.2. *Let the assumptions* (2.1)–(2.4) *and* (2.13)–(2.15) *hold. Then the unique solution to Problem* (P_2) *satisfies*

$$\mathbf{u}\in C^2([0,T];H)\cap C^1([0,T];V), \tag{2.16}$$

$$\mathbf{u}_\nu \in H^1(0,T;L^2(\partial\Omega;\mathbf{R}^n)), \tag{2.17}$$

and therefore equation (2.10) *is fulfilled almost everywhere in* Q_T. *Also, one can find a positive constant* Λ_1 *such that*

$$\begin{aligned}&\|\mathbf{u}\|_{C^2([0,T];H)} + \|\mathbf{u}\|_{C^1([0,T];V)} + \|\mathbf{u}_\nu\|_{H^1(0,T;L^2(\partial\Omega;\mathbf{R}^n))}\\ &\le \Lambda_1\left\{\|\mathbf{F}\|_{W^{1,1}(0,T;H)} + \|\mathbf{u}_0\|_V + \|\nabla\cdot\sigma(\mathbf{u}_0)\|_H + \|\mathbf{u}_1\|_V\right\}.\end{aligned} \tag{2.18}$$

Here Λ_1 *only depends on* Ω, T, ϱ, λ, *and* μ.

We now state some results from exact controllability via the Hilbert uniqueness method (HUM) (cf., e.g., [L]). The first one is a technical lemma whose proof can be found in section 7. This lemma generalizes the inverse inequality proved in [L, Chap. IV, section 1].

Let $x^0\in\mathbf{R}^n$ and set (cf. also [L, Chap. I, section 5])

$$\mathbf{m}(x) := x - x^0 \qquad \forall\, x\in\mathbf{R}^n, \tag{2.19}$$

$$\Gamma^{+}(x^0) := \{x \in \partial\Omega \ : \ (\mathbf{m}(x), \nu(x)) > 0\}, \tag{2.20}$$

$$R(x^0) := \|\mathbf{m}\|_{(L^\infty(\Omega))^n}, \tag{2.21}$$

$$\Omega_{\varrho}^{-}(x^0) := \{x \in \Omega \ : \ (\nabla\varrho(x), \mathbf{m}(x)) < 0\}, \tag{2.22}$$

$$\Omega_{\lambda}^{+}(x^0) := \{x \in \Omega \ : \ (\nabla\lambda(x), \mathbf{m}(x)) > 0\}. \tag{2.23}$$

For the sake of simplicity, we assume that μ is constant, i.e.,

$$\mu \equiv \mu_0 \qquad \text{in } \overline{\Omega}. \tag{2.24}$$

However, the result we are going to state can also be proved when μ fulfills (2.2)–(2.3) and is not necessarily constant, provided that condition (2.26) below is suitably modified (cf. [I]; see also [TB]).

Besides, indicate by $d\Sigma$ the $(n-1)$-dimensional Lebesgue measure, while $|\cdot|$ denotes either the euclidean norm in $\mathbf{R}^n$ or the norm defined by

$$|A| := \left(\sum_{i,j=1}^{n} |a_{ij}|^2\right)^{1/2},$$

where $A := [a_{ij}]$ is any $n \times n$ real matrix.

One can prove the following lemma.

LEMMA 2.3. *Let the assumptions* (2.1)–(2.4) *and* (2.6)–(2.7*) hold. Moreover, assume*

$$\mathbf{F} \equiv \mathbf{0}, \tag{2.25}$$

$$R_0 := R(x^0)\left(\left\|\frac{\nabla\lambda}{\lambda}\right\|_{(L^\infty(\Omega_{\lambda}^{+}(x^0)))^n} + \left\|\frac{\nabla\varrho}{\varrho}\right\|_{(L^\infty(\Omega_{\varrho}^{-}(x^0)))^n}\right) < 1, \tag{2.26}$$

$$T > 2\gamma^{-1}(\varrho_0\mu_0)^{-1/2}R(x^0)\|\varrho\|_{L^\infty(\Omega)}, \tag{2.27}$$

where

$$\gamma := 1 - R_0. \tag{2.28}$$

Then

$$\begin{aligned} R\,(x^0)\int_{(0,T)\times\Gamma^{+}(x^0)} &\left[\mu|\nabla\mathbf{u}|^2 + (\lambda+\mu)|\nabla\cdot\mathbf{u}|^2\right] dtd\Sigma \\ &\geq 2\left(\gamma T - 2(\varrho_0\mu_0)^{-1/2}R(x^0)\|\varrho\|_{L^\infty(\Omega)}\right)E_0, \end{aligned} \tag{2.29}$$

where $\mathbf{u}$ *is the unique solution to* (P_2) *and*

$$E_0 := \frac{1}{2}\int_{\Omega}\left[\mu|\nabla\mathbf{u}_0|^2 + (\lambda+\mu)|\nabla\cdot\mathbf{u}_0|^2 + \varrho|\mathbf{u}_1|^2\right] dx. \tag{2.30}$$

Taking advantage of Theorem 2.1, we obtain an adapted version of [L, Chap. I, section 4, Théo. 4.2], as follows.

THEOREM 2.4. *Let the assumptions* (2.1)–(2.3) *hold. Then, for any* $\{\mathbf{z}_0, \mathbf{z}_1, \tilde{\mathbf{v}}\} \in H \times V' \times L^2(0,T;L^2(\partial\Omega;\mathbf{R}^n))$, *there exists a unique* $\mathbf{z}$ *such that*

$$\mathbf{z} \in C^1([0,T];V') \cap C^0([0,T];H), \tag{2.31}$$

$$\begin{aligned} &\int_{Q_T}(\varrho\mathbf{z},\mathbf{r})dtdx = -(\mathbf{z}_0,\mathbf{q}_t(0))_H + \langle\mathbf{z}_1,\mathbf{q}(0)\rangle \\ &-\int_{(0,T)\times\partial\Omega}(\tilde{\mathbf{v}},\sigma(\mathbf{q}):\nu)d\Sigma dt \qquad \forall\,\mathbf{r} \in C^0([0,T];(H^2(\Omega))^n \cap V), \end{aligned} \tag{2.32}$$

where $\mathbf{q} \in C^0([0,T];(H^2(\Omega))^n \cap V)$ *is the unique solution to*

$$\varrho \mathbf{q}_{tt} = \nabla \cdot \sigma(\mathbf{q}) + \mathbf{r} \qquad \text{a.e. in } Q_T \tag{2.33}$$

$$\mathbf{q}(T) = \mathbf{q}_t(T) = \mathbf{0} \qquad \text{a.e. in } \Omega \tag{2.34}$$

$$\mathbf{q} = \mathbf{0} \qquad \text{a.e. on } (0,T) \times \partial\Omega. \tag{2.35}$$

Moreover, there is a positive constant Λ_2 *such that*

$$\begin{aligned} &\| \mathbf{z}\|_{C^0([0,T];H)} + \|\mathbf{z}_t\|_{C^0([0,T];V')} \\ &\le \Lambda_2 \left\{ \|\mathbf{F}\|_{W^{1,1}(0,T;H)} + \|\mathbf{z}_0\|_H + \|\mathbf{z}_1\|_{V'} + \|\mathbf{v}\|_{L^2(0,T;L^2(\partial\Omega;\mathbf{R}^n))} \right\}. \end{aligned} \tag{2.36}$$

Here Λ_2 *only depends on* Ω, T, ϱ, λ, *and* μ.

The proof is as in [L, Chap. I, section 4, pp. 47–50].

Finally, on account of Lemma 2.3 and Theorem 2.4, we can deduce an extension of [L, Chap. IV, section 1, Théo. 1.1] as follows.

THEOREM 2.5. *Let the assumptions* (2.1)–(2.3), (2.24)–(2.27) *hold. Then, there exists a mapping* $\widetilde{\Pi} : H \times V' \to L^2(0,T;L^2(\Gamma^+(x_0);\mathbf{R}^n))$ *such that the unique solution* $\mathbf{z}$ *satisfying* (2.31)–(2.32) *with*

$$\mathbf{v} := \begin{cases} \widetilde{\Pi}(\mathbf{u}_0, \mathbf{u}_1) & \text{on } (0,T) \times \Gamma^+(x_0), \\ 0 & \text{on } (0,T) \times \partial\Omega \backslash \Gamma^+(x_0) \end{cases}$$

fulfills

$$\mathbf{z}(T) \equiv \mathbf{z}_t(T) \equiv \mathbf{0}. \tag{2.37}$$

Moreover, there holds

$$\|\widetilde{\Pi}(\mathbf{z}_0, \mathbf{z}_1)\|_{L^2(0,T;L^2(\Gamma^+(x_0);\mathbf{R}^n))} \le \Lambda_3 \left\{ \|\mathbf{z}_0\|_H + \|\mathbf{z}_1\|_{V'} \right\} \tag{2.38}$$

for some positive constant Λ_3 *which only depends on* Ω, x^0, T, ϱ, λ, *and* μ.

Proof. The existence of a control $\mathbf{v} \in L^2(0,T;L^2(\Gamma^+(x^0);\mathbf{R}^n))$ such that (2.37) holds can be obtained via the HUM method as in [L, Chap. IV, section 1, pp. 227–228], by using our Lemma 2.3, which generalizes the *inverse inequality* proved in [L, Chap. IV, section 1, pp. 225–227]. Estimate (2.38) derives from the construction of $\mathbf{v}$ (see [L, Chap. I, section 2 and Chap. II, section 2]). □

Remark 2.1. All the previous results still hold when Ω is open, bounded, and convex (see [L, section 6, Chap. I]).

Remark 2.2. It is worth observing that, in some cases, condition (2.26) is not so restrictive on λ and ϱ. Consider, for instance, a layered medium with respect to x_3, which occupies a cube in $\mathbf{R}^3$ (cf. Remark 2.1). Then λ and ϱ only depend on the *depth* variable x_3. If, e.g., λ is nondecreasing and ϱ is nonincreasing, then x^0 can be chosen in such a way that $\Omega_\lambda^+(x^0) = \Omega_\varrho^-(x^0) \equiv \emptyset$. Therefore, we have $\gamma = 1$ (cf. (2.26) and (2.28)).

3. Main results. Taking advantage of the linear setting, we make a further harmless simplification letting (cf. (1.4))

$$\mathbf{u}_0 = \mathbf{u}_1 \equiv \mathbf{0}. \tag{3.1}$$

Moreover, we suppose

$$\varphi \in C^1([0,T]). \tag{3.2}$$

Then, from Theorems 2.1 and 2.2, one deduces the following proposition.

PROPOSITION 3.1. *Let the assumptions* (2.1)–(2.4), (3.1)–(3.2) *hold. Then, for any* $\mathbf{f} \in H$ *there exists a unique* $\mathbf{u}$ *which fulfills* (2.16)–(2.17) *and solves*

$$\varrho \mathbf{u}_{tt} = \nabla \cdot \sigma(\mathbf{u}) + \varphi \mathbf{f} \qquad a.e.\ in\ Q_T, \tag{3.3}$$

$$\mathbf{u}(0) = \mathbf{u}_t(0) = \mathbf{0} \qquad a.e.\ in\ \Omega. \tag{3.4}$$

In addition, one has

$$\begin{aligned} &\| \mathbf{u}\|_{C^2([0,T];H)} + \|\mathbf{u}\|_{C^1([0,T];V)} + \|\mathbf{u}_\nu\|_{H^1(0,T;L^2(\partial\Omega;\mathbf{R}^n))} \\ &\le \Lambda_1 \|\varphi\|_{W^{1,1}(0,T)} \|\mathbf{f}\|_H. \end{aligned} \tag{3.5}$$

Consider $\Gamma \subseteq \partial\Omega$ such that $meas\,\Gamma > 0$. Thanks to Proposition 3.1, we can introduce a mapping $\mathcal{G} : H \to H^1(0,T;L^2(\Gamma;\mathbf{R}^n))$ by setting

$$\mathcal{G}(\mathbf{f}) := \sigma(\mathbf{u}(\mathbf{f})) : \nu \qquad \text{on } (0,T) \times \Gamma, \tag{3.6}$$

where $\mathbf{u}(\mathbf{f})$ is the unique solution to (3.3)–(3.4) given by Proposition 3.1. Note that, owing to (2.4),

$$u^i_{x_j} = u^i_\nu \nu_j, \qquad 1 \le i,j \le n. \tag{3.7}$$

Hence $\mathcal{G}$ is Lipschitz continuous because of (3.5).

Our first result concerns the invertibility of $\mathcal{G}$.

THEOREM 3.2. *Let the assumptions* (2.1)–(2.4), (2.24), (2.26)–(2.27), (3.1)–(3.2) *hold. Assume, moreover,*

$$\varphi(0) \neq 0, \tag{3.8}$$

$$\Gamma^+(x^0) \subseteq \Gamma. \tag{3.9}$$

Then, for any $\mathbf{g} \in H^1(0,T;L^2(\Gamma;\mathbf{R}^n))$, *there exists at most a unique* $\mathbf{f} \in H$ *such that* $\mathcal{G}(\mathbf{f}) = \mathbf{g}$. *Also,*

$$\|\mathbf{f}\|_H \le \Lambda_4 \|\mathbf{g}\|_{H^1(0,T;L^2(\Gamma;\mathbf{R}^n))}, \tag{3.10}$$

where Λ_4 *is a positive constant depending only on* $\Omega, x^0, T, \varrho, \lambda, \mu, \varphi$.

In other words, Theorem 2.1 gives some sufficient conditions which ensure the existence of $\mathcal{G}^{-1}$ and its (Lipschitz) continuity. This result extends [Y, Thm. 1].

We are going to state now a generalization of [Y, Thm. 2], i.e., a reconstruction formula for the unknown function $\mathbf{f}$ in terms of $\mathbf{g}$ and the eigenfunctions of the linear operator (cf. (2.1)–(2.3))

$$\mathcal{E}(\mathbf{u}) := -\varrho^{-1} \nabla \cdot \sigma(\mathbf{u}), \qquad \mathbf{u} \in D(\mathcal{E}) := \{\mathbf{w} \in V \ :\ \mathcal{E}(\mathbf{w}) \in H\}. \tag{3.11}$$

Note that, owing to (2.2)–(2.4) and by the regularity of $\partial\Omega$, we have $D(\mathcal{E}) \equiv (H^2(\Omega))^n \cap V$.

To state our second result we need some preliminary considerations. Taking advantage of Theorem 2.4, one can construct a mapping $\Pi : H \to L^2(0,T;L^2(\Gamma;\mathbf{R}^n))$ by setting (see [Y])

$$\mathbf{v} = \Pi(\mathbf{z}_0) := \widetilde{\Pi}(\mathbf{z}_0, \mathbf{0}) \qquad \text{on } (0,T) \times \Gamma \tag{3.12}$$

for any $\mathbf{z}_0 \in H$.

Also, we introduce the linear bounded operator (see [Y]) $\Phi : L^2(0,T;L^2(\Gamma;\mathbf{R}^n)) \to H^1(0,T;L^2(\Gamma;\mathbf{R}^n))$ defined as

$$\xi := \Phi(\eta), \tag{3.13}$$

where ξ is the unique solution in $H^1(0,T;L^2(\Gamma;\mathbf{R}^n))$ to the linear system of Volterra integrodifferential equations

$$\varphi(0)\xi_t(t,\cdot) + \int_t^T [\varphi'(s-t)\xi_t(s,\cdot) + \varphi(s-t)\xi(s,\cdot)]\,ds = \eta(t,\cdot) \quad \text{a.e. in } (0,T)\times\Gamma \tag{3.14}$$

such that $\xi(0,\cdot) = \mathbf{0}$ almost everywhere on Γ.

Before stating the reconstruction formula, it is worth noting that the linear operator $\mathcal{E} : D(\mathcal{E}) \to H$ (see (3.11)) is V-coercive and self-adjoint with respect to the measure $\varrho(x)dx$; then its eigenvalues are all positive with finite multiplicity and they form a nondecreasing sequence $\{\lambda_k\}$, $k \in \mathbf{N}$, where any λ_k appears l times, l being its multiplicity. Besides, we recall that the corresponding eigenfunctions $\{\mathbf{w}_k\}$ can be chosen in order to constitute an orthonormal basis of $H_\varrho := (L^2(\Omega;\varrho(x)dx))^n$, that is,

$$(\mathbf{w}_k,\mathbf{w}_h)_{H_\varrho} := \int_\Omega (\mathbf{w}_k(x),\mathbf{w}_h(x))\varrho(x)dx = \delta_{kh}.$$

We have the following theorem (cf. [Y, Thm. 2]).

THEOREM 3.3. *Let the assumptions* (2.1)–(2.4), (2.24), (2.26)–(2.27), (3.1)–(3.2), *and* (3.8)–(3.9) *hold. Moreover, let*

$$\mathbf{g} \in H^1(0,T;L^2(\Gamma;\mathbf{R}^n)) \tag{3.15}$$

and set

$$\xi_k := -\Phi(\Pi(\mathbf{w}_k)) \qquad \forall\, k \in \mathbf{N}. \tag{3.16}$$

If there exists $\mathbf{f} \in H$ *such that* $\mathcal{G}(\mathbf{f}) = \mathbf{g}$, *then*

$$(\mathbf{f},\mathbf{w}_k)_{H_\varrho} = (\mathbf{g},\xi_k)_{H^1(0,T;L^2(\Gamma;\mathbf{R}^n))} \qquad \forall\, k \in \mathbf{N}. \tag{3.17}$$

Therefore,

$$\mathbf{f} = \sum_{h=1}^{\infty} (\mathbf{g},\xi_k)_{H^1(0,T;L^2(\Gamma;\mathbf{R}^n))}\mathbf{w}_k. \tag{3.18}$$

The final result concerns the characterization of the range $\mathcal{R}(\mathcal{G}^*)$ of the adjoint operator $\mathcal{G}^*$, when $\mathcal{G}$ is regarded as a linear operator from H to $L^2(0,T;L^2(\Gamma;\mathbf{R}^n))$. Indeed, we prove the following theorem (cf. [Y, Thm. 3]).

THEOREM 3.4. *Let the assumptions* (2.1)–(2.4), (2.24), (2.26)–(2.27), (3.1)–(3.2), *and* (3.8)–(3.9) *hold. Then*

$$V \subset \mathcal{R}(\mathcal{G}^*) := \left\{\mathcal{G}^*(\mathbf{g}), \quad \mathbf{g} \in L^2(0,T;L^2(\Gamma;\mathbf{R}^n))\right\} \subset (H^{1/2}(\Omega))^n. \tag{3.19}$$

Remark 3.1. Theorem 3.3 improves [Y, Thm. 3] and it turns out to be useful to construct a Tikhonov regularization procedure for determining *reasonable* approximations of $\mathbf{f}$ which converge to $\mathbf{f}$ in H (see [Y, section 3]).

4. Proof of Theorem 3.2. Suppose there is an $\mathbf{f} \in H$ corresponding to $\mathbf{g} \in H^1(0,T;L^2(\partial\Omega;\mathbf{R}^n))$ (cf. (3.6)).

In Problem (P_2) take

$$\mathbf{F} \equiv \mathbf{0}, \tag{4.1}$$

$$\mathbf{u}_0 \equiv \mathbf{0}, \tag{4.2}$$

$$\mathbf{u}_1 \equiv \mathbf{f}. \tag{4.3}$$

Then, by Theorem 2.1, we infer that (P_2) admits a unique solution $\mathbf{w}$. Moreover, by Lemma 2.3 and (3.7), we have

$$\|\mathbf{f}\|_H \leq \Lambda_5 \|\mathbf{w}_\nu\|_{L^2(0,T;L^2(\Gamma;\mathbf{R}^n))} \tag{4.4}$$

for some positive constant Λ_5 depending on Ω, x^0, T, ϱ, λ, μ.

Now set

$$\tilde{\mathbf{u}} := \varphi * \mathbf{w} \qquad \text{in } Q_T, \tag{4.5}$$

where $*$ stands for the usual time convolution product over the interval $(0,t)$. Thanks to (3.2), one has

$$\tilde{\mathbf{u}} \in C^2([0,T];H) \cap C^1([0,T];V). \tag{4.6}$$

Besides, one can easily check that $\tilde{\mathbf{u}}$ satisfies (2.16)–(2.17) and solves (3.3)–(3.4). Therefore, Proposition 3.1 implies $\tilde{\mathbf{u}} \equiv \mathbf{u}$.

Observe that (cf. (3.2) and (4.5))

$$\mathbf{u}_t = \varphi(0)\mathbf{w} + \varphi' * \mathbf{w} \qquad \text{in } Q_T. \tag{4.7}$$

Then, owing to (3.8), standard arguments (see [Y, section 4]) allow us to deduce from (4.7)

$$\|\mathbf{w}_\nu\|_{L^2(0,T;L^2(\Gamma;\mathbf{R}^n))} \leq \Lambda_6 \|(\mathbf{u}_t)_\nu\|_{L^2(0,T;L^2(\Gamma;\mathbf{R}^n))}, \tag{4.8}$$

where Λ_6 is a positive constant which only depends on $\varphi(0)$ and $\|\varphi'\|_{C^0([0,T])}$.

Combining (4.4) and (4.8), one gets

$$\|\mathbf{f}\|_H \leq \Lambda_7 \|(\mathbf{u}_t)_\nu\|_{L^2(0,T;L^2(\Gamma;\mathbf{R}^n))}, \tag{4.9}$$

where $\Lambda_7 := \Lambda_5\Lambda_6$.

On the other hand, recalling (2.1) and [L, Chap. IV, section 2, eq. (2.2)], from (3.6) and (3.7) we deduce the bound

$$\|\mathbf{u}_\nu\|_{H^1(0,T;L^2(\Gamma;\mathbf{R}^n))} \leq \Lambda_8 \|\mathbf{g}\|_{H^1(0,T;L^2(\Gamma;\mathbf{R}^n))} \tag{4.10}$$

for some positive constant Λ_8 depending only on λ and μ.

Finally, a combination of (4.9) and (4.10) yields estimate (3.10) which, in particular, implies the uniqueness of $\mathbf{f}$ in H.

5. Proof of Theorem 3.3. The argument is based on an identity for the eigenfunctions $\mathbf{w}_k$ of the linear operator $\mathcal{E}$ (cf. (3.11)). To be more precise, we have the following lemma (see [Y, section 5]).

LEMMA 5.1. *Let the assumptions* (2.1)–(2.4), (2.24), (2.26)–(2.27), *and* (3.9) *hold. Then*

$$\left(\lambda_k^{-1/2} \sin(\sqrt{\lambda_k}t)(\sigma(\mathbf{w}_k):\nu), -\Pi(\mathbf{w}_h)\right)_{L^2(0,T;L^2(\Gamma;\mathbf{R}^n))} = \delta_{kh} \tag{5.1}$$

for any k, $h \in \mathbf{N}$.

Let us postpone the proof of Lemma 5.1.

Consider the linear operator $K : L^2(0,T;L^2(\Gamma;\mathbf{R}^n)) \to H^1(0,T;L^2(\Gamma;\mathbf{R}^n))$ defined by (cf. also (3.2) and (4.5))

$$K(\xi) := \varphi * \xi. \tag{5.2}$$

One can easily check that the adjoint operator $K^* : \mathcal{R}(K) \subseteq H^1(0,T;L^2(\Gamma;\mathbf{R}^n)) \to L^2(0,T;L^2(\Gamma;\mathbf{R}^n))$ is given by (cf. (3.14))

$$K^*(\xi)(t,\cdot) = \varphi(0)\xi_t(t,\cdot) + \int_t^T [\varphi'(s-t)\xi_t(s,\cdot) + \varphi(s-t)\xi(s,\cdot)]ds \tag{5.3}$$

almost everywhere in $(0,T)\times\Gamma$.

Hence, recalling the linear operator Φ (see (3.13)–(3.14)), one has

$$K^*(\Phi(\eta)) = \eta \tag{5.4}$$

for any $\eta \in L^2(0,T;L^2(\Gamma;\mathbf{R}^n))$.

On account of (1.6), (3.15)–(3.16), (4.5), and (5.4), we obtain

$$\begin{aligned}(\mathbf{g},\xi_h)_{H^1(0,T;L^2(\Gamma;\mathbf{R}^n))} &= (\sigma(\mathbf{u}):\nu,\xi_h)_{H^1(0,T;L^2(\Gamma;\mathbf{R}^n))}\\ &= (K(\sigma(\mathbf{w}):\nu),\Phi(-\Pi(\mathbf{w}_h)))_{H^1(0,T;L^2(\Gamma;\mathbf{R}^n))}\\ &= (\sigma(\mathbf{w}):\nu,K^*(\Phi(-\Pi(\mathbf{w}_h))))_{L^2(0,T;L^2(\Gamma;\mathbf{R}^n))}\\ &= (\sigma(\mathbf{w}):\nu,-\Pi(\mathbf{w}_h))_{L^2(0,T;L^2(\Gamma;\mathbf{R}^n))}\end{aligned} \tag{5.5}$$

for any $h \in \mathbf{N}$.

Observe that (cf. (4.1)–(4.3))

$$\mathbf{w} = \sum_{k=1}^{\infty} (\mathbf{f},\mathbf{w}_k)_{H_\varrho}\lambda_k^{-1/2}\sin(\sqrt{\lambda_k}t)\mathbf{w}_k. \tag{5.6}$$

Then, recalling (2.1) and taking advantage of (3.7), from (5.6) we get

$$\sigma(\mathbf{w}):\nu = \sum_{k=1}^{\infty} (\mathbf{f},\mathbf{w}_k)_{H_\varrho}\lambda_k^{-1/2}\sin(\sqrt{\lambda_k}t)\,(\sigma(\mathbf{w}_k):\nu)\,, \tag{5.7}$$

where the series converges in $L^2(0,T;L^2(\Gamma;\mathbf{R}^n))$.

Thanks to (5.1) and (5.7), we derive

$$\begin{aligned}&(\ \sigma(\mathbf{w}):\nu,-\Pi(\mathbf{w}_h))_{L^2(0,T;L^2(\Gamma;\mathbf{R}^n))}\\ &= \sum_{k=1}^{\infty} (\mathbf{f},\mathbf{w}_k)_{H_\varrho}(\lambda_k^{-1/2}\sin(\sqrt{\lambda_k}t)(\sigma(\mathbf{w}_k):\nu),-\Pi(\mathbf{w}_h))_{L^2(0,T;L^2(\Gamma;\mathbf{R}^n))}\\ &= (\mathbf{f},\mathbf{w}_h)_{H_\varrho} \qquad\qquad \forall\, h \in \mathbf{N}.\end{aligned} \tag{5.8}$$

Finally, (5.5) and (5.8) yield (3.17) and, consequently, (3.18).

Proof of Lemma 5.1. Let us observe that, thanks to Theorem 2.4, for any given $\tilde{\mathbf{v}} \in L^2(0,T;L^2(\Gamma;\mathbf{R}^n))$, we can find a unique function $\tilde{\mathbf{z}}$ satisfying (2.31), (2.37), and (cf. (2.32)–(2.36))

$$\int_{Q_T} (\varrho\tilde{\mathbf{z}},\mathbf{r})dxdt = -\int_{(0,T)\times\Gamma} (\tilde{\mathbf{v}},\sigma(\tilde{\mathbf{q}}):\nu)d\Sigma dt \qquad \forall\ \mathbf{r} \in C^0([0,T];D(\mathcal{E})), \tag{5.9}$$

where $\tilde{\mathbf{q}} \in C^0([0,T]; D(\mathcal{E}))$ is the unique solution to

$$\varrho \tilde{\mathbf{q}}_{tt} = \nabla \cdot \sigma(\tilde{\mathbf{q}}) + \mathbf{r} \qquad \text{a.e. in } Q_T, \tag{5.10}$$

$$\tilde{\mathbf{q}}(0) = \tilde{\mathbf{q}}_t(0) = \mathbf{0} \qquad \text{a.e. in } \Omega, \tag{5.11}$$

$$\tilde{\mathbf{q}} = \mathbf{0} \qquad \text{a.e. on } (0,T) \times \partial\Omega. \tag{5.12}$$

Note that the roles of 0 and T are exchanged with respect to Theorem 2.4.

Then, provided that $\tilde{\mathbf{z}}$ is smooth enough and recalling that (cf. (3.11)) $\nabla \cdot \sigma(\mathbf{w}_k) = -\lambda_k \varrho \mathbf{w}_k$, we obtain the chain of equalities

$$\begin{aligned}
&\int_0^T \lambda_k^{-1/2} \sin(\sqrt{\lambda_k} t)(\mathbf{w}_k, \nabla \cdot \sigma(\tilde{\mathbf{z}}))_H dt \\
&= \int_0^T \lambda_k^{-1/2} \sin(\sqrt{\lambda_k} t)(\mathbf{w}_k, \tilde{\mathbf{z}}_{tt})_{H_\varrho} dt \\
&= \left(\mathbf{w}_k, \int_0^T \lambda_k^{-1/2} \sin(\sqrt{\lambda_k} t) \tilde{\mathbf{z}}_{tt} dt \right)_{H_\varrho} \\
&= \left(\mathbf{w}_k, -\int_0^T \cos(\sqrt{\lambda_k} t) \tilde{\mathbf{z}}_t dt \right)_{H_\varrho} \\
&= \left(\mathbf{w}_k, \tilde{\mathbf{z}}(0) - \lambda_k \int_0^T \lambda_k^{-1/2} \sin(\sqrt{\lambda_k} t) \tilde{\mathbf{z}} dt \right)_{H_\varrho} \\
&= (\mathbf{w}_k, \tilde{\mathbf{z}}(0))_{H_\varrho} + \int_0^T \lambda_k^{-1/2} \sin(\sqrt{\lambda_k} t)(\nabla \cdot \sigma(\mathbf{w}_k), \tilde{\mathbf{z}})_H dt.
\end{aligned} \tag{5.13}$$

From (5.13), we deduce the identity

$$\int_0^T \lambda_k^{-1/2} \sin(\sqrt{\lambda_k} t) \left[(\mathbf{w}_k, \nabla \cdot \sigma(\tilde{\mathbf{z}}))_H - (\tilde{\mathbf{z}}, \nabla \cdot \sigma(\mathbf{w}_k))_H \right] dt = (\mathbf{w}_k, \tilde{\mathbf{z}}(0))_{H_\varrho}. \tag{5.14}$$

Recalling Theorem 2.5 and (3.12), we can choose $\mathbf{z}_0 = \mathbf{w}_h$, $\mathbf{z}_1 = \mathbf{0}$, $\tilde{\mathbf{v}} = \Pi(\mathbf{w}_h)$. Then, denoting by $\tilde{\mathbf{z}}$ the solution to (2.32) associated with $\{\mathbf{z}_0, \mathbf{z}_1, \tilde{\mathbf{v}}\}$, we infer that $\tilde{\mathbf{z}}$ solves (5.9) as well. Hence identity (5.14) and Green's formula entail (5.1).

The present argument can be made rigorous by means of a standard approximation procedure based on well-known density results (cf. [Y, section 5]). □

6. Proof of Theorem 3.4. Let us consider the linear operator $\mathcal{F} : H \to L^2(0,T; L^2(\Gamma; \mathbf{R}^n))$ defined by

$$\mathcal{F}(\mathbf{f}) := \sigma(\mathbf{w}) : \nu, \tag{6.1}$$

where $\mathbf{w}$ is the unique solution to Problem (P_2) in the assumptions (4.1)–(4.3). Note that $\mathcal{F}$ is bounded owing to Theorem 2.1 and (3.7).

Recalling (3.6), (4.5), and (5.2), one realizes that

$$\mathcal{G}(\mathbf{f}) = K(\mathcal{F}(\mathbf{f})) \tag{6.2}$$

regarding K as a linear operator from $L^2(0,T; L^2(\Gamma; \mathbf{R}^n))$ into itself. Hence, from (6.2) it follows that

$$\mathcal{G}^* = \mathcal{F}^* K^* \tag{6.3}$$

so that

$$\mathcal{R}(\mathcal{G}^*) \equiv \{\mathcal{F}^*(\xi), \quad \xi \in \mathcal{R}(K^*)\}. \tag{6.4}$$

On the other hand, as $K : L^2(0,T;L^2(\Gamma;\mathbf{R}^n)) \to L^2(0,T;L^2(\Gamma;\mathbf{R}^n))$, we have

$$K^*(\eta)(t,\cdot) = \int_t^T \varphi(s-t)\eta(s,\cdot)ds \qquad \text{a.e. on } (0,T)\times\Gamma. \tag{6.5}$$

Therefore (cf. also (3.2)),

$$\mathcal{R}(K^*) \equiv \left\{\xi \in H^1(0,T;L^2(\Gamma;\mathbf{R}^n)) : \ \xi(T,\cdot) = \mathbf{0}\right\}. \tag{6.6}$$

Consequently, combining (6.4) and (6.6), we get

$$\mathcal{R}(\mathcal{G}^*) \equiv \left\{\mathcal{F}^*(\xi), \quad \xi \in H^1(0,T;L^2(\Gamma;\mathbf{R}^n)) : \ \xi(T,\cdot) = \mathbf{0}\right\}. \tag{6.7}$$

Let $\mathbf{v} \in L^2(0,T;L^2(\Gamma;\mathbf{R}^n))$ and consider the function $\tilde{\mathbf{z}} = \tilde{\mathbf{z}}(\mathbf{v})$ solving (5.9). Observe that, by formal computations (cf. (5.10)),

$$\begin{aligned}
\int_0^T (\nabla\cdot\sigma(\tilde{\mathbf{z}}),\mathbf{w})_H dt &= \int_0^T (\tilde{\mathbf{z}}_{tt},\mathbf{w})_{H_\varrho} dt \\
&= -\int_0^T (\tilde{\mathbf{z}}_t,\mathbf{w}_t)_{H_\varrho} dt \\
&= \int_0^T (\tilde{\mathbf{z}},\mathbf{w}_{tt})_{H_\varrho} dt + (\tilde{\mathbf{z}}(0),\mathbf{f})_{H_\varrho} \\
&= \int_0^T (\tilde{\mathbf{z}},\nabla\cdot\sigma(\mathbf{w}))_H dt + (\tilde{\mathbf{z}}(0),\mathbf{f})_{H_\varrho}.
\end{aligned} \tag{6.8}$$

Of course, as in the previous proof of Lemma 5.1, a suitable approximation argument is needed to make the above computations rigorous.

The chain of equalities (6.8) implies

$$\int_0^T \left[(\nabla\cdot\sigma(\tilde{\mathbf{z}}),\mathbf{w})_H - (\tilde{\mathbf{z}},\nabla\cdot\sigma(\mathbf{w}))_H\right] dt = (\tilde{\mathbf{z}}(0),\mathbf{f})_{H_\varrho}, \tag{6.9}$$

and via Green's formula, one derives (cf. also (6.1))

$$-(\mathbf{v},\mathcal{F}(\mathbf{f}))_{L^2(0,T;L^2(\Gamma;\mathbf{R}^n))} = (\tilde{\mathbf{z}}(0),\mathbf{f})_{H_\varrho} \tag{6.10}$$

for any $\mathbf{v} \in L^2(0,T;L^2(\Gamma;\mathbf{R}^n))$.

Identity (6.10) entails

$$\mathcal{F}^*(\mathbf{v}) = -\varrho\tilde{\mathbf{z}}(0). \tag{6.11}$$

Hence, a combination of (6.7) and (6.11) yields

$$\mathcal{R}(\mathcal{G}^*) \equiv \left\{\varrho\tilde{\mathbf{z}}(\mathbf{v})(0), \quad \mathbf{v} \in H^1(0,T;L^2(\Gamma;\mathbf{R}^n)) : \ \mathbf{v}(T,\cdot) = \mathbf{0} \text{ on } \Gamma\right\}. \tag{6.12}$$

The proof is now given by Lemma 6.1.

LEMMA 6.1. *Let the assumptions* (2.1)–(2.4), (2.24), (2.26)–(2.27), *and* (3.9) *hold. Then*

$$V \subset \left\{\varrho\tilde{\mathbf{z}}(\mathbf{v})(0), \quad \mathbf{v} \in H^1(0,T;L^2(\Gamma;\mathbf{R}^n)) : \ \mathbf{v}(T,\cdot) = \mathbf{0} \text{ on } \Gamma\right\} \subset (H^{1/2}(\Omega))^n. \tag{6.13}$$

Proof of Lemma 6.1. Let

$$ {}^0H^1(0,T;L^2(\Gamma;\mathbf{R}^n)) = \left\{ \mathbf{w} \in H^1(0,T;L^2(\Gamma;\mathbf{R}^n)) : \ \mathbf{w}(T,\cdot) = \mathbf{0} \text{ on } \Gamma \ \right\} $$

be a Hilbert space with the norm

$$ \|\mathbf{w}\|_{{}^0H^1} := \left(\int_0^T \int_\Gamma |\mathbf{w}_t|^2 d\Sigma dt \right)^{1/2}. \tag{6.14} $$

It is straightforward to observe that $\|\mathbf{w}\|_{{}^0H^1}$ is equivalent to

$$ \|\mathbf{w}\|_{H^1(0,T;L^2(\Gamma;\mathbf{R}^n))} := \left(\int_0^T \int_\Gamma |\mathbf{w}|^2 + |\mathbf{w}_t|^2 d\Sigma dt \right)^{1/2} $$

for any $\mathbf{w} \in {}^0H^1(0,T;L^2(\Gamma;\mathbf{R}^n))$. Denoting by X_1 the dual of ${}^0H^1(0,T;L^2(\Gamma;\mathbf{R}^n))$, we also have

$$ {}^0H^1(0,T;L^2(\Gamma;\mathbf{R}^n)) \hookrightarrow L^2(\Gamma\times(0,T);\mathbf{R}^n) \hookrightarrow X_1 $$

with dense injections, the dual of $L^2((0,T)\times\Gamma;\mathbf{R}^n)$ being identified with itself.

Indicating now by ${}_{X_1}\langle\cdot,\cdot\rangle_{{}^0H^1}$ the duality pairing between ${}^0H^1(0,T;L^2(\Gamma;\mathbf{R}^n))$ and X_1, we note that

$$ {}_{X_1}\langle \mathbf{v}, \mathbf{w}\rangle_{{}^0H^1} = (\mathbf{v},\mathbf{w})_{L^2((0,T)\times\Gamma;\mathbf{R}^n)} \tag{6.15} $$

for any $\mathbf{v} \in L^2((0,T)\times\Gamma;\mathbf{R}^n)$ and any $\mathbf{w} \in^0 H^1(0,T;L^2(\Gamma;\mathbf{R}^n))$.

On account of Theorem 2.4, take

$$ \mathbf{z}_0 \in H, \quad \mathbf{z}_1 \in V', \quad \tilde{\mathbf{v}} \equiv \mathbf{0}. \tag{6.16} $$

Then, Theorem 2.4 ensures the existence of a unique $\mathbf{z}$ satisfying (2.31)–(2.32).

Assume, for the moment,

$$ \mathbf{z}_0 \in \mathcal{D}(\Omega), \quad \mathbf{z}_1 \in \mathcal{J}(\Omega), \tag{6.17} $$

where

$$ \mathcal{J}(\Omega) := \{ \mathbf{z} = \mathcal{E}(\mathbf{w}) \quad \text{for some } \mathbf{w} \in \mathcal{D}(\Omega) \} \subset C_c^1(\Omega) $$

and $C_c^1(\Omega)$ indicates the space of C^1 functions taking values in $\mathbf{R}^n$ and having compact support in Ω.

Consider Problem (P_2) and take (cf. also (2.2) and (3.11))

$$ \mathbf{F} \equiv \mathbf{0}, \quad \mathbf{u}_0 := -\mathcal{E}^{-1}(\mathbf{z}_1) \in C_c^1(\Omega), \quad \mathbf{u}_1 := \mathbf{z}_0 \in \mathcal{D}(\Omega). \tag{6.18} $$

Consequently, Theorem 2.1 applies and Problem (P_2) admits a unique solution $\mathbf{u}$ fulfilling (2.8)–(2.12). Moreover, taking advantage of (2.12) and Lemma 2.3, one can find a pair of positive constants Λ_9, Λ_{10} such that (cf. also (1.6), (3.7), and (4.10))

$$ \begin{aligned} \Lambda_9 \, \|\sigma(\mathbf{u}) : \nu\|_{L^2((0,T)\times\Gamma;\mathbf{R}^n)} &\le \left(\|\mathbf{u}_0\|_V^2 + \|\mathbf{u}_1\|_H^2 \right)^{1/2} \\ &\le \Lambda_{10} \, \|\sigma(\mathbf{u}) : \nu\|_{L^2((0,T)\times\Gamma;\mathbf{R}^n)}. \end{aligned} \tag{6.19} $$

Here Λ_9 and Λ_{10} depend only on Ω, x^0, T, ϱ, λ, and μ.

A straightforward argument shows that

$$\sigma(\mathbf{z}) : \nu = (\sigma(\mathbf{u}) : \nu)_t \tag{6.20}$$

in the sense of distributions.

Let us prove that $\sigma(\mathbf{z}) : \nu \in X_1$. Indeed, by virtue of (6.18), we have (cf. also (6.15) and (6.20))

$$\begin{aligned} &{}_{X_1}\langle \sigma(\mathbf{z}) : \nu, \mathbf{w} \rangle_{{}_0H_1} \\ &= (\sigma(\mathbf{z}) : \nu, \mathbf{w})_{L^2((0,T)\times\Gamma;\mathbf{R}^n)} = ((\sigma(\mathbf{u}) : \nu)_t, \mathbf{w})_{L^2((0,T)\times\Gamma;\mathbf{R}^n)} \\ &= \left[\int_\Gamma ((\sigma(\mathbf{u}) : \nu)(t), \mathbf{w}(t)) d\Sigma \right]_0^T - (\sigma(\mathbf{u}) : \nu, \mathbf{w}_t)_{L^2((0,T)\times\Gamma;\mathbf{R}^n)} \\ &= - (\sigma(\mathbf{u}) : \nu, \mathbf{w}_t)_{L^2((0,T)\times\Gamma;\mathbf{R}^n)} \end{aligned} \tag{6.21}$$

for any $\mathbf{w} \in {}^0H^1(0,T;L^2(\Gamma;\mathbf{R}^n))$.

Therefore, we deduce

$$\begin{aligned} &\|\sigma(\mathbf{z}) : \nu\|_{X_1} \\ &\quad = \sup \left\{ |{}_{X_1}\langle \sigma(\mathbf{u}) : \nu, \mathbf{w}_t \rangle_{{}_0H^1}| \,;\ \mathbf{w} \in {}^0H^1(0,T;L^2(\Gamma;\mathbf{R}^n)),\ \|\mathbf{w}\|_{{}_0H^1} = 1 \right\}. \end{aligned} \tag{6.22}$$

Observe now that (cf. (6.14))

$$\begin{aligned} &\left\{ \mathbf{w}_t;\ \mathbf{w} \in {}^0H^1(0,T;L^2(\Gamma)),\ \|\mathbf{w}\|_{{}_0H^1} = 1 \right\} \\ &= \left\{ \widetilde{\mathbf{w}} \in L^2((0,T)\times\Gamma;\mathbf{R}^n) \,:\, \|\widetilde{\mathbf{w}}\|_{L^2((0,T)\times\Gamma;\mathbf{R}^n)} = 1 \right\}. \end{aligned}$$

Then, from (6.22) we infer

$$\|\sigma(\mathbf{z}) : \nu\|_{X_1} = \|\sigma(\mathbf{u}) : \nu\|_{L^2((0,T)\times\Gamma;\mathbf{R}^n)}. \tag{6.23}$$

Hence, taking (6.18) and (6.23) into account, from (6.19) we get

$$\Lambda_{11} \|\sigma(\mathbf{z}) : \nu\|_{X_1} \le \left(\|\mathbf{z}_0\|_H^2 + \|\mathbf{z}_1\|_{V'}^2 \right)^{1/2} \le \Lambda_{12} \|\sigma(\mathbf{z}) : \nu\|_{X_1}, \tag{6.24}$$

where Λ_{11} and Λ_{12} are positive constants which depend on Ω, x^0, T, ϱ, λ, and μ. Here we have used the fact that $\varrho\mathcal{E}$ is an isomorphism between V and V' (see (3.11)) along with the inequality

$$\Lambda_{13}\|\phi\|_V \le \|\rho\phi\|_V \le \Lambda_{14}\|\phi\|_V \quad \forall \phi \in V, \tag{6.25}$$

which holds for some positive constants Λ_{13}, Λ_{14} only depending on Ω and ϱ. In fact, the second inequality in (6.25) is straightforward by (2.2). As far as the first inequality is concerned, it suffices to check that

$$\|\phi^j_{x_i}\|_{L^2(\Omega)} \le \Lambda_{15}\|\rho\phi\|_V, \quad 1 \le i,j \le n, \quad \forall\, \phi \in V, \tag{6.26}$$

where Λ_{15} is a positive constant only depending on Ω and ϱ.

By (2.2) and (2.3), we have

$$\begin{aligned} &\| \ (\rho\phi^j)_{x_i}\|^2_{L^2(\Omega)} = \int_\Omega (\rho_{x_i}\phi^j + \rho\phi^j_{x_i})^2 dx \\ &= \int_\Omega \rho^2 (\phi^j_{x_i})^2 + \int_\Omega \rho^2_{x_i} (\phi^j)^2 dx + 2\int_\Omega \rho\rho_{x_i}\phi^j\phi^j_{x_i} dx \\ &\ge \rho_0^2 \|\phi^j_{x_i}\|^2_{L^2(\Omega)} - \Lambda_{16}\|\phi^j\|^2_{L^2(\Omega)} - \Lambda_{16}\int_\Omega 2|\phi^j||\phi^j_{x_i}| dx, \end{aligned}$$

where Λ_{16} is a positive constant depending on ρ. Observe now that

$$\int_\Omega 2|\phi^j||\phi^j_{x_i}|dx \le \frac{\rho_0^2}{2\Lambda_{16}}\int_\Omega |\phi^j_{x_i}|^2dx + \frac{2\Lambda_{16}}{\rho_0^2}\int_\Omega |\phi^j|^2dx.$$

Consequently, we get

$$\|\rho\phi\|_V^2 \ge \|(\rho\phi^j)_{x_i}\|^2_{L^2(\Omega)} \ge \frac{\rho_0^2}{2}\|\phi^j_{x_i}\|^2_{L^2(\Omega)} - \Lambda_{17}\|\phi^j\|^2_{L^2(\Omega)} \tag{6.27}$$

for some positive constant Λ_{17} which only depends on Ω and ϱ.. Then, using the Poincaré inequality and (2.3), from (6.27) we infer (6.26). Hence (6.25) holds.

Consider now $\mathbf{z}_0$, $\mathbf{z}_1$ satisfying (6.16). Recalling (6.17), since $\mathcal{D}(\Omega)$ and $\mathcal{J}(\Omega)$ are dense in H and in V', respectively, one can find two sequences $\{\mathbf{z}_0^m\} \subset \mathcal{D}(\Omega)$ and $\{\mathbf{z}_1^m\} \subset \mathcal{J}(\Omega)$ such that

$$\|\mathbf{z}_0^m - \mathbf{z}_0\|_H \to 0, \quad \|\mathbf{z}_1^m - \mathbf{z}_1\|_{V'} \to 0$$

as $m \nearrow +\infty$. Then, proceeding as in [LLT, Rem. 2.2 and Thm. 2.3], one obtains

$$\|\sigma(\mathbf{z}(\mathbf{z}_0^m, \mathbf{z}_1^m)) : \nu - \sigma(\mathbf{z}(\mathbf{z}_0, \mathbf{z}_1)) : \nu\|_{H^{-1}(0,T;L^2(\Gamma;\mathbf{R}^n))} \to 0$$

as $m \nearrow +\infty$.

Hence, (6.24) holds for any $\mathbf{z}_0 \in H$ and any $\mathbf{z}_1 \in V'$. This shows that the mapping

$$(\mathbf{z}_0, \mathbf{z}_1) \mapsto \sigma(\mathbf{z}) : \nu$$

is an isomorphism between $H \times V'$ and X_1. Then, arguing as in [L, Théo. 6.3, Chap. I, section 6], we deduce

$$V \subset \mathcal{R}(\mathcal{G}^*).$$

It remains to prove that

$$\mathcal{R}(\mathcal{G}^*) \subset (H^{1/2}(\Omega))^n. \tag{6.28}$$

Let $\mathbf{v} \in {}^0H^1(0,T;L^2(\Gamma;\mathbf{R}^n))$ and consider the unique solution $\tilde{\mathbf{w}} \in C^0([0,T];H) \cap C^1([0,T];V')$ to

$$\int_{Q_T} (\varrho\tilde{\mathbf{w}}, \mathbf{r})dtdx = -\int_{(0,T)\times\Gamma} (\mathbf{v}_t, \sigma(\tilde{\mathbf{q}}) : \nu)d\Sigma dt \qquad \forall\ \mathbf{r} \in C^0([0,T];D(\mathcal{E})), \tag{6.29}$$

where $\tilde{\mathbf{q}} \in C^0([0,T];D(\mathcal{E}))$ is the unique solution to (5.10)–(5.12) (see Theorem 2.4).

Then, for any $t \in [0,T]$, set

$$\mathbf{Z}(t) := \int_T^t \tilde{\mathbf{w}}(s)ds \qquad \text{a.e. in } \Omega. \tag{6.30}$$

On account of (6.29), one easily checks that $\mathbf{Z}$ fulfills (5.9). Hence $\mathbf{Z} \equiv \tilde{\mathbf{z}}(\mathbf{v})$. Then (cf. (6.30))

$$\tilde{\mathbf{z}}(\mathbf{v})_{tt} \in C^0([0,T];V')$$

and, consequently,

$$\nabla \cdot \sigma(\tilde{\mathbf{z}}(\mathbf{v})) \in C^0([0,T];V').$$

On the other hand, we have

$$\tilde{\mathbf{z}}(\mathbf{v}) = \begin{cases} \mathbf{v} & \text{on } [0,T]\times\Gamma, \\ 0 & \text{on } [0,T]\times\partial\Omega\setminus\Gamma, \end{cases}$$

and letting $t=0$, one infers

$$\nabla\cdot\sigma(\tilde{\mathbf{z}}(\mathbf{v}))(0)\in V',$$
$$\tilde{\mathbf{z}}(\mathbf{v})(0) = \begin{cases} \mathbf{v}(0) & \text{a.e. on } \Gamma, \\ 0 & \text{a.e. on } \partial\Omega\setminus\Gamma. \end{cases}$$

As $\mathbf{v}(0)\in(L^2(\Gamma))^n$, we conclude (cf., e.g., [KN, Part 1, Chap. 1, section 3.2, Thm. 3.2] and references therein) $\tilde{\mathbf{z}}(\mathbf{v})(0)\in(H^{1/2}(\Omega))^n$, i.e., (6.28). □

7. Proof of Lemma 2.3. Let us recall first the energy identity (cf. (2.29))

$$E(t) := \frac{1}{2}\int_\Omega \left[\mu|\nabla\mathbf{u}(t)|^2 + (\lambda+\mu)|\nabla\cdot\mathbf{u}(t)|^2 + \varrho|\mathbf{u}_t(t)|^2\right]dx = E_0 \tag{7.1}$$

for any $t\in[0,T]$. This identity is obtained by multiplying equation (2.10) by $\mathbf{u}_t$ and then integrating over Q_t, taking (2.4), (2.24)–(2.25) into account and using the divergence theorem. We recall that $|\nabla\mathbf{u}(t)|^2 := \sum_{i,j=1}^n |u^i_{x_j}(t)|^2$, where $\nabla\mathbf{u}(t) := [u^i_{x_j}(t)]$. Following [L, Chap. IV, pp. 225–227], we formally multiply both sides of equation (2.10) by $\nabla\mathbf{u}:\mathbf{m}$ and integrate over Q_T. Recalling (2.11), (2.24), (3.7) and integrating by parts in time, we obtain

$$\begin{aligned} &\left[\int_\Omega(\varrho\mathbf{u}_t(t),\nabla\mathbf{u}(t):\mathbf{m})dx\right]_0^T - \int_{Q_T}(\varrho\mathbf{u}_t,\nabla\mathbf{u}_t:\mathbf{m})dtdx \\ &= \int_{Q_T}(\nabla\cdot\sigma(\mathbf{u}),\nabla\mathbf{u}:\mathbf{m})dtdx. \end{aligned} \tag{7.2}$$

Observe now that

$$(\varrho\mathbf{u}_t,\nabla\mathbf{u}_t:\mathbf{m}) = \frac{1}{2}\nabla\cdot(\varrho\mathbf{m}|\mathbf{u}_t|^2) - \frac{1}{2}[(\nabla\varrho,\mathbf{m}) + n\varrho]|\mathbf{u}_t|^2. \tag{7.3}$$

Also, recalling (2.1) and (2.24), we have

$$\nabla\cdot\sigma(\mathbf{u}) = \mu\Delta\mathbf{u} + \nabla[(\lambda+\mu)\nabla\cdot\mathbf{u}],$$

from which we deduce

$$(\nabla\cdot\sigma(\mathbf{u}),\nabla\mathbf{u}:\mathbf{m}) = \nabla\cdot[A:(\nabla\mathbf{u}:\mathbf{m})] - (B,\nabla(\nabla\mathbf{u}:\mathbf{m})), \tag{7.4}$$

where

$$A := \mu(\nabla\mathbf{u})^{\mathbf{T}} + (\lambda+\mu)(\nabla\cdot\mathbf{u})\delta, \quad B := \mu\nabla\mathbf{u} + (\lambda+\mu)(\nabla\cdot\mathbf{u})\delta, \tag{7.5}$$

the superscript $\mathbf{T}$ denoting the transposition. In addition, note that

$$\begin{aligned} (B,\nabla(\nabla\cdot\mathbf{u}:\mathbf{m})) &= \frac{1}{2}[(\lambda+\mu)|\nabla\cdot\mathbf{u}|^2 + \mu|\nabla\mathbf{u}|^2] \\ &+ \frac{1}{2}\nabla\cdot\{[(\lambda+\mu)|\nabla\cdot\mathbf{u}|^2 + \mu|\nabla\mathbf{u}|^2]\mathbf{m}\} \\ &- \frac{1}{2}(\nabla\lambda,\mathbf{m})|\nabla\cdot\mathbf{u}|^2 + \frac{1-n}{2}[(\lambda+\mu)|\nabla\cdot\mathbf{u}|^2 + \mu|\nabla\mathbf{u}|^2]. \end{aligned} \tag{7.6}$$

Taking (7.1), (7.5)–(7.6) into account and using the divergence theorem, a combination of (7.2) and (7.3)–(7.4) yields

$$\begin{aligned}(7.7)\qquad &\left[\int_\Omega \left(\varrho \mathbf{u}_t(t), \nabla \mathbf{u}(t) : \mathbf{m} + \frac{n-1}{2}\mathbf{u}(t)\right) dx\right]_0^T + TE_0\\ &+\frac{1}{2}\int_{Q_T} (\nabla \varrho, \mathbf{m})|\mathbf{u}_t|^2 dtdx - \frac{1}{2}\int_{Q_T} (\nabla \lambda, \mathbf{m})|\nabla \cdot \mathbf{u}|^2 dtdx\\ &= \frac{1}{2}\int_0^T \int_{\partial\Omega} \left[\mu|\nabla \mathbf{u}|^2 + (\lambda+\mu)|\nabla \cdot \mathbf{u}|^2(\mathbf{m}, \nu)\right] d\Sigma dt.\end{aligned}$$

Here we have also used the formula

$$(A : (\nabla \mathbf{u} : \mathbf{m}), \nu) = [\mu|\nabla \mathbf{u}|^2 + (\lambda+\mu)|\nabla \cdot \mathbf{u}|^2](\mathbf{m}, \nu).$$

Arguing as in [L, Chap. I, section 5, pp. 58–59] and taking advantage of (2.3) and (7.1), we obtain

$$\begin{aligned}(7.8)\qquad &\left|\left[\int_\Omega \left(\varrho \mathbf{u}_t(t), \nabla \mathbf{u}(t) : \mathbf{m} + \frac{n-1}{2}\mathbf{u}(t)\right) dx\right]_0^T\right|\\ &\le 2\|\varrho\|_{L^\infty(\Omega)} \sup_{t\in(0,T)} \left|\left(\mathbf{u}_t(t), \nabla \mathbf{u}(t) : \mathbf{m} + \frac{n-1}{2}\mathbf{u}(t)\right)_H\right|\\ &\le \|\varrho\|_{L^\infty(\Omega)} \|\mathbf{u}_t\|_{L^\infty(0,T;H)} \left\|\nabla \mathbf{u} : \mathbf{m} + \frac{n-1}{2}\mathbf{u}\right\|_{L^\infty(0,T;H)}\\ &\le 2R(x^0)\|\varrho\|_{L^\infty(\Omega)} \|\mathbf{u}_t\|_{L^\infty(0,T;H)} \left(\sup_{t\in(0,T)} \int_\Omega |\nabla \mathbf{u}(t)|^2 dx\right)^{1/2}\\ &\le 2R(x^0)\|\varrho\|_{L^\infty(\Omega)} (\varrho_0\mu_0)^{-1/2} E_0.\end{aligned}$$

Here we have also taken into account that, using the divergence theorem, one has

$$\begin{aligned}&\left\|\nabla \mathbf{u}(t) : \mathbf{m} + \frac{n-1}{2}\mathbf{u}(t)\right\|_H^2 - \|\nabla \mathbf{u}(t) : \mathbf{m}\|_H^2\\ &= (n-1)(\nabla \mathbf{u}(t) : \mathbf{m}, \mathbf{u}(t)) + \frac{(n-1)^2}{4}\|\mathbf{u}(t)\|_H^2\\ &= \frac{(n-1)}{2}\sum_{i,j}\int_\Omega (u^i(t))_{x_j}^2 m^j dx + \frac{(n-1)^2}{4}\|\mathbf{u}(t)\|_H^2 = \frac{n(1-n)}{2}\|\mathbf{u}(t)\|_H^2 \le 0.\end{aligned}$$

Besides, by virtue of (7.1) (cf. also (2.21)–(2.23)), we have

$$(7.9)\qquad \frac{1}{2}\int_{Q_T} (\nabla \varrho, \mathbf{m})|\mathbf{u}_t|^2 dtdx \ge -R(x^0)TE_0 \left\|\frac{\nabla \varrho}{\varrho}\right\|_{(L^\infty(\Omega_\varrho^-(x^0)))^n},$$

$$(7.10)\qquad \frac{1}{2}\int_{Q_T} (\nabla \lambda, \mathbf{m})|\nabla \cdot \mathbf{u}|^2 dtdx \le R(x^0)TE_0 \left\|\frac{\nabla \lambda}{\lambda}\right\|_{(L^\infty(\Omega_\lambda^+(x^0)))^n}.$$

Finally, on account of (2.20), (2.26), and (2.27), the equality (7.7) and the inequalities (7.8)–(7.10) allow us to deduce (2.28).

REFERENCES

[BK] A. L. Bukhgeim and V. B. Kardakov, *Solution of the inverse problem for the equation of elastic waves by the method of spherical means*, Siberian Math. J., 19 (1978), pp. 528–535.

[I] M. Ikehata, *Private communication*, 1996.

[KN] H. Kardestuncer and D. H. Norrie, eds., *Finite Element Handbook*, McGraw-Hill, New York, 1987.

[LLT] I. Lasiecka, J.-L. Lions, and R. Triggiani, *Non homogeneous boundary value problems for second order hyperbolic operators*, J. Math. Pures Appl., 65 (1986), pp. 149–192.

[L] J.-L. Lions, *Contrôlabilité exacte perturbations et stabilisation de systèmes distribués*, Vol. 1, Coll. RMA 8, Masson, Paris, 1988.

[MH] J. E. Marsden and T. J. R. Hughes, *Mathematical Foundations of Elasticity*, Prentice-Hall, Englewood Cliffs, NJ, 1983.

[TB] J. J. Telega and W. R. Bielski, *Exact controllability of anisotropic elastic bodies*, in Modelling and Optimization Parameter Systems Applications to Engineering, K. Malanowski, Z. Nahorski, and M. Peszyńska, eds., Chapman and Hall, London, 1996, pp. 254–262.

[Y] M. Yamamoto, *Stability, reconstruction formula and regularization for an inverse source hyperbolic problem by a control method*, Inverse Problems, 11 (1995), pp. 481–496.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 4, pp. 1207–1224, July 1998

005

AN EXAMPLE OF A UNIVERSALLY OBSERVABLE FLOW ON THE TORUS*

ALISA DESTEFANO† AND G. R. HALL‡

Abstract. In this paper we examine the question of existence of a two-dimensional universally observable system, i.e., dynamics which are observable by every continuous nonconstant real-valued function on the state space. We are motivated by the work of D. McMahon, who proved that a class of three-dimensional manifolds with horocycle flow have this property. We examine this example and are able to give sufficient conditions for a flow to be universally observable. We then use these conditions to show the existence of a continuous universally observable flow on the torus. The proofs involve techniques and concepts from topological dynamics and dynamical systems on the torus.

Key words. universal observability, torus, dynamical system

AMS subject classifications. 93B07, 58F25, 34C35

PII. S0363012996308417

1. Introduction.

Determining the behavior of a dynamical system from some scalar observation of the system has been studied quite extensively in the literature (see [19], [4]). In particular, criteria for observability of nonlinear systems are given, that is, criteria about which systems will be observable by given observation functions and what types of observation functions observe a given system. We will consider the general setting of a Hausdorff space M with a continuous flow ϕ. We ask if a given real-valued function h of the space M distinguishes orbits. If the answer is "yes," then the system is observable under h.

The question arises as to whether there exist systems which are universally observable, i.e., dynamics which are observable by every continuous nonconstant real-valued function on the space. It seems unlikely that one could find such a system, but McMahon [18] proved that a class of three-dimensional manifolds ($SL(2, \mathbf{R})$ modulo a certain type of subgroup) with horocycle flow has this property.

The search for other examples, particularly for low-dimensional universally observable systems, has led to some interesting results but has produced no further examples. Most of the work thus far has focused on smooth dynamical systems, i.e., flows arising from smooth vector fields. Byrnes, Dayawansa, and Martin [3] determined necessary conditions for universally observable systems which lead to the conclusion that if there is a smooth low-dimensional universally observable system then it has to be a minimal flow on the torus.

In [9], it was shown that any universally observable low-dimensional system would be topologically equivalent to constant irrational flow on the torus. Using a property which is equivalent to universal observability developed by Wallace [25], it can be shown that constant irrational flow is not universally observable.

In this paper, we address the general question of existence of a universally observable flow on the torus (not necessarily smooth). We first examine the properties of McMahon's example which were sufficient for universal observability (see [25], [9]).

*Received by the editors August 23, 1996; accepted for publication (in revised form) May 29, 1997; published electronically May 15, 1998.
http://www.siam.org/journals/sicon/36-4/30841.html
†Department of Mathematics, College of the Holy Cross, Worcester, MA 01610 (alisad@math.holycross.edu).
‡Department of Mathematics, Boston University, Boston, MA 02215 (rockford@math.bu.edu).

We give a proof that any flow with these properties, referred to as property V, is universally observable. We then examine these properties, and using results from the field of topological dynamics, we further isolate some sufficient conditions for universal observability. Using these conditions, we are able to construct a continuous flow on the torus which provides the first example of a two-dimensional universally observable system.

2. Universal observability. In this section, we establish notation and give some basic definitions. We then discuss the properties of McMahon's example which are sufficient for universal observability.

Let $\phi(x,t)$ be a continuous flow on a Hausdorff space M. That is,

$$\phi : M \times \mathbf{R} \to M$$

is a continuous function and

$$\phi(x, t_1 + t_2) = \phi(\phi(x, t_1), t_2).$$

DEFINITION 2.1. *Let $h : M \to \mathbf{R}$ be a continuous nonconstant function. We say that h* observes *(M,ϕ) if $h(\phi(x_0,t)) = h(\phi(y_0,t))$ for all $t \geq 0$ implies that $x_0 = y_0$. In this case, we say that (M,ϕ,h) is* observable.

McMahon [18] proved that a class of three-dimensional manifolds ($SL(2,\mathbf{R})$ modulo a discrete, cocompact, nonarithmetic subgroup) with horocycle flow is observable by every nonconstant continuous function from the manifold to the real numbers. We refer to this phenomenon as universal observability. More precisely, we have the following definition.

DEFINITION 2.2. *(M,ϕ) is* universally observable *if (M,ϕ,h) is observable for every continuous nonconstant function $h : M \to \mathbf{R}$.*

McMahon's example had a very stong property which was sufficient for universal observability. Before we discuss this property, we develop some notation and state some definitions.

Given any fixed time t_*, let $f_{t_*}(x) = \phi(x, t_*)$. We consider the discrete flow generated by f_{t_*}, i.e., the iterates $\{f_{t_*}^n : n \in \mathbf{Z}\}$. We use the notation (M, f_{t_*}) to refer to this discrete flow.

DEFINITION 2.3. *The* orbit *of a point $x \in M$ under the flow (M,ϕ) is the set $O(x) = \{\phi(x,t) : t \in \mathbf{R}\}$. The orbit of a point $x \in M$ under the flow (M, f_{t_*}) is the set $O_{f_{t_*}}(x) = \{f_{t_*}^n(x) : n \in \mathbf{Z}\}$.*

DEFINITION 2.4. *The set K is* invariant *under the flow ϕ if $\{\phi(K,t) : t \in R\} \subset K$.*

DEFINITION 2.5. *The flow (M,ϕ) is* minimal *if M has no proper closed invariant sets under ϕ. Equivalently, every point of M has a dense orbit.*

DEFINITION 2.6. *If ϕ is a flow on M, then the* product flow *$\phi \times \phi$ on $M \times M$ is given by $\phi \times \phi((x,y),t) = (\phi(x,t), \phi(y,t))$.*

Now we describe the example of a universally observable system found by McMahon. Let $G = SL(2,\mathbf{R})$ and let Γ be a discrete subgroup of G with compact quotient space $M = \Gamma\backslash G$. Horocycle flow Φ on M is defined by

$$\Phi(\Gamma g, t) = \Gamma g \begin{pmatrix} 1 & 0 \\ t & 1 \end{pmatrix}.$$

In a paper by del Junco and Keane [7], it is remarked that any horocycle flow where Γ is a discrete cocompact maximal nonarithmetic subgroup has the property

that the past and future limit sets of (x, y) are $M \times M$ whenever x and y are on different orbits in (M, Φ). This follows from Theorem 4.5 in [6], which states that this flow has twofold minimal self-joinings, a strong property from ergodic theory; note that this is different from the definition of Markley [21]. Also, this flow has the property that the discrete flow induced by some time t_0 is minimal, so that the positive orbit of each point under the discrete flow is dense. These properties are sufficient for universal observability. Specifically, we now isolate the properties of McMahon's example which imply universal observability. We note that any flow with these properties will be universally observable.

DEFINITION 2.7. *A flow* (M, ϕ) *has* property V *if* (i) *the future limit set of* (x, y) *is* $M \times M$ *whenever* x *and* y *are on different orbits in* (M, ϕ), *and* (ii) *given any fixed time* t_*, *the positive orbit of each point in* (M, f_{t_*}) *is dense in* M.

THEOREM 2.8. *If* (M, ϕ) *has property* V, *then* (M, ϕ) *is universally observable.*

Remark. The following proof is a direct generalization of McMahon's proof (see [18]).

Proof. Consider any nonconstant continuous observation function $h : M \to \mathbf{R}$. First, consider any two points $x, y \in M$ on different orbits. Choose any two open sets $U, V \subset M$ such that the intersection of $h(U)$ and $h(V)$ is empty. Since the future limit set of (x, y) is $M \times M$, there exists some time t_0 such that $\phi(x, t_0) \in U$ and $\phi(y, t_0) \in V$. Therefore, any two points on different orbits can be distinguished by any nonconstant continuous observation function h; specifically, there exists a time t_0 where $h(\phi(x, t_0)) \neq h(\phi(y, t_0))$.

Next consider the case when x and y are on the same orbit, i.e., there is a time t_* such that $y = \phi(x, t_*)$. We assume that these points cannot be distinguished by the flow, i.e., $h(\phi(x, t)) = h(\phi(y, t))$ for all $t \geq 0$. Now consider the discrete flow induced by time t_* above. We denote this discrete flow by (M, f_{t_*}). Now we have $h(x) = h(y)$ and $h(y) = h(\phi(x, t_*))$. So we see that $h(x) = h(f_{t_*}(x))$. Now following the orbit of x under ϕ for time t_* again, we have that $h(f_{t_*}(x)) = h(f_{t_*}^2(x))$, and continuing in this fashion we see that

$$h(x) = h(f_{t_*}(x)) = h(f_{t_*}^2(x)) = h(f_{t_*}^3(x)) = \cdots.$$

So h must be constant since the positive orbit of the discrete flow (M, f_{t_*}) is dense. This contradicts our assumption that h is a nonconstant continuous function. Thus, (M, ϕ) is universally observable. ∎

McMahon's example exhibits very strong ergodic and topological properties. This leads to the question of what ergodic or topological properties are necessary for universal observability. Is there an equivalent notion in topological dynamics? The notion of primeness for flows from topological dynamics is closely related to universal observability. For more on these questions, see [10, 12, 17].

3. A sufficient condition for universal observability. In this section, we will further isolate sufficient conditions for a general system to be universally observable. Here, we consider the more general case of a continuous flow ϕ on a compact Hausdorff space. We first state some relevant definitions from topological dynamics.

DEFINITION 3.1. *The flow* (M, ϕ) *is* topologically ergodic *if every proper closed invariant set (under* ϕ*) is nowhere dense.*

DEFINITION 3.2. *The flow* (M, ϕ) *is* topologically weakly mixing *if the product flow,* $(M \times M, \phi \times \phi)$, *is topologically ergodic.*

Equivalently, the flow (M, ϕ) is topologically weakly mixing if every point of $M \times M$, with the possible exception of a set of first category, has an orbit which is dense (see [23, p. 152]).

Similarly, we can define these notions for the discrete flow as well.

LEMMA 3.3. *Given any topologically weakly mixing flow (M, ϕ) which is minimal, the discrete flow generated by a fixed time t_*, (M, f_{t_*}) is also minimal. That is, the iterates of $f_{t_*}(x)$ for any $x \in M$ form a dense subset of M.*

Remark. This lemma is often used in topological dynamics [5], [20]. For a proof, we refer the reader to the Appendix section 6.

Next we use this lemma to describe sufficient conditions for a flow to be universally observable.

DEFINITION 3.4. *A flow (M, ϕ) has* property W *if it is minimal and satisfies part* (i) *of property* V *(Definition* 2.7*).*

THEOREM 3.5. *If a flow (M, ϕ) has property* W*, then it is universally observable.*

Proof. If a flow satisfies part (i) of property V, then it is obviously topologically weakly mixing. Now by Lemma 3.3, the flow satisfies part (ii) of property V. Therefore, by Theorem 2.8, the flow is universally observable. □

4. Previous results. In this section, we review some results regarding universal observability. These give motivation for our main result, the construction of a universally observable flow on the torus.

First, we recall some results of Byrnes, Dayawansa, and Martin [3]. These give necessary conditions for universal observability.

Suppose X is a locally compact Hausdorff space and

$$\phi : \mathbf{R} \times M \to M$$

is a continuous flow on M.

THEOREM 4.1. *If ϕ is universally observable, then ϕ is minimal; i.e., all positive orbits are dense in M.*

For the next theorem we make the additional assumptions that M is a smooth manifold and f is a smooth vector field on M, so f should be complete.

THEOREM 4.2. *If f is a vector field on M which is universally observable, then M is compact with Euler characteristic zero.*

Now we focus on the question of existence of any low-dimensional systems which are universally observable. It is not hard to see that there cannot be a one-dimensional universally observable system. We consider the two-dimensional case below.

We consider smooth flows on two-dimensional manifolds without boundary. By the results of Byrnes, Dayawansa, and Martin above, we need only consider nonsingular vector fields with all positive orbits dense and manifolds which are compact with Euler characteristic zero. Now the classification of compact surfaces yields only two surfaces with vanishing Euler characteristic, the torus for orientable surfaces and the Klein bottle for nonorientable surfaces.

A result due to Kneser (see [13]) says that every smooth direction field on the Klein bottle has a periodic orbit. Therefore, there is no universally observable system on the Klein bottle. For more details on this, see [9].

Next we consider smooth flows on the torus. We note that any vector field on the torus can be represented by a set of differential equations in the plane which are periodic in the spatial variables. More specifically, let $\mathbf{X}$ be a vector field on the torus, $\mathbf{T}^2 = \mathbf{S}^1 \times \mathbf{S}^1$. Then $\mathbf{X}$ can be represented as a system of differential equations of the following form:

$$\frac{d\eta}{dt} = f(\eta, \theta),$$
$$\frac{d\theta}{dt} = g(\eta, \theta),$$

where $(\eta, \theta) \in \mathbf{T}^2$, and f, g are periodic in η, θ of period one.

A special important case of this type is the linear system

$$(*) \qquad \frac{d\eta}{dt} = \alpha,$$
$$\frac{d\theta}{dt} = 1.$$

This is just the constant vector field with orbits being the lines of slope $\frac{1}{\alpha}$. When α is irrational, we refer to this as constant irrational flow.

We need to discuss equivalence of flows (vector fields).

DEFINITION 4.3. *A C^r vector field $\mathbf{X}$ on M is called* C^k-equivalent *to a C^r vector field $\mathbf{X}'$ on M' if there is a diffeomorphism of class C^k of M onto M' which takes orbits of $\mathbf{X}$ to orbits of $\mathbf{X}'$ preserving orientation but not necessarily parametrization by time.*

The following proposition about vector fields on the torus stems from the work of Denjoy [8] on diffeomorphisms of the circle. This is a sort of classification for certain nonsingular vector fields on the torus.

PROPOSITION 4.4. *Any flow on the torus with no equilibrium points or closed orbits necessarily arises from a vector field C^0 equivalent (topologically equivalent) to the above linear system $(*)$ with α irrational.*

This follows from Denjoy's theorem and relevant facts about nonsingular vector fields on the torus with an irrational rotation number. For a more detailed discussion, see [14] or [2].

Using these facts, we get the following theorem.

THEOREM 4.5. *A universally observable system on the torus is topologically equivalent to system $(*)$ with α irrational.*

For a detailed proof, see [9].

As described by Wallace [25], it is well known that constant irrational flow, i.e., the flow arising from a constant vector field on the torus with orbits being winding lines of irrational slope, is not universally observable. To understand why this is so, we first discuss a topological criterion developed by Wallace [25] and DeStefano [9] which is equivalent to universal observability.

DEFINITION 4.6. *(M, ϕ) has* property U *if there do not exist $x_0, y_0 \in M$ and a one-parameter family of nontrivial open subsets $U_\alpha \subset M$, such that for all $\alpha \in [0,1]$, $cl(U_{\alpha_1}) \subset U_{\alpha_2}$ if and only if $\alpha_1 < \alpha_2$, $cl(U_\alpha) \neq M$, and for all $t \in \mathbf{R}$ and $\alpha \in [0,1]$, $\phi(x_0, t) \in U_\alpha$ if and only if $\phi(y_0, t) \in U_\alpha$.*

Remark. Note that the definition of property U above is slightly more restrictive than that given in [25] (see [9]).

THEOREM 4.7. *(M, ϕ) is universally observable if and only if (M, ϕ) has property* U.

For the proof, the reader is referred to [25] and [9].

Now we can see that this system does not have property U. This is done by constructing the sets U_α as pictured in Figure 4.1 and picking x_0, y_0 on the boundary of U_1.

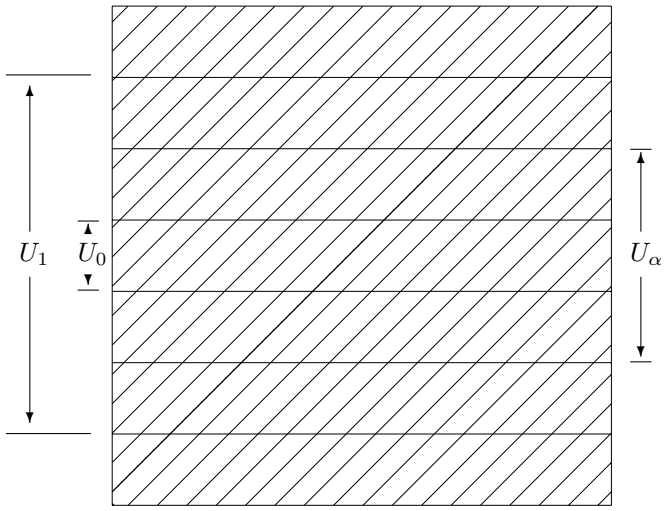


FIG. 4.1. *Construction of the U_α's on the torus.*

These results show that any universally observable flow on the torus is topologically equivalent to constant irrational flow, a flow which itself is not universally observable. So if there is a universally observable flow on the torus, its orbit structure is the same topologically as a flow which is not universally observable.

5. An example of a continuous universally observable flow on the torus. Now we are ready to discuss the existence of a universally observable flow on the torus.

THEOREM 5.1. *There exists a continuous flow on the torus with property* W.

By Theorem 3.5, we immediately obtain the following corollary.

COROLLARY 5.2. *There exists a continuous universally observable flow on the torus.*

The flow we construct is topologically equivalent to the flow on the torus $\mathbf{T}^2$ given by the vector field

$$(*) \qquad \begin{aligned} \frac{d\eta}{dt} &= \alpha, \\ \frac{d\theta}{dt} &= 1, \end{aligned}$$

where α is an irrational number fixed below. Hence the flow is minimal. To obtain a flow $\phi(x,t) : \mathbf{T}^2 \times \mathbf{R} \to \mathbf{T}^2$ with the other part of property W, that certain orbits of the product of the flow with itself, $\phi \times \phi$, are dense in $\mathbf{T}^2 \times \mathbf{T}^2$, we adjust the speed of orbits in the torus.

Interestingly, our example is only a continuous flow and not all irrationals α are admissible for the underlying flow on the torus. These limitations are required by our construction. We do not know if smoother torus flows exist with property W or if smoother universally observable flows on the torus exist. However, it is not unlikely that this is another example of the sort of "small-divisor" problem which commonly appears in the study of torus flows and circle maps (see Hermann [16]).

As noted above, the topology of our flow ϕ is that of a minimal "straight-line" flow. The speeds of the orbits are chosen so that any two points on different ϕ-orbits move with time so that they become arbitrarily far apart in (the lift of) the θ direction. By Lemma 5.4 below, this will suffice to give the density of orbits in $\mathbf{T}^2 \times \mathbf{T}^2$. Finally, we give the definition of the flow ϕ and show that it has the required properties.

We begin with some notation. The dynamics of flows on the torus given by the vector field $(*)$ are intimately connected to the continued fraction expansion of α which determines the slope of the vector field (see Hermann [16]). We use only the following basic facts concerning continued fractions. The reader is referred to Niven [22] or Hardy and Wright [15] for details and proofs.

Recall that each $\alpha \in (0,1)$ can be represented as a continued fraction

$$\alpha = \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cdots}}},$$

where each $a_i \in \mathbf{Z}^+ \cup \{0\}$. In the case of an infinite fraction this notation means that

$$\alpha = \lim_{n\to\infty} \frac{p_n}{q_n},$$

where p_n/q_n is the truncation of the continued fraction for α at the nth level, i.e.,

$$\frac{p_n}{q_n} = \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cdots + \cfrac{1}{a_n}}}}.$$

This rational p_n/q_n is called the nth convergent of α. The continued fraction expansion is infinite if and only if α is irrational, and the expansion is unique. For rational α, there is a slight ambiguity in the expansion in the last term; see the references cited above.

A general property of the continued fraction expansion of an irrational α with nth-convergent p_n/q_n is

$$\left|\alpha - \frac{p_n}{q_n}\right| < \frac{1}{q_n q_{n+1}},$$

where q_{n+1} is the denominator of the $(n+1)$st convergent. The size of these denominators is controlled by the a_i in the continued fraction. In fact, $q_{n+1} = a_{n+1}q_n + q_{n-1}$ and $p_{n+1} = a_{n+1}p_n + p_{n-1}$ for all n. This property is useful because it picks out the rational numbers with smallest denominator closest to the irrational. For more on this subject, see [15].

5.1. Density of curves in $\mathbf{T}^2$ and $\mathbf{T}^2 \times \mathbf{T}^2$. Since property W relates the topology of orbits in $\mathbf{T}^2$ and $\mathbf{T}^2 \times \mathbf{T}^2$, we need notation for measuring "how dense" a finite length curve is in these spaces. We let the usual metrics on $\mathbf{T}^2$ and $\mathbf{T}^2 \times \mathbf{T}^2$ be denoted by d and d_2, respectively.

DEFINITION 5.3. *A set $A \subset \mathbf{T}^2 \times \mathbf{T}^2$ is* δ-dense *for $\delta > 0$ if, for every $z \in \mathbf{T}^2 \times \mathbf{T}^2$, there exists $w \in A$ such that $d_2(z,w) \le \delta$.*

Finally, we need notation for the universal covers of $\mathbf{T}^2$ and $\mathbf{T}^2 \times \mathbf{T}^2$ and associated lifts. For $a \in \mathbf{R}$, we let $\langle a \rangle$ denote the fractional part of a, i.e., $\langle a \rangle \in [0,1)$ and $a - \langle a \rangle \in \mathbf{Z}$.

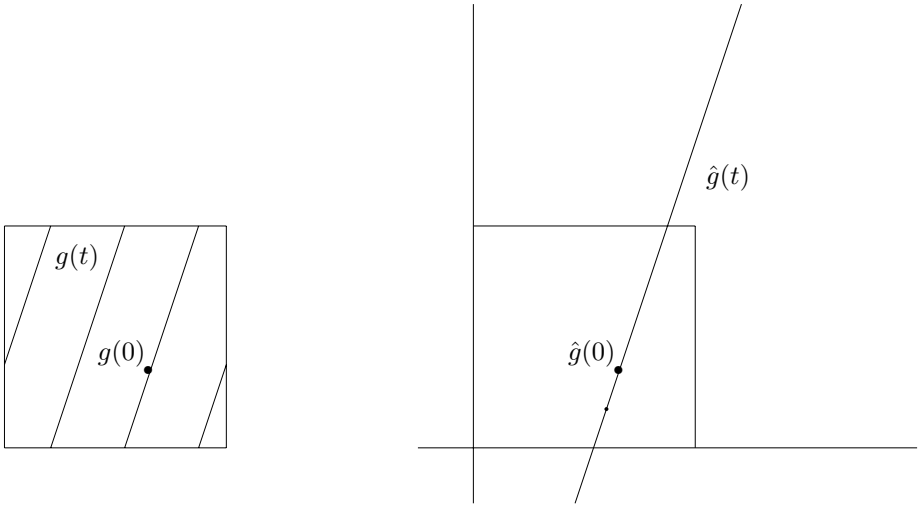


FIG. 5.1. *The lift of a curve on the torus.*

Let $\pi : \mathbf{R}^2 \to \mathbf{T}^2$ denote the projection

$$\pi(x, y) = (\langle x \rangle, \langle y \rangle).$$

This is the usual covering map from $\mathbf{R}^2$ to $\mathbf{T}^2$. Let $\mathbf{X}$ and $\mathbf{Y}$ denote the projections of $\mathbf{R}^2$ onto x and y coordinates, respectively. For $z \in \mathbf{T}^2$, let $\hat{z} \in \mathbf{R}^2$ be the lift of z satisfying $\pi(\hat{z}) = z$, $\mathbf{X}(\hat{z}), \mathbf{Y}(\hat{z}) \in [0, 1)$.

If $g : \mathbf{R} \to \mathbf{T}^2$ is a curve on the torus, then we let $\hat{g} : \mathbf{R} \to \mathbf{R}^2$ denote the continuous lift of g which satisfies $g = \pi \circ \hat{g}$ and $\hat{g}(0) \in [0, 1) \times [0, 1)$ (see Figure 5.1). For a flow ϕ on $\mathbf{T}^2$, we denote a lift by $\hat{\phi}$ where for each $z \in \mathbf{T}^2$, $\hat{\phi}(z, t)$ is the lift specified above of the curve $t \to \phi(z, t)$.

Next we prove a lemma which gives a key step in our construction. We know that a curve in $\mathbf{T}^2$ whose lift is a straight line with irrational slope wraps densely around $\mathbf{T}^2$. We can build a curve in $\mathbf{T}^2 \times \mathbf{T}^2$ by taking the product of two copies of such a line. The resulting curve will not be dense in $\mathbf{T}^2 \times \mathbf{T}^2$ because its image is restricted to the diagonal $\{(z, z) : z \in \mathbf{T}^2\} \subset \mathbf{T}^2 \times \mathbf{T}^2$. However, if we take two lines with slightly different speeds, i.e., slightly out of phase, but with equal irrational slopes, the curve formed in $\mathbf{T}^2 \times \mathbf{T}^2$ visits much more of the space. The following lemma embodies this observation. Its statement is somewhat technical since it is in the form we require below. See the remark below for an explanation of the qualitative meaning and implications of the following technical conditions.

Fix

$$\alpha = \frac{1}{a_1 + \frac{1}{a_2 + \frac{1}{a_3 + \cdots}}},$$

an irrational with $\frac{p_1}{q_1}, \frac{p_2}{q_2}, \ldots$ its convergents. Fix $m < n$ positive integers. Suppose $\sigma_1, \sigma_2 : \mathbf{R} \to \mathbf{T}^2$ are curves with lifts $\hat{\sigma}_1, \hat{\sigma}_2 : \mathbf{R} \to \mathbf{R}^2$ satisfying the following.

(1) Both $\hat{\sigma}_1$ and $\hat{\sigma}_2$ are lines with slope $1/\alpha$.

(2) For t fixed, j a positive integer, let $s_j(t)$ be such that

$$\mathbf{Y}(\hat{\sigma}_1(t + s_j(t))) - \mathbf{Y}(\hat{\sigma}_1(t)) = jq_n.$$

We call $s_j(t)$ the jth q_n return time; in particular, $s_1(t)$ is the q_n return time of $\sigma_1(t)$. Then we assume that there exists $N < \frac{a_{n+1}}{8}$ such that for all $t \leq s_{N+2}(0)$,

$$q_n < \mathbf{Y}(\hat{\sigma_2}(t + s_1(t))) - \mathbf{Y}(\hat{\sigma_2}(t)) < \frac{1}{q_m} + q_n.$$

(3) For N as in (2), we also require that

$$\mathbf{Y}(\hat{\sigma}_2(s_N(0))) - \mathbf{Y}(\hat{\sigma}_2(0)) > Nq_n + 2q_m.$$

Remark. Condition 2 states that the speed along σ_2 is slightly faster than that along σ_1. Condition 3 states that this slight difference eventually adds up to a large difference in positions along the two lines. In particular, it imposes a condition on the irrational α which requires that the a_n's in the continued fraction expansion grow fast enough.

Finally we state the lemma.

LEMMA 5.4. *Suppose σ_1, σ_2 are as above and $2q_m < q_n$. Then the curve given by*

$$(\sigma_1, \sigma_2) : [0, s_N(0)] \to \mathbf{T}^2 \times \mathbf{T}^2 : t \to (\sigma_1(t), \sigma_2(t))$$

is $4/q_m$ dense in $\mathbf{T}^2 \times \mathbf{T}^2$.

Proof. Fix a $t \in [0, s_1(0)]$ and consider the set

$$\{\sigma_1(t + s_j(t)), \sigma_2(t + s_j(t)) : 0 \leq j \leq N\}.$$

By the condition on N, the points $\sigma_1(t + s_j(t))$ are no more than $1/q_n$ from $\sigma_1(t)$. Moreover, $d(\sigma_2(t + s_j(t)), \sigma_2(t + s_{j+1}(t))) < 2/q_m$ for all j. Now for any $z \in \mathbf{T}^2$ there exists j such that $d(z, \sigma_2(t + s_j(t))) < \frac{2}{q_m} + \frac{1}{q_n} < \frac{5}{2q_m}$. Hence, for each point in the set $\{(\sigma_1(t), z) : z \in \mathbf{T}^2\}$ there is a point of the form $(\sigma_1(t + s_j(t)), \sigma_2(t + s_j(t)))$ within $3/q_m$ of it. Since every point of $\mathbf{T}^2$ is within $2/q_n$ of $\sigma_1(t)$ for some $t \in [0, s_1(0)]$, the set $\{(\sigma_1(t), \sigma_2(t)) : t \in [0, s_{N+2}(0)]\}$ is at least $4/q_m$ dense in $\mathbf{T}^2 \times \mathbf{T}^2$. □

5.2. The construction. Next we fix notation and make some choices of the parameters involved in our example. Where it is both possible and convenient, we make explicit choices rather than seeking optimal values.

Fix an irrational

$$\alpha = \frac{1}{a_1 + \frac{1}{a_2 + \frac{1}{a_3 + \dots}}},$$

where $a_i \to \infty$ very quickly as $i \to \infty$ (exactly how quickly will be set below). Let p_n/q_n denote the nth convergent of α.

Let $\rho : [0, 1] \to \mathbf{R}$ be a function given by

$$\rho(\eta) = \sum_{n=1}^{\infty} \beta_{q_n}[\sin(2\pi q_n \eta) + 1] = \sum_{n=1}^{\infty} \beta_{q_n} \left[\frac{e^{2\pi i q_n \eta} - e^{-2\pi i q_n \eta}}{2i} + 1 \right].$$

So $\rho(\eta)$ is given by a Fourier series on $[0,1]$ with nonzero terms having period $1/q_n$ for some n. Let the amplitudes β_{q_n} be given by

$$\beta_{q_n} = \frac{4^n}{2\pi q_n}.$$

Our first assumption on the growth rate of the a_n's (hence the q_n's) is that $q_n >> 4^n$ so that $\rho(\eta)$ converges uniformly to a C^0 function. We note that $\rho(\eta)$ is not C^1 because the q_nth coefficient of the formal Fourier series for $\rho'(\eta)$ grows like 4^n (see Tolstov [24, p. 129]). Since only the tail of $\rho(\eta)$ is important we may assume that $|\rho(\eta)| < 1$ for all η by removing the leading terms if necessary.

We next specify the flow $\phi(x,t) : \mathbf{T}^2 \times \mathbf{R} \to \mathbf{T}^2$. As noted above, we require that ϕ be topologically equivalent to the flow given by

$$(*) \qquad \begin{aligned} \frac{d\eta}{dt} &= \alpha, \\ \frac{d\theta}{dt} &= 1, \end{aligned}$$

where α is the irrational chosen above, so each orbit of ϕ lifts to a straight line with slope $1/\alpha$. Also, we require that for each $\eta \in [0,1)$, if

$$\mathbf{Y}(\hat{\phi}((\eta,0),t)) = 1,$$

then $t = 1 + \rho(\eta)$; that is, the time necessary for the ϕ-orbit of $(\eta,0)$ to loop once in the θ direction is $1 + \rho(\eta)$. With the bound on $|\rho(\eta)|$ above, this implies that we may assume the speed of each point under the flow is less than 4.

Next, we fix some notation concerning how close one point on $\mathbf{T}^2$ is to a segment of a ϕ-orbit of another point. Let

$$\varepsilon_n = \frac{3}{4q_{n+1}}.$$

For $z_1 \in \mathbf{T}^2$ with $\hat{z}_1$ a lift of z_1, let

$$\Sigma_n(z_1) = \pi\left\{\hat{z}_1 + (\alpha t, t) + (\delta, 0) : t \in \left[-\frac{q_n}{2}, \frac{q_n}{2}\right], \delta < \varepsilon_n\right\}.$$

That is, $\Sigma_n(z_1)$ is the set of points on $\mathbf{T}^2$ within ε_n in the η direction of a point in the set $\{\pi(\hat{z}_1 + (\alpha t, t)) : -\frac{q_n}{2} \le t \le \frac{q_n}{2}\}$. This is a narrow strip around the segment of the ϕ-orbit of z_1 which wraps around the torus q_n times in the θ direction.

In order to show that the flow ϕ has property W we proceed as follows.

Step 1. Show that if $z_2 \notin \Sigma_n(z_1)$, i.e., z_2 is not near the initial segment of the ϕ-orbit of z_1, then the average speeds along part of the ϕ-orbits of z_1 and z_2 are slightly different for a long period of time.

Step 2. Show that the difference in speeds of the ϕ-orbits of z_1 and z_2 implies a certain density for the initial segment of the orbit of (z_1, z_2) under $\phi \times \phi$; i.e., apply Lemma 5.4.

Step 3. Verify that if $z_2 \in \bigcap_{n \ge m} \Sigma_n(z_1)$ for some m, then z_2 is on the ϕ-orbit of z_1.

Step 4. Using Steps 1, 2, and 3, verify that if z_2 is not on the ϕ-orbit of z_1, then $\{\phi \times \phi((z_1, z_2), t) : t \ge 0\}$ is dense in $\mathbf{T}^2 \times \mathbf{T}^2$.

In the process of doing Steps 1, 2, and 4, we fix the rate at which the a_n's (hence q_n's) in the definition of α must tend to infinity.

Step 3 is the easiest, so we begin with it.

LEMMA 5.5. *If $z_2 \in \bigcap_{n\geq m} \Sigma_n(z_1)$ for some m, then z_2 is on the ϕ-orbit of z_1.*

Proof. First we note that for each n,

$$\Sigma_n(z_1) \bigcup \Sigma_{n+1}(z_1) = \pi\left\{\hat{z}_1 + (\alpha t, t) + (\delta, 0) : t \in \left[-\frac{q_n}{2}, \frac{q_n}{2}\right], \delta < \varepsilon_{n+1}\right\}.$$

That is, the intersection of the shorter (in the θ direction), wider (in the η direction) strip $\Sigma_n(z_1)$ with the longer, narrower strip $\Sigma_{n+1}(z_1)$ is a short, narrow strip. This follows from the fact that points in the set $\{\pi(\hat{z}_1 + (\alpha t, t)) : t \in [-\frac{q_{n+1}}{2}, \frac{q_{n+1}}{2}]\}$ with the same θ coordinate are approximately $1/q_{n+1}$ apart in the η direction.

By induction we see that

$$\bigcap_{n\geq m_1}^{m_2} \Sigma_n(z_1) = \pi\left\{\hat{z}_1 + (\alpha t, t) + (\delta, 0) : t \in \left[-\frac{q_{m_1}}{2}, \frac{q_{m_1}}{2}\right], \delta < \varepsilon_{m_2}\right\}$$

and hence

$$\bigcap_{n\geq m_1}^{\infty} \Sigma_n(z_1) \subset \pi\left\{\hat{z}_1 + (\alpha t, t) : t \in \mathbf{R}\right\};$$

that is, z_2 is in the ϕ-orbit of z_1. □

To show Step 1, we first fix z_1, z_2, and n so that $z_2 \notin \Sigma_n(z_1)$. Let $\hat{z}_1$ and $\hat{z}_2$ be lifts of z_1, z_2, respectively, and fix times s_j and r_j such that for $j = 1, 2, \ldots,$

$$\mathbf{Y}(\hat{\phi}(\hat{z}_1, s_j)) - \mathbf{Y}(\hat{z}_1) = jq_N,$$

$$\mathbf{Y}(\hat{\phi}(\hat{z}_2, r_j)) - \mathbf{Y}(\hat{z}_2) = jq_N.$$

In order to study the difference in speeds of z_1 and z_2 along their orbits as embodied in the difference in "q_n return times" s_j and r_j, we begin by showing that the effect of terms in $\rho(\eta)$ of period $1/q_m$, where $m \neq n$, $m \neq n+1$, can be safely ignored. We deal first with terms of period q_m for $m < n$.

LEMMA 5.6. *Let $|\cdot|$ denote the modulus of a complex number or absolute value where appropriate. With the notation as above,*

$$\left|\sum_{k=0}^{q_n-1} e^{2\pi i q_m \alpha k}\right| = \left|\frac{1 - e^{2\pi i q_m \alpha q_n}}{1 - e^{2\pi i q_m \alpha}}\right| < \frac{4q_m q_{m+1}}{q_{n+1}}.$$

Proof. From the theory of continued fractions, we know that

$$\left|\alpha - \frac{p_n}{q_n}\right| < \frac{1}{q_n q_{n+1}}$$

and

$$\left|\alpha - \frac{p_m}{q_m}\right| > \frac{1}{2q_m q_{m+1}}$$

provided $a_{m+1} \geq 2$. Hence

$$|q_m q_n \alpha - p_n q_m| < \frac{q_m}{q_{n+1}}$$

and

$$|q_m \alpha - p_m| > \frac{1}{2q_{m+1}}$$

and so

$$|e^{2\pi i q_m \alpha q_n} - 1| < \frac{q_m}{q_{n+1}},$$

$$|e^{2\pi i q_m \alpha} - 1| > \frac{1}{2}\frac{1}{2q_{m+1}}. \qquad \square$$

LEMMA 5.7. *With notation as above and $n \geq 2$,*

$$\left|\frac{d}{d\eta}\left(\sum_{k=0}^{q_n-1}\sum_{m=1}^{n-1}\beta_{q_m}\frac{(e^{2\pi i(\eta+\alpha k)q_m} - e^{-2\pi i(\eta+\alpha k)q_m})}{2i} + 1\right)\right| \leq 4\pi\left(\sum_{m=1}^{n-1} 4^m q_m q_{m+1}\right)/q_{n+1}.$$

Proof. Note that

$$\left|\frac{d}{d\eta}\left(\sum_{k=0}^{q_n-1}\sum_{m=1}^{n-1}\beta_{q_m}\frac{(e^{2\pi i(\eta+\alpha k)q_m} - e^{-2\pi i(\eta+\alpha k)q_m})}{2i} + 1\right)\right|$$

$$\leq \left|\sum_{m=1}^{n-1}\sum_{k=0}^{q_n-1}\pi q_m \beta_{q_m}(e^{2\pi i q_m(\eta+\alpha k)} + e^{-2\pi i q_m(\eta+\alpha k)})\right| \leq 4\pi\left(\sum_{m=1}^{n-1} 4^m q_m q_{m+1}\right)/q_{n+1}$$

by the previous lemma. $\square$

Hence, provided q_{n+1} is sufficiently large, we may assume that the slope of the part of $\rho(\eta)$ coming from period $1/q_m$ terms with $m < n$ is as small as we like, i.e., that we can neglect the differences in q_n return times of z_1 and z_2 caused by these terms.

Similarly, we can bound terms of $\rho(\eta)$ with period $1/q_m$ for $m > n+1$ as follows.

LEMMA 5.8.

$$\left|\sum_{m=n+2}^{\infty}\beta_{q_m}\left(\frac{e^{2\pi i q_m \alpha} - e^{-2\pi i q_m \alpha}}{2i} + 1\right)\right| \leq 2\sum_{m=n+2}^{\infty}\frac{4^m}{2\pi q_m}.$$

Proof. This follows immediately from the definition of $\beta_{q_m} = \frac{4^m}{2\pi q_m}$. Hence all the terms in the tail of the Fourier series of $\rho(\eta)$ tend to zero very quickly (provided the a_n's tend to infinity quickly). $\square$

We can collect these results as follows. Let

$$\rho_n(\eta) = \beta_{q_n}\left(\frac{e^{2\pi i q_n \eta} - e^{-2\pi i q_n \eta}}{2i} + 1\right).$$

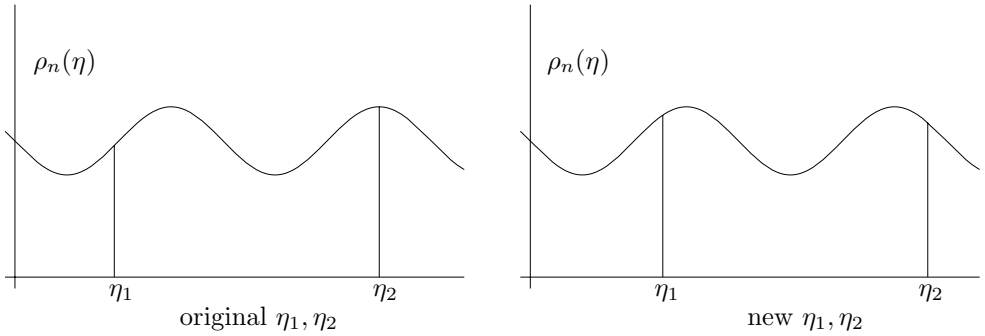


FIG. 5.2. *Evolution of η_1, η_2.*

For z_1, $z_2 \in \mathbf{T}^2$ with $z_2 \notin \Sigma_n(z_1)$, let η_1, η_2 be such that $(\eta_1, 0) + (\alpha\theta, \theta) = \hat{z}_1$ for some $\theta \in [0, 1)$, and similarly for η_2. The q_n return times of z_1 and $(\eta_1, 0)$ are at most order $1/q_{n+1}$ apart because their orbit segments bringing about the return differ only by segments of length order 1 and are at most $1/q_{n+1}$ apart. A similar statement holds for z_2 and $(\eta_2, 0)$. This difference can be ignored because the difference between r_1 and s_1 is at least q_n times as large since this difference is given by q_n times the difference in ρ values. (Similar estimates allow the difference in q_n return times of $\phi(z_1, t)$ and $\phi(z_1, t + k)$ for $k << q_n$ to be neglected.) Then the q_n return times s_1 and r_1 of z_1 and z_2 satisfy

$$s_1 = K + q_n(\rho_n(\eta_1) + \rho_{n+1}(\eta_1) + 1) + \zeta,$$

$$r_1 = K + q_n(\rho_n(\eta_2) + \rho_{n+1}(\eta_2) + 1) + \xi,$$

where K is a constant independent of η, and ζ, ξ tend to zero as the growth rate of the a_n's increases.

Remark. In fact, the same statement holds true for q_n return times if the ρ_{n+1} terms are omitted from the above since $|\rho_{n+1}(\eta)| \leq 2\frac{4^{n+1}}{2\pi q_{n+1}}$ for all η. However, since z_1 and z_2 can be only order $1/q_{n+1}$ apart, the difference in speed resulting from the $\rho_{n+1}(\eta)$ term can be significant; see below.

So we see that z_1 and z_2 have different q_n return times depending on $\rho(\eta_1)$ and $\rho(\eta_2)$. To apply Lemma 5.4 (Step 2 above) we must gain control of these return times. This can be accomplished by replacing z_1 and z_2 by $\phi(z_1, t)$ and $\phi(z_2, t)$, respectively, for a suitably chosen t. The points $\phi(z_1, s_j)$ and $\phi(z_2, r_j)$ are approximately j/q_{n+1} displaced from z_1 and z_2, respectively, in the η direction. Because q_{n+1} may be taken as large as we like, we may choose $t \in [0, q_{n+1}]$, so that if we replace z_1 and z_2 by $\phi(z_1, t)$ and $\phi(z_2, t)$, respectively, for the new η_1 and η_2,

$$0 < \rho_n(\eta_2) - \rho_n(\eta_1) < \frac{1}{q_n q_{n-2}};$$

see Figure 5.2.

First we note that because $z_2 \notin \Sigma_n(z_1)$ we know that the distance in the η direction from z_2 to a point in $\{\pi(\hat{z}_1 + (\alpha t, t)) : t \in [-\frac{q_n}{2}, \frac{q_n}{2}]\}$ is at least $\varepsilon_n = 3/4q_{n+1}$. Also, we know from continued fraction theory that

$$\left|\frac{p_n}{q_n} - \frac{p_{n+1}}{q_{n+1}}\right| = \frac{1}{q_n q_{n+1}}.$$

Hence, for $t \in [-\frac{q_n}{2}, \frac{q_n}{2}]$ the points $\pi(\hat{z}_1 + (\alpha t, t))$ are displaced in the η direction from $\pi(\hat{z}_1 + (\frac{p_n}{q_n}t, t))$ by less than $\frac{1}{2q_{n+1}} \bmod \frac{1}{q_n}$; that is,

$$\min\left\{\frac{|\langle q_n\eta_1 - q_n\eta_2\rangle|}{q_n}, \frac{1 - |\langle q_n\eta_1 - q_n\eta_2\rangle|}{q_n}\right\} \geq \frac{1}{4q_{n+1}}.$$

To apply Lemma 5.4 (Step 2 above) we must use this difference to compute estimates on the difference in q_n return times of z_1 and z_2.

We will deal with two cases depending on the distance between η_1 and η_2 mod $1/q_n$. First suppose η_1 and η_2 are relatively far apart, that is,

$$\min\left\{\frac{|\langle q_n\eta_1 - q_n\eta_2\rangle|}{q_n}, \frac{(1 - |\langle q_n\eta_1 - q_n\eta_2\rangle|)}{q_n}\right\} > \frac{2}{4^n q_n^2}.$$

Then we can choose $t \in [0, q_{n+1}]$ and replace z_1 and z_2 by $\phi(z_1, t)$ and $\phi(z_2, t)$ so that for the new η_1 and η_2,

$$\frac{1}{q_n^2} < \rho_n(\eta_2) - \rho_n(\eta_1) < \frac{1}{q_n q_{n-2}}.$$

Now, taking a_{n+1}, hence q_{n+1}, larger if necessary, we note that z_1 and z_2 are sufficiently close to being q_n periodic that the above inequality holds for $N > 2q_{n-2}q_n^2$ q_n return times of z_1 and z_2. Hence, Lemma 5.4 applies in this case with $m = n - 2$.

Next, suppose

$$\frac{1}{4q_{n+1}} < \min\left\{\frac{|\langle q_n\eta_1 - q_n\eta_2\rangle|}{q_n}, \frac{(1 - |\langle q_n\eta_1 - q_n\eta_2\rangle|)}{q_n}\right\} \leq \frac{2}{4^n q_n^2},$$

that is, η_1 and η_2 are very close mod $1/q_n$. In this case we must deal with the ρ_{n+1} term since the distance between the η's can be as small as $1/4q_{n+1}$.

The first step is to choose $t \in [0, q_{n+2}]$ so that if we replace z_1, z_2 with $\phi(z_1, t)$, $\phi(z_2, t)$, respectively, then for the new η_1, η_2 we have

$$\frac{4^n}{4q_{n+1}} < (\rho_n(\eta_2) + \rho_{n+1}(\eta_2)) - (\rho_n(\eta_1) + \rho_{n+1}(\eta_1)) < \frac{2}{q_n^2}.$$

Hence, the q_n return times are bounded below by $4^n q_n/4q_{n+1}$ and above by $2/q_n$. We must allow t's as large as q_{n+2} to accomplish this so that $\phi(z_1, t)$ and $\phi(z_2, t)$ can be properly placed relative to the period $1/q_{n+1}$ oscillations of $\rho_{n+1}(\eta)$. See Figure 5.3.

Recall that s_j and r_j are the jth q_n return times of z_1 and z_2, respectively. We fix N so that q_n return times of z_1 and z_2 are within 75% of the similar values for $\phi(z_1, s_j)$, $\phi(z_2, r_j)$ for $j \leq N$. How large we can take N depends on the difference between η_1 and η_2 mod $1/q_n$. If we let δ denote this difference, i.e.,

$$\delta = \frac{\min\{|\langle q_n\eta_1 - q_n\eta_2\rangle|, 1 - |\langle q_n\eta_1 - q_n\eta_2\rangle|\}}{q_n},$$

then we may take N to be the greatest integer less than $\frac{a_{n+1}}{8} - 2\delta q_{n+1}$ (recall, a_{n+1} is the number of $1/q_{n+1}$ periods per $1/q_n$ period, i.e., $q_{n+1} = a_{n+1}q_n + q_{n-1}$). Since $\delta < 2/4^n q_n^2$ (by assumption), we have that

$$2\delta q_{n+1} < \frac{4q_{n+1}}{4^n q_n^2} < \frac{8a_{n+1}q_n}{4^n q_n^2} = \frac{8a_{n+1}}{4^n q_n} < \frac{a_{n+1}}{16}.$$

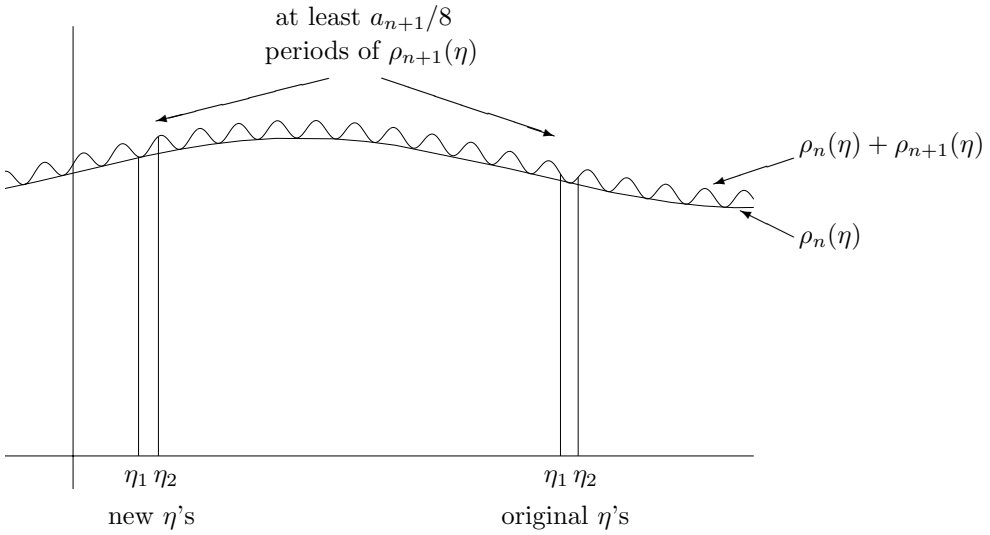


FIG. 5.3. *Proper placement of η_1 and η_2 attained by choosing the appropriate t for $\phi(z_1, t)$ and $\phi(z_2, t)$.*

(We have used $8/4^n q_n < 1/16$; since $q_n \to \infty$ rapidly, this is a very rough estimate.) Hence, in all cases, we may assume $N \geq a_{n+1}/16$. So the difference in q_n return times for $\phi(z_1, s_j)$ and $\phi(z_2, r_j)$ is bounded between $4^{n-1}q_n/4q_{n+1}$ and $2/q_n^2$. Hence,

$$s_N - r_N > N\frac{4^{n-1}q_n}{4q_{n+1}} > 4^{n-3}.$$

So we have (using the bound on $|\rho(\eta)|$) that

$$\mathbf{Y}(\hat{\phi}(\hat{z}_2, s_N)) - \mathbf{Y}(\hat{\phi}(\hat{z}_1, s_N)) > 4^{n-4},$$

and we may apply Lemma 5.4 using any q_m such that $2q_m < 4^{n-4}$.

Collecting this together, we see that if $z_2 \notin \Sigma_n(z_1)$, then for some $t \in [0, q_{n+2}]$ the points $\phi(z_1, t)$, $\phi(z_2, t)$ satisfy the hypotheses of Lemma 5.4 for any $q_m < 4^{n-5}$, and hence

$$\phi((z_1, z_2), [0, 2q_{n+2}]) \subset \mathbf{T}^2 \times \mathbf{T}^2$$

is at least $4/q_m > 4^{-n+6}$ dense in $\mathbf{T}^2 \times \mathbf{T}^2$. Letting $n \to \infty$ we see that if z_2 is not in the ϕ-orbit of z_1, then

$$\phi((z_1, z_2), [0, \infty))$$

is dense in $\mathbf{T}^2 \times \mathbf{T}^2$ and the construction is complete.

Remarks. (1) This construction actually gives a class of continuous flows with property W, since the flow is topologically equivalent to constant irrational flow and is specified by the crossing times and not the particular speed at any given point.

(2) Note that not all irrationals α are admissible. The necessary rate of growth of the a_i's in the continued fraction expansion is specified throughout the proof.

6. Appendix. The proof of Lemma 3.3 which we give below is based on a sketch given to the authors by Markley [20].

Before we give the proof, we will need some definitions, lemmas, and a proposition. For more details, we refer the reader to Auslander [1] or Ellis [11].

DEFINITION 6.1. *Let (M, ϕ) and (N, ψ) be two real continuous flows. A* homomorphism *from M to N is a continuous map $h : M \to N$ such that $h(\phi(x,t)) = \psi(h(x), t)$. If there is a homomorphism h from M onto N, we say that N is a* factor *of M.*

LEMMA 6.2. *A factor of a minimal flow is minimal.*

LEMMA 6.3. *If a minimal flow has a periodic factor, it is not topologically weakly mixing.*

DEFINITION 6.4. *A point $x \in M$ is an* almost periodic point *if for every neighborhood U of x, there is a relatively dense subset $A \subset \mathbf{R}$ such that $\phi(x, A) \subset U$.*

LEMMA 6.5. *Let M be a compact Hausdorff space and ϕ be a flow on M. Then M is a disjoint union of minimal subsets if and only if every point of M is almost periodic. In this case we say that M is pointwise almost periodic.*

Remark. See Corollary 1.10 of [1].

LEMMA 6.6. *Suppose (M, ϕ) is a flow. Consider the discrete flow generated by some time t_*; that is, consider the dynamical system given by iteration of f_{t_*}. Then the original flow is pointwise almost periodic if and only if the discrete flow is pointwise almost periodic.*

Remark. See Corollary 1.13 of [1].

LEMMA 6.7. *Given a flow on a locally compact Hausdorff space M (discrete or continuous), then $x \in M$ is an almost periodic point if and only if the orbit closure of x is a compact minimal set.*

Remark. See Theorem 1.7 of [1].

PROPOSITION 6.8. *The orbit closures for the action of a fixed time from a minimal flow are minimal sets.*

Proof. If the flow $\phi(x,t)$ is minimal, then M is a minimal set. Lemma 6.5 implies that (M, ϕ) is pointwise almost periodic. Now using Lemma 6.6, we have that the discrete flow generated by f_{t_*} is pointwise almost periodic. But by Lemma 6.7, the orbit closure $\overline{O_{f_{t_*}}(x)}$ is a minimal set. □

Proof of Lemma 3.3. Let (M, ϕ) be a minimal flow and consider the discrete flow generated by the fixed time t_*. That is, consider the flow which consists of the iterates of $f_{t_*}(x) = \phi(x, t_*)$. The orbit closure under the discrete flow at each point $x \in M$, $\overline{O_{f_{t_*}}(x)}$, is a minimal set.

Now we define an equivalence relation on M as follows: two points are equivalent if they are in the same orbit closure for the action of a fixed time. More precisely, let

$$R = \{(x, y) \in M \times M : \text{ there exists } \quad z \in M \quad \text{such that } \quad x, y \in \overline{O_{f_{t_*}}(z)}\}.$$

R is an invariant closed equivalence relation. We will denote the equivalence class of a point x by $[x]$. If we mod out by this equivalence relation, we obtain the quotient flow $\hat{\phi}$ on $\tilde{X} = X/R$ given by $\hat{\phi}([x], t) = [\phi(x,t)]$. Now $\hat{\phi}([x], t_*) = [\phi(x, t_*)] = [f_{t_*}(x)] = [x]$, so $\hat{\phi}$ is periodic of period t_*. So by Lemma 6.2, the quotient flow consists of a single periodic orbit. Therefore, Lemma 6.3 implies that the single periodic orbit must be a fixed point. Hence there is only one equivalence class, and the orbit closure for the action of a fixed time is all of M. This shows that the discrete flow generated by the fixed time t_* is minimal.

7. Conclusion. We have shown the existence of a continuous universally observable flow on the torus. Our construction actually gives a class of flows with property W which is sufficient for universal observability. These flows have orbit structure equivalent to constant irrational flow (with an appropriate irrational slope) and are obtained by defining the crossing times in such a way that any pair of points under the product flow get as dense as necessary. The construction sets conditions on how fast the denominators of the convergents in the continued fraction expansion of the irrational determining the slope of the orbits grow in order to achieve the necessary density. The choice of Fourier coefficients in the crossing time function is limited by the dynamics of the flow, and so this construction cannot be made smoother. It is still unknown if there are smoother universally observable flows on the torus.

Because of the connection between universal observability and topological dynamics, the example constructed here exhibits many topological properties of interest. This is explored by DeStefano and Markley in [10].

Acknowledgements. The authors would like to thank Dorothy Wallace and Clyde Martin for introducing them to this interesting and challenging problem.

REFERENCES

[1] J. Auslander, *Minimal Flows and Their Extensions*, North-Holland, New York, 1988.
[2] D. Anosov and V. Arnold, *Dynamical Systems* I, Springer-Verlag, Berlin, 1988.
[3] C. Byrnes, W. Dayawansa, and C. Martin, *On the topology and geometry of universally observable systems*, in Proceedings of the 26th IEEE Conference on Decision and Control, Los Angeles, 1987, pp. 963–965.
[4] C. Byrnes and C. Martin, *Global observability and detectability: An overview*, in Modeling and Adaptive Control, C. Byrnes and A. Kurzhanski, eds., Lecture Notes in Control and Inform. Sci. 105, Springer-Verlag, Berlin, 1988, pp. 71–89.
[5] A. Del Junco, personal communication, 1990.
[6] A. Del Junco, *On minimal self-joinings in topological dynamics*, Ergodic Theory Dynam. Systems, 7 (1987), pp. 211–227.
[7] A. Del Junco and M. Keane, *On generic points in the Cartesian square of Chacon's transformation*, Ergodic Theory Dynam. Systems, 5 (1985), pp. 59–69.
[8] A. Denjoy, *Sur les courbes définies par les équations différentielles à la surface du tore*, J. Math. Pures Appl., 11 (1932), pp. 333–375.
[9] A. DeStefano, *Universal observability*, in Computation and Control, Proceedings of the Bozeman Conference, Bozeman, MT, 1990, K. Bowers and J. Lund, eds., Progress in Systems and Control Theory, Birkhäuser, Boston, MA, 1991, pp. 85–94.
[10] A. DeStefano, *The relationship between universally observable flows and prime flows*, in preparation.
[11] R. Ellis, *Lectures on Topological Dynamics*, W.A. Benjamin, New York, 1969.
[12] H. Furstenberg, H. Keynes, and L. Shapiro, *Prime flows in topological dynamics*, Israel J. Math., 14 (1973), pp. 26–38.
[13] C. Godbillon, *Dynamical Systems on Surfaces*, Springer-Verlag, New York, 1983.
[14] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag, New York, 1983.
[15] G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers*, Oxford University Press, London, UK, 1956.
[16] M. Hermann, *Sur la conjugaison différentiable des difféomorphismes du cercle à des rotations*, Inst. Hautes Etudes Sci., 49 (1979), pp. 5–233.
[17] H. B. Keynes and D. Newton, *Real prime flows*, Trans. Amer. Math. Soc., 217 (1976), pp. 237–255.
[18] D. McMahon, *An example of a universally observable dynamical system*, System Control Lett., 8 (1987), pp. 247–248.
[19] C. Martin, *Observability, interpolation and related topics*, in Computation and Control: Proceedings of the Bozeman Conference, Bozeman, MT, August 1988, K. Bowers and J. Lund, eds., Birkhauser, Boston, MA, 1989, pp. 209–232.
[20] N. Markley, personal communication, 1995.

[21] N. G. Markley, *Toplogical minimal self-joinings*, Ergodic Theory Dynam. Systems, 3 (1983), pp. 579–599.
[22] I. Niven, *Irrational Numbers*, The Carus Mathematical Monographs: Number II, Mathematical Association of America, Rahway, NJ, 1967.
[23] K. Petersen, *Ergodic Theory*, Cambridge University Press, New York, 1983.
[24] G. P. Tolstov, *Fourier Series*, trans. from Russian by R. A. Silverman, Prentice-Hall, Englewood Cliffs, NJ, 1962.
[25] D. Wallace, *Observability, predictability and chaos*, in Computation and Control: Proceedings of the Bozeman Conference, Bozeman, MT, August 1988, K. Bowers and J. Lund, eds., Birkhauser, Boston, MA, 1989, pp. 365–374.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 4, pp. 1225–1239, July 1998

006

DIFFERENTIAL FLATNESS AND ABSOLUTE EQUIVALENCE OF NONLINEAR CONTROL SYSTEMS *

M. VAN NIEUWSTADT†, M. RATHINAM†, AND R. M. MURRAY†

Abstract. This paper presents a formulation of differential flatness—a concept originally introduced by Fliess, Levine, Martin, and Rouchon—in terms of absolute equivalence between exterior differential systems. Systems that are differentially flat have several useful properties that can be exploited to generate effective control strategies for nonlinear systems. The original definition of flatness was given in the context of differential algebra and required that all mappings be meromorphic functions. The formulation of flatness presented here does not require any algebraic structure and allows one to use tools from exterior differential systems to help characterize differentially flat systems. In particular, it is shown that, under regularity assumptions and in the case of single input control systems (i.e., codimension 2 Pfaffian systems), a system is differentially flat if and only if it is feedback linearizable via static state feedback. In higher codimensions our approach does not allow one to prove that feedback linearizability about an equilibrium point and flatness are equivalent: one must be careful with the role of time as well as the use of prolongations that may not be realizable as dynamic feedback in a control setting. Applications of differential flatness to nonlinear control systems and open questions are also discussed.

Key words. exterior differential systems, flatness, prolongations, trajectory generation

AMS subject classifications. 93C10, 93B29, 93A05

PII. S0363012995274027

1. Introduction.

The problem of equivalence of nonlinear systems (in particular to linear systems, that is, feedback linearization) is traditionally approached in the context of differential geometry [16, 17, 22]. A complete characterization of static feedback linearizability in the multi-input case is available, and for single input systems it has been shown that static and dynamic feedback linearizability are equivalent [5]. Some special results have been obtained for dynamic feedback linearizability of multi-input systems, but the general problem remains unsolved. Typically, the conditions for feedback linearizability are expressed in terms of the involutivity of distributions on a manifold.

More recently it has been shown that the conditions on distributions have a natural interpretation in terms of exterior differential systems [14, 26]. In exterior differential systems, a control system is viewed as a Pfaffian module. Some of the advantages of this approach are the wealth of tools available and the fact that implicit equations and nonaffine systems can be treated in a unified framework. For an extensive treatment of exterior differential systems we refer to [1].

Fliess and coworkers [7, 12, 8, 9, 18] studied the feedback linearization problem in the context of differential algebra and introduced the concept of *differential flatness*. In differential algebra, a system is viewed as a differential field generated by a set of variables (states and inputs). The system is said to be differentially flat if one can find a set of variables, called the flat outputs, such that the system is (nondifferential)

*Received by the editors August 21, 1995; accepted for publication (in revised form) June 11, 1997; published electronically May 15, 1998. A preliminary version of this paper, "Differential Flatness and Absolute Equivalence," appeared in the *Proceedings of the* 1994 *Control and Decision Conference*, IEEE Control Systems Society, Piscataway, NJ, 1994, pp. 326–332. This research was partially supported by NASA, NSF grant CMS-9502224, and AFOSR grant F49620-95-1-0419.

http://www.siam.org/journals/sicon/36-4/27402.html.

†Division of Engineering and Applied Science, California Institute of Technology, Pasadena, CA 91125 (vannieuw@indra.caltech.edu, muruhan@ama.caltech.edu, murray@indra.caltech.edu).

algebraic over the differential field generated by the set of flat outputs. Roughly speaking, a system is flat if we can find a set of outputs (equal in number to the number of inputs) such that all states and inputs can be determined from these outputs without integration. More precisely, if the system has states $x \in \mathbb{R}^n$ and inputs $u \in \mathbb{R}^p$, then the system is flat if we can find outputs $y \in \mathbb{R}^p$ of the form

$$y = y(x, u, \dot{u}, \ldots, u^{(l)}) \tag{1.1}$$

such that

$$\begin{aligned} x &= x(y, \dot{y}, \ldots, y^{(q)}), \\ u &= u(y, \dot{y}, \ldots, y^{(q)}). \end{aligned} \tag{1.2}$$

Differentially flat systems are useful in situations where explicit trajectory generation is required. Since the behavior of flat systems is determined by the flat outputs, we can plan trajectories in output space and then map these to appropriate inputs. A common example is the kinematic car with trailers, where the xy position of the last trailer provides flat outputs [20]. This implies that all feasible trajectories of the system can be determined by specifying only the trajectory of the last trailer. Unlike other approaches in the literature (such as converting the kinematics into a normal form), this approach is intrinsic.

A limitation of the differential algebraic setting is that it does not provide tools for regularity analysis. The results are given in terms of meromorphic functions in the variables and their derivatives, without characterizing the solutions. In particular, solutions to the differential polynomials may not exist. For example, the system

$$\begin{aligned} \dot{x}_1 &= u, \\ \dot{x}_2 &= x_1^2, \end{aligned} \tag{1.3}$$

is flat in the differential algebraic sense with flat output $y = x_2$. However, it is clear that the derivative of x_2 always has to be positive, and therefore we cannot follow an arbitrary trajectory in y space.

In the beginning of this century, the French geometer E. Cartan developed a set of powerful tools for the study of equivalence of systems of differential equations [3, 4, 26]. Equivalence need not be restricted to systems of equal dimensions. In particular a system can be *prolonged* to a bigger system on a bigger manifold, and equivalence between these prolongations can be studied. This is the concept of *absolute equivalence* of systems. Prolonging a system corresponds to dynamic feedback, and it is clear that we can benefit from the tools developed by Cartan to study the feedback linearization problem. The connections between Cartan prolongations and feedback linearizability for single input systems were studied in [24].

In this paper we reinterpret flatness in this differential geometric setting. We make extensive use of the tools offered by exterior differential systems and the ideas of Cartan. This approach allows us to study some of the regularity issues and also to give an explicit treatment of time dependence. Moreover, we can easily make connections to the extensive body of theory that exists in differential geometry. We show how to recover the differential algebraic definition and give an exterior differential systems proof for a result proven by Martin [18, 19] in differential algebra: a flat system can be put into Brunovsky normal form by dynamic feedback in an open and dense set (this set need not contain an equilibrium point).

We also give a complete characterization of flatness for systems with a single input. In this case, flatness in the neighborhood of an equilibrium point is equivalent to linearizability by static state feedback around that point. This result is stronger

than linearizability by endogenous feedback as indicated by Martin et al. [9, 18], since the latter only holds in an open and dense set. We also treat the case of time varying versus time invariant flat outputs and show that in the case of a single input time invariant system, the flat output can always be chosen time independent. In exterior differential systems, the special role of the time coordinate is expressed as an independence condition, i.e., a one-form that is not allowed to vanish on any of the solution curves. A fundamental problem with exterior differential systems is that most results only hold on open dense sets [15]. It requires extra effort to obtain results in the neighborhood of a point; see for example [21]. In this paper, too, we can only get local results by introducing regularity assumptions, typically in the form of rank conditions.

Recently, Fliess and coworkers have proposed a geometric framework using Lie–Bäcklund morphisms on infinite-dimensional jet spaces for studying flatness [10, 11]. In the latter paper Fliess and coworkers also introduce a more general notion of flatness called "orbital (or topological) flatness" where the transformations do not necessarily preserve the independent variable. Pomet has also proposed a related approach to differential flatness in [23]. Both approaches closely capture some features of the differential algebra while providing a geometric framework that enables local analysis. These approaches differ from ours in that they do not make extensive use of tools from exterior differential systems.

The organization of the paper is as follows. In section 2 we introduce the definitions pertaining to absolute equivalence and their interpretation in control theory. In section 3 we introduce our definition of differential flatness and show how to recover the differential algebraic results. In section 4 we study the connections between flatness and feedback linearizability. In section 5 we present our main theorems characterizing flatness for single input systems, and in section 6 we summarize our results and point out some open questions.

2. Prolongations and control theory. This section introduces the concept of prolongations and states some basic theorems. It relates these concepts to control theory. Proofs of most of these results can be found in [26]. We assume that all manifolds and mappings are smooth (C^∞) unless explicitly stated otherwise.

DEFINITION 2.1 (Pfaffian system). *A* Pfaffian system I *on a manifold* M *is a submodule of the module of differential one-forms* $\Omega^1(M)$ *over the commutative ring of smooth functions* $C^\infty(M)$. *A set of one-forms* $\omega^1, \dots, \omega^n$ *generates a Pfaffian system* $I = \{\omega^1, \dots, \omega^n\} = \{\Sigma f_k \omega^k | f_k \in C^\infty(M)\}$.

In this paper, we restrict attention to finitely generated Pfaffian systems on finite-dimensional manifolds. It is important to distinguish between a Pfaffian system and its set of generators or the algebraic ideal $\mathcal{I}$ in $\Lambda(M)$ generated by I. Since we are only dealing with Pfaffian systems the term *system* will henceforth mean a Pfaffian system.

For a Pfaffian system I we can define its *derived system* $I^{(1)}$ as $I^{(1)} = \{\omega \in I | d\omega \equiv 0 \bmod \mathcal{I}\}$, where $\mathcal{I}$ is the algebraic ideal generated by I. The derived system is itself a Pfaffian system, so we can define the sequence $I, I^{(1)}, I^{(2)}, \dots$, which is called the *derived flag* of I.

ASSUMPTION 1 (regularity of Pfaffian systems). *Unless stated explicitly otherwise, we will assume throughout this paper that the system is* regular*; i.e.,*

1. *the system and all its derived systems have constant rank;*
2. *for each* k, *the exterior differential system generated by* $I^{(k)}$ *has a degree-*2 *part with constant rank.*

If the system is regular the derived flag is decreasing, so there will be an N such that $I^{(N)} = I^{(N+1)}$. This $I^{(N)}$ is called the *bottom derived system.*

When one studies the system of one-forms corresponding to a system of differential equations, the independent variable time becomes just another coordinate on the manifold along with the dependent variables. Hence the notion of an independent variable is lost. If x denotes the dependent variables, a solution to such a system $c : s \to (t(s), x(s))$ is a curve on the manifold. But we are only interested in solution curves that correspond to graphs of functions $x(t)$. Hence we need to reject solutions for which $\frac{dt}{ds}$ vanishes at some point. This is done by introducing dt as an *independence condition*, i.e., a one-form that is not allowed to vanish on any of the solution curves. An independence condition is well defined only up to a nonvanishing multiple and modulo I. We will write a system with independence condition τ as (I, τ). The form τ is usually exact, but it does not have to be. In this paper we shall always take τ exact, in agreement with its physical interpretation as time.

DEFINITION 2.2 (control system). *A Pfaffian system with independence condition (I, dt) is called a* control system *if $\{I, dt\}$ is integrable.*

In local coordinates, control systems can be written in the form

$$I = \{dx_1 - f_1(x, u, t)dt, \ldots, dx_n - f_n(x, u, t)dt\} \tag{2.1}$$

with states $\{x_1, \ldots, x_n\}$ and inputs $\{u_1, \ldots, u_p\}$. Note that a control system is always assumed to have independence condition dt. If the functions f are independent of time then we speak of a *time invariant* control system.

DEFINITION 2.3 (Cartan prolongation). *Let (I, dt) be a Pfaffian system on a manifold M. Let B be a manifold such that $\pi : B \to M$ is a fiber bundle. A Pfaffian system $(J, \pi^* dt)$ on B is a* Cartan prolongation *of the system (I, dt) if the following conditions hold:*

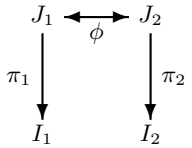
1. *$\pi^*(I) \subset J$;*
2. *for every integral curve of I, $c : (-\epsilon, \epsilon) \to M$, there is a unique lifted integral curve of J, $\tilde{c} : (-\epsilon, \epsilon) \to B$ with $\pi \circ \tilde{c} = c$.*

ASSUMPTION 2 (regularity of Cartan prolongations). *In this paper we only consider Cartan prolongations that preserve codimension.*

Note that all prolongations are required to preserve the independence condition of the original system. The above definition implies that there is a smooth one-to-one correspondence between the integral curves of a system and its Cartan prolongation. Cartan prolongations are useful to study equivalence between systems of differential equations that are defined on manifolds of different dimensions. This occurs in dynamic feedback extensions of control systems. We increase the dimension of the state by adding dynamic feedback, but the extended system is still in some sense equivalent to the original system.

This allows us to define the concept of absolute equivalence introduced by Elie Cartan as follows.

DEFINITION 2.4 (absolute equivalence). *Two systems I_1 and I_2 are* absolutely equivalent *if they have Cartan prolongations J_1 and J_2, respectively, that are equivalent in the usual sense; i.e., there exists a diffeomorphism ϕ such that $\phi^*(J_2) = J_1$. This is illustrated in the following diagram.*



An interesting subclass of Cartan prolongations is formed by *prolongations by differentiation*: if (I, dt) is a system with independence condition on M, if du is an exact one-form on M that is independent of $\{I, dt\}$, and if y is a fiber coordinate of B, then $\{I, du - ydt\}$ is called a *prolongation by differentiation* of I. Note that we have not written $\pi^*(du - ydt)$ where $\pi : B \to M$ is the surjective submersion. We will make this abuse in the rest of the paper for notational convenience. Prolongations by differentiation correspond to adding integrators to a system. In the context of control systems, the coordinate u is the input that is differentiated.

If we add integrators to all controls, we obtain a *total prolongation*: let (I, dt) be a system with independence condition, where $\dim\ I = n$. Let $\dim M = n + p + 1$. Let $u_1, \ldots, u_p$ be coordinates such that $du_1, \ldots, du_p$ are independent of $\{I, dt\}$, and let $y_1, \ldots, y_p$ be fiber coordinates of B; then $\{I, du_1 - y_1 dt, \ldots, du_p - y_p dt\}$ is called a *total prolongation* of I. Total prolongations can be defined independent of coordinates and are therefore intrinsic geometric objects. It can be shown that in codimension 2 (i.e., a system with n generators on an $n + 2$-dimensional manifold), all Cartan prolongations are locally equivalent to total prolongations [26].

We define a *dynamic feedback* to be a feedback of the form

$$\begin{aligned} \dot{z} &= a(x, z, v, t), \\ u &= b(x, z, v, t). \end{aligned}$$

If t does not appear in (a, b) we call (a, b) a *time invariant* dynamic feedback. The dynamic feedback is called *regular* if for each fixed x and t the map $b(x, ., ., t) : (z, v) \mapsto u$ is a submersion. An important question is what type of dynamic feedback corresponds to what type of prolongation. Clearly, prolongations by differentiation correspond to dynamic extension (adding integrators to the inputs).

Cartan prolongations provide an intrinsic, geometric way to study dynamic feedback. We shall show that Cartan prolongations that extend a control system to another control system can be expressed as dynamic feedback in local coordinates. The following example shows that not every dynamic feedback corresponds to a Cartan prolongation.

EXAMPLE 1 (dynamic feedback versus Cartan prolongation). *Consider the control system*

$$\dot{x}_1 = u$$

with feedback

$$\begin{aligned} \dot{z}_1 &= z_2, \\ \dot{z}_2 &= -z_1, \\ u &= g(z)v. \end{aligned}$$

This dynamic feedback introduces harmonic components that can be used to asymptotically stabilize nonholonomic systems (see [6] *for a description of how this might be done). It is not a Cartan prolongation since (z, v) cannot be uniquely determined from (x, u).*

The feedback in Example 1 is somewhat unusual, in that most theorems concerning dynamic feedback are restricted to adding some type of integrator to the inputs of the system. The particular property that this feedback is missing is defined in the following definition.

DEFINITION 2.5 (endogenous feedback). *Let $\dot{x} = f(x, u, t)$ be a control system.*

A dynamic feedback

$$\begin{aligned} \dot{z} &= a(x, z, v, t), \\ u &= b(x, z, v, t), \end{aligned} \tag{2.2}$$

is said to be endogenous *if z and v satisfying* (2.2) *can be expressed as functions of x, u, t, and a finite number of their derivatives*

$$\begin{aligned} z &= \alpha(x, u, \dots, u^{(l)}, t), \\ v &= \beta(x, u, \dots, u^{(l)}, t). \end{aligned} \tag{2.3}$$

An endogenous feedback is called regular *if for each fixed x and t the map $b(x, ., ., t) : (z, v) \mapsto u$ is a submersion.*

Note that this differs slightly from the definition given in [18, 19] due to the explicit time dependence used here. The feedback in Example 1 is not endogenous. The relationship between Cartan prolongations and endogenous dynamic feedback is given by the following two theorems. The first says that a regular endogenous feedback corresponds to a Cartan prolongation.

THEOREM 2.6 (endogenous feedback are Cartan prolongations). *Let I be a control system on an open set $T \times X \times U$ which in coordinates (t, x, u) is given by $\dot{x} = f(x, u, t)$. Let J denote the control system on the open set $T \times X \times Z \times V$ which is obtained from the above system by adding a regular endogenous dynamic feedback. Then J is a Cartan prolongation of I.*

Proof. Define the mapping $F : T \times X \times Z \times V \to T \times X \times U$ by $F(t, x, z, v) = (t, x, b(x, z, v, t))$. Since b is regular, F is a submersion. Furthermore b is surjective since the feedback is endogenous. Therefore F is surjective too. Since F is a surjective submersion, $T \times X \times Z \times V$ is fibered over $T \times X \times U$. Hence we have that solutions $(t, x(t), z(t), v(t))$ of J project down to solutions $(t, x(t), b(x(t), z(t), v(t), t))$ of I. Therefore the first requirement of being a Cartan prolongation is satisfied. The second requirement of unique lifting is trivially satisfied by the fact that z and v are obtained uniquely by equation (2.3). ∎

Conversely, a Cartan prolongation can be realized by endogenous dynamic feedback in an open and dense set if the resulting prolongation is a control system as described in Theorem 2.7.

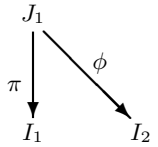
THEOREM 2.7 (Cartan prolongations are locally endogenous feedback). *Let I be a control system on a manifold M with p inputs $\{u_1, \dots, u_p\}$. Every Cartan prolongation $J = \{I, \omega_1, \dots, \omega_r\}$ on B with independence condition dt such that J is again a control system is realizable by endogenous regular feedback on an open and dense set of B.*

Proof. Let r denote the fiber dimension of B over M, and let $\{w_1, \dots, w_r\}$ denote the fiber coordinates. Since I is a control system, $\{I, dt\}$ is integrable, and we can find n first integrals $x_1, \dots, x_n$. Preservation of the codimension and integrability of $\{J, dt\}$ means that we can find r extra functions $a_1, \dots, a_r$ such that $J = \{I, dz_1 - a_1 dt, \dots, dz_r - a_r dt\}$. Here the z_i are first integrals of $\{J, dt\}$ that are not first integrals of $\{I, dt\}$. Pick p coordinates $v(u, w)$ such that $\{t, x, z, v\}$ form a set of coordinates of B. The v coordinates are the new control inputs. Clearly $a_i = a_i(x, z, v, t)$ since we have no other coordinates. Also since $\{t, x, z, v\}$ form coordinates for B, and u is defined on B, there has to be a function b such that $u = b(x, z, v, t)$. Since both (t, x, u, w) and (t, x, z, v) form coordinates on B, there has to be a diffeomorphism ϕ between the two. From the form of the matrix $\frac{\partial \phi}{\partial (t,x,z,v)}$ it can be seen that $\frac{\partial b}{\partial (z,v)}$ is full rank, and hence b is regular. This recovers the form of equation (2.2). Since J

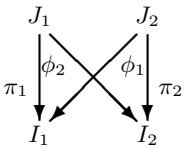
is a Cartan prolongation, every (x, u, t) is lifted to a unique (x, z, v, t). From Lemma 3.5, to be presented in the next section, it then follows that we can express (z, v) as functions of x and u and its derivatives in an open and dense set. We thus obtain the form of equation (2.3). □

3. Differentially flat systems. In this section we present a definition of flatness in terms of prolongations. Our goal is to establish a definition of flatness in terms of differential geometry while capturing the essential features of flatness in differential algebra [8, 9]. We build our definition on the minimal requirements needed to recover these features, namely, the one-to-one correspondence between solution curves of the original system and an unconstrained system, while maintaining regularity of the various mappings. Our definition makes use of the concept of an absolute morphism [26].

DEFINITION 3.1 (absolute morphism). *An* absolute morphism *from a system (I_1, dt) on M_1 to a system (I_2, dt) on M_2 consists of a Cartan prolongation (J_1, dt) on $\pi : B_1 \to M_1$ together with surjective submersion $\phi : B_1 \to M_2$ such that $\phi^*(I_2) \subset J_1$. This is illustrated below.*



DEFINITION 3.2 (invertibly absolutely morphic systems). *Two systems (I_1, dt) and (I_2, dt) are said to be* absolutely morphic *if there exist absolute morphisms from (I_1, dt) to (I_2, dt) and from (I_2, dt) to (I_1, dt). This is illustrated below.*



Two systems (I_1, dt) and (I_2, dt) are said to be invertibly absolutely morphic *if they are absolutely morphic and the following inversion property holds: let $c_1(t)$ be an integral curve of I_1 with $\tilde{c}_1$ the (unique) integral curve of J_1 such that $c_1 = \pi \circ \tilde{c}_1$, and let $\gamma(t) = \phi_2 \circ \tilde{c}_1(t)$ be the projection of $\tilde{c}_1$. Then we require that $c_1(t) = \phi_1 \circ \tilde{\gamma}(t)$, where $\tilde{\gamma}(t)$ is the lift of γ from I_2 to J_2. The same property must hold for solution curves of I_2.*

If two systems are invertibly absolutely morphic, then the integral curves of one system map to the integral curves of the other, and this process is invertible in the sense described above. If two systems are absolutely equivalent then they are also absolutely morphic, since they can both be prolonged to systems of the same dimension which are diffeomorphic to each other. However, for two systems to be absolutely morphic we do not require that any of the systems have the same dimension.

A differentially flat system is one in which the "flat outputs" completely specify the integral curves of the system. More precisely, we have Definition 3.3.

DEFINITION 3.3 (differential flatness). *A system (I, dt) is* differentially flat *if it is invertibly absolutely morphic to the trivial system $I_t = (\{0\}, dt)$.*

Notice that we require that the independence condition be preserved by the absolute morphisms, and hence our notion of time is the same for both systems. Since an independence condition is only well defined up to nonvanishing multiples and modulo

the system, we do allow time scalings between the systems. We also allow time to enter into the absolute morphisms that map one system onto the other.

If the system (I, dt) is defined on a manifold M, then we can restrict the system to a neighborhood around a point in M, which is again itself a manifold. We will call a system flat in that neighborhood if the restricted system is flat.

The following discussion leans heavily on a theorem due to Shadwick and Sluis [25] and Sluis [26], which we recall here for completeness.

THEOREM 3.4. *Let I be a system on a manifold M and let J be a Cartan prolongation of I on $\pi : B \to M$. On an open and dense subset of B, there exists a prolongation by differentiation of J that is also a prolongation by differentiation of I.*

Proof. For the proof see [26, Theorem 24]. ∎

In order to establish the relationship between our definition and the differential algebraic notion of flatness, we need the following straightforward corollary to Theorem 3.4. This lemma expresses the dependence of the fiber coordinates of a Cartan prolongation on the coordinates of the base space.

LEMMA 3.5. *Let (I, dt) be a system on a manifold M with local coordinates $(t, x) \in \mathbb{R}^1 \times \mathbb{R}^n$, and let (J, dt) be a Cartan prolongation on the manifold B with fiber coordinates $y \in \mathbb{R}^r$. Assume the regularity Assumptions 1 and 2 hold. Then on an open dense set, each y_i can be uniquely determined from t, x, and a finite number of derivatives of x.*

Proof. By Theorem 3.4 there is a prolongation by differentiation, on an open and dense set, say, I_2, of J, with fiber coordinates z_i, that is also a prolongation by differentiation of the original system I, say, with fiber coordinates w_i. This means that the (x, y, z, t) are diffeomorphic to (x, w, t): $y = y(x, w, t)$. The w are derivatives of x, and therefore the claim is proven. ∎

This lemma allows us to explicitly characterize differentially flat systems in a local coordinate chart. Let a system in local coordinates (t, x) be differentially flat, and let the corresponding trivial system have local coordinates (t, y). Then on an open and dense set there are surjective submersions h and g with the following property: given any curve $y(t)$, then

$$x(t) = g(t, y(t), \dots, y^{(q)}(t))$$

is a solution of the original system and furthermore the curve $y(t)$ can be obtained from $x(t)$ by

$$y(t) = h(t, x(t), \dots, x^{(l)}(t)).$$

This follows from using definitions of absolute morphisms, the invertibility property, and Lemma 3.5, stating that fiber coordinates are functions of base coordinates and their derivatives and the independent coordinate.

This local characterization of differential flatness corresponds to the differential algebraic definition except that h and g need not be algebraic or meromorphic. Also, we do not require the system equations to be algebraic or meromorphic. The explicit time dependence corresponds to the differential algebraic setting where the differential ground field is a field of functions and is not merely a field of constants. The functions g and h now being surjective submersions enable us to link the concept of flatness to geometric nonlinear control theory where we usually impose regularity. We emphasize that we only required a one-to-one correspondence of solution curves a priori for our definition of flatness, and not that this dependence was in the form of derivatives. The particular form of this dependence followed from our analysis.

Finally, the following theorem allows us to characterize the notion of flatness in terms of absolute equivalence.

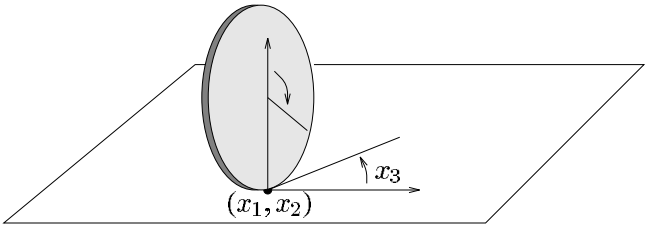


FIG. 3.1. *Rolling penny.*

THEOREM 3.6. *Two systems are invertibly absolutely morphic if and only if they are absolutely equivalent.*

Proof. Sufficiency is trivial. We shall prove necessity. For convenience we shall not mention independence conditions, but they are assumed to be present and do not affect the proof. Let I_1 on M_1 and I_2 on M_2 be invertibly absolutely morphic. Let J_1 on B_1 be the prolongation of I_1 with $\pi_1 : B_1 \to M_1$ and similarly J_2 on B_2 be the prolongation of I_2 with $\pi_2 : B_2 \to M_2$. Let the absolute morphisms be $\phi_1 : B_2 \to M_1$ and $\phi_2 : B_1 \to M_2$.

We now argue that J_2 is a Cartan prolongation of I_1 (and hence I_1 and I_2 are absolutely equivalent). By assumption ϕ_1 is a surjective submersion, and every solution $\tilde{c}_2$ of J_2 projects down to a solution c_1 of I_1 on M_1. The only extra requirement for J_2 on $\phi_1 : B_2 \to M_1$ to be a (Cartan) prolongation is that every solution c_1 of I_1 has a unique lift $\tilde{c}_2$ (on B_2) which is a solution of J_2.

To show existence of a lift, observe that for any given c_1 that is a solution of I_1, we can obtain its unique lift $\tilde{c}_1$ on B_1 (which solves J_1) and get its projection c_2 on M_2 (which solves I_2) and then consider its unique lift $\tilde{c}_2$ on B_2. Now it follows from the invertibility property that $\phi_1 \circ \tilde{c}_2 = c_1$. In other words, $\tilde{c}_2$ projects down to c_1.

To see the uniqueness of this lift, suppose $\tilde{c}_2$ and $\tilde{c}_3$, which are solutions of J_2 on B_2, both project down to c_1 on M_1. Consider their projections c_2 and c_3 (respectively) on M_2. When we lift c_2 or c_3 to B_2 and project down to M_1 we get c_1, which when lifted to B_1 gives, say, $\tilde{c}_1$. By the requirement that the absolute morphisms be invertible, $\tilde{c}_1$ should project down to (via ϕ_2) c_2 as well as c_3. Then uniqueness of projection implies that c_2 and c_3 are the same, which implies $\tilde{c}_2$ and $\tilde{c}_3$ are the same.

Hence J_2 is a Cartan prolongation of I_1 as well. Hence I_1 and I_2 are absolutely equivalent. □

Using this theorem we can completely characterize differential flatness in terms of absolute equivalence as follows.

COROLLARY 3.7. *A system (I, dt) is differentially flat if and only if it is absolutely equivalent to the trivial system $I_t = (\{0\}, dt)$.*

Note that we require the feedback equivalence to preserve time, since both systems have the same independence condition. In the classical feedback equivalence we only consider diffeomorphisms of the form $(t, x, u) \mapsto (t, \phi(x), \psi(x, u))$. For flatness we allow diffeomorphisms of the form $(t, x, u) \mapsto (t, \phi(t, x, u), \psi(t, x, u))$. We could allow time scalings of the form $t \mapsto s(t)$, but this does not change the independence condition and does therefore not gain any generality. In Cartan's notion of equivalence all diffeomorphisms are completely general. This is akin to the notion of *orbital* flatness presented in [10], where one allows time scalings dependent on all states and inputs.

EXAMPLE 2. *Consider the motion of a rolling penny, as shown in Figure* 3.1. *Let (x_1, x_2) represent the xy position of the penny on the plane, let x_3 represent the heading angle of the penny relative to a fixed line on the plane, and let x_4 represent*

the rotational velocity of the angle of Lincoln's head, i.e., the rolling velocity. We restrict $x_3 \in [0, \pi)$ *since we cannot distinguish between a positive rolling velocity* x_4 *at a heading angle* x_3 *and a negative rolling velocity at a heading angle* $x_3 + \pi$.

The dynamics of the penny can be written as a Pfaffian system described by

$$\begin{aligned} \omega^1 &= \sin x_3\, dx_1 - \cos x_3\, dx_2, \\ \omega^2 &= \cos x_3\, dx_1 + \sin x_3\, dx_2 - x_4 dt, \\ \omega^3 &= dx_3 - x_5 dt, \\ \omega^4 &= dx_4 - u_1 dt, \\ \omega^5 &= dx_5 - u_2 dt, \end{aligned} \tag{3.1}$$

where $x_5 = \dot{x}_3$ *is the velocity of the heading angle. The controls* u_1 *and* u_2 *correspond to the torques around the rolling and heading axes. We take* dt *as the independence condition.*

This system is differentially flat away from $x_4 = 0$ *using the outputs* x_1 *and* x_2 *plus knowledge of time. If* dx_1 *and* dx_2 *are not both zero, we can solve for* x_3 *using* ω_1. *Given these three variables plus time, we can solve for all other variables in the system by differentiation with respect to time. This argument also shows that the system is differentially flat, since we only need to know* (x_1, x_2) *and their derivatives up to order three in order to solve for all of the states of the system.*

Often we will be interested in a more restricted form of flatness that eliminates the explicit appearance of time that appears in the general definition.

DEFINITION 3.8. *An absolute morphism from a time invariant control system* (I_1, dt) *to a time invariant control system* (I_2, dt) *is a* time-independent absolute morphism *if locally the maps* $\pi : B_1 \to M_1$ *and* $\phi : B_1 \to M_2$ *in Definition* 3.1 *have the form* $(t, x, u) \mapsto (t, \eta(x, u), \psi(x, u))$*; i.e., the mappings between states and inputs do not depend on time. A system* (I, dt) *is* time-independent differentially flat *if it is differentially flat using time-independent absolute morphisms.*

Note that Example 2 is time-independent differentially flat. We may be tempted to think that if the control system I is time invariant and we know that the trivial system is time invariant, we can assume that the absolute morphism $x = \phi(t, y, y^{(1)}, \ldots, y^{(q)})$ has to be time independent as well. That this is not true is illustrated by the following example.

EXAMPLE 3. *Consider the system* $\dot{y} = ay$ *and the coordinate transformation* $y = x^2 e^{t+x}$. *Then* $\dot{x} = \frac{(a-1)x}{2+x}$. *Both systems are time invariant, but the coordinate transformation depends on time.*

4. Linear systems and linearizability. Conforming to the established literature [2], we define a *linear time-invariant* system as a system of the form

$$\begin{aligned} \dot{x} &= Ax + Bu, \\ y &= Cx + Du. \end{aligned} \tag{4.1}$$

Here (A, B, C, D) are matrices of appropriate dimensions. If the system is controllable, we can put it in *Brunovsky* normal form by a linear coordinate transformation on the states and a static feedback. A system is in Brunovsky normal form if we can write the dynamics as

$$z_i^{(l_i)} = u_i \tag{4.2}$$

for some new outputs z_i.

DEFINITION 4.1 (feedback linearizability). *The time invariant nonlinear system*

$$\dot{x} = f(x, u) \tag{4.3}$$

is feedback linearizable *if there is a* regular endogenous *dynamic feedback*

$$\begin{aligned} \dot{z} &= \alpha(x, z, v), \\ u &= \beta(x, z, v), \end{aligned} \tag{4.4}$$

and new coordinates $\xi = \phi(x, z)$ *and* $\eta = \psi(x, z, v)$ *such that in the new coordinates the system has the form*

$$\dot{\xi} = A\xi + B\eta \tag{4.5}$$

and the mapping ϕ *maps onto a neighborhood of the origin. If* $\dim z = 0$ *then we say the system is* static *feedback linearizable.*

REMARK 1. *It may seem overly restrictive to require the feedback* (4.4) *to be regular endogenous. It is however common practice to impose regularity on the feedback* (*see Definition* 6.1 *in* [22])*, and it avoids anomalies like setting inputs equal to a constant.*

The form in equation (4.5) is the standard form in linear systems theory. It is useful if one wants to design controllers for nonlinear systems around equilibrium points.

It might be that the system can be put in the form (4.5) but that the coordinate transformation is not valid in a neighborhood of the origin of the target system. In that case we can shift the origin of the linear system to put it in the form

$$\dot{\xi} = A\xi + B\eta + E. \tag{4.6}$$

This form is called linear in [20], but most results in linear systems theory cannot be applied since the origin is not an equilibrium point. However, it is still useful in the context of trajectory generation. For example, a nonholonomic system in chained form [21] can be transformed to this state space affine form.

It is clear that systems that are feedback linearizable (by regular endogenous feedback) are flat, since we can put them into Brunovsky normal form. The following theorem shows that the converse is also true in an open and dense subset. An analogous result was proven by Martin in a differential algebraic setting [18, 19] and has also been derived using the formalism of Lie–Bäcklund transformations [10, 11].

THEOREM 4.2. *Every differentially flat system can be put in Brunovsky normal form in an open and dense set through regular endogenous feedback.*

Proof. Let J, J_t be the Cartan prolongations of I, I_t, respectively. Then by Theorem 3.4, on an open and dense set there is a prolongation by differentiation of J_t that is also a prolongation by differentiation of I_t, say, J_{t1}. Let J_1 be the corresponding Cartan prolongation of J. Then J_1 is equivalent to J_{t1}, which is in Brunovsky normal form. In particular, since J_1 is a Cartan prolongation, it can be realized by regular endogenous feedback. □

This proof relies on Theorem 3.4, which restricts its validity to an open and dense set. We conjecture that the result holds everywhere, but the above proof technique does not allow us to conclude that. The obstruction lies in certain prolongations that we cannot prove to be regular.

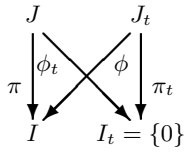
5. Flatness for single input systems. For single input control systems, the corresponding differential system has codimension 2. There are a number of results available in codimension 2 that allow us to give a complete characterization of differentially flat single input control systems. In codimension 2 every Cartan prolongation is a total prolongation around every point of the fibered manifold [26], given our regularity assumptions 1, 2. This allows us to prove the following theorem.

THEOREM 5.1. *Let I be a time invariant control system*

$$I = \{dx_1 - f_1(x,u)dt, \ldots, dx_n - f_n(x,u)dt\},$$

where u is a scalar control; i.e., the system has codimension 2. If I is time-independent differentially flat around an equilibrium point, then I is feedback linearizable by static time invariant feedback at that equilibrium point.

Proof. Let I be defined on M with coordinates (x,u,t), let the trivial system I_t be defined on B_t with coordinates (y_0,t), let the prolongation of I_t be J_t, and let J_t be defined on M_t. This is illustrated below.



First we show that J_t can be taken as a Goursat normal form around the equilibrium point. In codimension 2, every Cartan prolongation is a repeated total prolongation in a neighborhood of every point of the fibered manifold [26, Theorem 5]. Let $I_{t0} = I_t, I_{t1}, I_{t2}, \ldots$ denote the total prolongations starting at I_t, defined on fibered manifolds $B_{t0} = B_t, B_{t1}, \ldots$. If y_1 denotes the fiber coordinate of B_{t1} over B_{t0}, then I_{t1} has the form $\lambda dt + \mu dy_0$, where either λ or μ depends nontrivially on y_1. Since the last derived system of I does not drop rank at the equilibrium, neither does I_{t1} and we have that not both λ and μ vanish at the equilibrium. Now, $\mu \neq 0$ at the equilibrium point. This can be seen as follows: $y_0 \equiv c$ is a solution curve to I_t, and would not have a lift to I_{t1} if $\mu = 0$, since dt is required to remain the independence condition of all Cartan prolongations. From continuity we then have $\mu \neq 0$ around the equilibrium point. So we can define $y_1 := -\lambda/\mu$, and I_{t1} can be written as $dy_0 - y_1 dt$. We can continue this process for every Cartan prolongation, both of I_t and of I. This brings J_t in Goursat normal form in a neighborhood of the equilibrium point.

Now we will argue that we don't need to prolong I to establish equivalence. Since J is a Cartan prolongation, and therefore a total prolongation, its first derived system will be equivalent to the first derived system of J_t. Continuing this we establish equivalence between I and I_{tn}, where $I_{tn} = \{dy_0 - y_1 dt, \ldots, dy_{n-1} - y_n dt\}$. So we have $y = (y_0, \ldots, y_n) = y(x,u,t)$.

Next we will show that $y_0, \ldots, y_n$ are independent of time and that $y_0, \ldots, y_{n-1}$ are independent of u. By assumption y_0 is independent of time. Since the corresponding derived systems on each side are equivalent, $dy_0 - y_1 dt$ is equivalent to the last one-form in the derived flag of I. Since the differential du does not appear in this one-form, y_0 is independent of u. Analogously, $y_i, i = 1, \ldots, n-1$ are all independent of u. Since the $y_i, i = 1, \ldots, n$ are repeated derivatives of y_0, and since I is time invariant, these coordinates are also independent of time.

We still have to show that the mapping $x \mapsto y$ is a valid coordinate transformation. Suppose $dy_0, \ldots, dy_{n-1}$ are linearly dependent at the equilibrium. Then, J_t drops rank at the equilibrium, and since we have equivalence, so would I. But from the form of I we can see this is not the case.

Therefore $y_i = y_i(x), i = 0, \ldots, n-1, y_n = y_n(x,u)$ and the system J_t is just a chain of integrators with input y_n. The original system I is equivalent to this linear system by a coordinate transformation on the states and a state dependent and time invariant feedback. This coordinate transformation is well defined around the equilibrium point. It is therefore feedback linearizable by a static feedback that is

time invariant. Note that $\partial y_n/\partial u \neq 0$ because y_n is the only one of the y variables that depends on u. □

The conclusion that the feedback can be taken static goes back to [5, 23].

EXAMPLE 4. *Notice that in our definition the system*

$$\begin{aligned} \dot{x}_2 &= u, \\ \dot{x}_1 &= x_2^3, \\ y &= x_1, \end{aligned} \tag{5.1}$$

is not flat around the origin, because we get $u = \frac{\ddot{y}}{3\,\dot{y}^{2/3}}$ *so that curves with* $\dot{y} = 0$ *and* $\ddot{y} \neq 0$ *have no lift. It is also not feedback linearizable at the origin.*

We will now show that in the case of a time invariant system we don't need the assumption of time invariant flatness to conclude static feedback linearizability. We will require the following preliminary result, which appeared in a proof in [24].

LEMMA 5.2. *Given a one-form* $\alpha = A_i(x,u)dx_i - A_0(x,u)dt$ *(using implicit summation) on a manifold* M *with coordinates* (x,u,t)*, and suppose we can write* $\alpha = dX(x,u,t) - U(x,u,t)dt$*. Then we can also write* α *as* $\alpha = dY(x) - V(x,u)dt$*; i.e., we can take the function* X *independent of time and the input, and we can take* U *independent of time. If we know in addition that* $\alpha = A_i(x)dx_i - A_0(x)dt$*, then we can scale* α *as* $\alpha = dY(x) - V(x)dt$*; i.e., we can take* V *independent of* u *as well.*

Proof. For the proof see [24]. □

The following theorem seems to be implied in [24], but the proof there refers to a general discussion of Cartan's method of equivalence as applied to control systems in [13]. We work out the proof for this special case.

THEOREM 5.3. *A controllable single input time invariant control system is differentially flat if and only if it is feedback linearizable by static, time invariant feedback.*

Proof. The proof of sufficiency follows trivially from controllability, so we shall only prove necessity. Let the control system be $I = \{dx_1 - f_1(x,u)dt, \ldots, dx_n - f_n(x,u)dt\}$, where u is a scalar control; i.e., the system has codimension 2. Let $\{\alpha^i, i = 1,\ldots,n\}$ and $\{\alpha_t^i, i = 1,\ldots,n\}$ be one-forms adapted to the derived flag of I, I_t, respectively. Thus $I^{(i)} = \{\alpha^1, \ldots, \alpha^{n-i}\}$ and $I_t^{(i)} = \{\alpha_t^1, \ldots, \alpha_t^{n-i}\}$. Since I does not contain the differential du, the forms $\alpha^1, \ldots, \alpha^{n-1}$ can be taken to be independent of u. Since I is time invariant, the forms $\alpha_1, \ldots, \alpha_n$ can be chosen to be independent of time. We can thus invoke the second part of Lemma 5.2 for the forms $\alpha^1, \ldots, \alpha^{n-1}$.

Assume $n \geq 2$. As in Theorem 5.1 we have equivalence between α^1 and $\alpha_t^1 = dy_0(x,t) - y_1(x,t)dt$ (if $n = 1$ we have $y_n = y_n(x,u,t)$, which we will reach eventually). Since I is time invariant we can choose α^1 time independent: $\alpha^1 = A_i(x)dx_i - A_0(x)dt$. From Lemma 5.2 we know that we can write α^1 as $dY_0 - Y_1 dt$ where Y_0, Y_1 are functions of x only.

Again according to Lemma 5.2, we can write $\alpha^2 = dV(x) - W(x)dt$. Now from

$$\begin{aligned} 0 &= d\alpha^1 \wedge \alpha^1 \wedge \alpha^2 \\ &= -dY_1 \wedge dt \wedge dY_0 \wedge dV \end{aligned}$$

we know $V = V(Y_1, Y_0)$. And from

$$\begin{aligned} 0 &\neq d\alpha^2 \wedge \alpha^1 \wedge \alpha^2 \\ &= -dW \wedge dt \wedge dY_0 \wedge dV \end{aligned}$$

we know that $\gamma_1 := \partial V/\partial Y_1 \neq 0$. Then writing $\gamma_0 := \partial V/\partial Y_0$ (and $\simeq$ denotes

equivalence in the sense that both systems generate the same ideal),

$$\begin{aligned}\{\alpha^1, \alpha^2\} &\simeq \{dY_0 - Y_1 dt, \gamma_1 dY_1 + \gamma_0 dY_0 - W dt\} \\ &\simeq \{dY_0 - Y_1 dt, \gamma_1 dY_1 + \gamma_0 Y_1 dt - W dt\} \\ &\simeq \{dY_0 - Y_1 dt, dY_1 - (-\gamma_0 Y_1 + W)/\gamma_1 dt\} \\ &:= \{dY_0 - Y_1 dt, dY_1 - Y_2 dt\}, \end{aligned} \tag{5.2}$$

where Y_2, defined to be $Y_2 = (-\gamma_0 Y_1 + W)/\gamma_1$, is independent of (t, u) since $(\gamma_1, \gamma_0, Y_1, W)$ are. One can continue this procedure, at each step defining a new coordinate Y_i. In the last step the variable $W = W(x, u)$ (this will also be the first step if $n = 1$), and therefore Y_n depends on u nontrivially. Hence we obtain equivalence between I and $\{dY_0 - Y_1 dt, \ldots, dY_{n-1} - Y_n dt\}$ with $Y_i = Y_i(x), i = 0, \ldots, n-1$, and $Y_n = Y_n(x, u)$, i.e., feedback linearizability by static time invariant feedback. □

COROLLARY 5.4. *If a time invariant single input system is differentially flat we can always take the flat output as a function of the states only: $y = y(x)$.*

None of these results extend easily to higher codimensions. The reason for this is that only in codimension 2 can we find regularity assumptions on the original system such that every Cartan prolongation is a total prolongation. This is related to the well-known fact that for SISO systems static linearizability is equivalent to dynamic linearizability. For MIMO systems we cannot express these regularity conditions on the original system: we have to check regularity on the prolonged systems.

6. Concluding remarks. We have presented a definition of flatness in terms of the language of exterior differential systems and prolongations. Our definition remains close to the original definition due to Fliess [8, 9], but it involves the notion of a preferred coordinate corresponding to the independent variable (usually time).

Using this framework we were able to recover all results in the differential algebra formulation. In particular we showed that flat systems can be put in linear form in an open and dense set. This set need not contain an equilibrium point, and this linearizability therefore does not allow one to use most methods from linear systems theory. In other words, although flatness implies a linear *form*, it does not necessarily imply a linear *structure*. For a SISO flat system we resolved the regularity issue and established feedback linearizability around an equilibrium point. We also resolved the time dependence of flat outputs in the SISO case.

The most important open question is a characterization of flatness in codimension higher than 2.

Acknowledgments. The authors would like to thank Willem Sluis for many fruitful and inspiring discussions and for introducing us to Cartan's work and its applications to control theory. We also thank Shankar Sastry for valuable comments on this paper, and Philippe Martin for several useful discussions which led to a more complete understanding of the relationship between endogenous feedback and differential flatness. The reviewers provided many valuable comments which helped improve several specific results and the overall presentation.

REFERENCES

[1] R. BRYANT, S. CHERN, R. GARDNER, H. GOLDSCHMIDT, AND P. GRIFFITHS, *Exterior Differential Systems*, Springer-Verlag, New York, 1991.

[2] F. CALLIER AND C. DESOER, *Linear System Theory*, Springer-Verlag, New York, 1991.

[3] E. CARTAN, *Sur l'équivalence absolue de certains systèmes d'équations différentielles et sur certaines familles de courbes*, in Œuvres Complètes, Vol. II, Gauthier–Villars, Paris, 1953, pp. 1133–1168.

[4] E. Cartan, *Sur l'intégration de certains systèmes indéterminés d'équations différentielles*, in Œuvres Complètes, Vol. II, Gauthier–Villars, Paris, 1953, pp. 1169–1174.
[5] B. Charlet, J. Lévine, and R. Marino, *On dynamic feedback linearization*, Systems Control Lett., 13 (1989), pp. 143–151.
[6] J. Coron, *Global stabilization for controllable systems without drift*, Math. Control Signals Systems, 5 (1992), pp. 295–312.
[7] M. Fliess, *Generalized controller canonical forms for linear and nonlinear dynamics*, IEEE Trans. Automat. Control, 35 (1990), pp. 994–1001.
[8] M. Fliess, J. Lévine, P. Martin, and P. Rouchon, *On differentially flat nonlinear systems*, in IFAC Symposium on Nonlinear Control Systems Design (NOLCOS), IFAC, Laxenburg, Austria, 1992, pp. 408–412.
[9] M. Fliess, J. Lévine, P. Martin, and P. Rouchon, *Sur les systèmes non linéaires différentiellement plats*, C.R. Acad. Sci. Paris Sér. I, 315 (1992), pp. 619–624.
[10] M. Fliess, J. Lévine, P. Martin, and P. Rouchon, *Linéarisation par bouclage dynamique et transformations de Lie-Bäcklund*, C.R. Acad. Sci. Paris Sér. I, 317 (1993), pp. 981–986.
[11] M. Fliess, J. Lévine, P. Martin, and P. Rouchon, *Nonlinear control and Lie-Bäcklund transformations: Toward a new differential geometric standpoint*, in Proc. of the 1994 IEEE Control and Decision Conference, IEEE Control Systems Society, Piscataway, NJ, 1994, pp. 339–344.
[12] M. Fliess, J. Lévine, P. Martin, and P. Rouchon, *Flatness and defect of non-linear systems: Introductory theory and examples*, Internat. J. Control, 61 (1995), pp. 1327–1361.
[13] R. Gardner and W. Shadwick, *Symmetry and the implementation of feedback linearization*, Systems Control Lett., 15 (1991), pp. 25–33.
[14] R. Gardner and W. Shadwick, *The GS algorithm for exact linearization to Brunovsky normal form*, IEEE Trans. Automat. Control, 37 (1992), pp. 224–230.
[15] A. Giaro, A. Kumpera, and C. Ruiz, *Sur la lecture correcte d'un résultat d'Elie Cartan*, C.R. Acad. Sci. Paris Sér. A, 287 (1978), pp. 241–244.
[16] A. Isidori, *Nonlinear Control Systems*, Springer-Verlag, New York, 1989.
[17] B. Jacubczyk, *Invariants of dynamic feedback and free systems*, in Proc. European Control Conference, IFAC, Laxenburg, Austria, 1993, pp. 1510–1513.
[18] P. Martin, *Contribution à l'étude des systèmes différentiellement plats*, Ph.D. thesis, L'Ecole Nationale Supérieure des Mines de Paris, 1993.
[19] P. Martin, *Endogenous feedbacks and equivalence*, in Mathematical Theory of Networks and Systems, Regensburg, Germany, 1993.
[20] P. Martin and P. Rouchon, *Feedback linearization and driftless systems*, Math. Control Signals Systems, 7 (1994), pp. 235–254.
[21] R. Murray, *Nilpotent bases for a class of non-integrable distributions with applications to trajectory generation for nonholonomic systems*, Math. Control Signals Systems, 7 (1994), pp. 58–75.
[22] H. Nijmeijer and A. van der Schaft, *Nonlinear Dynamical Control Systems*, Springer-Verlag, New York, 1990.
[23] J. Pomet, *A differential geometric setting for dynamic linearization*, in Geometry in Nonlinear Control and Differential Inclusions 32, W. R. B. Jacubczyk and T. Rzezuchowski, eds., Banach Center Publications, Warsaw, 1995, pp. 319–339.
[24] W. Shadwick, *Absolute equivalence and dynamic feedback linearization*, Systems Control Lett., 15 (1990), pp. 35–39.
[25] W. Shadwick and W. Sluis, *On E. Cartan's absolute equivalence of differential systems*, C.R. Acad. Sci. Paris Sér. I, 313 (1991), pp. 455–459.
[26] W. Sluis, *Absolute Equivalence and Its Applications to Control Theory*, Ph.D. thesis, University of Waterloo, Waterloo, Ontario, 1992.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 4, pp. 1240–1249, July 1998

007

GLOBAL ERROR BOUNDS FOR CONVEX INEQUALITY SYSTEMS IN BANACH SPACES*

SIEN DENG†

Abstract. We study conditions under which a global error bound in terms of a natural residual exists for a convex inequality system. Specifically, we obtain an error bound result, which unifies many existing results assuming a Slater condition. We also derive two characterizations for a convex inequality system to possess a global error bound; one is in terms of metric regularity, and the other is in terms of an associated convex inequality system. As a consequence, we show that in $\mathbb{R}^n$ a global error bound holds for such a system under the assumption of the zero vector in the relative interior of the domain of an associated conjugate function along with metric regularity at every point of the feasible set defined by the system. Finally, we discuss some applications of these results to convex programs.

Key words. error bounds, metric regularity, relative interior, Hausdorff distance, weak sharp minima

AMS subject classifications. 90C31, 90C25, 49J52

PII. S0363012995293645

1. Introduction.

This paper deals with the convex inequality system

$$f(x) \le 0, \quad x \in C \subset X, \tag{1}$$

where X is a Banach space, C is a nonempty closed convex set, and f is a continuous convex function on X. Let S be the set of solutions to (1). We assume throughout that S is nonempty. In this paper, we are interested in knowing conditions under which the following global error bound holds for S: there is a positive constant τ such that

$$d(x, S) \le \tau [f(x)]_+ \quad \text{for all } x \in C, \tag{2}$$

where $d(x, S) = \inf_{y \in S} \|x - y\|$, $[f(x)]_+ = \max\{f(x), 0\}$, and $\|\cdot\|$ denotes the norm on X. When $X = \mathbb{R}^n$, $\|\cdot\|$ denotes the usual Euclidean norm on $\mathbb{R}^n$.

Error bounds in the form of (2), which are expressed in terms of a constant multiple of a natural residual, have found many important applications in sensitivity analysis of convex programs and complementarity problems, and in the convergence analysis of some descent methods. See [21, 24, 25, 26, 28, 5] for more details.

The Slater condition, which postulates the existence of an $x_0 \in C$ such that $f(x_0) < 0$, plays an important role in establishing the global error bound (2). In a normed linear space setting, Robinson [29] proved that (2) holds when S is bounded and (1) satisfies the Slater condition. In a finite-dimensional space setting, Mangasarian [23] established (2) when f is a pointwise maximum of finitely many convex differentiable functions, and (1) satisfies the Slater condition and an asymptotic constraint qualification. Auslender and Crouzeix [1] extended Mangasarian's result by relaxing the differentiability assumptions. Luo and Luo [19] established (2) for convex quadratic systems under the Slater condition. In a Banach space setting, Deng [6]

* Received by the editors October 25, 1995; accepted for publication (in revised form) June 23, 1997; published electronically May 15, 1998.
http://www.siam.org/journals/sicon/36-4/29364.html
† Department of Mathematical Sciences, Northern Illinois University, DeKalb, IL 60115 (deng@math.niu.edu).

established (2) under a Slater condition on the associated recession functions. Klatte [13, 14] established (2) under certain Hausdorff continuity assumptions. Li [17] derived an interesting characterization of metric regularity of a convex differentiable inequality system in terms of the Abadie constraint qualification and showed that the Abadie constraint qualification holding at every point of S is a necessary and sufficient condition for (2) to hold when f is a pointwise maximum of finitely many convex quadratic functions. A very recent paper of Lewis and Pang [15] gives an excellent survey on this active research area. By establishing an interesting necessary and sufficient condition involving directional derivatives first, Lewis and Pang [15] derived many interesting results. For other related error bound results, see [3, 8, 9, 12, 16, 20, 22, 27, 32] and references therein.

The purpose of this note is twofold. First, we establish an error bound result involving level set conditions of associated functions. This result unifies and extends many existing results assuming a Slater condition. Second, we derive two characterizations for the existence of the global error bound (2). This part of the research is partly motivated by the recent work of Li [17]. In a Banach space setting, we show that the global error bound (2) holds if and only if (1) is metrically regular at S; see section 2.2 for the definition of metric regularity at a set. When $X = \mathbb{R}^n$ and the affine hull of the domain of the conjugate function f^* is a proper subset of $\mathbb{R}^n$, we obtain another characterization by relating (1) to an associated system. We show that the existence of the global error bound (2) for (1) is equivalent to that for the associated system. As a consequence, we show that (2) holds when the zero vector is in the relative interior of the domain of the conjugate function f^* and (1) is metrically regular at every point of S (see section 2.2 for more details). In section 3, by applying the aforementioned results, we obtain some sufficient conditions for weak sharp minima of convex programs. Our results are complementary to the work done by Lewis and Pang [15] and Li [17].

We now describe the notation and some of the basic concepts used in this paper.

For a nonempty convex set U in $\mathbb{R}^n$, we denote by int (U), ri (U), aff (U), and cl (U), the interior of U, the relative interior of U, the affine hull of U, and the closure of U, respectively. For a nonempty closed convex set U in X, we denote by U^∞ the recession cone of U. For any two nonempty closed sets U_1 and U_2 in X, we define the Hausdorff distance between them as

$$\text{haus}\,(U_1, U_2) = \max\left\{\sup_{x\in U_1} d(x, U_2), \sup_{x\in U_2} d(x, U_1)\right\}.$$

A closed proper convex function g on X is an everywhere defined function with values in $(-\infty, +\infty]$, not identically $+\infty$, such that epi g is a closed convex set in $X \times \mathbb{R}$, where epi g denotes the epigraph of g. Its *effective* domain is the nonempty convex set

$$\text{dom}\, g = \{x \in X \mid g(x) < +\infty\}.$$

For a closed proper convex function g, we denote by $\partial g(x)$, g^*, and g^∞ the subdifferential of g at x $(\in \text{dom}\, g)$, the conjugate function of g, and the recession function of g, respectively.

2. Main results. We study the existence of the global error bound (2). Section 2.1 deals with the case where the Slater condition is satisfied. Section 2.2 deals with the case without such an assumption.

2.1. A level set condition result. We begin with an error bound result assuming certain level set conditions on f for the system (1).

PROPOSITION 1. *Suppose that X is a Banach space, C is a nonempty closed convex set in X, f is continuous and convex on X, and $S = \{x \in C \mid f(x) \le 0\}$. Suppose that there are positive scalars δ and Δ such that*

(a) $\tilde{S} = \{x \in C \mid f(x) \le -\delta\}$ *is nonempty, and*

(b) haus $(\tilde{S}, S) \le \Delta$.

Then

$$d(x, S) \le \delta^{-1}\Delta[f(x)]_+ \qquad \text{for all } x \in C.$$

Before proving Proposition 1, we would like to point out that conditions (a) and (b) hold whenever S is nonempty and bounded, and the system $f(x) \le 0$ satisfies the Slater condition. Moreover, when S is bounded, Δ can be chosen as the diameter of S. Thus Proposition 1 extends Robinson's result [29] in this setting. The proof technique is a refinement of that used in [29].

Proof. We only have to show that (2) holds for any $x \in C$, but $x \notin S$. Given $x \notin S$ and $x \in C$, since the projection of x onto S may not exist, we have to use a limit argument. For any $n > 0$, there is some $y(n) \in \tilde{S}$ such that

$$\|x - y(n)\| \le d(x, \tilde{S}) + 1/n.$$

By a similar argument, the distance between x and $\tilde{S}$ can be bounded by $d(x, S) + \Delta + 1/n$; that is,

$$d(x, \tilde{S}) \le d(x, S) + \Delta + 1/n.$$

Consequently,

$$\|x - y(n)\| \le d(x, S) + \Delta + 2/n. \tag{3}$$

With this given $y(n)$, let

$$x - z(n) = f(x)(f(x) + \delta)^{-1}(x - y(n)). \tag{4}$$

Then $z(n) \in S$ by the convexity of f and C. Since

$$\|x - y(n)\| = \|x - z(n)\| + \|z(n) - y(n)\|, \quad \text{and } d(x, S) \le \|z(n) - x\|, \tag{5}$$

it follows that

$$\begin{aligned} \|z(n) - y(n)\| &= \|x - y(n)\| - \|z(n) - x\| \\ &\le d(x, S) + \Delta + 2/n - d(x, S) \qquad \text{(by (3) and (5))} \\ &\le \Delta + 2/n. \end{aligned} \tag{6}$$

By (4), $\delta(x - z(n)) = f(x)[(x - y(n)) - (x - z(n))]$. Therefore,

$$\begin{aligned} d(x, S) &\le \|x - z(n)\| \\ &\le \delta^{-1} f(x)\|y(n) - z(n)\| \\ &\le \delta^{-1}(\Delta + 2/n)[f(x)]_+ \qquad \text{(by (6))}. \end{aligned}$$

By letting $n \to +\infty$, we obtain the desired inequality. ∎

Remarks. In an early version of this paper, the above result was proved under the assumption that the Banach space X is reflexive. A suggestion of J. Burke led to the present formulation and its proof.

Many known global error bounds assuming the Slater condition can be derived from Proposition 1. By invoking Proposition 1, we can prove the following corollaries. Corollary 1 can be used to show that the global error bound (2) holds for S when $C = X = \mathbb{R}^n$, f is well-posed, and (1) satisfies the Slater condition. Recall [1, 15] that f is well posed if every stationary sequence of f is a minimizing sequence; that is,

$$\left[\lim_{k\to\infty} u^k = 0,\ u^k \in \partial f(x^k), \forall k\right] \Rightarrow \left[\ \lim_{k\to\infty} f(x^k) = \inf_{x\in X} f(x)\right].$$

COROLLARY 1. *For the system* (1)*, suppose that X is a Banach space, C is a nonempty closed convex set in X, S is nonempty, and the global error bound* (2) *holds. Then, for any $\lambda > 0$,*

$$d(x, S_\lambda) \le \tau[f(x) - \lambda]_+ \qquad \text{for all } x \in C,$$

where $S_\lambda = \{x \in C \mid f(x) \le \lambda\}$ and τ is the multiplicative constant in (2).

COROLLARY 2 (see [6]). *Suppose that X is a Banach space, C is a nonempty closed convex set in X, and there is a unit vector $u \in C^\infty$ such that $f^\infty(u) \le -\tau^{-1}$. Then*

$$d\,(x, S) \le \tau[f(x)]_+ \qquad \text{for all } x \in C.$$

Proof. We only need to show that conditions (a) and (b) in Proposition 1 hold with haus $(\tilde{S}, S) \le \tau$, where $\tilde{S} = \{x \in C \mid f(x) \le -1\}$. For any $x \in S$, since

$$f(x + \tau u) \le f(x + \tau u) - f(x) \le \tau f^\infty(u) \le -1,$$

haus $(\tilde{S}, S) \le \tau$ follows from the fact that $x + \tau u \in \tilde{S}$. □

As a referee noted, a direct proof of Corollary 2 can be given by using the inequality $f(x + \lambda u) \le f(x) - \tau^{-1}\lambda$ $(\forall x,\ \lambda \ge 0)$ and putting $\lambda = \tau f(x)$ for $x \in C \backslash S$.

COROLLARY 3. *Suppose that $C = X = \mathbb{R}^n$. Let F be a vector-valued mapping from $\mathbb{R}^n$ to $\mathbb{R}^m$ with each component of F a finite convex function. Let $S = \{x \mid F(x) \le 0\}$. Suppose that there is a vector $b \in -\text{int}\,(\mathbb{R}^m_+)$ and a positive scalar $\tilde{\Delta}$ such that*

(a) $\tilde{S} = \{x \mid F(x) \le b\}$ *is nonempty, and*

(b) haus $(\tilde{S}, S) \le \tilde{\Delta}$.

Then

$$d(z, S) \le \tilde{\delta}^{-1}\tilde{\Delta}\|[F(z)]_+\|_\infty \qquad \text{for all } z \in \mathbb{R}^n,$$

where $\tilde{\delta} = \min_{1\le i\le m}\{-b_i\}$ and b_i is the ith component of b, and $\|\cdot\|_\infty$ denotes the ∞-norm on $\mathbb{R}^m$.

Proof. Let $f(x) = \max_{1\le i\le m}\{f_i(x)\}$, where f_i are components of F for $1 \le i \le m$. Clearly $\{x \mid f(x) \le 0\} = S$. Since $-\tilde{\delta} = -\min_{1\le i\le m}\{-b_i\} = \max_{1\le i\le m}\{b_i\}$, it follows that, for any $x \in X$, $f(x) \le -\tilde{\delta}$ whenever $f_i(x) \le b_i$ for all $1 \le i \le m$. That is, $\bar{S} = \{x \mid f(x) \le -\tilde{\delta}\} \supset \tilde{S}$. Consequently, haus $(\bar{S}, S) \le$ haus $(\tilde{S}, S) \le \tilde{\Delta}$. Thus the system $f(x) \le 0$ satisfies conditions (a) and (b) in Proposition 1. Therefore

$$d(z, S) \le \tilde{\delta}^{-1}\tilde{\Delta}[f(z)]_+ = \tilde{\delta}^{-1}\tilde{\Delta}\|[F(z)]_+\|_\infty \qquad \text{for all } z \in \mathbb{R}^n. \qquad \square$$

The following example illustrates an application of Proposition 1.

Example 1. Let $X = \mathbb{R}^2$, and let $P = \left\{ (x_1, x_2) \in \mathbb{R}^2 \mid x_2 \geq x_1^2 \right\}$. Consider $f(x) = (d(x,P))^2 - 1$, where $d(x,P)$ is the Euclidean distance from a point $x \in \mathbb{R}^2$ to the set P. Let $S = \{x \,|\, f(x) \leq 0\}$. Since $(d(x,P))^2$ is the inf-convolution of $\|\cdot\|^2$ and $\delta_P(\cdot)$, it is easy to see that $f^*(x^*) = 1/4\|x^*\|^2 + \sigma_P(x^*) + 1$ (Theorem 12.3 of [30]), where σ_P is the support function of P. Consequently, dom f^* = dom σ_P. Clearly, S is unbounded, and Robinson's result [29] is not applicable. It is evident that $\{x \,|\, f(x) \leq -1\} = P$. An easy computation shows that

$$\text{haus}(P, S) = 1.$$

By Proposition 1, we have

$$d(z, S) \leq [f(z)]_+ \qquad \text{for all } z \in \mathbb{R}^2.$$

2.2. Metric regularity and global error bounds. For simplicity, throughout the rest of the paper we assume that $C = X$.

The study of error bounds of (1) is closely related to that of metric regularity of (1). By extending the concept of metric regularity at a point to that of metric regularity at a set, we first obtain a useful characterization of the existence of a global error bound, and then study its implications. Note that the Slater condition is not assumed in this section.

DEFINITION 1. *Suppose that the solution set S of* (1) *is nonempty. We say that the system* (1) *is metrically regular at a nonempty set $\hat{S} \subset S$ if there exist positive constants δ and $\tau(\delta)$ such that*

$$d(x, S) \leq \tau(\delta)[f(x)]_+ \qquad \textit{when} \qquad d(x, \hat{S}) \leq \delta.$$

When $\hat{S} = \{z\}$, we say that the system (1) *is metrically regular at z. We say that the system* (1) *is metrically regular at every point of S if the system* (1) *is metrically regular at z for all $z \in S$.*

It is clear that if the system (1) has the global error bound (2), then the system (1) is metrically regular at S. It is the convexity of f that implies the reverse implication. We state this observation as the following theorem.

THEOREM 1. *Suppose that X is a Banach space and f is continuous and convex on X. For the system* (1)*, suppose that S is nonempty. Consider the following statements.*

(a) *The global error bound* (2) *holds.*

(b) *The system* (1) *is metrically regular at any nonempty set $\hat{S} \subset S$.*

(c) *The system* (1) *is metrically regular at S.*

(d) *The system* (1) *is metrically regular at every point of S.*

(e) *The system* (1) *satisfies the Slater condition.*

Then the following implications hold:

$$(a) \Leftrightarrow (b) \Leftrightarrow (c) \Rightarrow (d) \Leftarrow (e).$$

Proof. The implications $(a) \Rightarrow (b) \Rightarrow (c) \Rightarrow (d)$ are trivial.

$[(e) \Rightarrow (d)]$. This is a consequence of the inequality (3) in [29, p. 272].

$[(c) \Rightarrow (a)]$. Since (1) is metrically regular at S, it follows that there are some positive scalars δ and $\tau(\delta)$ such that

$$d(x, S) \leq \tau(\delta)[f(x)]_+ \qquad \text{when } d(x, S) \leq \delta.$$

For any $x \notin S$, let $d(x, S) = r$. Suppose that $r > \delta$; otherwise, the result holds trivially. For any $\epsilon > 0$, let $\bar{x} \in S$ such that $\|\bar{x} - x\| \leq d(x, S) + \epsilon$, and let $y = \lambda x + (1 - \lambda)\bar{x}$ with $\lambda = \delta/(r + \epsilon)$. Then $\|y - \bar{x}\| = \lambda\|x - \bar{x}\| \leq \delta$. Since $d(y, S) + \|y - x\| \geq d(x, S)$,

$$\begin{aligned}\|y - \bar{x}\| &= \|\bar{x} - x\| - \|y - x\| \\ &\leq d(x, S) + \epsilon - (d(x, S) - d(y, S)) \\ &\leq d(y, S) + \epsilon.\end{aligned}$$

Therefore,

$$\begin{aligned}d(x, S) \leq \|x - \bar{x}\| &= (\lambda)^{-1}\|y - \bar{x}\| = (\lambda)^{-1}\,(d(y, S) + \epsilon) \\ &\leq (\lambda)^{-1}\,(\tau(\delta)[f(y)]_+ + \epsilon) \qquad (\text{since } d(y, S) \leq \delta) \\ &\leq (\lambda)^{-1}\,(\tau(\delta)\lambda f(x) + \epsilon) \qquad (\text{by the convexity of } f) \\ &\leq \tau(\delta)[f(x)]_+ + \epsilon(\lambda)^{-1}.\end{aligned}$$

We obtain the desired inequality by letting $\epsilon \downarrow 0$. □

Remarks. A characterization of the existence of the global error bound (2) in terms of directional derivatives can be found in [15]. In [4], directional derivatives are used to characterize weak sharp minima for convex programs. It is not difficult to see that metric regularity at a set can also be characterized in terms of directional derivatives. For $X = \mathbb{R}^n$ and f being a pointwise maximum of finitely many convex differentiable functions, Li [17] is able to characterize metric regularity at a point of S in terms of the Abadie constraint qualification.

In the rest of this paper, we suppose that $X = \mathbb{R}^n$. Since any nonempty bounded closed set in $\mathbb{R}^n$ is compact, an immediate consequence of Theorem 1 is the following corollary.

COROLLARY 4. *Suppose that f is a finite convex function on $\mathbb{R}^n$, S is nonempty, and the system* (1) *is metrically regular at every point of S. If S is bounded, then the global error bound* (2) *holds.*

Proof. By metric regularity at every point of S and the compactness of S, there exist positive scalars τ_i, δ_i, and $\mathbb{B}(x_i, \delta_i)$ for $i = 1, ..., m$ with $x_i \in S$ such that $d(x, S) \leq \tau_i[f(x)]_+$ when $x \in \mathbb{B}(x_i, \delta_i)$ and $S \subset \cup_{i=1}^m \text{int}\,(\mathbb{B}(x_i, \delta_i))$, where $\mathbb{B}(x_i, \delta_i)$ denotes the closed Euclidean ball centered at x_i with radius δ_i. Again, by the compactness of S, there is a $\delta > 0$ such that

$$\{x \mid d(x, S) \leq \delta\} \subset \cup_{i=1}^m \text{int}\,(\mathbb{B}(x_i, \delta_i)).$$

Let $\tau = \max_{1 \leq i \leq m}\{\tau_i\}$. It follows that

$$d(x, S) \leq \tau[f(x)]_+ \qquad \text{when } d(x, S) \leq \delta.$$

The desired result follows by invoking Theorem 1. □

Without the boundedness assumption on S, metrical regularity at every point of S does not guarantee that (2) holds. Examples 1 and 2 in [15] show that the global error bound (2) does not hold even under the Slater condition, which implies metric regularity at every point of S by Theorem 1.

The next issue on our agenda is to obtain conditions along with metric regularity at every point of S under which (2) holds. In [6], we have shown that if $0 \notin \text{cl}\,[\text{dom}\, f^*]$ (which is equivalent to the system $f^\infty(u) < 0$ being solvable since

$f^\infty(u) = \sup_{v\in \mathrm{cl}\,[\mathrm{dom}\, f^*]}\langle v, u\rangle)$ [31], then the system (1) has the global error bound (2). Thus we only need to consider the case when

$$0 \in \mathrm{cl}\,[\mathrm{dom}\,(f^*)].$$

We will show that the existence of the global error bound (2) for S is completely determined by an associated convex inequality system. First, some notation: let $E = \mathrm{aff}(\mathrm{dom}\, f^*)$, which is a subspace since $\mathrm{cl}\,[\mathrm{dom}\,(f^*)] \subset E$ (p. 44 of [30]) and $0 \in \mathrm{cl}\,[\mathrm{dom}\,(f^*)]$. Let $E^\perp$ be the orthogonal complement of E, and $\Pi_E : \mathbb{R}^n \to E$ be the orthogonal projector. Then $\mathbb{R}^n = E \bigoplus E^\perp$. Define

$$f_E^*(x^*) = f^*(\Pi_E(x^*)) \qquad\qquad \forall x^* \in \mathbb{R}^n.$$

Since E is a subspace, by (b) of Lemma 2.1 in [2],

$$f(x) = f_E(\Pi_E(x)) \qquad\qquad \forall x \in \mathbb{R}^n. \tag{7}$$

In view of (7), f_E is convex and continuous on E. Now we consider the following auxiliary inequality system:

$$f_E(x) \le 0, \qquad \text{and } x \in E. \tag{8}$$

Let $S_E \subset E$ be the set of solutions to (8). We list below some basic properties associated with f_E as the following proposition.

PROPOSITION 2. *With the previous notation, we have*

(a) $S = S_E + E^\perp$;

(b) $\mathrm{dom}\ {f_E}^* = \mathrm{dom}\ f^* + E^\perp$ *and* $\mathrm{int}\ [\mathrm{dom}\ {f_E}^*] = \mathrm{ri}\ [\mathrm{dom}\ f^*] + E^\perp$.

Proof. Part (a) follows from the fact that $S_E = \Pi_E(S)$ and the relation (7), and Part (b) is proved in [2, Lemma 2.1]. □

It follows from (a) in Proposition 2 that S_E is nonempty if and only if S is nonempty. By invoking Proposition 2, we have the following lemma.

LEMMA 1. *For the system* (1)*, suppose that S is nonempty, and that E and f_E are defined as above. Then*

$$d(z, S) = d(\Pi_E(z), S_E) \qquad \textit{for all } z \in \mathbb{R}^n.$$

Proof. Since $S = S_E + E^\perp$, for all $x \in S_E$ and $y \in E^\perp$,

$$\begin{aligned}\|z - (x + y)\|^2 &= \|(\Pi_E(z) - x) + [(z - \Pi_E(z)) - y]\|^2 \\ &= \|\Pi_E(z) - x\|^2 + \|(z - \Pi_E(z)) - y\|^2.\end{aligned}$$

It follows that $d(z, S) = d(\Pi_E(z), S_E)$ for all $z \in \mathbb{R}^n$. □

Now we are in a position to state another main result of this paper.

THEOREM 2. *For the system* (1)*, suppose that f is a finite convex function on $\mathbb{R}^n$, S is nonempty, and* $0 \in \mathrm{cl}\ [\mathrm{dom}\ (f^*)]$. *Let* $E = \mathrm{aff}\ [\mathrm{dom}\ (f^*)]$. *Then the following holds:*

$$\Big[\ d(x, S) \le \tau[f(x)]_+\ \forall x \in \mathbb{R}^n\ \Big] \Leftrightarrow \Big[\ d(x, S_E) \le \tau[f_E(x)]_+\ \forall x \in E\ \Big].$$

Proof. ($\Rightarrow$) For any $x \in E$, by Lemma 1 and the relation (7), we have

$$d(x, S_E) = d(x, S) \le \tau[f(x)]_+ = \tau[f_E(x)]_+.$$

($\Leftarrow$) For any $x \in \mathbb{R}^n$, by Lemma 1 and the relation (7), we have

$$d(x, S) = d(\Pi_E(x), S_E) \leq \tau[f_E(\Pi_E(x))]_+ = \tau[f(x)]_+.$$

This completes the proof of Theorem 2. □

As a consequence of Theorem 2, we obtain a useful sufficient condition for (2) to hold.

COROLLARY 5. *For the system* (1)*, suppose that f is a finite convex function on $\mathbb{R}^n$, and S is nonempty. Suppose that the system* (1) *is metrically regular at every point of S, and* $0 \in \text{ri } [\text{dom } f^*]$. *Then the global error bound* (2) *holds.*

Proof. Since the system (1) is metrically regular at every point of S, and $S_E \subset S$, the system (8) is metrically regular at every point of S_E (its proof is similar to that for the necessity part in Theorem 2). By (b) of Proposition 2, $0 \in \text{int } [\text{dom } f_E^*]$. Hence S_E is bounded. The result follows by invoking Corollary 4 and Theorem 2. □

Remarks. The class of functions f with $0 \in \text{ri } [\text{dom } f^*]$ has been extensively studied by Auslender, Cominetti, and Crouzeix in [2].

3. Applications. In this section, we give an application of Corollary 5 in section 2 to convex programs.

Consider

$$(\mathcal{P}) \qquad \begin{aligned} &\text{minimize} \quad h(x) \\ &\text{subject to} \quad g(x) \leq 0, \end{aligned}$$

where h and g are finite convex functions on $\mathbb{R}^n$.

Let $\tilde{S}$ be the set of optimal solutions to $(\mathcal{P})$, and suppose that $\tilde{S}$ is nonempty. Let h_{min} be the optimal value of $(\mathcal{P})$. Following [4], we say that $\tilde{S}$ is a set of weak sharp minima if there exists some positive scalar γ such that

$$h(x) \geq h_{min} + \gamma d(x, \tilde{S}) \qquad \forall x \quad \text{with } g(x) \leq 0.$$

Let $f(x) = \max\{h(x) - h_{min}, g(x)\}$. By invoking Corollary 5, we obtain a sufficient condition for $(\mathcal{P})$ possessing weak sharp minima.

PROPOSITION 3. *Consider $(\mathcal{P})$. With the previous notation, if the system $f(x) \leq 0$ is metrically regular at every point of $\tilde{S}$ and* $0 \in \text{ri } [\text{dom } f^*]$, *then $\tilde{S}$ is a set of weak sharp minima. In particular, the relative interiority condition holds when $\tilde{S}$ is bounded.*

Proof. By Corollary 5, there is a positive scalar τ such that $d(x, \tilde{S}) \leq \tau[f(x)]_+$ for all $x \in \mathbb{R}^n$. In particular, $d(x, \tilde{S}) \leq \tau(h(x) - h_{min})$ for all x with $g(x) \leq 0$. This is what we needed to prove. □

Remarks. According to Theorem 2.4.7 (p. 68) in [10], $\text{dom } f^* = \text{co } [\text{dom } h^* \cup \text{dom } g^*]$. By Theorem 6.9 in [30],

$$\text{ri } [\text{dom } f^*] = \cup\{\lambda \text{ ri } [\text{dom } h^*] + (1 - \lambda) \text{ ri } [\text{dom } g^*] \mid 0 < \lambda < 1\}.$$

In [7], inspired by Corollary 2 in [15], Deng and Hu proved the following error bound result.

PROPOSITION 4. *Consider the system $\tilde{f}(x) \leq 0,\ x \in C$, where C is a nonempty closed convex set in $\mathbb{R}^n$ and $\tilde{f}$ is a finite convex function on $\mathbb{R}^n$. Suppose that $\bar{S}$ is the set of solutions to the system $\tilde{f}(x) \leq 0,\ x \in C$, and $\bar{S}$ is nonempty. Suppose that*

$\tilde{f}$ is Lipschitz continuous on $\mathbb{R}^n$ with a Lipschitz constant l, and there is a positive scalar γ such that

$$d(x, \bar{S}) \leq \gamma[\tilde{f}(x)]_+ \quad \textit{for all } x \in C. \tag{9}$$

Then

$$d(x, \bar{S}) \leq (\gamma l + 1)d(x, C) + \gamma[\tilde{f}(x)]_+ \quad \textit{for all } x \in \mathbb{R}^n. \tag{10}$$

In view of Propositions 3 and 4, we obtain a new characterization of weak sharp minima when g is a finite convex polyhedral function in $(\mathcal{P})$.

COROLLARY 6. *Consider $(\mathcal{P})$. With the same notation as in Proposition 3, suppose that $g(x) = \max_{1\leq i\leq m}\{a_i^T x + b_i\}$, where $a_i \in \mathbb{R}^n$ and $b_i \in \mathbb{R}$, that h is Lipschitz continuous on $\mathbb{R}^n$, and that $0 \in \text{ri}\,[\text{dom}\, f^*]$. Then the following statements are equivalent.*

(a) *$\tilde{S}$ is the set of weak sharp minima for $(\mathcal{P})$;*

(b) *the system $f(x) \leq 0$ is metrically regular at every point of $\tilde{S}$.*

Proof. $[(a) \Rightarrow (b)]$. This is a consequence of Proposition 4 and Hoffman's theorem [11].

$[(b) \Rightarrow (a)]$. This follows from Proposition 3. □

Notes added in revision. The results in section 2.2 were added in a revision of this paper (July 1996). After the paper was resubmitted, the author received a paper of Li and Singer [18], where a result similar to that of the equivalence of (a) and (c) in Theorem 1 was proved for convex multifunctions.

Acknowledgments. The author thanks Professor R. F. Wheeler for some helpful comments on an earlier version of this paper, Professor James Burke and the associate editor for some useful suggestions, and two referees for their insightful comments.

REFERENCES

[1] A. A. AUSLENDER AND J.-P. CROUZEIX, *Global regularity theorems*, Math. Oper. Res., 13 (1988), pp. 243–253.

[2] A. A. AUSLENDER, R. COMINETTI, AND J.-P. CROUZEIX, *Convex functions with unbounded level sets and applications to duality theory*, SIAM J. Optim., 3 (1993), pp. 669–687.

[3] J. V. BURKE AND P. TSENG, *A unified analysis of Hoffman's bound via Fenchel duality*, SIAM J. Optim., 6 (1996), pp. 265–282.

[4] J. V. BURKE AND M. C. FERRIS, *Weak sharp minima in mathematical programming*, SIAM J. Control Optim., 31 (1993), pp. 1340–1359.

[5] C. C. CHOU, K. F. NG, AND J. S. PANG, *Minimizing and stationary sequences of optimization problems*, SIAM J. Control Optim., to appear.

[6] S. DENG, *Computable error bounds for convex inequality systems in reflexive Banach spaces*, SIAM. J. Optim., 7 (1997), pp. 274–279.

[7] S. DENG AND H. HU, *Computable error bounds for semidefinite programming*, J. Global Optim., submitted.

[8] M. C. FERRIS AND J. S. PANG, *Nondegenerate solutions and related concepts in affine variational inequalities*, SIAM J. Control Optim., 34 (1996), pp. 244–263.

[9] M. S. GOWDA, *An analysis of zero set and global error bound properties of a piecewise affine function via its recession function*, SIAM J. Matrix Anal. Appl., 17(1996), pp. 594–609.

[10] J.-B. HIRIART-URRUTY AND C. LEMARÉCHAL, *Convex Analysis and Minimization Algorithms*, Springer-Verlag, Berlin, Heidelberg, 1993.

[11] A. J. HOFFMAN, *On approximate solutions of systems of linear inequalities*, J. Res. Nat. Bur. Standards, 49 (1952), pp. 263–265.

[12] H. Hu and Q. Wang, *On approximate solutions of infinite systems of linear inequalities*, Linear Algebra Appl., 114/115 (1989), pp. 429–438.
[13] D. Klatte, *Lipschitz stability and Hoffman's error bounds for convex inequality systems*, in Parametric Optimization and Related Topics IV, Approx. Optim. 9, Lang, Frankfurt am Main, 1977, pp. 201–212.
[14] D. Klatte, *Hoffman's error bound for systems of convex inequalities*, Preprint, April 1996, revised July 1996; in Mathematical Programming with Data Perturbations, A. V. Fiacco, ed., Marcel Dekker, New York, 1997.
[15] A. Lewis and J. S. Pang, *Error bounds for convex inequality systems*, in Generalized Convexity, Generalized Monotonicity: Recent Results, Proceedings of the Fifth Symposium on Generalized Convexity, Luminy, June 1996, J.-P. Crouzeix, J.-E. Martinez-Legaz, and M. Volle, eds., Kluwer Academic Publishers, Dordrecht, 1997, pp. 75–100.
[16] W. Li, *Error bounds for piecewise convex quadratic programs and applications*, SIAM J. Control Optim., 33 (1995), pp. 1510–1529.
[17] W. Li, *Abadie's constraint qualification, metric regularity, and error bounds for differentiable convex inequalities*, SIAM J. Optim., 7(1997), pp. 966–978.
[18] W. Li and I. Singer, *Global error bounds for convex multifunctions and applications*, Math. Oper. Res., to appear.
[19] X. D. Luo and Z. Q. Luo, *Extension of Hoffman's error bound to polynomial systems*, SIAM J. Optim., 4 (1994), pp. 383–392.
[20] Z. Q. Luo and J. S. Pang, *Error bounds for analytic systems and their applications*, Math. Programming, 67 (1995), pp. 1–28.
[21] Z. Q. Luo and P. Tseng, *On the linear convergence of descent methods for convex essentially smooth minimization*, SIAM J. Control Optim., 30 (1992), pp. 408–425.
[22] A. D. Ioffe, *Regular points of Lipschitz functions*, Trans. Amer. Math. Soc., 251 (1979), pp. 61–69.
[23] O. L. Mangasarian, *A condition number for differentiable convex inequalities*, Math. Oper. Res., 10 (1985), pp. 175–179.
[24] O. L. Mangasarian, *Simple computable bounds for solutions of linear complementarity problems and linear programs*, Math. Programming Stud., 25 (1985), pp. 1–12.
[25] O. L. Mangasarian and R. De Leone, *Error bounds for strongly convex programs and (super)linearly convergent iterative schemes for the least* 2*-norm solution of linear programs*, Appl. Math. Optim., 17 (1989), pp. 1–14.
[26] O. L. Mangasarian and T.-H. Shiau, *Lipschitz continuity of solutions of linear inequalities, programs and complementarity problems*, SIAM J. Control Optim., 25 (1987), pp. 583–595.
[27] R. Mathias and J. S. Pang, *Error bounds for the linear complementarity problem with a P-matrix*, Linear Algebra Appl., 132 (1990), pp. 123–136.
[28] J. S. Pang, *A posteriori error bounds for the linearly-constrained variational inequality problem*, Math. Oper. Res., 12 (1987), pp. 474–484.
[29] S. M. Robinson, *An application of error bounds for convex programming in a linear space*, SIAM J. Control, 13 (1975), pp. 271–273.
[30] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[31] R. T. Rockafellar, *Level sets and continuity of conjugate convex functions*, Trans. Amer. Math. Soc., 123 (1966), pp. 46–63.
[32] T. Wang and J. S. Pang, *Global error bounds for convex quadratic inequality systems*, Optimization, 31 (1994), pp. 1–12.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 4, pp. 1250–1267, July 1998

008

OPTIMAL CONTROL OF A COERCIVE DIRICHLET PROBLEM*

DARIUSZ IDCZAK†

Abstract. In this paper, the maximum principle for some n-dimensional coercive Dirichlet problem of the second order is proved and sufficient conditions for the existence of an optimal solution are given. The results obtained generalize, in the sense of the dimension of the state space, some special case of the maximum principle for the one-dimensional Dirichlet problem, derived in [M. Goebel and V. Raitums, *J. Global Optim.*, 4 (1994), 367–395].

Key words. Dirichlet problem, variational method, maximum principle, existence of an optimal solution

AMS subject classifications. 34B15, 49B45, 49B15

PII. S0363012997296341

1. Introduction.

In this paper we consider a nonlinear system of ordinary differential equations of the second order with functional parameters (controls) and Dirichlet-type boundary conditions. This system is of the form

$$\frac{d}{dx}(D_{z'}F(x,z(x),z'(x),u(x))) = D_zF(x,z(x),z'(x),u(x)), \quad x \in I=[0,\pi] \text{ a.e.}, \tag{1}$$

$$z(0)=z(\pi)=0, \tag{2}$$

where $z=(z_1,\dots,z_n)$, $u=(u_1,\dots,u_r)$, $F: I\times\mathbb{R}^n\times\mathbb{R}^n\times\mathbb{R}^r\to\mathbb{R}$, $D_zF=(D_{z_1}F,\dots,D_{z_n}F)$, $D_{z'}F=(D_{z'_1}F,\dots,D_{z'_n}F)$, $n\geq 1$, $r\geq 1$.

For more than a hundred years, systems of this type have played an essential role in mathematical models of physical and technical phenomena. This is connected with the principle of minimal action which holds true universally in nature (cf. [10]). So, it seems to be purposeful to investigate optimal processes for system (1)–(2).

The existence of a solution to system (1)–(2) was considered in many papers and monographs (see [2] and [9] and its references).

In our paper we consider system (1)–(2) with a constraint on controls

$$u(x)\in M, \tag{3}$$

and with the performance index

$$\mathcal{F}_0(z,u)=\int_I F_0(x,z(x),z'(x),u(x))\,dx\to\min, \tag{4}$$

where $M\subset\mathbb{R}^r$, $F_0: I\times\mathbb{R}^n\times\mathbb{R}^n\times\mathbb{R}^r\to\mathbb{R}$.

The control problem (1)–(4) is investigated in the spaces

$$H_0^1(I,\mathbb{R}^n)=\{z: I\to\mathbb{R}^n;\ \ z \text{ is absolutely continuous on } I,\ z(0)=z(\pi)=0,\ \ z'\in L^2(I,\mathbb{R}^n)\}$$

*Received by the editors October 17, 1996; accepted for publication (in revised form) June 23, 1997; published electronically May 15, 1998.
http://www.siam.org/journals/sicon/36-4/29634.html

†Faculty of Mathematics, University of Łódź, ul. Stefana Banacha 22, 90-238, Łódź, Poland (idczak@imul.uni.lodz.pl). This work was supported by grants 2P03A05910 and 8T11A01109 of the State Committee for Scientific Research, Poland.

of trajectories and

$$\mathcal{U}_M = \{u \in L^2(I, \mathbb{R}^r); \quad u(x) \in M \quad \text{for} \quad x \in I \quad \text{a.e.}\}$$

of controls.

The scalar problem ($n = 1$) of type (1)–(4) was studied by German and Lithuanian mathematicians in many papers (see [4] and its references). In [4] the authors considered this problem with additional inequality and equality constraints, but without u on the left-hand side of (1). They derived the maximum principle by using McShane variations and assuming the existence of a solution to system (1)–(2).

In general, our original control problem cannot be reformulated into a control problem for a normal system of two first-order differential equations. We use the direct method of the calculus of variations to study the original problem.

In this paper we prove the existence of a solution in $H_0^1(I, \mathbb{R}^n)$ to system (1)–(2) for any control $u \in L^2(I, \mathbb{R}^n)$. Next, we derive the maximum principle for problem (1)–(4) in the case of $D_{z'}F$ affine in z' by using the extremum principle for a smooth-convex problem (cf. [6, section 1.1.3], [14, section 1.1.4]). Finally, we obtain a theorem on the existence of an optimal solution for some special form of system (1) by using the results concerning the continuous dependence of solutions to system (1)–(2) on controls, obtained in [15].

2. Existence of a solution to system (1)–(2). To assert the existence of a solution to system (1)–(2) in the space $H_0^1(I, \mathbb{R}^n)$, we shall apply the variational method for solving differential equations (cf. [2], [9]).

We shall say that a function $z \in H_0^1(I, \mathbb{R}^n)$ satisfies system (1) (with a fixed control u) if

$$\int_I D_{z'}F(x, z(x), z'(x), u(x))h'(x)\,dx + \int_I D_z F(x, z(x), z'(x), u(x))h(x)\,dx = 0$$

for any $h \in H_0^1(I, \mathbb{R}^n)$.

From the Du Bois–Reymond lemma (cf. [9]) it follows that if z is a solution of (1) in the above sense, then the function $D_{z'}F(x, z(x), z'(x), u(x))$ is equal almost everywhere (a.e.) on I to an absolutely continuous function

$$\int_0^x D_z F(x, z(x), z'(x), u(x))\,dx + c,$$

where $c \in \mathbb{R}^n$ is some constant.

Consequently, if we identify these functions, we may write

$$\frac{d}{dx}(D_{z'}F(x, z(x), z'(x), u(x))) = D_z F(x, z(x), z'(x), u(x))$$

a.e. on I.

It is easy to see that the space $H_0^1(I, \mathbb{R}^n)$ with the inner product

$$(z, w) = \int_I z'(x)w'(x)\,dx$$

is a Hilbert space; the corresponding norm is given by

$$\|z\| = \Big(\int_I |z'(x)|^2 dx\Big)^{\frac{1}{2}}.$$

Below, we shall use the following estimates:

$$\int_I |z(x)|^2\,dx \le \int_I |z'(x)|^2 dx,$$

$$\max\{|z(x)|;\quad x \in I\} \le \sqrt{\pi}\|z\|$$

for $z \in H_0^1(I,\mathbb{R}^n)$. The first of the above inequalities is called the Poincaré inequality (cf. [9]). The second one is obvious.

Below, we shall use the following lemma.

LEMMA 1 (see [9]). *If*

$$z_n \underset{n\to\infty}{\rightharpoonup} z_0$$

weakly in $H_0^1(I,\mathbb{R}^n)$, *then*

$$z_n \underset{n\to\infty}{\rightrightarrows} z_0$$

uniformly on I.

LEMMA 2. *If*

$$z_n \underset{n\to\infty}{\rightharpoonup} z_0$$

weakly in $H_0^1(I,\mathbb{R}^n)$, *then*

$$z_n' \underset{n\to\infty}{\rightharpoonup} z_0'$$

weakly in $L^2(I,\mathbb{R}^n)$.

Proof. The assertion follows from the fact that the operator

$$H_0^1(I,\mathbb{R}^n) \ni z \mapsto z' \in L^2(I,\mathbb{R}^n)$$

is linear and continuous. So, it preserves weak convergence. □

Below, $\mathbb{R}_0^+$ means $[0,\infty)$ and $\mathbb{R}^+$ means $(0,\infty)$.

In an analogous way to [11] one can prove Theorem 1.

THEOREM 1. *Suppose F satisfies the following conditions.*

(5) *For $x \in I$ a.e., $u \in \mathbb{R}^r$, the function $F(x,\cdot,\cdot,u)$ is continuously differentiable in the Fréchet sense on $\mathbb{R}^n \times \mathbb{R}^n$.*

(6) *For $z, z' \in \mathbb{R}^n$, the function $F(\cdot,z,z',\cdot)$ is of Carathéodory type; i.e. for $x \in I$ a.e., $F(x,z,z',\cdot)$ is continuous on $\mathbb{R}^r$ and, for $u \in \mathbb{R}^r$, $F(\cdot,z,z',u)$ is measurable on I.*

(7) *For $z, z' \in \mathbb{R}^n$, the functions $D_z F(\cdot,z,z',\cdot)$ and $D_{z'}F(\cdot,z,z',\cdot)$ are of Carathéodory type.*

(8) *There exist functions $a(\cdot) \in C(\mathbb{R}_0^+,\mathbb{R}_0^+)$, $b(\cdot) \in L^1(I,\mathbb{R}_0^+)$, $c(\cdot), d(\cdot) \in L^2(I,\mathbb{R}_0^+)$ such that, for $x \in I$ a.e., $z, z' \in \mathbb{R}^n$, $u \in \mathbb{R}^r$, one has*

   (a) $|F(x,z,z',u)| \le a(|z|)(b(x) + |z'|^2 + d(x)|z'| + |u|^2)$,

   (b) $|D_z F(x,z,z',u)| \le a(|z|)(b(x) + |z'|^2 + |z'| + |u|^2)$,

   (c) $|D_{z'}F(x,z,z',u)| \le a(|z|)(c(x) + |z'| + |u|)$.

Then, for any control $u(\cdot) \in L^2(I, \mathbb{R}^n)$*, the functional*

$$f_u : H_0^1(I, \mathbb{R}^n) \ni z \mapsto \int_I F(x, z(x), z'(x), u(x))\, dx \in \mathbb{R}$$

is continuously Fréchet differentiable and

$$f_u'(z) : H_0^1(I, \mathbb{R}^n) \ni h \mapsto \int_I (D_z F(x, z(x), z'(x), u(x))h(x) + D_{z'} F(x, z(x), z'(x), u(x))h'(x))\, dx \in \mathbb{R}$$

for $z \in H_0^1(I, \mathbb{R}^n)$.

We also have Theorem 2.

THEOREM 2. *Let us assume that a function* F *satisfies* (6), (8a), *and*

(9) *for* $x \in I$ *a.e.,* $u \in \mathbb{R}^r$*, the function* $F(x, \cdot, \cdot, u)$ *is continuous on* $\mathbb{R}^n \times \mathbb{R}^n$*,*

(10) *for* $x \in I$ *a.e.,* $z \in \mathbb{R}^n$*,* $u \in \mathbb{R}^r$*, the function* $F(x, z, \cdot, u)$ *is convex on* $\mathbb{R}^n$*,*

(11) *there exists* $\varepsilon > 0$ *such that, for any* $\lambda > 0$*,*

$$|F(x, z_1, z', u) - F(x, z_2, z', u)| \le a_\lambda(x) + \beta_\lambda |z'|^{2-\varepsilon} + \gamma_\lambda |u|^2$$

for $x \in I$ *a.e.,* $|z_1| < \lambda$*,* $|z_2| < \lambda$*,* $z' \in \mathbb{R}^n$*,* $u \in \mathbb{R}^r$*, with some (depending on* λ*) function* $a_\lambda(\cdot) \in L^1(I, \mathbb{R}_0^+)$ *and constants* $\beta_\lambda > 0$*,* $\gamma_\lambda > 0$.

Then, for any control $u(\cdot) \in L^2(I, \mathbb{R}^r)$*, the functional* f_u *is weakly lower semi continuous (l.s.c) on* $H_0^1(I, \mathbb{R}^n)$.

Proof. Let $z_n \rightharpoonup z_0$ weakly in $H_0^1(I, \mathbb{R}^n)$. From Lemmas 1, 2 it follows that $z_n \rightharpoonup z_0$ weakly in $C(I, \mathbb{R}^n)$ and $z_n' \rightharpoonup z_0'$ weakly in $L^2(I, \mathbb{R}^n)$. Applying [13, Thm. 12], we obtain the assertion. □

Remark 1. It is easily seen that the condition

(12) there exists $\varepsilon > 0$, and the functions $a(\cdot) \in C(\mathbb{R}_0^+, \mathbb{R}_0^+)$, $b(\cdot) \in L^1(I, \mathbb{R}_0^+)$, such that for $x \in I$ a.e., $z, z' \in \mathbb{R}^n$, $u \in \mathbb{R}^n$ one has

$$|F(x, z, z', u)| \le a(|z|)(b(x) + \min\{|z'|^2, |z'|^{2-\varepsilon}\} + |u|^2)$$

implies (8a) and (11).

Remark 2. The term $\gamma_\lambda |u|^2$ in condition (11) may be replaced by $\gamma_\lambda(x)|u|$ with $\gamma_\lambda(\cdot) \in L^2(I, \mathbb{R}_0^+)$.

THEOREM 3. *Let us assume that* F *satisfies* (9), (6), (8a) *and suppose that*

(13) *there exist constants* $\alpha_1, \gamma_2, \gamma_3 \in \mathbb{R}^+$ *and functions* $\alpha_2(\cdot)$*,* $\gamma_1(\cdot) \in L^2(I, \mathbb{R}_0^+)$*,* $\beta_1(\cdot)$*,* $\beta_2(\cdot)$*,* $\delta_0(\cdot) \in L^1(I, \mathbb{R}_0^+)$*, such that*

$$\alpha_1 - \pi \int_I \beta_1(x) dx - \sqrt{\pi} \Big(\int_I |\gamma_1(x)|^2 dx \Big)^{\frac{1}{2}} > 0$$

and

$$F(x, z, z', u) \ge \alpha_1 |z'|^2 - \alpha_2(x)|z'| - \beta_1(x)|z|^2 - \beta_2(x)|z| - \gamma_1(x)|z||z'| - \gamma_2 |u|^2 - \gamma_3 |z||u|^2 - \delta_0(x)$$

for $x \in I$ *a.e.,* $z, z' \in \mathbb{R}^n$*,* $u \in \mathbb{R}^r$.

Then, for any control $u \in L^2(I, \mathbb{R}^r)$*, the functional* f_u *is coercive; i.e.,*

$$f_u(z) \to +\infty \quad as \quad \|z\| \to +\infty.$$

Proof. We have

$$\begin{aligned}
f_u(z) = \int_I F(x, z(x), z'(x), u(x))dx &\geq \alpha_1\|z\|^2 - \Big(\int_I |\alpha_2(x)|^2 dx\Big)^{\frac{1}{2}}\|z\| \\
&- \max\{|z(x)|^2; x \in I\}\int_I \beta_1(x)dx - \max\{|z(x)|; x \in I\}\int_I \beta_2(x)dx \\
&- \max\{|z(x)|; x \in I\}\Big(\int_I |\gamma_1(x)|^2 dx\Big)^{\frac{1}{2}}\|z\| - \gamma_2\int_I |u(x)|^2 dx \\
&- \gamma_3 \max\{|z(x)|;\ x \in I\}\int_I |u(x)|^2 dx - \int_I \delta_0(x)dx \\
&\geq \Big(\alpha_1 - \pi\int_I \beta_1(x)dx - \sqrt{\pi}\Big(\int_I |\gamma_1(x)|^2 dx\Big)^{\frac{1}{2}}\Big)\|z\|^2 \\
&- \Big(\Big(\int_I |\alpha_2(x)|^2 dx\Big)^{\frac{1}{2}} + \sqrt{\pi}\int_I \beta_2(x)dx + \gamma_3\int_I |u(x)|^2 dx\Big)\|z\| \\
&- \Big(\int_I \delta_0(x)dx + \gamma_2\int_I |u(x)|^2 dx\Big).
\end{aligned}$$

The proof is completed. □

Now, we can prove Theorem 4.

THEOREM 4. *If F satisfies conditions* (5), (6), (7), (8a), (8b), (8c), (10), (11), *and* (13), *then for any control $u(\cdot) \in L^2(I, \mathbb{R}^r)$, system* (1) *possesses a solution in $H_0^1(I, \mathbb{R}^n)$.*

Proof. The assumptions imply that the conditions of Theorems 2, 3 are satisfied. So, for any control $u(\cdot) \in L^2(I, \mathbb{R}^r)$, there exists a $z_u \in H_0^1(I, \mathbb{R}^n)$ such that

$$f_u(z_u) = \min\{f_u(z) : z \in H_0^1(I, \mathbb{R}^n)\}.$$

Consequently, Theorem 1 yields the equality

$$f_u'(z_u) = 0;$$

i.e.,

$$\int_I D_z F(x, z_u(x), {z_u}'(x), u(x))h(x)dx + \int_I D_2 F(x, z_u(x), {z_u}'(x), u(x))h'(x)dx = 0$$

for any $h \in H_0^1(I, \mathbb{R}^n)$. Thus, as was mentioned,

$$\frac{d}{dx}\Big(D_{z'}F(x, z_u(x), z_u'(x), u(x))\Big) = D_z F(x, z_u(x), z_u'(x), u(x))$$

a.e. on I. Of course, z_u satisfies (2) as a function from $H_0^1(I, \mathbb{R}^n)$ and the proof is completed. □

Remark 3. From Remark 1 it follows that the assumptions of Theorem 4 may be replaced by (5), (6), (7), (8b), (8c), (10), (12), (13).

Remark 4. From convex analysis it follows that z_u will be a unique solution of (1) in $H_0^1(I, \mathbb{R}^n)$ if f_u is strictly convex, i.e., if

$$f_u(\alpha z_1 + \beta z_2) < \alpha f_u(z_1) + \beta f_u(z_2)$$

for $z_1, z_2 \in H_0^1(I, \mathbb{R}^n)$, $z_1 \neq z_2$, $\alpha, \beta > 0$, $\alpha + \beta = 1$.

Remark 5. It is easily seen (cf. also [15]) that the boundary condition (2) in the assertion of Theorem 4 may be replaced by

$$z(0) = a, \quad z(\pi) = b$$

with arbitrary fixed $a, b \in \mathbb{R}^n$. (It suffices to observe that if F satisfies the assumptions of Theorem 4, then $G : I \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^r \to \mathbb{R}$ given by

$$G(x, z, z', u) = F\left(x, z + \frac{1}{\pi}(b-a)x + a, z' + \frac{1}{\pi}(b-a), u\right)$$

also satisfies these assumptions.)

3. Necessary conditions for optimality. In this section we shall assume that F satisfies the assumptions of Theorem 4 and F_0 satisfies the assumptions of Theorem 1 (with F replaced by F_0). Consequently, the functional $\mathcal{F}_0$ is continuously differentiable with respect to $z \in H_0^1(I, \mathbb{R}^n)$ for any fixed control $u(\cdot) \in \mathcal{U}_M$, and

$$(\partial_z \mathcal{F}_0(z,u), h) = \int_I \Big(D_z F_0(x, z(x), z'(x), u(x))h(x) + D_{z'} F_0(x, z(x), z'(x), u(x))h'(x) \Big) dx$$

for $z(\cdot)$, $h(\cdot) \in H_0^1(I, \mathbb{R}^n)$.

Let us define the following operator:

$$\mathcal{F} : H_0^1(I, \mathbb{R}^n) \times \mathcal{U}_M \longrightarrow \Big[H_0^1(I, \mathbb{R}^n)\Big]^*,$$

$$(\mathcal{F}(z,u), h) = \int_I \Big(D_z F(x, z(x), z'(x), u(x))h(x) + D_{z'} F(x, z(x), z'(x), u(x))h'(x) \Big) dx.$$

The optimization problem (1)–(4) may be formulated as

$$(S) \begin{cases} \mathcal{F}_0(z,u) \longrightarrow \min, \\ \mathcal{F}(z,u) = 0, \\ u \in \mathcal{U}_M, \end{cases}$$

where $\mathcal{F}_0 : H_0^1(I, \mathbb{R}^n) \times \mathcal{U}_M \longrightarrow \mathbb{R}$, $\mathcal{F} : H_0^1(I, \mathbb{R}^n) \times \mathcal{U}_M \longrightarrow [H_0^1(I, \mathbb{R}^n)]^*$.

Before we formulate the maximum principle, we shall prove some lemmas.

LEMMA 3. *Let us assume that*

(14) *for $x \in I$ a.e., $u \in \mathbb{R}^r$, the functions $D_z F(x, \cdot, \cdot, u)$, $D_{z'} F(x, \cdot, \cdot, u)$ are continuously differentiable in the Fréchet sense on $\mathbb{R}^n \times \mathbb{R}^n$;*

(15) *for $z, z' \in \mathbb{R}^n$, the functions $D_{zz} F(\cdot, z, z', \cdot)$, $D_{zz'} F(\cdot, z, z', \cdot) = D_{z'z} F(\cdot, z, z', \cdot)$, $D_{z'z'} F(\cdot, z, z', \cdot)$ are of Carathéodory type;*

(16) *there exist functions $\bar{a}(\cdot) \in C(\mathbb{R}_0^+, \mathbb{R}_0^+)$, $\bar{b}(\cdot) \in L^2(I, \mathbb{R}_0^+)$, $\bar{c}(\cdot) \in L^\infty(I, \mathbb{R}_0^+)$ such that, for $x \in I$ a.e., $z, z' \in \mathbb{R}^n$, $u \in \mathbb{R}^r$, one has*

(a) $|D_{zz} F(x, z, z', u)| \le \bar{a}(|z|)(\bar{b}(x) + |z'| + |u|)$,

(b) $|D_{zz'} F(x, z, z', u)| \le \bar{a}(|z|)\bar{c}(x)$,

(c) $|D_{z'z'} F(x, 0, 0, u)| \le \bar{c}(x)$;

(17) *for any* $\mu > 0$, *there exists* $\bar{c}_\mu(\cdot) \in L^\infty(I, \mathbb{R}^+_0)$ *such that, for* $x \in I$ *a.e.,* $z, w \in \mathbb{R}^n$, $|z| < \mu$, $|w| < \mu$, $z', w' \in \mathbb{R}^n$, $u \in \mathbb{R}^r$,

$$|D_{z'z'}F(x, z, z', u) - D_{z'z'}F(x, w, w', u)| \le \bar{c}_\mu(x)|z - w|.$$

Then, the operator $\mathcal{F}$ *is continuously Gâteaux differentiable (thus continuously Fréchet differentiable) with respect to* $z \in H^1_0(I, \mathbb{R}^n)$.

Remark 6. From (16(c)) and (17) it follows that

$$\begin{aligned}|D_{z'z'}F(x, z, z', u)| &\le |D_{z'z'}F(x, z, z', u) - D_{z'z'}F(x, 0, 0, u)| \\ &\quad + |D_{z'z'}F(x, 0, 0, u)| \le \bar{c}_\mu(x)|z| + \bar{c}(x) \le \bar{c}_\mu(x)\mu + \bar{c}(x)\end{aligned}$$

for $x \in I$ a.e., $z \in \mathbb{R}^n$, $|z| < \mu$, $z' \in \mathbb{R}^n$, $u \in \mathbb{R}^r$.

Moreover, from (17) it follows ($z = w$) that the function $D_{z'}F(x, z, \cdot, u)$ is affine.

Proof of Lemma 3. Let us consider the following operator:

$$\partial_z \mathcal{F}(z, u) : H^1_0(I, \mathbb{R}^n) \longrightarrow [H^1_0(I, \mathbb{R}^n)]^*, \tag{18}$$

$$\begin{aligned}\Big((\partial_z \mathcal{F}(z, u), w), h\Big) &= \int_I \Big(D_{zz}F(x, z(x), z'(x), u(x))w(x)\Big)h(x)dx \\ &\quad + \int_I \Big(D_{zz'}F(x, z(x), z'(x), u(x))w'(x)\Big)h(x)dx \\ &\quad + \int_I \Big(D_{z'z}F(x, z(x), z'(x), u(x))w(x)\Big)h'(x)dx \\ &\quad + \int_I \Big(D_{z'z'}F(x, z(x), z'(x), u(x))w'(x)\Big)h'(x)dx,\end{aligned}$$

where $z \in H^1_0(I, \mathbb{R}^n)$, $u \in \mathcal{U}_M$ are fixed and w, $h \in H^1_0(I, \mathbb{R}^n)$.

Its linearity is obvious. Moreover,

$$\begin{aligned}&\|\big((\partial_z \mathcal{F}(z, u), w), \cdot\big)\|_{[H^1_0(I,\mathbb{R}^n)]^*} \\ &\quad \le \sqrt{\pi} \int_I |D_{zz}F(x, z(x), z'(x), u(x))w(x)|dx \\ &\qquad + \sqrt{\pi} \int_I |D_{zz'}F(x, z(x), z'(x), u(x))w'(x)|dx \\ &\qquad + \Big(\int_I |D_{z'z}F(x, z(x), z'(x), u(x))w(x)|^2 dx\Big)^{\frac{1}{2}} \\ &\qquad + \Big(\int_I |D_{z'z'}F(x, z(x), z'(x), u(x))w'(x)|^2 dx\Big)^{\frac{1}{2}} \qquad (19) \\ &\quad \le \pi \cdot \int_I |D_{zz}F(x, z(x), z'(x), u(x))|dx \cdot \|w\| \\ &\qquad + \sqrt{\pi}\Big(\int_I |D_{zz'}F(x, z(x), z'(x), u(x))|^2 dx\Big)^{\frac{1}{2}} \cdot \|w\| \\ &\qquad + \sqrt{\pi}\Big(\int_I |D_{z'z}F(x, z(x), z'(x), u(x))|^2 dx\Big)^{\frac{1}{2}} \|w\| \\ &\qquad + \operatorname{ess}\sup_{x \in I} |D_{z'z'}F(x, z(x), z'(x), u(x))| \cdot \|w\|.\end{aligned}$$

This implies the continuity of (18), i.e., $\partial_z \mathcal{F}(z, u) \in \mathcal{L}(H^1_0(I, \mathbb{R}^n), [H^1_0(I, \mathbb{R}^n)]^*)$.

Now, we shall show that, for any $w \in H_0^1(I,\mathbb{R}^n)$,

$$\lim_{\lambda\to 0}\Big\|\frac{\mathcal{F}(z+\lambda w,u)-\mathcal{F}(z,u)}{\lambda}-(\partial_z\mathcal{F}(z,u),w)\Big\|_{[H_0^1(I,\mathbb{R}^n)]^*}=0, \tag{20}$$

i.e., that $\partial_z\mathcal{F}(z,u)$ is a Gâteaux differential of $\mathcal{F}$ with respect to $z\in H_0^1(I,\mathbb{R}^n)$ at the point (z,u).

Indeed, for any $h\in H_0^1(I,\mathbb{R}^n)$ and $\lambda\in\mathbb{R}\backslash\{0\}$, $|\lambda|\le 1$, we have

$$\begin{aligned}
&\Big|\frac{(\mathcal{F}(z+\lambda w,u),h)-(\mathcal{F}(z,u),h)}{\lambda}-((\partial_z\mathcal{F}(z,u),w),h)\Big|\\
&\le\int_I\Big|\frac{D_zF(x,z(x)+\lambda w(x),z'(x)+\lambda w'(x),u(x))-D_zF(x,z(x),z'(x),u(x))}{\lambda}\\
&\quad-\Big(D_{zz}F(x,z(x),z'(x),u(x))w(x)+D_{zz'}F(x,z(x),z'(x),u(x))w'(x)\Big)\Big|dx\cdot\sqrt{\pi}\|h\|\\
&\quad+\Big(\int_I\Big|\frac{D_{z'}F\big(x,z(x)+\lambda w(x),z'(x)+\lambda w'(x),u(x)\big)-D_{z'}F(x,z(x),z'(x),u(x))}{\lambda}\\
&\quad-\big(D_{z'z}F(x,z(x),z'(x),u(x))w(x)+D_{z'z'}F(x,z(x),z'(x),u(x))w'(x)\big)\Big|^2dx\Big)^{\frac12}\cdot\|h\|.
\end{aligned}$$

So,

$$\begin{aligned}
&\Big|\Big|\frac{\mathcal{F}(z+\lambda w,u)-\mathcal{F}(z,u)}{\lambda}-(\partial_z\mathcal{F}(z,u),w)\Big|\Big|_{[H_0^1(I,\mathbb{R}^n)]^*}\\
&\quad\le\sqrt{\pi}\int_I\Big|\frac{D_zF(x,z(x)+\lambda w(x),z'(x)+\lambda w'(x),u(x))-D_zF(x,z(x),z'(x),u(x))}{\lambda}\\
&\qquad-\Big(D_{zz}F(x,z(x),z'(x),u(x))w(x)+D_{zz'}F(x,z(x),z'(x),u(x))w'(x)\Big)\Big|dx\\
&\qquad+\Big(\int_I\Big|\frac{D_{z'}F(x,z(x)+\lambda w(x),z'(x)+\lambda w'(x),u(x))-D_{z'}F(x,z(x),z'(x),u(x))}{\lambda}\\
&\qquad-\big(D_{z'z}F(x,z(x),z'(x),u(x))w(x)+D_{z'z'}F(x,z(x),z'(x),u(x))w'(x)\big)\Big|^2dx\Big)^{\frac12}.
\end{aligned}$$

From the differentiability of D_zF, $D_{z'}F$ with respect to $(z,z')\in\mathbb{R}^n\times\mathbb{R}^n$ it follows that the above integrands tend (with fixed z, $w\in H_0^1(I,\mathbb{R}^n)$, $u\in\mathcal{U}_M$) pointwise to 0 as $\lambda\to 0$.

Moreover, from the mean value theorem we have

$$\begin{aligned}
&\Big|\frac{D_zF(x,z(x)+\lambda w(x),z'(x)+\lambda w'(x),u(x))-D_zF(x,z(x),z'(x),u(x))}{\lambda}\\
&\qquad-(D_{zz}F(x,z(x),z'(x),u(x))w(x)+D_{zz'}F(x,z(x),z'(x),u(x))w'(x))\Big|\\
&\quad\le\max\{\bar a(|z(x)+t\lambda w(x)|);(x,t,\lambda)\in[0,\pi]\times[0,1]\times[-1,1]\}\\
&\qquad\cdot(\bar b(x)+|z'(x)|+|w'(x)|+|u(x)|)(|w(x)|+|w'(x)|)\\
&\qquad+\ \max\{\bar a(|z(x)+t\lambda w(x)|);(x,t,\lambda)\in[0,\pi]\times[0,1]\times[-1,1]\}\\
&\qquad\cdot\bar c(x)(|w(x)|+|w'(x)|)+\bar a(|z(x)|)(\bar b(x)+|z'(x)|+|u(x)|)(|w(x)|+|w'(x)|)\\
&\qquad+\bar a(|z(x)|)\bar c(x)(|w(x)|+|w'(x)|)
\end{aligned}$$

and, analogously,

$$\begin{aligned}
&\Big|\frac{D_{z'}F(x,z(x)+\lambda w(x),z'(x)+\lambda w'(x),u(x))-D_{z'}F(x,z(x),z'(x),u(x))}{\lambda}\\
&\quad-\big(D_{z'z}F(x,z(x),z'(x),u(x))w(x)+D_{z'z'}F(x,z(x),z'(x),u(x))\big)\Big|^2\\
&\le\Big(\max\{\bar a(|z(x)+t\lambda w(x)|);(x,t,\lambda)\in[0,\pi]\times[0,1]\times[-1,1]\}\bar c(x)(|w(x)|+|w'(x)|)\\
&\quad+(\bar c_{\mu+\nu}(x)(\mu+\nu)+\bar c(x))(|w(x)|+|w'(x)|)+\bar a(|z(x)|)\bar c(x)(|w(x)|+|w'(x)|)\\
&\quad+(\bar c_\mu(x)\mu+\bar c(x))(|w(x)|+|w'(x)|)\Big)^2,
\end{aligned}$$

where $\mu=\max\{|z(x)|:x\in I\}$, $\nu=\max\{|w(x)|:x\in I\}$.

Since the right-hand sides of the above estimates are integrable functions, using the Lebesgue theorem, we therefore obtain (20).

Now, we shall show that (with a fixed $u\in\mathcal{U}_M$) the mapping

$$\partial_z\mathcal{F}(\cdot,u):H_0^1(I,\mathbb{R}^n)\ni z\mapsto\partial_z\mathcal{F}(z,u)\in\mathcal{L}\big(H_0^1(I,\mathbb{R}^n),[H_0^1(I,\mathbb{R}^n)]^*\big)$$

is continuous.

Indeed, in an analogous way as in (19), we assert that

(21)
$$\begin{aligned}
&\big|\big|\partial_z\mathcal{F}(z_n,u)-\partial_z\mathcal{F}(z_0,u)\big|\big|_{\mathcal{L}(H_0^1(I,\mathbb{R}^n),[H_0^1(I,\mathbb{R}^n)]^*)}\\
&\quad\le\pi\cdot\int_I\big|D_{zz}F(x,z_n(x),z_n'(x),u(x))-D_{zz}F(x,z_0(x),z_0'(x),u(x))\big|dx\\
&\qquad+2\sqrt{\pi}\Big(\int_I\big|D_{zz'}F(x,z_n(x),z_n'(x),u(x))-D_{zz'}F(x,z_0(x),z_0'(x),u(x))\big|^2dx\Big)^{\frac12}\\
&\qquad+\operatorname{ess}\sup_{x\in I}\big|D_{z'z'}F(x,z_n(x),z_n'(x),u(x))-D_{z'z'}F(x,z_0(x),z_0'(x),u(x))\big|
\end{aligned}$$

for any z_n, $z_0\in H_0^1(I,\mathbb{R}^n)$.

Now, let us assume that $z_n\xrightarrow[n\to\infty]{}z_0$ in $H_0^1(I,\mathbb{R}^n)$. So,

$$z_n\underset{n\to\infty}{\rightrightarrows}z_0$$

uniformly on I and $z_n'\xrightarrow[n\to\infty]{}z_0'$ in $L^2(I,\mathbb{R}^n)$.

To prove that the first of the above integrals tends to 0, let us observe that its integrand converges in measure to the zero-function. This follows from the fact that the sequence $(z_n(\cdot),z_n'(\cdot))_{n\in\mathbb{N}}$ tends in measure to $(z_0(\cdot),z_0'(\cdot))$ and from the properties of the Nemytskii operator (cf. [8]). Moreover, the sequence of integrands has equi-absolutely continuous integrals (cf. [12]):

(22)
$$\begin{aligned}
&\big|D_{zz}F(x,z_n(x),z_n'(x),u(x))-D_{zz}F(x,z_0(x),z_0'(x),u(x))\big|\\
&\quad\le\bar a(|z_n(x)|)(\bar b(x)+|z_n'(x)|+|u(x)|)\\
&\qquad+\ \bar a(|z_0(x)|)(\bar b(x)+|z_0'(x)|+|u(x)|)\\
&\quad\le\text{const }(2\bar b(x)+2|u(x)|+|z_0'(x)|+|z_n'(x)|).
\end{aligned}$$

Since $z_n'\to z_0'$ in $L^2(I,\mathbb{R}^n)$ and, consequently, in $L^1(I,\mathbb{R}^n)$, the sequence $(|z_n'(x)|)_{n\in\mathbb{N}}$ has equi-absolutely continuous integrals. So, the sequence

$$\Big(\big|D_{zz}F(x,z_n(x),z_n'(x),u(x))-D_{zz}F(x,z_0(x),z_0'(x),u(x))\big|\Big)_{n\in\mathbb{N}}$$

has equi-absolutely continuous integrals. Applying the Vitali theorem on the convergence of an integral [12], we assert that the first term in (21) tends to 0 as $n \to +\infty$.

In an analogous way, replacing (22) by

$$\begin{aligned}&\left|D_{zz'}F(x,z_n(x),z_n'(x),u(x)) - D_{zz'}F(x,z_0(x),z_0'(x),u(x))\right|^2\\ &\quad\le 4\Big(\left|D_{zz'}F(x,z_n(x),z_n'(x),u(x))\right|^2 + \left|D_{zz'}F(x,z_0(x),z_0'(x),u(x))\right|^2\Big)\\ &\quad\le 4\Big((\bar a(|z_n(x)|))^2(\bar c(x))^2 + (\bar a(|z_0(x)|))^2(\bar c(x))^2\Big) \le \text{const},\end{aligned}$$

we assert that the second term in (21) tends to 0 as $n \to +\infty$.

The convergence of the third term follows from the fact that

$$z_n \underset{n\to\infty}{\rightrightarrows} z_0$$

uniformly on I and from (17) because

$$\left|D_{z'z'}F(x,z_n(x),z_n'(x),u(x)) - D_{z'z'}F(x,z_0(x),z_0(x),u(x))\right| \le \bar c_\mu(x)|z_n(x) - z_0(x)|$$

where $|z_n(x)| < \mu$, $n = 1,2,\dots,|z_0(x)| < \mu$ for all $x \in I$.

The proof of the lemma is completed. □

Now, let us denote

$$\begin{aligned}&A(x) = D_{z'z}F(x,z_*(x),z_*'(x),u_*(x)) = D_{zz'}F(x,z_*(x),z_*'(x),u_*(x)),\\ &B(x) = D_{z'z'}F(x,z_*(x),z_*'(x),u_*(x)),\\ &C(x) = D_{zz}F(x,z_*(x),z_*'(x),u_*(x)),\end{aligned} \tag{23}$$

where $z_* \in H_0^1(I,\mathbb{R}^n)$, $u_* \in \mathcal{U}_M$ are fixed.

From (16), (17) it follows that $A(\cdot)$, $B(\cdot) \in L^\infty(I,\mathbb{R}^{n\times n})$, $C(\cdot) \in L^2(I,\mathbb{R}^{n\times n})$.

LEMMA 4. *Let $z_* \in H_0^1(I,\mathbb{R}^n)$, $u_* \in \mathcal{U}_M$ be such that*

(24) *the matrix $B(x)$ is positive for $x \in I$ a.e. (i.e., there exists a set $S \subset I$ of full measure such that, for $x \in S$, $B(x)z'z' > 0$ for $z' \in \mathbb{R}^n - \{0\}$), and*

$$\inf\{B(x)z'z';\ |z'| = 1,\ \ x \in S\} - \pi\int_I |C(x)|dx - 2\pi(\operatorname{ess\ sup}_I |A(\cdot)|) > 0. \tag{25}$$

Then, for any $\Lambda \in [H_0^1(I,\mathbb{R}^n)]^*$, there exists a $w \in H_0^1(I,\mathbb{R}^n)$ such that

$$(\partial_z\mathcal{F}(z_*,u_*),w) = \Lambda;$$

i.e.,

$$\begin{aligned}&\int_I (C(x)w(x) + A(x)w'(x))h(x)dx\\ &\qquad + \int_I (A(x)w(x) + B(x)w'(x))h'(x)dx = (\Lambda,h)\end{aligned}$$

for any $h \in H_0^1(I,\mathbb{R}^n)$.

Proof. Since $H_0^1(I,\mathbb{R}^n)$ is a Hilbert space, there exists an $a \in H_0^1(I,\mathbb{R}^n)$ such that

$$(\Lambda,h) = \int_I a'(x)h'(x)dx$$

for all $h \in H_0^1(I, \mathbb{R}^n)$. So, we have to show the existence of $w \in H_0^1(I, \mathbb{R}^n)$ such that

$$\int_I (A(x)w(x) + B(x)w'(x) - a'(x))h'(x)dx + \int_I (C(x)w(x) + A(x)w'(x))h(x)dx = 0$$

for all $h \in H_0^1(I, \mathbb{R}^n)$; i.e.,

$$\frac{d}{dx}(A(x)w(x) + B(x)w'(x) - a'(x)) = C(x)w(x) + A(x)w'(x)$$

a.e. on I.

Let us denote

$$G(x, w, w') = A(x)ww' - a'(x)w' + \frac{1}{2}B(x)w'w' + \frac{1}{2}C(x)ww$$

for $x \in I$ a.e., w, $w' \in \mathbb{R}^n$.

Of course,

$$D_w G(x, w, w') = A(x)w' + C(x)w,$$
$$D_{w'} G(x, w, w') = A(x)w - a'(x) + B(x)w'$$

for $x \in I$ a.e., w, $w' \in \mathbb{R}^n$.

It is easy to check that G satisfies (5), (6), (7), (10). Now, we show that it satisfies (8a), (8b), (8c), (11), and (13).

Indeed, we have

$$|G(x, w, w')| \le |A(x)\|w\|w'| + |a'(x)||w'| + \frac{1}{2}|B(x)||w'|^2 + \frac{1}{2}|C(x)||w|^2$$
$$\le \max\{1, |w|, |w|^2\}\left((\operatorname{ess\,sup}_I |A(\cdot)|)|w'| + |a'(x)||w'| + \frac{1}{2}(\operatorname{ess\,sup}_I |B(\cdot)|)|w'|^2 + \frac{1}{2}|C(x)|\right)$$
$$= \max\{1, |w|, |w|^2\}\left(\frac{1}{2}|C(x)| + ((\operatorname{ess\,sup}_I |A(\cdot)|) + |a'(x)|)|w'| + \frac{1}{2}(\operatorname{ess\,sup}_I |B(\cdot)|)|w'|^2\right),$$

$$|D_w G(x, w, w')| \le |A(x)||w'| + |C(x)||w| \le (\operatorname{ess\,sup}_I |A(\cdot)|)|w'| + |C(x)||w|$$
$$\le \max\{|w|, 1\}\big(|c(x)| + (\operatorname{ess\,sup}_I |A(\cdot)|)|w'|\big),$$

$$|D_{w'} G(x, w, w')| \le |A(x)||w| + |a'(x)| + |B(x)||w'|$$
$$\le \max\{|w|, 1\}(|A(x)| + |a'(x)| + |B(x)||w'|)$$
$$\le \max\{|w|, 1\}(|A(x)| + |a'(x)| + (\operatorname{ess\,sup}_I |B(\cdot)|)|w'|)$$

for $x \in I$ a.e., w, $w' \in \mathbb{R}^n$. So, (8a), (8b), (8c) are satisfied. Moreover, for any $\lambda > 0$,

$$\Big|G(x, w_1, w') - G(x, w_2, w')\Big|$$
$$= \Big|A(x)w_1w' + \frac{1}{2}C(x)w_1w_1 - A(x)w_2w' - \frac{1}{2}C(x)w_2w_2\Big|$$
$$\le \Big|A(x)(w_1 - w_2)w'\Big| + \frac{1}{2}|C(x)w_1w_1 - C(x)w_2w_2|$$
$$\le (\operatorname{ess\,sup}_I |A(\cdot)|)2\lambda|w'|^{2-1} + |c(x)|\lambda^2$$

for $x \in I$ a.e., $|w_1| < \lambda$, $|w_2| < \lambda$, $w' \in \mathbb{R}^n$. So, (11) is satisfied with $\varepsilon = 1$.

To prove that (13) holds, let us denote $\alpha = \inf\{B(x)z'z', |z'| = 1,\ x \in S\}$ and observe that (24) implies (cf. [5, I.4])

$$B(x)w'w' \geq \alpha|w'|^2$$

for all $x \in S$ and $w' \in \mathbb{R}^n$. Consequently,

$$\begin{aligned} G(x,w,w') &= A(x)ww' - a'(x)w' + \frac{1}{2}B(x)w'w' + \frac{1}{2}C(x)ww \\ &\geq \frac{1}{2}\alpha|w'|^2 - (\operatorname{ess\ sup}_I |A(\cdot)|)|w||w'| - |a'(x)||w'| - \frac{1}{2}|C(x)||w|^2 \end{aligned}$$

and, by (25),

$$\begin{aligned} &\frac{1}{2}\alpha - \pi\int_I \frac{1}{2}|C(x)|dx - \sqrt{\pi}\Big(\int_I (\operatorname{ess\ sup}_I |A(\cdot)|)^2 dx\Big)^{\frac{1}{2}} \\ &\quad = \frac{1}{2}\alpha - \frac{1}{2}\pi\int_I |C(x)|dx - \pi\Big(\operatorname{ess\ sup}_I |A(\cdot)|\Big) > 0. \end{aligned}$$

Hence (13) holds.

Thus, Theorem 4 implies the assertion of the lemma. □

We also have Lemma 5.

LEMMA 5. *If*

(26) *$M \subset \mathbb{R}^r$ is compact and for $x \in I$ a.e., $z, z' \in \mathbb{R}^n$, the set $\{(D_zF(x,z,z',u), D_{z'}F(x,z,z',u), F^0(x,z,z',u)) \in \mathbb{R}^{n+n+1},\ \ u \in M\}$ is convex,*

then, for any $z \in H_0^1(I,\mathbb{R}^n)$, the operator $\mathcal{F}$ and the functional $\mathcal{F}_0$ satisfy the following convexity condition: for any $u_1, u_2 \in \mathcal{U}_M$, $z \in H_0^1(I,\mathbb{R}^n)$, and $\alpha \in [0,1]$ there exists $\bar{u} \in \mathcal{U}_M$ such that

(27) $\mathcal{F}(z,\bar{u}) = \alpha\mathcal{F}(z,u_1) + (1-\alpha)\mathcal{F}(z,u_2)$,

(28) $\mathcal{F}_0(z,\bar{u}) = \alpha\mathcal{F}_0(z,u_1) + (1-\alpha)\mathcal{F}_0(z,u_2)$.

Proof. Let us fix $z \in H_0^1(I,\mathbb{R}^n)$, $u_1, u_2 \in \mathcal{U}_M$, $\alpha \in [0,1]$. For $x \in I$ a.e. there exists a point $u_0(x) \in M$ such that

$$\begin{aligned} &\alpha D_zF(x,z(x),z'(x),u_1(x)) + (1-\alpha)D_zF(x,z(x),z'(x),u_2(x)) \\ &\quad = D_zF(x,z(x),z'(x),u_0(x)), \end{aligned}$$

$$\begin{aligned} &\alpha D_{z'}F(x,z(x),z'(x),u_1(x)) + (1-\alpha)D_{z'}F(x,z(x),z'(x),u_2(x)) \\ &\quad = D_{z'}F(x,z(x),z'(x),u_0(x)), \end{aligned}$$

$$\begin{aligned} &\alpha F^0(x,z(x),z'(x),u_1(x)) + (1-\alpha)F^0(x,z(x),z'(x),u_2(x)) \\ &\quad = F^0(x,z(x),z'(x),u_0(x)). \end{aligned}$$

From the above it follows that the functions $D_zF(x,z(x),z'(x),u_0(x))$, $D_{z'}F(x,z(x),z'(x),u_0(x))$, $F^0(x,z(x),z'(x),u_0(x))$ are measurable. Of course,

$$\begin{aligned} &(D_zF(x,z(x),z'(x),u_0(x)), D_{z'}F(x,z(x),z'(x),u_0(x)), F^0(x,z(x),z'(x),u_0(x))) \\ &\quad \in \{(D_zF(x,z(x),z'(x),u), D_{z'}F(x,z(x),z'(x),u), F^0(x,z(x),z'(x),u));\ u \in M\}. \end{aligned}$$

Applying the implicit function theorem for a set-valued function associated with a Carathéodory function (cf. [7, Chap. II, Thm. 3.12]), we obtain an existence of a measurable function $\bar{u} : I \to M$ such that

$$\begin{aligned} D_z F(x, z(x), z'(x), u_0(x)) &= D_z F(x, z(x), z'(x), \bar{u}(x)), \\ D_{z'} F(x, z(x), z'(x), u_0(x)) &= D_{z'} F(x, z(x), z'(x), \bar{u}(x)), \\ F^0(x, z(x), z'(x), u_0(x)) &= F^0(x, z(x), z'(x), \bar{u}(x)) \end{aligned}$$

for $x \in I$ a.e. The above equalities imply the assertion of the lemma. □

The last lemma is a generalization of [3, ex. 10.5] to the case of a nonlinear functional.

LEMMA 6. *Let $u_* \in \mathcal{U}_M$ and $\varphi : I \times \mathbb{R}^r \to \mathbb{R}$ be of Carathéodory type; i.e., $\varphi(\cdot, u)$ is measurable on I for any $u \in \mathbb{R}^r$ and $\varphi(x, \cdot)$ is continuous on $\mathbb{R}^r$ for any $x \in I$.*

If

$$-\infty < \int_I \varphi(x, u_*(x))dx \leq \int_I \varphi(x, u(x))dx < +\infty$$

for any $u \in \mathcal{U}_M$, then

$$\varphi(x, u_*(x)) \leq \varphi(x, u)$$

for $x \in I$ a.e. and $u \in M$.

Proof. Suppose that the assertion of the lemma is false. Then, the set

$$\begin{aligned} \mathcal{R} &= \{x \in I; \exists_{u \in M} \quad \varphi(x, u) < \varphi(x, u_*(x))\} \\ &= \{x \in I; \exists_{u \in M_0} \quad \varphi(x, u) < \varphi(x, u_*(x))\} \\ &= \bigcup_{u \in M_0} \bigcup_{m \in \mathbb{N}} \Big\{x \in I; \quad \varphi(x, u) - \varphi(x, u_*(x)) < -\frac{1}{m}\Big\}, \end{aligned}$$

where M_0 is a countable, everywhere-dense subset of M, is measurable (in the Lebesgue sense) and $|\mathcal{R}| > 0$. From this we conclude that there exist $\bar{u} \in M_0$, $\bar{m} \in \mathbb{N}$ such that the set

$$\mathcal{R}_{\bar{u}, \bar{m}} = \Big\{x \in I; \quad \varphi(x, \bar{u}) < \varphi(x, u_*(x)) < \frac{-1}{\bar{m}}\Big\}$$

has a positive measure, i.e. $|\mathcal{R}_{\bar{u}, \bar{m}}| > 0$. □

Now, let us define the following element of $\mathcal{U}_M$:

$$\hat{u} : I \ni x \mapsto \begin{cases} \bar{u}, & x \in \mathcal{R}_{\bar{u}, \bar{m}}, \\ u_*(x), & x \notin \mathcal{R}_{\bar{u}, \bar{m}}. \end{cases}$$

We have

$$\begin{aligned} \int_I \varphi(x, \hat{u}(x))dx &= \int_{\mathcal{R}_{\bar{u}, \bar{m}}} \varphi(x, \bar{u})dx + \int_{I \setminus \mathcal{R}_{\bar{u}, \bar{m}}} \varphi(x, u_*(x))dx \\ &\quad + \int_{\mathcal{R}_{\bar{u}, \bar{m}}} \varphi(x, u_*(x))dx - \int_{\mathcal{R}_{\bar{u}, \bar{m}}} \varphi(x, u_*(x))dx \\ &= \int_I \varphi(x, u_*(x))dx + \int_{\mathcal{R}_{\bar{u}, \bar{m}}} (\varphi(x, \bar{u}) - \varphi(x, u_*(x)))dx \\ &\leq \int_I \varphi(x, u_*(x))dx + |\mathcal{R}_{\bar{u}, \bar{m}}| \cdot \left(\frac{-1}{\bar{m}}\right) < \int_I \varphi(x, u_*(x))dx. \end{aligned}$$

This contradicts our assumption. □

We shall say that a pair $(z_*, u_*) \in H_0^1(I, \mathbb{R}^n) \times \mathcal{U}_M$ is a local optimal solution of problem (S) if it satisfies the equality $\mathcal{F}(z_*, u_*) = 0$ and there exists a neighborhood V of z_* in $H_0^1(I, \mathbb{R}^n)$ such that

$$\mathcal{F}_0(z_*, u_*) \leq \mathcal{F}(z, u)$$

for any pair $(z, u) \in V \times \mathcal{U}_M$ satisfying the equality $\mathcal{F}(z, u) = 0$.

In the case when $V = H_0^1(I, \mathbb{R}^n)$, we shall say that (z_*, u_*) is a global optimal solution.

Now, we have Theorem 5.

THEOREM 5 (maximum principle). *Let us assume that the following conditions are satisfied:*

- —(5), (6), (7), (8a), (8b), (8c), (10), (11), *and* (13) *(guaranteeing the existence of a solution in* $H_0^1(I, \mathbb{R}^n)$ *of the equation* $\mathcal{F}(z, u) = 0$ *for any* $u \in \mathcal{U}_M$*);*
- —(14), (15), (16), *and* (17) *(guaranteeing the continuous Fréchet differentiability of* $\mathcal{F}$ *with respect to* $z \in H_0^1(I, \mathbb{R}^n)$*);*
- —*the function* $\mathcal{F}_0$ *satisfies the assumptions of Theorem* 1 *with* $\mathcal{F}$ *replaced by* $\mathcal{F}_0$ *(guaranteeing the continuous Fréchet differentiability of* $\mathcal{F}_0$ *with respect to* $z \in H_0^1(I, \mathbb{R}^n)$ *);*
- —(26) *(guaranteeing the fulfillment of the convexity conditions* (27), (28) *by* $\mathcal{F}$ *and* $\mathcal{F}_0$*).*

If a pair $(z_*, u_*) \in H_0^1(I, \mathbb{R}^n) \times \mathcal{U}_M$ *is such that* (24) *and* (25) *hold with* $A(\cdot)$, $B(\cdot)$, $C(\cdot)$ *given by* (23)*, then the local optimality of the pair* (z_*, u_*) *implies the existence of a function* $\lambda \in H_0^1(I, \mathbb{R}^n)$ *such that*

$$\begin{aligned}\frac{d}{dx}&(D_{z'}F_0(x, z_*(x), z_*'(x), u_*(x)) + A(x)\lambda(x) + B(x)\lambda'(x))\\ &= D_z F_0(x, z_*(x), z_*'(x), u_*(x)) + C(x)\lambda(x) + A(x)\lambda'(x)\end{aligned}$$

for $x \in I$ *a.e. and*

$$\begin{aligned}F_0(x, z^*(x), z_*'(x), u_*(x)) &+ D_z F(x, z_*(x), z_*'(x), u_*(x))\lambda(x)\\ &+ D_{z'}F(x, z_*(x), z_*'(x), u_*(x))\lambda'(x)\\ &= \min_{u \in M}\{F_0(x, z_*(x), z_*'(x), u) + D_z F(x, z_*(x), z_*'(x), u)\lambda(x)\\ &\qquad + D_{z'}F(x, z_*(x), z_*'(x), u)\lambda'(x)\}\end{aligned}$$

for $x \in I$ *a.e.*

Proof. From Lemmas 3, 4, 5 it follows that the assumptions of the extremum principle for a smooth-convex problem are satisfied.

So, there exist a constant $\lambda_0 \geq 0$ and $\Lambda \in [H_0^1(I, \mathbb{R}^n)]^{**}$ (not all zero) such that

$$\lambda_0 \partial_z \mathcal{F}_0(z_*, u_*) + (\partial_z \mathcal{F}(z_*, u_*))^* \Lambda = 0$$

and

$$\lambda_0 \mathcal{F}_0(z_*, u_*) + \Lambda(\mathcal{F}(z_*, u_*)) = \min_{u \in \mathcal{U}_M}\{\lambda_0 \mathcal{F}(z_*, u) + \Lambda(\mathcal{F}(z_*, u))\}.$$

Since the operator

$$\partial_z \mathcal{F}(z^*, u^*) : H_0^1(I, \mathbb{R}^n) \to [H_0^1(I, \mathbb{R}^n)]^*$$

is onto and, consequently, the regularity assumptions of the extremum principle are satisfied, therefore $\lambda_0 > 0$ and it may be assumed without loss of generality that $\lambda_0 = 1$.

The first of the above equalities may be written down as

$$(\partial_z \mathcal{F}_0(z_*, u_*), h) + \Lambda((\partial_z \mathcal{F}(z_*, u_*), h), \cdot) = 0$$

for any $h \in H_0^1(I, \mathbb{R}^n)$. Using the fact that

$$[H_0^1(I, \mathbb{R}^n)]^{**} \cong H_0^1(I, \mathbb{R}^n),$$

we assert that there exists a $\lambda \in H_0^1(I, \mathbb{R}^n)$ such that

$$(\partial_z \mathcal{F}_0(z_*, u_*), h) + ((\partial_z \mathcal{F}(z_*, u_*), h), \lambda) = 0$$

for any $h \in H_0^1(I, \mathbb{R}^n)$; i.e.,

$$\begin{aligned}\int_I &(D_{z'} F_0(x, z_*(x), z_*'(x), u_*(x)) \\ &+ D_{zz'} F(x, z_*(x), z_*'(x), u_*(x))\lambda(x) \\ &+ D_{z'z'} F(x, z_*(x), z_*'(x), u_*(x))\lambda'(x))h'(x)dx \\ &+ \int_I (D_z F_0(x, z_*(x), z_*'(x), u_*(x)) \\ &+ D_{zz} F(x, z_*(x), z_*'(x), u_*(x))\lambda(x) \\ &+ D_{zz'} F(x, z_*(x), z_*'(x), u_*(x))\lambda'(x))h(x)dx = 0\end{aligned}$$

for any $h \in H_0^1(I, \mathbb{R}^n)$; i.e.,

$$\begin{aligned}&\frac{d}{dx}(D_{z'} F_0(x, z^*(x), z_*'(x), u_*(x)) + D_{zz'} F(x, z^*(x), z_*'(x), u_*(x))\lambda(x) \\ &\quad + D_{z'z'} F(x, z^*(x), z_*'(x), u_*(x))\lambda^{'}(x)) = D_z F_0(x, z^*(x), z_*'(x), u_*(x)) \\ &\quad + D_{zz} F(x, z^*(x), z_*'(x), u_*(x))\lambda(x) + D_{zz'} F(x, z^*(x), z_*'(x), u_*(x))\lambda^{'}(x)\end{aligned}$$

for $x \in I$ a.e.

Analogously, we get

$$\begin{aligned}&\int_I (F_0(x, z^*(x), z_*'(x), u_*(x)) + D_z F(x, z^*(x), z_*'(x), u_*(x))\lambda(x) \\ &\quad + D_{z'} F(x, z^*(x), z_*'(x), u_*(x))\lambda^{'}(x))dx = \min_{u \in \mathcal{U}_M} \left\{ \int_I (F_0(x, z^*(x), z_*'(x), u(x)) \right. \\ &\quad \left. + D_z F(x, z^*(x), z_*'(x), u(x))\lambda(x) + D_{z'} F(x, z^*(x), z_*'(x), u(x))\lambda^{'}(x))dx \right\}.\end{aligned}$$

Consequently, from Lemma 6 we obtain

$$\begin{aligned}&F_0(x, z_*(x), z_*'(x), u_*(x)) + D_z F(x, z_*(x), z_*'(x), u_*(x))\lambda(x) \\ &\quad + D_{z'} F(x, z_*(x), z_*'(x), u_*(x))\lambda'(x) = \min_{u \in M} \{F_0(x, z_*(x), z_*'(x), u) \\ &\quad + D_z F(x, z_*(x), z_*'(x), u)\lambda(x) + D_{z'} F(x, z_*(x), z_*'(x), u)\lambda^{'}(x)\}\end{aligned}$$

for $x \in I$ a.e.

The proof is completed. □

4. The existence of an optimal solution. In this section we shall consider control problem (1)–(4) in the case when $M \subset \mathbb{R}^r$ is compact-convex and

$$F(x,z,z',u) = F_1(x,z,z') + F_2(x,z)u \tag{29}$$

for $x \in I$ a.e., z, $z' \in \mathbb{R}^n$, $u \in \mathbb{R}^r$.

We assume that F satisfies the assumptions of Theorem 4 guaranteeing the existence of a solution to system (1)–(2). Additionally, we assume that F_2 is of Carathéodory type and such that, for any $\lambda > 0$, there exists a function $g_\lambda(\cdot) \in L^2(I,\mathbb{R}_0^+)$ such that

$$|F_2(x,z)| \le g_\lambda(x) \tag{30}$$

for $x \in I$ a.e., $|z| < \lambda$.

Moreover, we shall assume that F is convex with respect to $(z,z') \in \mathbb{R}^n \times \mathbb{R}^n$ (cf. (10)), i.e., that

$$F(x,\alpha z + \beta w, \alpha z' + \beta w', u) \le \alpha F(x,z,z',u) + \beta F(x,w,w',u) \tag{31}$$

for $x \in I$ a.e., z, w, z', $w' \in \mathbb{R}^n$, $u \in \mathbb{R}^r$, α, $\beta \ge 0$, $\alpha + \beta = 1$.

The above assumption guarantees that (with a fixed control $u(\cdot)$) the set of all minimizers of f_u and the set of all solutions to (1)–(2) are identical.

Remark 7. One can omit assumption (31) and minimize functional (4) on the set $\{(z,u) \in H_0^1(I,\mathbb{R}^n) \times \mathcal{U}_M;\ \ z \ \text{ is a minimizer of } \ f_u\}$.

Let us denote by V_u the set of all solutions to (1)–(2) corresponding to a control $u(\cdot)$.

In the same way as in [15] one can show that there exists a constant $\rho > 0$ such that

$$\|z\|_{H_0^1(I,\mathbb{R}^n)} \le \rho$$

for all $z \in V_u$ and $u \in \mathcal{U}_M$.

Consequently, as in [15] one can prove the following lemma.

LEMMA 7. *If F, being of form* (29)*, satisfies the assumptions of Theorem* 4*,* (31)*, and F_2 satisfies* (30)*, $u_k \rightharpoonup u_0$ weakly in $L^2(I,\mathbb{R}^r)$, $u_k \in \mathcal{U}_M$, $k = 0,1,\ldots$, z_0 is a weak limit in $H_0^1(I,\mathbb{R}^n)$ of a subsequence of $(z_k)_{k\in\mathbb{N}}$, $z_k \in V_{u_k}$, then $z_0 \in V_{u_0}$.*

On the function $F_0 : I \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^r \to \mathbb{R}$ we assume in this section that

(32) $F_0(\cdot,z,z',u)$ is measurable on I for all z, $z' \in \mathbb{R}^n$, $u \in \mathbb{R}^r$, and $F_0(x,(\cdot,\cdot,\cdot))$ is continuous on $\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^r$ for $x \in I$ a.e.

(33) $F_0(x,z,(\cdot,\cdot))$ is convex on $\mathbb{R}^n \times M$ for $x \in I$ a.e., $z \in \mathbb{R}^n$,

(34) for any $\lambda > 0$, there exists a function $\Psi_\lambda \in L^1(I,\mathbb{R}_0^+)$ such that

$$F_0(x,z,z',u) \ge -\Psi_\lambda(x)$$

for $x \in I$ a.e., $|z| < \lambda$, $z' \in \mathbb{R}^n$, $u \in M$.

The above assumptions imply that

$$-\infty < \mathcal{F}_0(z,u) \le +\infty \tag{35}$$

for any $z \in H_0^1(I,\mathbb{R}^n)$, $u \in \mathcal{U}_M$.

Moreover, from [1, Thm. 10.8.i] it follows that if $u_k \rightharpoonup u_0$ weakly in $L^2(I,\mathbb{R}^r)$, $u_k \in \mathcal{U}_M$, $k = 0,1,\ldots$, $z_k \rightharpoonup z_0$ weakly in $H_0^1(I,\mathbb{R}^n)$, then

$$\liminf_{k\to\infty} \mathcal{F}_0(z_k,u_k) \ge \mathcal{F}_0(z_0,u_0). \tag{36}$$

Now, we can prove the following theorem.

THEOREM 6. *If the assumptions of Theorem* 4 *and conditions* (29)–(34) *are satisfied and* M *is compact-convex, then there exists a global optimal solution of problem* (1)–(4) *in* $H_0^1(I,\mathbb{R}^n)\times\mathcal{U}_M$.

Proof. Let us denote

$$\mathcal{A}=\bigcup_{u\in\mathcal{U}_M}(V_u\times\{u\})$$

and

$$m=\inf\{\mathcal{F}_0(z,u):(z,u)\in\mathcal{A}\}.$$

If $m=+\infty$, then the existence of an optimal solution is obvious.

So, assume that $-\infty\le m<+\infty$.

Let $(z_k,u_k)_{k\in\mathbb{N}}$ be a minimizing sequence. Of course, without loss of generality we may assume that $u_k\rightharpoonup u_0\in\mathcal{U}_M$ weakly in $L^2(I,\mathbb{R}^r)$ and $z_k\rightharpoonup z_0\in H_0^1(I,\mathbb{R}^n)$ weakly in $H_0^1(I,\mathbb{R}^n)$ (the latter follows from the boundedness of the set $\bigcup_{u\in\mathcal{U}_M}V_u$). From Lemma 7 it follows that $(z_0,u_0)\in\mathcal{A}$. Moreover, by (36),

$$m\le\mathcal{F}_0(z_0,u_0)\le\liminf_{k\to\infty}\mathcal{F}_0(z_k,u_k)=\lim_{k\to\infty}\mathcal{F}_0(z_k,u_k)=m.$$

This means that

$$m=\mathcal{F}_0(z_0,u_0);$$

i.e., in view of (35) $m>-\infty$ and (z_0,u_0) is a global optimal solution of (1)–(4).

The proof is completed. □

Acknowledgment. The author expresses his gratitude to unknown referees for their very valuable remarks and improvements.

REFERENCES

[1] L. CESARI *Optimization - Theory and Applications*, Springer-Verlag, New York, 1983.
[2] S. FUČIK AND A. KUFNER, *Nonlinear Differential Equations*, Elsevier, Amsterdam, 1980.
[3] I. V. GIRSANOV, *Lectures on Mathematical Theory of Extremum Problems*, Lectures Notes in Econom. and Math. Systems 67, Springer-Verlag, Berlin, Heidelberg, 1972.
[4] M. GOEBEL AND U. RAITUMS, *Constrained control of a nonlinear two point boundary value problem,* I, J. Global Optim., 4 (1994), pp. 367–395.
[5] M. R. HESTENES, *Calculus of Variations and Optimal Control Theory*, J. Wiley & Sons, New York, 1966.
[6] A. D. IOFFE AND V. M. TIKHOMIROV, *Theory of Extremal Problems*, North-Holland, Amsterdam, 1979.
[7] M. KISIELEWICZ, *Differential Inclusions and Optimal Control*, Kluwer Academic Publishers, Dordrecht, the Netherlands, and PWN - Polish Scientific Publishers, Warsaw, Poland, 1991.
[8] M. A. KRASNOSELSKII, *Topological Methods in the Theory of Nonlinear Integral Equations*, Pergamon Press, Elmsford, NY, 1964.
[9] J. MAWHIN, *Problèmes de Dirichlet Variationnels Non-Linéaires*, L'Université de Montréal, 1987.
[10] J. MAWHIN, *Le principe de moindre action: de la théologie au calcul*, Bulletin de la Classe des Sciences de l'Academie Royale de Belgique, 6 (1992), pp. 413–427.
[11] J. MAWHIN AND M. WILLEM, *Critical Point Theory and Hamiltonian Systems*, Springer-Verlag, New York, 1989.
[12] I. P. NATANSON *Theory of the Functions of Real Variable*, Moscow, 1950 (in Russian).

[13] B. T. Poljak, *Semicontinuity of integral functionals and existence theorems in extremal problems*, Mat. Sb., 78 (1969), pp. 65–84.
[14] V. M. Tikhomirov, *Fundamental Principles of the Theory of Extremal Problems*, J. Wiley & Sons, Chichester, UK, 1986.
[15] S. Walczak, *On the continuous dependence on parameters of solutions of the Dirichlet problem*, Bulletin de la Classe des Sciences de l'Academie Royale de Belgique, 6, 7–12 (1995), pp. 247–261.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 4, pp. 1268–1292, July 1998

009

COPRIME FACTORIZATIONS AND WELL-POSED LINEAR SYSTEMS*

OLOF J. STAFFANS†

Abstract. We study the basic notions related to the stabilization of an infinite-dimensional well-posed liner system in the sense of Salamon and Weiss. We first introduce an appropriate stabilizability and detectability notion and show that if a system is jointly stabilizable and detectable then its transfer function has a doubly coprime factorization in H^∞. The converse is also true: every function with a doubly coprime factorization in H^∞ is the transfer function of a jointly stabilizable and detectable well-posed linear system. We show further that a stabilizable and detectable system is stable if and only if its input/output map is stable. Finally, we construct a dynamic, possibly non-well-posed, stabilizing compensator. The notion of stability that we use is the natural one for the quadratic cost minimization problem, and it does not imply exponential stability.

Key words. stabilizability, detectability, input/output stability

AMS subject classifications. 93A05, 93B05, 93B07

PII. S0363012995285417

1. Introduction.

Although the theory of well-posed linear systems in the sense of Salamon and Weiss has been around for some time, applications of this theory to "real" control problems are scarce. This is in sharp contrast to the widespread use of the theory of Pritchard–Salamon systems; see, e.g., [2], [12], [15], [16], and [26] for discussions of different aspects of this theory. A fair number of recent pure frequency domain results for H^∞ transfer functions do exist (some of these are listed in the References), but they have not been connected to the theory of well-posed linear systems. The few connections known to us in the spring of 1995 when the first version of this paper was written were the discussion of the Lyapunov equation in [10], the discussion on feedback and estimation of well-posed systems in [13], the discussion of balanced realizations in [14], the discussion of the connection between internal and external stability in [17], and the discussions of the quadratic cost minimization problem in [19] and [24] (not to mention the basic papers [20], [30], and [31] and the nice review [1]); the list above is certainly not complete. In particular, at that time we were not able to find any reasonably complete results on the stabilizability and detectability of general well-posed linear systems, and the connection of these notions to the notion of a (doubly) coprime factorization of the transfer function of the system.[1] We needed these results in order to solve the quadratic cost minimization problem for unstable systems, and we were forced to develop the needed stabilization theory ourselves.[2]

Subsequently the situation changed significantly with the appearance of the preprints [4], [5], [6], [11], [32], and [33]. Out of these [4], [5], and [6] are fairly closely related to our work. Several of the results that we prove here are also found in [5], in a slightly less general setting. The results given in [4] and [6] overlap those that we

*Received by the editors May 3, 1995; accepted for publication (in revised form) July 10, 1997; published electronically May 15, 1998.

http://www.siam.org/journals/sicon/36-4/28541.html

†Department of Mathematics, Åbo Akademi University, FIN-20500 Åbo, Finland (Olof.Staffans@abo.fi).

[1] Throughout [13] Morris takes the observation operator to be bounded.

[2] Our solution to the unstable quadratic cost minimization problem is based on the present work, and it is presented in [25].

present in section 5.[3] These preprints have also had a certain influence on a revision carried out in late 1996: we were able to use the ideas presented in [5] to simplify our original proofs of Theorems 4.4 and 5.3, and Remark 5.5 was inspired by [5] and [6]. In addition we added Lemma 3.17 and Corollary 3.18 as an answer to a question asked by George Weiss in the summer of 1996, and we redrew all the diagrams.

The notion of stability that we use was forced upon us by our solution to the quadratic cost minimization problem for stable well posed linear systems. In particular, it does not imply exponential stability. The most important notion is what we call "joint stabilizability and detectability." The word "joint" refers to the fact that in our setting the notions of stabilizability and detectability are not decoupled from each other as they are in the Pritchard–Salamon theory; roughly speaking, the "feedback operator and the output injection operator must be compatible." This is a problem that is not present in a Pritchard–Salamon system, due to the fact that for such systems the admissibility of a particular control operator together with the admissibility of a particular observation operator implies the well posedness of the corresponding input/output map, something that is not true for general well posed linear systems. Maybe the main individual result of this paper is the statement that if a system is jointly stabilizable and detectable, then its transfer function has a doubly coprime factorization in H^∞.[4] The converse is also true: every function that has a doubly coprime factorization in H^∞ is the transfer function of a jointly strongly stabilizable and detectable well-posed linear system.

Having explained the relation between joint stabilizability and detectability on one hand and doubly coprime factorizations on the other hand, we continue with a short discussion of how to use an observer as a stabilizing dynamic compensator. This theory parallels the classical theory, apart from the fact that the well posedness of the observer is not automatically guaranteed.

For the convenience of the reader, we start with a short presentation of well-posed linear systems.

We use the following notation.

$\mathcal{L}(U;Y)$, $\mathcal{L}(U)$: The set of bounded linear operators from U into Y or from U into itself, respectively.

I: The identity operator.

A^*: The (Hilbert space) adjoint of the operator A.

$\mathbf{R}$, $\mathbf{R}^+$, $\mathbf{R}^-$: $\mathbf{R} = (-\infty, \infty)$, $\mathbf{R}^+ = [0, \infty)$, and $\mathbf{R}^- = (-\infty, 0]$.

$L^2(J;U)$: The set of U-valued L^2-functions on the interval J.

$L^2_\omega(J;U)$: $L^2_\omega(J;U) = \left\{ u \in L^2_{\mathrm{loc}}(J;U) \;\middle|\; (t \mapsto \mathrm{e}^{-\omega t}u(t)) \in L^2(J;U) \right\}$.

$H^\infty_\omega(U;Y)$: The set of bounded analytic $\mathcal{L}(U;Y)$-valued functions over the half-plane $\Re z > \omega$, with the sup-norm.

$TI_\omega(U;Y)$, $TI_\omega(U)$: The set of bounded linear time-invariant operators from $L^2_\omega(\mathbf{R};U)$ into $L^2_\omega(\mathbf{R};Y)$ or from $L^2_\omega(\mathbf{R};U)$ into itself.

$TIC_\omega(U;Y)$, $TIC_\omega(U)$: The set of causal operators in $TI_\omega(U;Y)$ or $TI_\omega(U)$.

$\langle \cdot, \cdot \rangle_H$: The inner product in the Hilbert space H.

[3] Curtain, Weiss, and Weiss call the compensator in Theorem 5.3 a "controller with internal loop."

[4] As we mentioned above, the main purpose of this paper is not to prove any particular result but to develop a general theory that can be used in the study of the quadratic cost minimization problem, to which we return in [25].

$\tau(t)$: The time shift group $\tau(t)u(s) = u(t+s)$ (this is a left shift when $t > 0$ and a right shift when $t < 0$).

π_J: $(\pi_J u)(s) = u(s)$ if $s \in J$ and $(\pi_J u)(s) = 0$ if $s \notin J$. Here J is a subset of $\mathbf{R}$.

$\pi_+,\ \pi_-$: $\pi_+ = \pi_{\mathbf{R}^+}$ and $\pi_- = \pi_{\mathbf{R}^-}$.

We extend an L^2_ω-function u defined on a subinterval J of $\mathbf{R}$ to the whole real line by requiring u to be zero outside of J, and we denote the extended function by $\pi_J u$. Thus we use the same symbol π_J both for the embedding operator $L^2_\omega(J) \to L^2_\omega(\mathbf{R})$ and for the corresponding projection operator $L^2_\omega(\mathbf{R}) \to L^2_\omega(J)$. With this interpretation, $\pi_J L^2_\omega(\mathbf{R};U) = L^2_\omega(J;U) \subset L^2_\omega(\mathbf{R};U)$ for each interval $J \subset \mathbf{R}$.

2. A review of well-posed linear systems and time-invariant operators. In order to fix the notation and describe the basic setting we first give a brief presentation of the theory of the Salamon–Weiss class of well-posed linear systems. This theory has been developed in [18], [19], [20], [3], [7], and [27], [28], [30], [31] (and many other papers), and we refer the reader to these sources for additional reading.[5] A recent contribution is found in [24], and the setting that we use here is a slight extension of the one in [24]. The difference is that the discussion in [24] is restricted to the case of (externally) stable systems; here we also need to consider unstable systems. The major parts of this setting are found in [20], too.

In order to formulate the axioms satisfied by a well-posed linear system we introduce exponentially weighted L^2-spaces. For each Hilbert space U and each $\omega \in \mathbf{R}$ we let $L^2_\omega(\mathbf{R};U)$ be the weighted L^2-space

$$L^2_\omega(\mathbf{R};U) = \big\{\, u \in L^2_{\mathrm{loc}}(\mathbf{R};U) \;\big|\; \big(t \mapsto \mathrm{e}^{-\omega t}u(t)\big) \in L^2(\mathbf{R};U) \,\big\}.$$

This is a Hilbert space with the natural norm $\|\mathrm{e}^{-\omega\cdot}u(\cdot)\|_{L^2(\mathbf{R};U)}$. We also need the "past time" projection π_-, the "future time" projection π_+, and the "time shift" group $\tau(t)$ that operate on functions $u \in L^2_\omega(\mathbf{R};U)$ in the following way:

$$(\pi_- u)(s) = \begin{cases} u(s), & s \in \mathbf{R}^-, \\ 0, & s \in \mathbf{R}^+, \end{cases}$$

$$(\pi_+ u)(s) = \begin{cases} u(s), & s \in \mathbf{R}^+, \\ 0, & s \in \mathbf{R}^-, \end{cases}$$

$$(\tau(t)u)(s) = u(t+s), \quad t, s \in \mathbf{R}.$$

DEFINITION 2.1. *Let U, H, and Y be Hilbert spaces, and let $\omega \in \mathbf{R}$. A (causal) ω-stable well-posed linear system on (U,H,Y) is a quadruple $\Psi = \left[\begin{smallmatrix} \mathcal{A} & \mathcal{B} \\ \mathcal{C} & \mathcal{D} \end{smallmatrix}\right]$, where $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, and $\mathcal{D}$ are bounded linear operators of the following type:*

(i) *$\mathcal{A}(t)\colon H \to H$ is a strongly continuous semigroup of bounded linear operators on H satisfying $\sup_{t\in\mathbf{R}^+} \|\mathrm{e}^{-\omega t}\mathcal{A}(t)\| < \infty$;*

(ii) *$\mathcal{B}\colon L^2_\omega(\mathbf{R};U) \to H$ satisfies $\mathcal{A}(t)\mathcal{B}u = \mathcal{B}\tau(t)\pi_- u$ for all $u \in L^2_\omega(\mathbf{R};U)$ and $t \in \mathbf{R}^+$;*

(iii) *$\mathcal{C}\colon H \to L^2_\omega(\mathbf{R};Y)$ satisfies $\mathcal{C}\mathcal{A}(t)x = \pi_+\tau(t)\mathcal{C}x$ for all $x \in H$ and $t \in \mathbf{R}^+$;*

(iv) *$\mathcal{D}\colon L^2_\omega(\mathbf{R};U) \to L^2_\omega(\mathbf{R};Y)$ satisfies $\tau(t)\mathcal{D}u = \mathcal{D}\tau(t)u$, $\pi_-\mathcal{D}\pi_+ u = 0$, and $\pi_+\mathcal{D}\pi_- u = \mathcal{C}\mathcal{B}u$ for all $u \in L^2_\omega(\mathbf{R};U)$ and $t \in \mathbf{R}$.*

[5] In the early literature these systems were called "well-posed semigroup control systems" by Salamon and "abstract linear systems" by Weiss.

If, in addition, $\mathrm{e}^{-\omega t}\mathcal{A}(t)x \to 0$ *as* $t \to \infty$ *for all* $x \in H$*, then* Ψ *is strongly* ω*-stable. The system* Ψ *is [strongly]*[6] *stable iff it is [strongly]* ω*-stable with* $\omega = 0$*, and it is exponentially stable iff it is* ω*-stable for some* $\omega < 0$*.*

The different components of Ψ *are named as follows:* U *is the input space,* H *is the state space,* Y *is the output space,* $\mathcal{A}$ *is the semigroup,* $\mathcal{B}$ *is the controllability map,* $\mathcal{C}$ *is the observability map, and* $\mathcal{D}$ *is the input/output map of* Ψ*.*

This is the same definition as [24, Definition 2.1], except that throughout we took $\omega = 0$ and did not put any restrictions on the growth rate of the semigroup $\mathcal{A}$.

The axioms listed above describe standard properties of the corresponding maps induced by systems with bounded control and observation operators. Whenever when we refer to an ω-stable "classical" system, we mean a system of the following type: we let A be the generator of a semigroup $\mathcal{A}$ on a Hilbert space H satisfying $\sup_{t\in\mathbf{R}^+} \|\mathrm{e}^{\epsilon t}\mathrm{e}^{-\omega t}\mathcal{A}(t)\| < \infty$ for some $\epsilon > 0$, let U and Y be Hilbert spaces, let $B \in \mathcal{L}(U;H)$, $C \in \mathcal{L}(H;Y)$, and $D \in \mathcal{L}(U;Y)$, and consider the system

$$\begin{aligned} x'(t) &= Ax(t) + Bu(t), \\ y(t) &= Cx(t) + Du(t), \qquad t \geq T, \\ x(T) &= x_T, \end{aligned} \tag{2.1}$$

where T is a given initial time and x_T is a given initial value. We call u the control, x the state, y the output (or observation), A the generator, B the control operator, C the observation operator, and D the feed-through operator of this classical system. The state x is required to be a strong solution of (2.1); i.e., the state x and output y are given by

$$x(t) = \mathcal{A}(t)x_T + \int_T^t \mathcal{A}(t-s)Bu(s)\,ds, \quad t \geq T, \tag{2.2}$$

$$y(t) = C\mathcal{A}(t)x_T + \int_T^t C\mathcal{A}(t-s)Bu(s)\,ds + Du(t), \quad t \geq T. \tag{2.3}$$

In this case we define $\mathcal{B}$, $\mathcal{C}$, and $\mathcal{D}$ by

$$\mathcal{B}u = \int_{-\infty}^0 \mathcal{A}(-s)Bu(s)\,ds, \tag{2.4}$$

$$\mathcal{C}x = \left(t \mapsto C\mathcal{A}(t)x,\ t \in \mathbf{R}^+\right), \tag{2.5}$$

$$\mathcal{D}u = \left(t \mapsto \int_{-\infty}^t C\mathcal{A}(t-s)Bu(s)\,ds + Du(t),\ t \in \mathbf{R}\right). \tag{2.6}$$

Thus $\mathcal{B}$ is the mapping from the control $u \in L^2_\omega(\mathbf{R}^-;U)$ to the final state $x(0) \in H$ (take $T = -\infty$, $x_T = 0$, and $t = 0$), $\mathcal{C}$ is the mapping from the initial state $x_0 \in H$ to the output $y \in L^2_\omega(\mathbf{R}^+;Y)$ (take $T = 0$ and $u = 0$), and $\mathcal{D}$ is the mapping from the control $u \in L^2_\omega(\mathbf{R};U)$ to the output $y \in L^2_\omega(\mathbf{R};Y)$ (take $T = -\infty$ and $x_T = 0$). We leave the easy proof of the fact that these operators indeed are bounded linear operators between the given spaces to the reader.

Each well-posed linear system Ψ has a *controlled state* and an *output*. Depending on whether the initial time is finite or infinite these are defined in two slightly different ways as follows.

[6]Square brackets represent optional parts of a sentence. Statements in square brackets are supposed to be true (a) if you omit all square brackets (single or double), (b) if you keep the single brackets, (c) if you keep the double brackets.

DEFINITION 2.2. *Let $\Psi = \left[\begin{smallmatrix} \mathcal{A} & \mathcal{B} \\ \mathcal{C} & \mathcal{D} \end{smallmatrix}\right]$ be an ω-stable well-posed linear system on (U, H, Y), and let $u \in L^2_\omega(\mathbf{R}; U)$. In the time-invariant setting (corresponding to the initial time $-\infty$) the controlled state $x(t)$ at time $t \in \mathbf{R}$ and the output y of Ψ with control u are given by*

$$\begin{bmatrix} x(t) \\ y \end{bmatrix} = \begin{bmatrix} \mathcal{B}\tau(t)u \\ \mathcal{D}u \end{bmatrix},$$

and in the initial value setting with initial time s, initial value $x(s)$, and control u, the controlled state $x(t)$ at time $t \geq s$ and the output y of Ψ are given by

$$\begin{bmatrix} x(t) \\ y \end{bmatrix} = \begin{bmatrix} \mathcal{A}(t-s) & \mathcal{B}\tau(t) \\ \tau(-s)\mathcal{C} & \mathcal{D} \end{bmatrix} \begin{bmatrix} x(s) \\ \pi_{[s,\infty)}u \end{bmatrix} = \begin{bmatrix} \mathcal{A}(t-s)x(s) + \mathcal{B}\tau(t)\pi_{[s,\infty)}u \\ \tau(-s)\mathcal{C}x(s) + \mathcal{D}\pi_{[s,\infty)}u \end{bmatrix},$$

where $\pi_{[s,\infty)} = \tau(-s)\pi_+\tau(s)$ is given by

$$(\pi_{[s,\infty)}u)(t) = \begin{cases} u(t), & t \geq s, \\ 0, & t < s. \end{cases}$$

In particular, in the initial value setting with initial time zero, initial value x_0, and control u, the controlled state $x(t)$ at time $t \in \mathbf{R}^+$ and the output y of Ψ are given by

$$\begin{bmatrix} x(t) \\ y \end{bmatrix} = \begin{bmatrix} \mathcal{A}(t) & \mathcal{B}\tau(t) \\ \mathcal{C} & \mathcal{D} \end{bmatrix} \begin{bmatrix} x_0 \\ \pi_+u \end{bmatrix} = \begin{bmatrix} \mathcal{A}(t)x_0 + \mathcal{B}\tau(t)\pi_+u \\ \mathcal{C}x_0 + \mathcal{D}\pi_+u \end{bmatrix}.$$

Let us remark that the most commonly studied problem is the initial value problem with initial time zero, and in most papers this is the only one that is treated.

REMARK 2.3. *Because of Definition 2.2, we shall frequently use the alternative notation $\left[\begin{smallmatrix} \mathcal{A} & \mathcal{B}\tau \\ \mathcal{C} & \mathcal{D} \end{smallmatrix}\right]$ for the well-posed linear system $\left[\begin{smallmatrix} \mathcal{A} & \mathcal{B} \\ \mathcal{C} & \mathcal{D} \end{smallmatrix}\right]$.*

In the case of the classical ω-stable system (2.1) with bounded control operator B, bounded observation operator C, and control $u \in L^2_\omega(\mathbf{R}; U)$, in the time-invariant setting the state x and output y of Ψ are given by

$$x(t) = \int_{-\infty}^{t} \mathcal{A}(t-s)Bu(s)\,ds, \quad t \in \mathbf{R}, \tag{2.7}$$

$$y(t) = \int_{-\infty}^{t} C\mathcal{A}(t-s)Bu(s)\,ds + Du(t), \quad t \in \mathbf{R}, \tag{2.8}$$

and in the initial value setting with initial time T, initial value x_T, and control u, the state and output are given by (2.2) and (2.3).

An important fact is that the number ω in Definition 2.1 is not uniquely determined as described in Lemma 2.4.

LEMMA 2.4. *Let $\Psi = \left[\begin{smallmatrix} \mathcal{A} & \mathcal{B} \\ \mathcal{C} & \mathcal{D} \end{smallmatrix}\right]$ be an ω-stable well-posed linear system on (U, H, Y). Then Ψ is also an α-stable well-posed linear system on (U, H, Y) for every $\alpha > \omega$. If instead $\alpha < \omega$, then Ψ has a unique extension to an α-stable well-posed linear system on (U, H, Y) iff $\sup_{t \in \mathbf{R}^+} \|\mathrm{e}^{-\alpha t}\mathcal{A}(t)\| < \infty$ and the operators $\mathcal{B}$, $\mathcal{C}$, and $\mathcal{D}$ can be extended to bounded linear operators in $\mathcal{L}(L^2_\alpha(\mathbf{R}; U); H)$, $\mathcal{L}(H; L^2_\alpha(\mathbf{R}; Y))$, and $\mathcal{L}(L^2_\alpha(\mathbf{R}; U); L^2_\alpha(\mathbf{R}; Y))$, respectively.*

Proof. The first claim follows from Lemma 2.9 below and the fact that if $\alpha > \omega$, then $L^2_\alpha(\mathbf{R}^-; U) \subset L^2_\omega(\mathbf{R}^-; U)$ and $L^2_\omega(\mathbf{R}^+; Y) \subset L^2_\alpha(\mathbf{R}^+; Y)$. To prove the second claim it suffices to observe that $L^2_\omega \cap L^2_\alpha$ is dense in L^2_α. ∎

DEFINITION 2.5. *We call Ψ a well-posed linear system on (U, H, Y) iff it is an ω-stable well-posed linear system on (U, H, Y) for some $\omega \in \mathbf{R}$. The infimum of all the numbers ω for which Ψ is ω-stable is the exponential growth rate of Ψ. Thus, Ψ is exponentially stable iff its exponential growth rate is negative.*

As Salamon [20] and Weiss [27], [28], [30] have shown, the growth rate of a system Ψ is equal to the growth rate of its semigroup as explained in Lemma 2.6.

LEMMA 2.6. *The exponential growth rate of a well-posed linear system Ψ is equal to the exponential growth rate $\omega = \lim_{t\to\infty} t^{-1}\log(\|\mathcal{A}(t)\|)$ of its semigroup. In particular, Ψ is exponentially stable iff its semigroup is exponentially stable.*

See [20, Lemma 2.1] or [27, Proposition 2.5], [28, Proposition 2.3], and [30, Proposition 4.1] for proofs.

One of the required properties of the input/output operator $\mathcal{D}$ of Ψ is that it is time invariant.

DEFINITION 2.7. *Let U and Y be two Hilbert spaces. A bounded linear operator $\mathcal{D}\colon L^2_\omega(\mathbf{R};U) \to L^2_\omega(\mathbf{R};Y)$ is time invariant iff it commutes with time shifts; i.e., $\tau(t)\mathcal{D}u = \mathcal{D}\tau(t)u$ for all $u \in L^2_\omega(\mathbf{R};U)$ and all $t \in \mathbf{R}$. We denote this class of operators by $TI_\omega(U;Y)$. The Hankel operator induced by $\mathcal{D}$ is the operator $\pi_+\mathcal{D}\pi_-$, and the anti-Hankel operator induced by $\mathcal{D}$ is the operator $\pi_-\mathcal{D}\pi_+$. The Toeplitz operator induced by $\mathcal{D}$ is the operator $\pi_+\mathcal{D}\pi_+$, and the anti-Toeplitz operator induced by $\mathcal{D}$ is the operator $\pi_-\mathcal{D}\pi_-$.*

The word "causal" that we have included in the definition of a well-posed linear system relates to the fact that all the components of Ψ in Definition 2.1 are causal as follows.

DEFINITION 2.8. *An operator $\mathcal{B}\colon L^2_\omega(\mathbf{R};U) \to H$ is causal [anticausal] if $\mathcal{B}\pi_+ = 0$ [$\mathcal{B}\pi_- = 0$]. An operator $\mathcal{C}\colon H \to L^2_\omega(\mathbf{R};Y)$ is causal [anticausal] if $\pi_-\mathcal{C} = 0$ [$\pi_+\mathcal{C} = 0$]. A time-invariant operator $\mathcal{D}\colon L^2_\omega(\mathbf{R};U) \to L^2_\omega(\mathbf{R};Y)$ is causal [anticausal] if $\pi_-\mathcal{D}\pi_+ = 0$ [$\pi_+\mathcal{D}\pi_- = 0$], and it is static if it is both causal and anticausal. We denote the class of bounded linear time invariant causal operators by $TIC_\omega(U;Y)$.*

Thus, the condition imposed on $\mathcal{D}$ in Definition 2.1 requires that $\mathcal{D} \in TIC_\omega(U;Y)$ (i.e., $\mathcal{D}$ is time invariant and causal) and that the Hankel operator induced by $\mathcal{D}$ is equal to $\mathcal{CB}$. Intuitively, a causal controllability map $\mathcal{B}$ maps past inputs into the present state, a causal observability map $\mathcal{C}$ maps the present state into future outputs, and the past output of a causal input/output map $\mathcal{D}$ does not depend on future inputs.

As is well known, there is a one-to-one correspondence between $TIC_\omega(U;Y)$ and the set of $\mathcal{L}(U;Y)$-valued H^∞-functions over the half-plane $\Re z > \omega$. We denote this set of functions by $H^\infty_\omega(U;Y)$. The norm in this space is the usual H^∞-norm.

LEMMA 2.9. *The two spaces $TIC_\omega(U;Y)$ and $H^\infty_\omega(U;Y)$ are isometrically isomorphic. More precisely, to each operator $\mathcal{D} \in TIC_\omega(U;Y)$ there corresponds a unique function $\widehat{\mathcal{D}} \in H^\infty_\omega(U;Y)$ such that for each $u \in L^2_\omega(\mathbf{R}^+;U)$, the Laplace transform of $\mathcal{D}u$ is given by $\widehat{\mathcal{D}}(z)\widehat{u}(z)$, $\Re z > \omega$, where $\widehat{u}$ is the Laplace transform of u. The function $\widehat{\mathcal{D}}$ is called the transfer function (or symbol) of $\mathcal{D}$.*

Thus, intuitively, $\widehat{\mathcal{D}}$ is the Laplace transform of $\mathcal{D}$. This result is classic; see, for example, [8] or [29].

By a result due to Salamon [20, Section 4], a time-invariant operator $\mathcal{D}$ can be interpreted as the input/output operator of a well-posed linear system Ψ iff it belongs to $TIC_\omega(U;Y)$ for some $\omega \in \mathbf{R}$. Such a system is called a *realization* of $\mathcal{D}$. Two particular realizations are described below.

DEFINITION 2.10. *Let $\omega \in \mathbf{R}$ and $\mathcal{D} \in TIC_\omega(U;Y)$, and let τ be the left-shift group. Then the ω-stable well-posed linear system*

$$\begin{bmatrix} \tau\pi_- & \pi_- \\ \pi_+\mathcal{D}\pi_- & \mathcal{D} \end{bmatrix}$$

on $(U, L^2_\omega(\mathbf{R}^-;U), Y)$ *is called the exactly controllable realization of* $\mathcal{D}$*, and the strongly* ω*-stable well-posed linear system*

$$\begin{bmatrix} \pi_+\tau\pi_+ & \pi_+\mathcal{D}\pi_- \\ I & \mathcal{D} \end{bmatrix}$$

on $(U, L^2_\omega(\mathbf{R}^+;Y), Y)$ *is called the continuously observable realization of* $\mathcal{D}$*.*

Indeed, it is obvious that the two systems defined above satisfy the requirements of Definition 2.1. See [20, section 4] for an explanation of the names of these realizations.

Occasionally we shall need to discuss the stability of the different parts of Ψ separately, and for this purpose we further introduce the following natural terminology.

DEFINITION 2.11. *Let* $\Psi = \left[\begin{smallmatrix} \mathcal{A} & \mathcal{B} \\ \mathcal{C} & \mathcal{D} \end{smallmatrix}\right]$ *be a well-posed linear system on* (U, H, Y)*. Then*

(i) $\mathcal{A}$ *is* ω*-stable iff* $\sup_{t\in\mathbf{R}^+} \|\mathrm{e}^{-\omega t}\mathcal{A}(t)\| < \infty$ *and strongly* ω*-stable iff* $\mathrm{e}^{-\omega t}\mathcal{A}(t)x \to 0$ *as* $t \to \infty$ *for all* $x \in H$*,*

(ii) $\mathcal{B}$ *is* ω*-stable iff* $\mathcal{B} \in \mathcal{L}(L^2_\omega(\mathbf{R};U);H)$*,*

(iii) $\mathcal{C}$ *is* ω*-stable iff* $\mathcal{C} \in \mathcal{L}(H;L^2_\omega(\mathbf{R};Y))$*,*

(iv) $\mathcal{D}$ *is* ω*-stable iff* $\mathcal{D} \in TIC_\omega(U;Y)$*.*

As before, stability of a component of Ψ *means* ω*-stability with* $\omega = 0$*, and exponential stability means* ω*-stability for some* $\omega < 0$*.*

REMARK 2.12. *Almost all results presented below remain valid if we throughout drop the assumption that* $\mathcal{A}$ *is* ω*-stable in condition* (i) *of Definition* 2.1*. Thus for our purposes it suffices if the system is input* ω*-stable* (*condition* (ii))*, output* ω*-stable* (*condition* (iii))*, and input/output* ω*-stable* (*condition* (iv) *of Definition* 2.11)*. In* [24] *this situation was referred to as* external ω*-stability.*

One gets the *adjoint* of a system Ψ by replacing each operator by its adjoint and exchanging the controllability and observability maps with each other. In the computation of the adjoints of $\mathcal{B}$, $\mathcal{C}$, and $\mathcal{D}$ we use the ordinary (unweighted) L^2-inner product. This means that the resulting operators are bounded linear operators on $L^2_{-\omega}$ instead of bounded linear operators on L^2_ω. Moreover, causality is replaced by anticausality. The resulting system is an $(-\omega)$-stable anticausal system of the following type.

DEFINITION 2.13. *Let* Y*,* H*, and* U *be Hilbert spaces. An anticausal* ω*-stable well-posed linear* L^2*-system on* (Y, H, U) *is a quadruple* $\Psi^* = \left[\begin{smallmatrix} \mathcal{A}^* & \mathcal{C}^* \\ \mathcal{B}^* & \mathcal{D}^* \end{smallmatrix}\right]$*, where* $\mathcal{A}^*$*,* $\mathcal{C}^*$*,* $\mathcal{B}^*$*, and* $\mathcal{D}^*$ *are bounded linear operators of the following type:*

(i) $\mathcal{A}^*(t)\colon H \to H$ *is a strongly continuous semigroup of bounded linear operators on* H *satisfying* $\sup_{t\in\mathbf{R}^+} \|\mathrm{e}^{\omega t}\mathcal{A}^*(t)\| < \infty$*;*

(ii) $\mathcal{C}^*\colon L^2_\omega(\mathbf{R};Y) \to H$ *satisfies* $\mathcal{A}^*(-s)\mathcal{C}^*y^* = \mathcal{C}^*\tau(s)\pi_+y^*$ *for all* $y^* \in L^2_\omega(\mathbf{R};Y)$ *and* $s \in \mathbf{R}^-$*;*

(iii) $\mathcal{B}^*\colon H \to L^2_\omega(\mathbf{R};U)$ *satisfies* $\mathcal{B}^*\mathcal{A}^*(-s)x^* = \pi_-\tau(s)\mathcal{B}^*x^*$ *for all* $x^* \in H$ *and* $s \in \mathbf{R}^-$*;*

(iv) $\mathcal{D}^*\colon L^2(\mathbf{R};Y) \to L^2_\omega(\mathbf{R};U)$ *satisfies* $\tau(s)\mathcal{D}^*y^* = \mathcal{D}^*\tau(s)y^*$*,* $\pi_+\mathcal{D}^*\pi_-y^* = 0$*, and* $\pi_-\mathcal{D}^*\pi_+y^* = \mathcal{B}^*\mathcal{C}^*y^*$ *for all* $y^* \in L^2_\omega(\mathbf{R};Y)$ *and* $s \in \mathbf{R}$*.*

If, in addition, $\mathrm{e}^{\omega t}\mathcal{A}^*(t)x^* \to 0$ *as* $t \to \infty$ *for all* $x^* \in H$*, then* Ψ^* *is strongly* ω*-stable. The system* Ψ^* *is [strongly] stable iff it is [strongly]* ω*-stable with* $\omega = 0$*, and it is exponentially stable iff it is* ω*-stable for some* $\omega > 0$*.*

The different components of Ψ^* *are named as follows:* Y *is the input space,* H *is the state space,* U *is the output space,* $\mathcal{A}^*$ *is the semigroup,* $\mathcal{C}^*$ *is the controllability*

map, $\mathcal{B}^*$ is the observability map, and $\mathcal{D}^*$ is the input/output map of Ψ^*.

The controlled *state* and *output* of Ψ^* are defined as follows.

DEFINITION 2.14. *Let $\Psi^* = \left[\begin{smallmatrix} \mathcal{A}^* & \mathcal{C}^* \\ \mathcal{B}^* & \mathcal{D}^* \end{smallmatrix}\right]$ be an ω-stable anticausal well-posed linear system on (Y, H, U) and let $y^* \in L^2_\omega(\mathbf{R}; Y)$. In the time-invariant setting the controlled state $x^*(s)$ at time $s \in \mathbf{R}$ and the output u^* of Ψ^* with control y^* are given by*

$$\begin{bmatrix} x^*(s) \\ u^* \end{bmatrix} = \begin{bmatrix} \mathcal{C}^*\tau(s)y^* \\ \mathcal{D}^*y^* \end{bmatrix},$$

and in the initial value setting with initial time t, initial value $x^*(t)$, and control y^*, the controlled state $x^*(s)$ at time $s \le t$ and the output u^* of Ψ^* are given by

$$\begin{bmatrix} x^*(s) \\ u^* \end{bmatrix} = \begin{bmatrix} \mathcal{A}^*(t-s) & \mathcal{C}^*\tau(s) \\ \tau(-t)\mathcal{B}^* & \mathcal{D}^* \end{bmatrix} \begin{bmatrix} x^*(t) \\ \pi_{(-\infty,t]}y^* \end{bmatrix} = \begin{bmatrix} \mathcal{A}^*(t-s)x^*(t) + \mathcal{C}^*\tau(s)\pi_{(-\infty,t]}y^* \\ \tau(-t)\mathcal{B}^*x^*(t) + \mathcal{D}^*\pi_{(-\infty,t]}y^* \end{bmatrix},$$

where $\pi_{(-\infty,t]} = \tau(-t)\pi_-\tau(t)$ is given by

$$(\pi_{(-\infty,t]}y^*)(s) = \begin{cases} y^*(s), & s \le t, \\ 0, & s > t. \end{cases}$$

In particular, in the initial value setting with initial time zero, initial value x_0^*, and control y^*, the controlled state $x^*(s)$ at time $s \in \mathbf{R}^-$ and the output u^* of Ψ^* are given by

$$\begin{bmatrix} x^*(s) \\ u^* \end{bmatrix} = \begin{bmatrix} \mathcal{A}^*(-s) & \mathcal{C}^*\tau(s) \\ \mathcal{B}^* & \mathcal{D}^* \end{bmatrix} \begin{bmatrix} x_0^* \\ \pi_-y^* \end{bmatrix} = \begin{bmatrix} \mathcal{A}^*(-s)x_0^* + \mathcal{C}^*\tau(s)\pi_-y^* \\ \mathcal{B}^*x_0^* + \mathcal{D}^*\pi_-y^* \end{bmatrix}.$$

The formulas in Definitions 2.2 and 2.14 are chosen in such a way that Ψ and Ψ^* interact in the following way.

LEMMA 2.15. *Let Ψ be a well-posed linear system on (U, H, Y) and Ψ^* its adjoint (with respect to the ordinary unweighted inner product in L^2). Let $-\infty < s < t < \infty$, let x and y be the state and output of Ψ with initial time s and control u, and let x^* and u^* be the state and output of Ψ^* with initial time t and control y^*. Then*

$$\langle x^*(t), x(t)\rangle_H + \int_s^t \langle y^*(v), y(v)\rangle_Y \, dv = \langle x^*(s), x(s)\rangle_H + \int_s^t \langle u^*(v), u(v)\rangle_U \, dv.$$

We leave the easy proof of this lemma to the reader.

We use diagrams of the type drawn in Figure 2.1 to represent the relation between the state x, the output y, the initial value x_0, and the control u of Ψ in the initial value setting with initial time zero. In our diagrams we use the following conventions.

(i) Initial states and controls enter at the top or bottom, and they are acted on by all the operators located in the column to which they are attached. In particular, note that x_0 is attached to the first column and u to the second.

(ii) Final states and outputs leave to the left or right, and they are the sums of all the elements in the row to which they are attached. In particular, note that x is attached to the top row and y to the bottom row.

A similar diagram is used to describe the adjoint system Ψ^*.

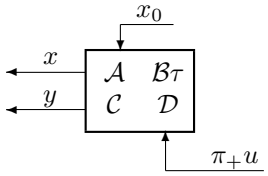


FIG. 2.1. *Input/state/output diagram of* Ψ.

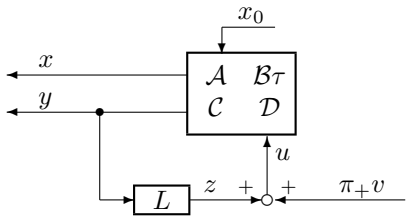


FIG. 3.1. *Static output feedback.*

3. Feedback, stabilizability, and detectability. The notions of stabilizability and detectability deal with the possibility of stabilizing a well-posed linear system by the use of either a state feedback or an output injection. Therefore, before we can study these notions, we must first look at different kinds of feedback connections.

We start with the most basic feedback connection, namely, the notion of a (static) output feedback, drawn in Figure 3.1. Here L is a bounded linear operator from the output space into the input space. Thus if we consider this feedback configuration in the initial value setting with initial time zero, initial value x_0, and control v, we find that the effective input u, the state $x(t)$ at time $t \geq 0$, the output y, and the feedback signal z satisfy the equations

$$\begin{aligned} u &= z + \pi_+ v, \\ x(t) &= \mathcal{A}(t)x_0 + \mathcal{B}\tau(t)u, \\ y &= \mathcal{C}x_0 + \mathcal{D}u, \\ z &= Ly, \end{aligned} \tag{3.1}$$

which formally can be solved as

$$\begin{aligned} u &= (I - L\mathcal{D})^{-1}(L\mathcal{C}x_0 + \pi_+ v), \\ x(t) &= \big(\mathcal{A}(t) + \mathcal{B}\tau(t)L(I - \mathcal{D}L)^{-1}\mathcal{C}\big)x_0 + \mathcal{B}(I - L\mathcal{D})^{-1}\tau(t)\pi_+ v, \\ y &= (I - \mathcal{D}L)^{-1}(\mathcal{C}x_0 + \mathcal{D}\pi_+ v), \\ z &= (I - L\mathcal{D})^{-1}L(\mathcal{C}x_0 + \mathcal{D}\pi_+ v). \end{aligned} \tag{3.2}$$

We say that the feedback operator L is admissible whenever these equations are valid.

DEFINITION 3.1. *Let* $\Psi = \left[\begin{smallmatrix} \mathcal{A} & \mathcal{B} \\ \mathcal{C} & \mathcal{D} \end{smallmatrix}\right]$ *be a well-posed linear system on* (U, H, Y). *Then* $L \in \mathcal{L}(Y;U)$ *is called an admissible output feedback operator for* Ψ *iff the operator* $I - L\mathcal{D}$ *has an inverse in* $TIC_\alpha(U)$ *for some* $\alpha \in \mathbf{R}$ *or, equivalently, iff the operator* $I - \mathcal{D}L$ *has an inverse in* $TIC_\alpha(Y)$ *for some* $\alpha \in \mathbf{R}$.

As Weiss [31, section 6] proved, x and y in (3.2) can be interpreted as the state and output of another well-posed linear system as follows.

PROPOSITION 3.2. *Let $\Psi = \left[\begin{smallmatrix} \mathcal{A} & \mathcal{B} \\ \mathcal{C} & \mathcal{D} \end{smallmatrix}\right]$ be a well-posed linear system on (U, H, Y), and let $L \in \mathcal{L}(Y;U)$ be an admissible output feedback operator for Ψ. Then the system*

$$\begin{aligned}\Psi_L &= \begin{bmatrix} \mathcal{A}_L & \mathcal{B}_L\tau \\ \mathcal{C}_L & \mathcal{D}_L \end{bmatrix} = \begin{bmatrix} \mathcal{A} + \mathcal{B}\tau L\left(I - \mathcal{D}L\right)^{-1}\mathcal{C} & \mathcal{B}\left(I - L\mathcal{D}\right)^{-1}\tau \\ \left(I - \mathcal{D}L\right)^{-1}\mathcal{C} & \mathcal{D}\left(I - L\mathcal{D}\right)^{-1} \end{bmatrix} \\ &= \begin{bmatrix} \mathcal{A} & \mathcal{B}\tau \\ \mathcal{C} & \mathcal{D} \end{bmatrix} + \begin{bmatrix} \mathcal{B}\tau \\ \mathcal{D} \end{bmatrix} L\left(I - \mathcal{D}L\right)^{-1}\begin{bmatrix} \mathcal{C} & \mathcal{D} \end{bmatrix} \\ &= \begin{bmatrix} \mathcal{A} & \mathcal{B}\tau \\ \mathcal{C} & \mathcal{D} \end{bmatrix} + \begin{bmatrix} \mathcal{B}\tau \\ \mathcal{D} \end{bmatrix} L\begin{bmatrix} \mathcal{C}_L & \mathcal{D}_L \end{bmatrix} \\ &= \begin{bmatrix} \mathcal{A} & \mathcal{B}\tau \\ \mathcal{C} & \mathcal{D} \end{bmatrix} + \begin{bmatrix} \mathcal{B}_L\tau \\ \mathcal{D}_L \end{bmatrix} L\begin{bmatrix} \mathcal{C} & \mathcal{D} \end{bmatrix}\end{aligned}$$

is another well-posed linear system on (U, H, Y). We call this system the closed loop system with feedback operator L. In the initial value setting with initial time zero, initial value x_0, and control v, the controlled state $x(t)$ at time t and the output y of Ψ_L are given by (3.2).

See [31, section 6] for a proof. (The major part of this proposition is also contained in [19, Theorem 4.2].)

We remark that if in the classical system (2.1) we replace u by $u = Ly + v$, then we get a new well defined system of the same type iff $I - DL$ is invertible or, equivalently, iff $I - LD$ is invertible. In the new system the operators $\left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right]$ have been replaced by

$$\begin{aligned}\begin{bmatrix} A_L & B_L \\ C_L & D_L \end{bmatrix} &= \begin{bmatrix} A + BL\left(I - DL\right)^{-1}C & B\left(I - LD\right)^{-1} \\ \left(I - DL\right)^{-1}C & D\left(I - LD\right)^{-1} \end{bmatrix} \\ &= \begin{bmatrix} A & B \\ C & D \end{bmatrix} + \begin{bmatrix} B \\ D \end{bmatrix} L\left(I - DL\right)^{-1}\begin{bmatrix} C & D \end{bmatrix} \\ &= \begin{bmatrix} A & B \\ C & D \end{bmatrix} + \begin{bmatrix} B \\ D \end{bmatrix} L\begin{bmatrix} C_L & D_L \end{bmatrix} \\ &= \begin{bmatrix} A & B \\ C & D \end{bmatrix} + \begin{bmatrix} B_L \\ D_L \end{bmatrix} L\begin{bmatrix} C & D \end{bmatrix}.\end{aligned} \tag{3.3}$$

Observe the striking similarity between this formula and the one given in Proposition 3.2.[7]

Repeated feedback behaves in the expected way.

PROPOSITION 3.3. *Let $L \in \mathcal{L}(U;Y)$ be an admissible output feedback operator for Ψ. Then $K \in \mathcal{L}(U;Y)$ is an admissible output feedback operator for the closed loop system Ψ_L iff $L + K$ is an admissible output feedback operator for Ψ, and $\Psi_{L+K} = (\Psi_L)_K$. In particular, $-L$ is always an admissible feedback operator for Ψ_L, and $(\Psi_L)_{-L} = \Psi$.*

See [31, Remark 6.4] for the straightforward proof.

DEFINITION 3.4. *The operator $L \in \mathcal{L}(Y;U)$ is a (strongly) ω-stabilizing [stabilizing] [[exponentially stabilizing]] output feedback operator for the well-posed linear system Ψ on (U, H, Y) iff L is an admissible output feedback operator for Ψ and the resulting closed loop system Ψ_L is (strongly) ω-stable [stable] [[exponentially stable]].*

We observe the following basic facts.

LEMMA 3.5. *Let $\Psi = \left[\begin{smallmatrix} \mathcal{A} & \mathcal{B} \\ \mathcal{C} & \mathcal{D} \end{smallmatrix}\right]$ be ω-stable, and let $L \in \mathcal{L}(Y;U)$.*

[7] Usually the feed-through operator D is taken to be zero, in which case this formula simplifies significantly and the invertibility condition on $I - DL$ drops out.

(i) *L is an admissible output feedback operator for Ψ iff there is some $\alpha \geq \omega$ for which the diagram in Figure* 3.1 *(i.e., the set of equations* (3.1)*) with $x_0 = 0$ defines a continuous linear mapping from the external input $v \in L^2_\alpha(\mathbf{R}^+; U)$ to the internal input $u \in L^2_\alpha(\mathbf{R}^+; U)$ or, equivalently, iff the operator $I - L\mathcal{D}$ has an inverse in $TIC_\alpha(U)$ or, equivalently, iff the operator $I - \mathcal{D}L$ has an inverse in $TIC_\alpha(Y)$. The resulting closed loop Ψ_L system is α-stable.*

(ii) *L is ω-stabilizing iff any one of the three equivalent conditions in part* (i) *is true with $\omega = \alpha$ (hence all of them are true with $\omega = \alpha$). In this case the closed loop system Ψ_L is strongly ω-stable iff the open loop system Ψ is strongly ω-stable.*

REMARK 3.6. *Thus if a system is ω-stable but not strongly so, then it is impossible to make it strongly ω-stable by using our notion of admissible output feedback.*

Proof of Lemma 3.5. (i) Most of this follows immediately from Definition 3.1. To see that the resulting system is α-stable we observe that if $I - L\mathcal{D}$ has an inverse in $TIC_\alpha(U)$ (or, equivalently, the operator $I - \mathcal{D}L$ has an inverse in $TIC_\alpha(Y)$), then the formulas for Ψ_L given in Proposition 3.2 imply that Ψ_L is α-stable. (Here we use the fact that ω-stability implies α-stability for every $\alpha \geq \omega$; see Lemma 2.4.)

(ii) Clearly, if the conditions in part (i) are true with $\alpha = \omega$, then L is ω-stabilizing. Conversely, if L is ω-stabilizing, then $\mathcal{D}_L = \mathcal{D}(I - L\mathcal{D})^{-1} \in TIC_\omega(U; Y)$, and this implies that $(I - L\mathcal{D})^{-1} = I + L\mathcal{D}(I - L\mathcal{D})^{-1} = I + L\mathcal{D}_L \in TIC_\omega(U)$.

To prove the second claim in part (ii) it suffices to show that

$$\mathrm{e}^{-\omega t}\mathcal{B}\tau(t)L(I - \mathcal{D}L)^{-1}\mathcal{C}x \to 0 \text{ as } t \to \infty$$

for every $x \in H$, since $\mathcal{A}_L - \mathcal{A} = \mathcal{B}\tau L\,(I - \mathcal{D}L)^{-1}\,\mathcal{C}$. Fix $x \in H$ and split the expression above into

$$\begin{aligned}\mathrm{e}^{-\omega t}\mathcal{B}\tau(t)L(I - \mathcal{D}L)^{-1}\mathcal{C}x &= \mathrm{e}^{-\omega t}\mathcal{B}\tau(t - T)(\pi_+ + \pi_-)\tau(T)L(I - \mathcal{D}L)^{-1}\mathcal{C}x \\ &= \mathcal{B}\mathrm{e}^{-\omega t}\tau(t)\pi_{[T,\infty)}L(I - \mathcal{D}L)^{-1}\mathcal{C}x \\ &\quad + \mathrm{e}^{-\omega t}\mathcal{A}(t - T)\mathcal{B}\tau(T)L(I - \mathcal{D}L)^{-1}\mathcal{C}x.\end{aligned}$$

Here the first term tends to zero as $T \to \infty$, uniformly in $t \geq T$, and the second term tends to zero as $t \to \infty$ and T is fixed. □

REMARK 3.7. *The same proof shows that (with the same terminology as in Remark* 2.12*) if Ψ is input ω-stable and output ω-stable and $L \in \mathcal{L}(Y; U)$ is an admissible output feedback operator, then L is input ω-stabilizing, output ω-stabilizing, and input/output ω-stabilizing iff $\mathcal{D}(I - L\mathcal{D})^{-1} \in TIC_\omega(U)$. Moreover, in this case L is [strongly] ω-stabilizing iff the semigroup of Ψ is [strongly] ω-stable.*

The notion of a *state feedback* can be reduced formally to the notion of an output feedback. Intuitively, a state feedback means that an additional output is created, and this output is then fed back into the input, as shown in Figure 3.2. In this figure the original system is represented by $\left[\begin{smallmatrix}\mathcal{A} & \mathcal{B}\\ \mathcal{C} & \mathcal{D}\end{smallmatrix}\right]$. We find two additional components, namely, a new observability map $\mathcal{K}$ (from the initial state to the new output) and a new input/output map $\mathcal{F}$ (from the original input to the new output). The pair $\begin{bmatrix}\mathcal{K} & \mathcal{F}\end{bmatrix}$ is *admissible* if the resulting system is well posed, i.e., if $\begin{bmatrix}0 & I\end{bmatrix}$ is an admissible output feedback operator for the extended system defined in Definition 3.8.

DEFINITION 3.8. *Let $\Psi = \left[\begin{smallmatrix}\mathcal{A} & \mathcal{B}\\ \mathcal{C} & \mathcal{D}\end{smallmatrix}\right]$ be a well-posed linear system on (U, H, Y). The*

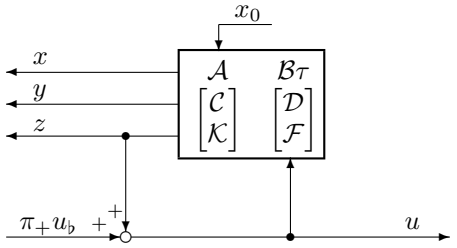


FIG. 3.2. *State feedback connection.*

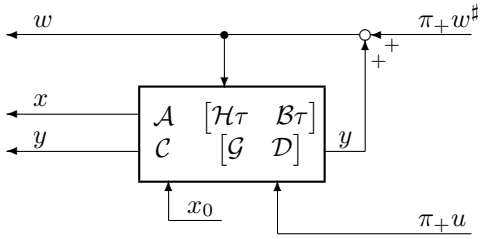


FIG. 3.3. *Output injection connection.*

pair $\begin{bmatrix}\mathcal{K} & \mathcal{F}\end{bmatrix}$ *is an admissible state feedback pair for* Ψ *iff the extended system*

$$\Psi_{\mathrm{SF}} = \begin{bmatrix} \mathcal{A} & \mathcal{B} \\ \begin{bmatrix}\mathcal{C}\\ \mathcal{K}\end{bmatrix} & \begin{bmatrix}\mathcal{D}\\ \mathcal{F}\end{bmatrix} \end{bmatrix}$$

is a well-posed linear system on $(U, H, Y \times U)$ *and* $\begin{bmatrix}0 & I\end{bmatrix}$ *is an admissible output feedback operator for* Ψ_{SF}*; i.e.,* $I - \mathcal{F}$ *has an inverse in* $T\dot{I}C_\omega(U)$ *for some* $\omega \in \mathbf{R}$.[8] *It is (strongly)* ω*-stabilizing [stabilizing] [[exponentially stabilizing]] if the resulting closed loop system is (strongly)* ω*-stable [stable] [[exponentially stable]].*

REMARK 3.9. *We shall frequently regard the signal* u *in Figure* 3.2 *(i.e., the input to the open loop system) as an additional output of the closed loop system (although it is not part of the official definition). This output has the same observability map* $(I - \mathcal{F})^{-1}\mathcal{K}$ *as the output* z.[9] *Its input/output map, given by* $(I - \mathcal{F})^{-1}$*, differs from the input/output map from* $u_\flat$ *to* z *by an identity operator (see Lemma* 3.13*). Similar remarks apply to the signals* w *in Figure* 3.3*,* $u_\flat$ *and* $w^\sharp$ *in Figure* 3.4*,* u *and* $w^\sharp$ *in Figure* 3.5*,* $u_\flat$ *and* w *in Figure* 3.6*, etc.*

The notion of an *output injection* is analogous. In this case a new input is created into which we feed the original output y plus a new perturbation $w^\sharp$, as shown in Figure 3.3. The original system is still represented by $\left[\begin{smallmatrix}\mathcal{A} & \mathcal{B}\\ \mathcal{C} & \mathcal{D}\end{smallmatrix}\right]$. In this figure we find a new controllability map $\mathcal{H}$ (from the new input to the state) and a new input/output map $\mathcal{G}$ (from the new input to the original output). The pair $\left[\begin{smallmatrix}\mathcal{H}\\ \mathcal{G}\end{smallmatrix}\right]$ is *admissible* if the resulting system is well posed.

DEFINITION 3.10. *Let* $\Psi = \left[\begin{smallmatrix}\mathcal{A} & \mathcal{B}\\ \mathcal{C} & \mathcal{D}\end{smallmatrix}\right]$ *be a well-posed linear system on* (U, H, Y).

[8] The input of this system is the signal $u_\flat$ in Figure 3.2, and its outputs are y and z. See also Remark 3.9.

[9] See Lemma 3.13.

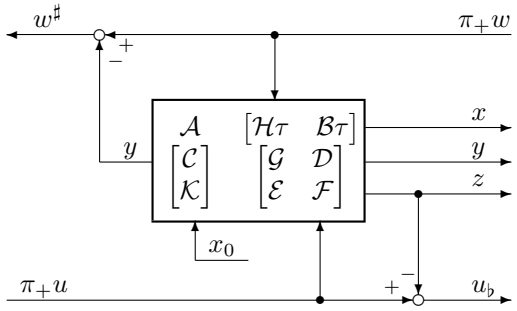


FIG. 3.4. *The extended system.*

The pair $\left[\begin{smallmatrix}\mathcal{H}\\ \mathcal{G}\end{smallmatrix}\right]$ *is an admissible output injection pair for* Ψ *iff the extended system*

$$\Psi_{\mathrm{OI}} = \begin{bmatrix}\mathcal{A} & \begin{bmatrix}\mathcal{H} & \mathcal{B}\end{bmatrix}\\ \begin{bmatrix}\mathcal{C}\\ \mathcal{K}\end{bmatrix} & \begin{bmatrix}\mathcal{G} & \mathcal{D}\\ \mathcal{E} & \mathcal{F}\end{bmatrix}\end{bmatrix}$$

is a well-posed linear system on $(Y\times U, H, Y)$ *and* $\left[\begin{smallmatrix}I\\ 0\end{smallmatrix}\right]$ *is an admissible output feedback operator for* Ψ_{OI}*; i.e.,* $I-\mathcal{G}$ *has an inverse in* $TIC_\omega(Y)$ *for some* $\omega\in\mathbf{R}$.[10] *It is (strongly)* ω*-stabilizing [stabilizing] [[exponentially stabilizing]] if the resulting closed loop system is (strongly)* ω*-stable [stable] [[exponentially stable]].*

In the sequel we shall need to study a case where at the same time we want to add both a state feedback pair $\begin{bmatrix}\mathcal{K} & \mathcal{F}\end{bmatrix}$ and an output injection pair $\left[\begin{smallmatrix}\mathcal{H}\\ \mathcal{G}\end{smallmatrix}\right]$ to a given system $\left[\begin{smallmatrix}\mathcal{A} & \mathcal{B}\\ \mathcal{C} & \mathcal{D}\end{smallmatrix}\right]$. If we try to write a figure similar to Figures 3.2 and 3.3, we immediately observe that we need one more input/output map $\mathcal{E}$ (from the output injection input to the state feedback output); see Figure 3.4. This operator need not always exist,[11] and this forces us to introduce still another definition.

DEFINITION 3.11. *Let* $\Psi = \left[\begin{smallmatrix}\mathcal{A} & \mathcal{B}\\ \mathcal{C} & \mathcal{D}\end{smallmatrix}\right]$ *be a well-posed linear system on* (U, H, Y)*. The pairs* $\begin{bmatrix}\mathcal{K} & \mathcal{F}\end{bmatrix}$ *and* $\left[\begin{smallmatrix}\mathcal{H}\\ \mathcal{G}\end{smallmatrix}\right]$ *are called jointly admissible state feedback and output injection pairs for* Ψ *iff* $\begin{bmatrix}\mathcal{K} & \mathcal{F}\end{bmatrix}$ *is an admissible state feedback pair for* Ψ*,* $\left[\begin{smallmatrix}\mathcal{H}\\ \mathcal{G}\end{smallmatrix}\right]$ *is an admissible output injection pair for* Ψ*, and in addition, there exists a operator* $\mathcal{E}$*, called the interaction operator, such that and the combined extended system*

$$\Psi_{\mathrm{ext}} = \begin{bmatrix}\mathcal{A} & \begin{bmatrix}\mathcal{H} & \mathcal{B}\end{bmatrix}\\ \begin{bmatrix}\mathcal{C}\\ \mathcal{K}\end{bmatrix} & \begin{bmatrix}\mathcal{G} & \mathcal{D}\\ \mathcal{E} & \mathcal{F}\end{bmatrix}\end{bmatrix}$$

is a well-posed linear system on $(Y\times U, H, Y\times U)$*.*

LEMMA 3.12. *Let* $\Psi = \left[\begin{smallmatrix}\mathcal{A} & \mathcal{B}\\ \mathcal{C} & \mathcal{D}\end{smallmatrix}\right]$ *be a well-posed linear system on* (U, H, Y)*. Then the following conditions are equivalent:*

(i) *the pairs* $\begin{bmatrix}\mathcal{K} & \mathcal{F}\end{bmatrix}$ *and* $\left[\begin{smallmatrix}\mathcal{H}\\ \mathcal{G}\end{smallmatrix}\right]$ *are jointly admissible state feedback and output injection pairs with interaction operator* $\mathcal{E}$*;*

(ii) *the system* Ψ_{ext} *in Definition* 3.11 *is a well-posed linear system on* $(Y\times U, H, Y\times U)$*, and both* $\left[\begin{smallmatrix}0 & 0\\ 0 & I\end{smallmatrix}\right]$ *and* $\left[\begin{smallmatrix}I & 0\\ 0 & 0\end{smallmatrix}\right]$ *are admissible output feedback operators for* Ψ_{ext}*.*

[10] The inputs of this system are the signals u and $w^\sharp$ in Figure 3.3, and its output is y. See also Remark 3.9.

[11] More precisely, it need not be a bounded operator. The operator $\mathcal{E}$, if it exists, is determined uniquely modulo a static operator; this follows from [24, Corollary 7].

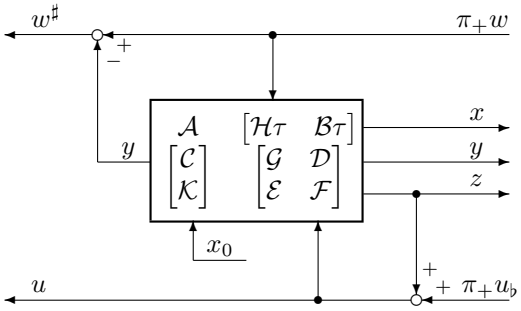


FIG. 3.5. *Right coprime factor.*

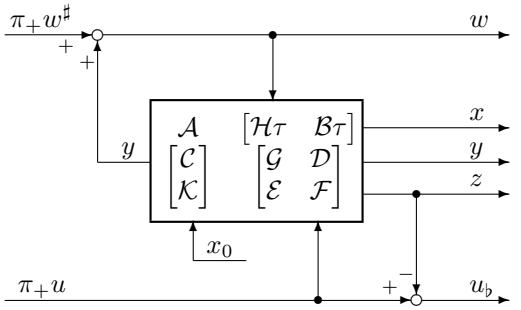


FIG. 3.6. *Left coprime factor.*

(iii) *the system* Ψ_{ext} *in Definition* 3.11 *is a well-posed linear system on* $(Y \times U, H, Y \times U)$, *and* $I - \mathcal{F}$ *and* $I - \mathcal{G}$ *have inverses in* $TIC_\omega(U)$, *respectively,* $TIC_\omega(Y)$ *for some* $\omega \in \mathbf{R}$.

To prove this lemma it suffices to make a straightforward calculation based on Proposition 3.2. As a part of this calculation we get the following expressions for the two closed loop systems in part (ii), drawn in Figures 3.5 and 3.6, respectively.[12]

LEMMA 3.13. *Let* $\Psi = \left[\begin{smallmatrix} \mathcal{A} & \mathcal{B} \\ \mathcal{C} & \mathcal{D} \end{smallmatrix}\right]$ *be a well-posed linear system on* (U, H, Y), *and let* $\begin{bmatrix} \mathcal{K} & \mathcal{F} \end{bmatrix}$ *and* $\left[\begin{smallmatrix} \mathcal{H} \\ \mathcal{G} \end{smallmatrix}\right]$ *be jointly admissible state feedback and output injection pairs for* Ψ *with interaction operator* $\mathcal{E}$. *Then the closed loop system* $\Psi_\flat$ *that we get by using* $\left[\begin{smallmatrix} 0 & 0 \\ 0 & I \end{smallmatrix}\right]$ *as a output feedback operator for* Ψ_{ext} *(see Figure* 3.5*) is given by*[13]

$$
\begin{aligned}
\Psi_\flat &= \begin{bmatrix} \mathcal{A}_\flat & \begin{bmatrix} \mathcal{H}_\flat \tau & \mathcal{B}_\flat \tau \end{bmatrix} \\ \begin{bmatrix} \mathcal{C}_\flat \\ \mathcal{K}_\flat \end{bmatrix} & \begin{bmatrix} \mathcal{G}_\flat & \mathcal{D}_\flat \\ \mathcal{E}_\flat & \mathcal{F}_\flat \end{bmatrix} \end{bmatrix} \\
&= \begin{bmatrix} \mathcal{A} + \mathcal{B}\tau\,(I - \mathcal{F})^{-1}\,\mathcal{K} & \begin{bmatrix} \mathcal{H}\tau + \mathcal{B}\,(I - \mathcal{F})^{-1}\,\mathcal{E}\tau & \mathcal{B}\,(I - \mathcal{F})^{-1}\,\tau \end{bmatrix} \\ \begin{bmatrix} \mathcal{C} + \mathcal{D}\,(I - \mathcal{F})^{-1}\,\mathcal{K} \\ (I - \mathcal{F})^{-1}\,\mathcal{K} \end{bmatrix} & \begin{bmatrix} \mathcal{G} + \mathcal{D}\,(I - \mathcal{F})^{-1}\,\mathcal{E} & \mathcal{D}\,(I - \mathcal{F})^{-1} \\ (I - \mathcal{F})^{-1}\,\mathcal{E} & (I - \mathcal{F})^{-1} - I \end{bmatrix} \end{bmatrix} \\
&= \begin{bmatrix} \mathcal{A} & \begin{bmatrix} \mathcal{H}\tau & \mathcal{B}\tau \end{bmatrix} \\ \begin{bmatrix} \mathcal{C} \\ \mathcal{K} \end{bmatrix} & \begin{bmatrix} \mathcal{G} & \mathcal{D} \\ \mathcal{E} & \mathcal{F} \end{bmatrix} \end{bmatrix} + \begin{bmatrix} \mathcal{B}\tau \\ \mathcal{D} \\ \mathcal{F} \end{bmatrix} (I - \mathcal{F})^{-1} \begin{bmatrix} \mathcal{K} & \mathcal{E} & \mathcal{F} \end{bmatrix},
\end{aligned}
$$

[12] See section 4 for an explanation of the captions of these figures.

[13] The inputs of this system are the signals $u_\flat$ and w in Figure 3.5, and its outputs are y and z. See also Remark 3.9.

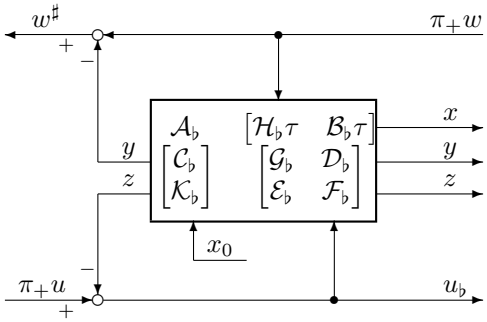


FIG. 3.7. *Cancelation of state feedback.*

and the closed loop system $\Psi^\sharp$ *that we get by using* $\left[\begin{smallmatrix} I & 0\\ 0 & 0\end{smallmatrix}\right]$ *as a output feedback operator for* Ψ_{ext} *(see* Figure 3.6*) is given by*[14]

$$\begin{aligned}\Psi^\sharp &= \begin{bmatrix} \mathcal{A}^\sharp & \begin{bmatrix}\mathcal{H}^\sharp\tau & \mathcal{B}^\sharp\tau\end{bmatrix}\\ \begin{bmatrix}\mathcal{C}^\sharp\\ \mathcal{K}^\sharp\end{bmatrix} & \begin{bmatrix}\mathcal{G}^\sharp & \mathcal{D}^\sharp\\ \mathcal{E}^\sharp & \mathcal{F}^\sharp\end{bmatrix}\end{bmatrix}\\ &= \begin{bmatrix} \mathcal{A}+\mathcal{H}\tau\left(I-\mathcal{G}\right)^{-1}\mathcal{C} & \begin{bmatrix}\mathcal{H}\left(I-\mathcal{G}\right)^{-1}\tau & \mathcal{B}\tau+\mathcal{H}\left(I-\mathcal{G}\right)^{-1}\mathcal{D}\tau\end{bmatrix}\\ \begin{bmatrix}\left(I-\mathcal{G}\right)^{-1}\mathcal{C}\\ \mathcal{K}+\mathcal{E}\left(I-\mathcal{G}\right)^{-1}\mathcal{C}\end{bmatrix} & \begin{bmatrix}\left(I-\mathcal{G}\right)^{-1}-I & \left(I-\mathcal{G}\right)^{-1}\mathcal{D}\\ \mathcal{E}\left(I-\mathcal{G}\right)^{-1} & \mathcal{F}+\mathcal{E}\left(I-\mathcal{G}\right)^{-1}\mathcal{D}\end{bmatrix}\end{bmatrix}\\ &= \begin{bmatrix} \mathcal{A} & \begin{bmatrix}\mathcal{H}\tau & \mathcal{B}\tau\end{bmatrix}\\ \begin{bmatrix}\mathcal{C}\\ \mathcal{K}\end{bmatrix} & \begin{bmatrix}\mathcal{G} & \mathcal{D}\\ \mathcal{E} & \mathcal{F}\end{bmatrix}\end{bmatrix} + \begin{bmatrix}\mathcal{H}\tau\\ \mathcal{G}\\ \mathcal{E}\end{bmatrix}\left(I-\mathcal{G}\right)^{-1}\begin{bmatrix}\mathcal{C} & \mathcal{G} & \mathcal{D}\end{bmatrix}.\end{aligned}$$

REMARK 3.14. *According to Proposition* 3.3*, it is possible to recover the extended system* Ψ_{ext} *from either of the systems* $\Psi_\flat$ *or* $\Psi^\sharp$ *by using negative feedback. For example, the feedback connection drawn in Figure* 3.7 *is equivalent to* Ψ_{ext}*.*

So far we have only looked at the *joint admissibility* of state feedback and output injection pairs. If the resulting closed loop systems drawn in Figures 3.5 and 3.6 are ω-stable, then we call these pairs *jointly* ω*-stabilizing* as follows.

DEFINITION 3.15. *The pairs* $\begin{bmatrix}\mathcal{K} & \mathcal{F}\end{bmatrix}$ *and* $\left[\begin{smallmatrix}\mathcal{H}\\ \mathcal{G}\end{smallmatrix}\right]$ *are called jointly (strongly)* ω*-stabilizing [stabilizing] [[exponentially stabilizing]] state feedback and output injection pairs for* Ψ *if they are jointly admissible state feedback and output injection pairs with some interaction operator* $\mathcal{E}$*, and both the closed loop systems* $\Psi_\flat$ *and* $\Psi^\sharp$ *in Lemma* 3.13 *are (strongly)* ω*-stable [stable] [[exponentially stable]].*

Observe that if the two pairs in Definition 3.15 are ω-stabilizing (but not "jointly" ω-stabilizing), then we know that the operators in the left and right columns of $\Psi_\flat$ and in the top and middle rows of $\Psi^\sharp$ are ω-stable (see the formulas in Lemma 3.13), but we do not know anything about the operators in the middle column of $\Psi_\flat$ and in the bottom row of $\Psi^\sharp$.

DEFINITION 3.16. *Let* Ψ *be a well-posed linear system.*

(i) Ψ *is (strongly)* ω*-stabilizable [stabilizable] [[exponentially stabilizable]] iff there exists a (strongly)* ω*-stabilizing [stabilizing] [[exponentially stabilizing]] state feedback pair for* Ψ*.*

[14]The inputs of this system are the signals u and $w^\sharp$ in Figure 3.6, and its outputs are y and z. See also Remark 3.9.

(ii) Ψ *is (strongly)* ω*-detectable [detectable] [[exponentially detectable]] iff there exists a (strongly)* ω*-stabilizing [stabilizing] [[exponentially stabilizing]] output injection pair for* Ψ*.*

(iii) Ψ *is jointly (strongly)* ω*-stabilizable [stabilizable] [[exponentially stabilizable]] and detectable iff there exist some jointly (strongly)* ω*-stabilizing [stabilizing] [[exponentially stabilizing]] state feedback and output injection pairs for* Ψ*.*

We do not know if it is possible for a system to be both stabilizable and detectable without being jointly stabilizable and detectable.

There is a simple connection between ω-stability, ω-detectability, and input/output ω-stability as shown in Lemma 3.17.

LEMMA 3.17. *Let* $\Psi = \left[\begin{smallmatrix} \mathcal{A} & \mathcal{B} \\ \mathcal{C} & \mathcal{D} \end{smallmatrix}\right]$ *be input/output* ω*-stable [exponentially stable] (i.e., let* $\mathcal{D}$ *be* ω*-stable [exponentially stable]).*

(i) *If* Ψ *is* ω*-stabilizable [exponentially stabilizable], then the observability map* $\mathcal{C}$ *is* ω*-stable [exponentially stable].*

(ii) *If* Ψ *is* ω*-detectable [exponentially detectable], then the controllability map* $\mathcal{B}$ *is* ω*-stable [exponentially stable].*

(iii) *If* Ψ *is both (strongly)* ω*-stabilizable [exponentially stabilizable] and* ω*-detectable [exponentially detectable] (not necessarily jointly), then* Ψ *is (strongly)* ω*-stable [exponentially stable].*

Proof. Introduce the same notation as in Lemma 3.13.

(i) By Lemma 3.13, the observability map $\mathcal{C}$ is given by $\mathcal{C} = \mathcal{C}_\flat - \mathcal{D}\mathcal{K}_\flat$. Thus $\mathcal{C}$ is ω-stable [exponentially stable] whenever $\mathcal{C}_\flat$, $\mathcal{D}$, and $\mathcal{K}_\flat$ are so.

(ii) By the same lemma, the controllability map $\mathcal{B}$ is given by $\mathcal{B} = \mathcal{B}^\sharp - \mathcal{H}^\sharp\mathcal{D}$. Thus $\mathcal{B}$ is ω-stable [exponentially stable] whenever $\mathcal{B}^\sharp$, $\mathcal{H}^\sharp$, and $\mathcal{D}$ are so.

(iii) Again, by Lemma 3.13, $\mathcal{A} = \mathcal{A}_\flat - \mathcal{B}\tau\mathcal{K}_\flat$. Thus $\mathcal{A}$ is (strongly) ω-stable [exponentially stable] whenever $\mathcal{A}_\flat$, $\mathcal{B}$, and $\mathcal{K}_\flat$ are so.[15] □

By adding the trivial converse to part (iii) of Lemma 3.17 we get the following corollary.

COROLLARY 3.18. *A (strongly)* ω*-stabilizable and* ω*-detectable [exponentially stabilizable and detectable] well-posed linear system is (strongly)* ω*-stable [exponentially stable] iff it is input/output* ω*-stable [exponentially stable].*

This result generalizes most other results in this direction such as [17, Corollary 1.8].

Finally, let us present two lemmas concerning exponential stability. Both of these follow directly from Lemma 2.6.

LEMMA 3.19. *Let* Ψ *be a well-posed linear system.*

(i) *The state feedback pair* $\begin{bmatrix} \mathcal{K} & \mathcal{F} \end{bmatrix}$ *is exponentially stabilizing iff it is admissible and the closed loop semigroup* $\mathcal{A}_\flat$ *in Lemma* 3.13 *is exponentially stable.*

(ii) *The output injection pair* $\left[\begin{smallmatrix} \mathcal{H} \\ \mathcal{G} \end{smallmatrix}\right]$ *is exponentially stabilizing iff it is admissible and the closed loop semigroup* $\mathcal{A}^\sharp$ *in Lemma* 3.13 *is exponentially stable.*

(iii) *The state feedback pair* $\begin{bmatrix} \mathcal{K} & \mathcal{F} \end{bmatrix}$ *and the output injection pair* $\left[\begin{smallmatrix} \mathcal{H} \\ \mathcal{G} \end{smallmatrix}\right]$ *are jointly exponentially stabilizing iff they are jointly admissible and both the closed loop semigroups* $\mathcal{A}_\flat$ *and* $\mathcal{A}^\sharp$ *in Lemma* 3.13 *are exponentially stable.*

LEMMA 3.20. *If* $\begin{bmatrix} \mathcal{K} & \mathcal{F} \end{bmatrix}$ *is an exponentially stabilizing state feedback pairs for the system* $\Psi = \left[\begin{smallmatrix} \mathcal{A} & \mathcal{B} \\ \mathcal{C} & \mathcal{D} \end{smallmatrix}\right]$*, then the same feedback pair is exponentially stabilizing for every well-posed extension of* Ψ *of the type*

$$\begin{bmatrix} \mathcal{A} & \mathcal{B} \\ \begin{bmatrix} \mathcal{C} \\ \mathcal{C}_1 \end{bmatrix} & \begin{bmatrix} \mathcal{D} \\ \mathcal{D}_1 \end{bmatrix} \end{bmatrix}.$$

[15] See also Remark 3.6.

4. Coprime factorizations. As is well known, classical (lumped) transfer functions always have doubly coprime factorizations, and these factorizations can be computed by a state space method through the use of a stabilizable and detectable realization. Arbitrary H^∞ transfer functions do not always have coprime factorizations even in the single-input single-output case [12, p. 108], but transfer functions that can be stabilized by a dynamic output feedback do, at least in the case where the input and output spaces are finite dimensional [9, 21]. Below we extend these results and show that the transfer function of every jointly stabilizable and detectable well-posed linear system has a doubly coprime factorization that can be computed by the standard state space method. Conversely, every transfer function with a doubly coprime factorization has a strongly stabilizable and detectable realization.

According to Lemma 2.9, there is a one-to-one correspondence between the set of transfer functions in $H^\infty_\omega(U;Y)$ and the set of causal time-invariant operators in $TIC_\omega(U;Y)$. Rather than switching over to the frequency domain we continue to work in the time domain and leave the transformation of our results to the frequency domain to the reader.

DEFINITION 4.1. *Let U, Y, and Z be Hilbert spaces, and let $\omega \in \mathbf{R}$.*

(i) *The operators $\mathcal{N} \in TIC_\omega(U;Y)$ and $\mathcal{M} \in TIC_\omega(U;Z)$ are right ω-coprime iff there exist operators $\widetilde{\mathcal{Y}} \in TIC_\omega(Y;U)$ and $\widetilde{\mathcal{X}} \in TIC_\omega(Z;U)$ that together with $\mathcal{N}$ and $\mathcal{M}$ satisfy the Bezout identity*

$$\widetilde{\mathcal{Y}}\mathcal{N} + \widetilde{\mathcal{X}}\mathcal{M} = I$$

in $TIC_\omega(U)$. In the case where $\omega = 0$ we call $\mathcal{N}$ and $\mathcal{M}$ right coprime, and in the case where $\omega < 0$ we call $\mathcal{N}$ and $\mathcal{M}$ exponentially right coprime.

(ii) *The operators $\widetilde{\mathcal{N}} \in TIC_\omega(U;Y)$ and $\widetilde{\mathcal{M}} \in TIC_\omega(Z;Y)$ are left ω-coprime iff there exist operators $\mathcal{Y} \in TIC_\omega(Y;U)$ and $\mathcal{X} \in TIC_\omega(Y;Z)$ that together with $\widetilde{\mathcal{N}}$ and $\widetilde{\mathcal{M}}$ satisfy the Bezout identity*

$$\widetilde{\mathcal{N}}\mathcal{Y} + \widetilde{\mathcal{M}}\mathcal{X} = I$$

in $TIC_\omega(Y)$. In the case where $\omega = 0$ we call $\widetilde{\mathcal{N}}$ and $\widetilde{\mathcal{M}}$ left coprime, and in the case where $\omega < 0$ we call $\widetilde{\mathcal{N}}$ and $\widetilde{\mathcal{M}}$ exponentially left coprime.

Thus $\mathcal{N}$ and $\mathcal{M}$ are right ω-coprime iff $\left[\begin{smallmatrix}\mathcal{N}\\ \mathcal{M}\end{smallmatrix}\right]$ has a left inverse in $TIC_\omega(Y \times Z;U)$. $\widetilde{\mathcal{N}}$ and $\widetilde{\mathcal{M}}$ are left ω-coprime iff $\begin{bmatrix}\widetilde{\mathcal{N}} & \widetilde{\mathcal{M}}\end{bmatrix}$ has a right inverse in $TIC_\omega(Y;U \times Z)$.

DEFINITION 4.2. *Let U and Y be Hilbert spaces, let $\omega, \alpha \in \mathbf{R}$ with $\omega \le \alpha$, and let $\mathcal{D} \in TIC_\alpha(U;Y)$.*

(i) *The pair $(\mathcal{N},\mathcal{M})$ is a right ω-coprime factorization of $\mathcal{D}$ if $\mathcal{N} \in TIC_\omega(U;Y)$ and $\mathcal{M} \in TIC_\omega(U)$ are right ω-coprime, $\mathcal{M}$ has an inverse in $TIC_\alpha(U)$, and $\mathcal{D} = \mathcal{N}\mathcal{M}^{-1}$.*

(ii) *The pair $(\widetilde{\mathcal{M}},\widetilde{\mathcal{N}})$ is a left ω-coprime [coprime] [[exponentially coprime]] factorization of $\mathcal{D}$ if $\widetilde{\mathcal{M}} \in TIC_\omega(Y)$ and $\widetilde{\mathcal{N}} \in TIC_\omega(U;Y)$ are left ω-coprime [coprime] [[exponentially coprime]], $\widetilde{\mathcal{M}}$ has an inverse in $TIC_\alpha(Y)$, and $\mathcal{D} = \widetilde{\mathcal{M}}^{-1}\widetilde{\mathcal{N}}$.*

(iii) *A doubly ω-coprime factorization of $\mathcal{D}$ consists of eight operators in TIC_ω (of the appropriate dimensions) satisfying*

$$\begin{bmatrix}\widetilde{\mathcal{M}} & \widetilde{\mathcal{N}}\\ -\widetilde{\mathcal{Y}} & \widetilde{\mathcal{X}}\end{bmatrix}\begin{bmatrix}\mathcal{X} & -\mathcal{N}\\ \mathcal{Y} & \mathcal{M}\end{bmatrix} = \begin{bmatrix}\mathcal{X} & -\mathcal{N}\\ \mathcal{Y} & \mathcal{M}\end{bmatrix}\begin{bmatrix}\widetilde{\mathcal{M}} & \widetilde{\mathcal{N}}\\ -\widetilde{\mathcal{Y}} & \widetilde{\mathcal{X}}\end{bmatrix} = I \tag{4.1}$$

in $TIC_\omega(U \times Y;U \times Y)$, and, in addition, we require that $(\mathcal{N},\mathcal{M})$ is a right ω-coprime and $(\widetilde{\mathcal{M}},\widetilde{\mathcal{N}})$ a left ω-coprime factorization of $\mathcal{D}$. In the case where

$\omega = 0$ we call this a doubly coprime factorization, and in case where $\omega < 0$ we call it a doubly exponentially coprime factorization.

Our definition of coprimeness is slightly nonstandard. It is possible to study coprime factorizations in the quotient field of TIC_ω without our additional assumption that $\mathcal{D}$ belongs to $TIC_\alpha(U;Y)$ and that $\mathcal{M}$ and $\widetilde{\mathcal{M}}$ are invertible in $TIC_\alpha(U)$, respectively, $TIC_\alpha(Y)$ for some $\alpha > \omega$; see, e.g., [9], [12], or [21]. Usually, one only assumes the transfer functions of $\mathcal{M}$ and $\widetilde{\mathcal{M}}$ to be invertible in at least one point in the half-plane $\Re z > \omega$. Observe that if $\mathcal{M}$ is invertible in any reasonable sense, then $\mathcal{D} \in TIC_\alpha(U;Y)$ iff $\mathcal{M}^{-1} \in TIC_\alpha(U)$ because $\mathcal{D} = \mathcal{N}\mathcal{M}^{-1}$ and $\mathcal{M}^{-1} = \widetilde{\mathcal{X}} + \widetilde{\mathcal{Y}}\mathcal{D}$. Likewise, if $\widetilde{\mathcal{M}}$ is invertible in any reasonable sense, then $\mathcal{D} \in TIC_\alpha(U;Y)$ iff $\widetilde{\mathcal{M}}^{-1} \in TIC_\alpha(Y)$ because $\mathcal{D} = \widetilde{\mathcal{M}}^{-1}\mathcal{N}$ and $\widetilde{\mathcal{M}}^{-1} = \mathcal{X} + \mathcal{D}\mathcal{Y}$. According to [20] or [31], if $\mathcal{D}$ does not belong to $TIC_\alpha(U;Y)$ for any $\alpha > \omega$, then $\mathcal{D}$ cannot be realized as the input/output map of a well-posed linear system on a triple of Hilbert spaces.

A coprime factorization is unique, modulo a unit as shown in Lemma 4.3.

LEMMA 4.3. *Let U and Y be Hilbert spaces, let $\omega, \alpha \in \mathbf{R}$ with $\omega \le \alpha$, and let $\mathcal{D} \in TIC_\alpha(U;Y)$.*

(i) *Let $(\mathcal{N},\mathcal{M})$ be a right ω-coprime factorization of $\mathcal{D}$. Then the set of all possible right ω-coprime factorizations of $\mathcal{D}$ can be parameterized in the form $(\mathcal{N}\mathcal{U},\mathcal{M}\mathcal{U})$, where $\mathcal{U}$ is an invertible operator in $TIC_\omega(U)$.*

(ii) *Let $(\widetilde{\mathcal{M}},\widetilde{\mathcal{N}})$ be a left ω-coprime factorization of $\mathcal{D}$. Then the set of all possible left ω-coprime factorizations of $\mathcal{D}$ can be parameterized in the form $(\widetilde{\mathcal{U}}\widetilde{\mathcal{M}},\widetilde{\mathcal{U}}\widetilde{\mathcal{N}})$, where $\widetilde{\mathcal{U}}$ is an invertible operator in $TIC_\omega(Y)$.*

(iii) *If $\mathcal{D}$ has both a right ω-coprime factorization $(\mathcal{N},\mathcal{M})$ and a left ω-coprime factorization $(\widetilde{\mathcal{M}},\widetilde{\mathcal{N}})$, then these two factorizations can be extended to a doubly ω-coprime factorization (i.e., a factorization that contains the given operators $\mathcal{N}$, $\mathcal{M}$, $\widetilde{\mathcal{M}}$, and $\widetilde{\mathcal{N}}$).*

Proof. (i) If $(\mathcal{N},\mathcal{M})$ is a right ω-coprime factorization of $\mathcal{D}$ and $\mathcal{U} \in TIC_\omega(U)$ is invertible in $TIC_\omega(U)$, then it is obvious that $(\mathcal{N}\mathcal{U},\mathcal{M}\mathcal{U})$ is another right ω-coprime factorization. Conversely, suppose that we have two right ω-coprime factorizations $(\mathcal{N},\mathcal{M})$ and $(\mathcal{N}_1,\mathcal{M}_1)$ satisfying the Bezout identities

$$\widetilde{\mathcal{Y}}\mathcal{N} + \widetilde{\mathcal{X}}\mathcal{M} = \widetilde{\mathcal{Y}}_1\mathcal{N}_1 + \widetilde{\mathcal{X}}_1\mathcal{M}_1 = I.$$

Then $\mathcal{M}^{-1} = \widetilde{\mathcal{X}} + \widetilde{\mathcal{Y}}\mathcal{D}$ and $\mathcal{M}_1^{-1} = \widetilde{\mathcal{X}}_1 + \widetilde{\mathcal{Y}}_1\mathcal{D}$ in $TIC_\alpha(U)$, so

$$\mathcal{M}^{-1}\mathcal{M}_1 = \widetilde{\mathcal{X}}\mathcal{M}_1 + \widetilde{\mathcal{Y}}\mathcal{N}_1, \qquad \mathcal{M}_1^{-1}\mathcal{M} = \widetilde{\mathcal{X}}_1\mathcal{M} + \widetilde{\mathcal{Y}}_1\mathcal{N}$$

in $TIC_\alpha(U)$. Define $\mathcal{U} = \mathcal{M}^{-1}\mathcal{M}_1$. We know that $\mathcal{U}$ is invertible in $TIC_\alpha(U)$. However, since $L^2_\omega(\mathbf{R};U) \cap L^2_\alpha(\mathbf{R};U)$ is dense in $L^2_\alpha(\mathbf{R};U)$, the two equations above imply that $\mathcal{U}$ can be extended to an invertible operator in $TIC_\omega(U)$. Moreover, $\mathcal{M}_1 = \mathcal{M}\mathcal{U}$ and $\mathcal{N}_1 = \mathcal{D}\mathcal{M}_1 = \mathcal{D}\mathcal{M}\mathcal{U} = \mathcal{N}\mathcal{U}$. Thus $(\mathcal{N}_1,\mathcal{M}_1) = (\mathcal{N}\mathcal{U},\mathcal{M}\mathcal{U})$, as claimed.

(ii) The proof of (ii) is completely analogous to the proof of (i).

(iii) Choose some operators $\widetilde{\mathcal{Y}}$, $\widetilde{\mathcal{X}}$, $\mathcal{X}$, and $\mathcal{Y}$ in TIC_ω that together with the given operators satisfy the Bezout identities $\widetilde{\mathcal{Y}}\mathcal{N} + \widetilde{\mathcal{X}}\mathcal{M} = I$ and $\widetilde{\mathcal{N}}\mathcal{Y} + \widetilde{\mathcal{M}}\mathcal{X} = I$. Then a direct computation shows that

$$\begin{bmatrix} \widetilde{\mathcal{M}} & \widetilde{\mathcal{N}} \\ -\widetilde{\mathcal{Y}} & \widetilde{\mathcal{X}} \end{bmatrix} \begin{bmatrix} \mathcal{X} - \mathcal{N}(\widetilde{\mathcal{Y}}\mathcal{X} - \widetilde{\mathcal{X}}\mathcal{Y}) & -\mathcal{N} \\ \mathcal{Y} + \mathcal{M}(\widetilde{\mathcal{Y}}\mathcal{X} - \widetilde{\mathcal{X}}\mathcal{Y}) & \mathcal{M} \end{bmatrix} = I.$$

By using the invertibility of $\mathcal{M}$ and $\widetilde{\mathcal{M}}$ in TIC_α, we get

$$\begin{aligned}\mathcal{X}\widetilde{\mathcal{M}}+\mathcal{D}\mathcal{Y}\widetilde{\mathcal{M}}&=I, \qquad \mathcal{X}\widetilde{\mathcal{N}}+\mathcal{D}\mathcal{Y}\widetilde{\mathcal{N}}=\mathcal{D},\\ \mathcal{M}\widetilde{\mathcal{X}}+\mathcal{M}\widetilde{\mathcal{Y}}\mathcal{D}&=I, \qquad \mathcal{N}\widetilde{\mathcal{X}}+\mathcal{N}\widetilde{\mathcal{Y}}\mathcal{D}=\mathcal{D},\end{aligned}$$

and by using these identities we find that

$$\begin{bmatrix}\mathcal{X}-\mathcal{N}(\widetilde{\mathcal{Y}}\mathcal{X}-\widetilde{\mathcal{X}}\mathcal{Y}) & -\mathcal{N}\\ \mathcal{Y}+\mathcal{M}(\widetilde{\mathcal{Y}}\mathcal{X}-\widetilde{\mathcal{X}}\mathcal{Y}) & \mathcal{M}\end{bmatrix}\begin{bmatrix}\widetilde{\mathcal{M}} & \widetilde{\mathcal{N}}\\ -\widetilde{\mathcal{Y}} & \widetilde{\mathcal{X}}\end{bmatrix}=I$$

in TIC_α (as opposed to TIC_ω). However, since all the operators above belong to TIC_ω, and since $L^2_\alpha\cap L^2_\omega$ is dense in L^2_ω, we find that the same identity must be true in TIC_ω, too. Thus, we have a doubly ω-coprime factorization. ∎

As the following theorem shows, if a well-posed linear system is jointly stabilizable and detectable, then its input/output map has a doubly coprime factorization. A converse to this statement is true as well.

THEOREM 4.4.

(i) *Let* $\Psi=\left[\begin{smallmatrix}\mathcal{A} & \mathcal{B}\\ \mathcal{C} & \mathcal{D}\end{smallmatrix}\right]$ *be a jointly* ω*-stabilizable [stabilizable] [[exponentially stabilizable]] and detectable well-posed linear system* (*in the sense of Definition* 3.16). *Then, with the notations of Lemma* 3.13 *and Definition* 4.2,

$$\begin{bmatrix}\widetilde{\mathcal{M}} & \widetilde{\mathcal{N}}\\ -\widetilde{\mathcal{Y}} & \widetilde{\mathcal{X}}\end{bmatrix}\begin{bmatrix}\mathcal{X} & -\mathcal{N}\\ \mathcal{Y} & \mathcal{M}\end{bmatrix}=\begin{bmatrix}I+\mathcal{G}^\sharp & \mathcal{D}^\sharp\\ -\mathcal{E}^\sharp & I-\mathcal{F}^\sharp\end{bmatrix}\begin{bmatrix}I-\mathcal{G}_\flat & -\mathcal{D}_\flat\\ \mathcal{E}_\flat & I+\mathcal{F}_\flat\end{bmatrix}$$

is a doubly ω*-coprime [coprime] [[exponentially coprime]] factorization of* $\mathcal{D}$.

(ii) *Conversely, every* $\mathcal{D}$ *that belongs to* $TIC_\alpha(U;Y)$ *for some* $\alpha\in\mathbf{R}$ *and has a doubly* ω*-coprime [coprime] [[exponentially coprime]] factorization can be realized as the input/output map of a jointly strongly* ω*-stabilizable [stabilizable] [[exponentially stabilizable]] and detectable well-posed linear system* $\Psi=\left[\begin{smallmatrix}\mathcal{A} & \mathcal{B}\\ \mathcal{C} & \mathcal{D}\end{smallmatrix}\right]$.

Proof. $\Psi=\left[\begin{smallmatrix}\mathcal{A} & \mathcal{B}\\ \mathcal{C} & \mathcal{D}\end{smallmatrix}\right]$ be jointly ω-stabilizable and detectable. Then both the systems drawn in Figures 3.5 and 3.6 are ω-stable. In particular, both the input/output map from $\left[\begin{smallmatrix}w\\ u_\flat\end{smallmatrix}\right]$ to $\left[\begin{smallmatrix}w^\sharp\\ u\end{smallmatrix}\right]$ in Figure 3.5, and the input/output map from $\left[\begin{smallmatrix}w^\sharp\\ u\end{smallmatrix}\right]$ to $\left[\begin{smallmatrix}w\\ u_\flat\end{smallmatrix}\right]$ in Figure 3.6 are ω-stable. The former one is given by $\left[\begin{smallmatrix}I-\mathcal{G}_\flat & -\mathcal{D}_\flat\\ \mathcal{E}_\flat & I+\mathcal{F}_\flat\end{smallmatrix}\right]$ (cf. Remark 3.9), and the latter one is given by $\left[\begin{smallmatrix}I+\mathcal{G}^\sharp & \mathcal{D}^\sharp\\ -\mathcal{E}^\sharp & I-\mathcal{F}^\sharp\end{smallmatrix}\right]$. Moreover, by comparing the two figures with each other we immediately realize that they are equivalent in the sense that the relationships between the different signals with the same names are identical in the two diagrams. This means that the input/output map given above are inverses of each other; i.e.,

$$\begin{bmatrix}I-\mathcal{G}_\flat & -\mathcal{D}_\flat\\ \mathcal{E}_\flat & I+\mathcal{F}_\flat\end{bmatrix}=\begin{bmatrix}I+\mathcal{G}^\sharp & \mathcal{D}^\sharp\\ -\mathcal{E}^\sharp & I-\mathcal{F}^\sharp\end{bmatrix}^{-1}.$$

Moreover, as is easily seen, $(\mathcal{D}_\flat,(I+\mathcal{F}_\flat))$ is a right ω-coprime factorization of $\mathcal{D}$, and $((I+\mathcal{G}^\sharp),\mathcal{D}^\sharp)$ is a left ω-coprime factorization of $\mathcal{D}$. This proves part (i) of the theorem.

Conversely, suppose that there exists a doubly coprime factorization of $\mathcal{D}$. Our construction below starts with a realization of the closed loop system $\Psi_\flat$; another equally good choice would be to start with a realization of $\Psi^\sharp$. Motivated by the formula that we found above, we pick the input/output map of $\Psi_\flat$ to be given by

$$\begin{bmatrix}\mathcal{G}_\flat & \mathcal{D}_\flat\\ \mathcal{E}_\flat & \mathcal{F}_\flat\end{bmatrix}=\begin{bmatrix}I-\mathcal{X} & \mathcal{N}\\ \mathcal{Y} & \mathcal{M}-I\end{bmatrix}$$

and choose an arbitrary strongly ω-stable realization of this input/output map, for example, the continuously observable realization presented in Definition 2.10. Since $\mathcal{M}$ is supposed to have an inverse in TIC_α for some $\alpha > \omega$, the operator $\left[\begin{smallmatrix} 0 & 0 \\ 0 & -I \end{smallmatrix}\right]$ is an admissible feedback operator for $\Psi_\flat$. Denote the resulting α-stable closed loop system by Ψ_{ext} and the system that we get by dropping the state feedback row and the output injection column from Ψ_{ext} by Ψ. As a straightforward computation shows, the input/output map of Ψ_{ext} is

$$\begin{aligned} \begin{bmatrix} \mathcal{G} & \mathcal{D} \\ \mathcal{E} & \mathcal{F} \end{bmatrix} &= \begin{bmatrix} I - \mathcal{X} - \mathcal{N}\mathcal{M}^{-1}\mathcal{Y} & \mathcal{N}\mathcal{M}^{-1} \\ \mathcal{M}^{-1}\mathcal{Y} & I - \mathcal{M}^{-1} \end{bmatrix} \\ &= \begin{bmatrix} I - \widetilde{\mathcal{M}}^{-1} & \widetilde{\mathcal{M}}^{-1}\widetilde{\mathcal{N}} \\ \widetilde{\mathcal{Y}}\widetilde{\mathcal{M}}^{-1} & I - \widetilde{\mathcal{X}} - \widetilde{\mathcal{Y}}\widetilde{\mathcal{M}}^{-1}\widetilde{\mathcal{N}} \end{bmatrix}. \end{aligned} \tag{4.2}$$

Observe, in particular, that the input/output map of Ψ is the desired $\mathcal{D} = \mathcal{N}\mathcal{M}^{-1} = \widetilde{\mathcal{M}}^{-1}\widetilde{\mathcal{N}}$. It follows from Proposition 3.3 that the system Ψ that we get in this way is strongly ω-stabilizable (and that the closed loop state feedback system is $\Psi_\flat$). Moreover, by Proposition 3.3 and Lemma 3.5, Ψ is strongly ω-detectable if the operator $\left[\begin{smallmatrix} I & 0 \\ 0 & -I \end{smallmatrix}\right]$ is an ω-stabilizing output feedback operator for $\Psi_\flat$ or, equivalently, if

$$\begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} - \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix} \begin{bmatrix} I - \mathcal{X} & \mathcal{N} \\ \mathcal{Y} & \mathcal{M} - I \end{bmatrix} = \begin{bmatrix} \mathcal{X} & -\mathcal{N} \\ \mathcal{Y} & \mathcal{M} \end{bmatrix}$$

has an inverse in $TIC_\omega(U \times Y, U \times Y)$. But this is true because of the doubly coprimeness assumption. Thus Ψ is jointly strongly ω-stabilizable and detectable. □

The notion of a coprime factorization makes it possible to refine Lemma 3.20 as follows.

LEMMA 4.5. *Assume that both* $\begin{bmatrix} \mathcal{K}^1 & \mathcal{F}^1 \end{bmatrix}$ *and* $\begin{bmatrix} \mathcal{K}^2 & \mathcal{F}^2 \end{bmatrix}$ *are* ω*-stabilizing state feedback pairs for the system* $\Psi = \left[\begin{smallmatrix} \mathcal{A} & \mathcal{B} \\ \mathcal{C} & \mathcal{D} \end{smallmatrix}\right]$. *Then the following conditions are equivalent:*

(i) $\begin{bmatrix} \mathcal{K}^2 & \mathcal{F}^2 \end{bmatrix}$ *is an* ω*-stabilizing state feedback pair for the extended system*

$$\begin{bmatrix} \mathcal{A} & \mathcal{B} \\ \begin{bmatrix} \mathcal{C} \\ \mathcal{K}^1 \end{bmatrix} & \begin{bmatrix} \mathcal{D} \\ \mathcal{F}^1 \end{bmatrix} \end{bmatrix};$$

(ii) *the pair*

$$\begin{bmatrix} (\mathcal{K}^2 - (I - \mathcal{F}^2)(I - \mathcal{F}^1)^{-1}\mathcal{K}^1) & (I - (I - \mathcal{F}^2)(I - \mathcal{F}^1)^{-1}) \end{bmatrix}$$

is an ω*-stabilizing state feedback pair for the closed loop system*

$$\begin{aligned} \Psi_\flat &= \begin{bmatrix} \mathcal{A}_\flat & \mathcal{B}_\flat \\ \begin{bmatrix} \mathcal{C}_\flat \\ \mathcal{K}^1_\flat \end{bmatrix} & \begin{bmatrix} \mathcal{D}_\flat \\ \mathcal{F}^1_\flat \end{bmatrix} \end{bmatrix} \\ &= \begin{bmatrix} \mathcal{A} + \mathcal{B}\tau \left(I - \mathcal{F}^1\right)^{-1} \mathcal{K}^1 & \mathcal{B}\left(I - \mathcal{F}^1\right)^{-1} \\ \begin{bmatrix} \mathcal{C} + \mathcal{D}\left(I - \mathcal{F}^1\right)^{-1} \mathcal{K}^1 \\ \left(I - \mathcal{F}^1\right)^{-1} \mathcal{K}^1 \end{bmatrix} & \begin{bmatrix} \mathcal{D}\left(I - \mathcal{F}^1\right)^{-1} \\ \left(I - \mathcal{F}^1\right)^{-1} - I \end{bmatrix} \end{bmatrix} \end{aligned}$$

that one gets from Ψ *by using the state feedback pair* $\begin{bmatrix} \mathcal{K}^1 & \mathcal{F}^1 \end{bmatrix}$;

(iii) $\mathcal{F}^1(I - \mathcal{F}^2)^{-1}$ *and* $\mathcal{K}^1 + \mathcal{F}^1(I - \mathcal{F}^2)^{-1}\mathcal{K}^2$ *are* ω*-stable.*

A sufficient condition for (i)–(iii) *to hold is that* $\mathcal{D}(I-\mathcal{F}^1)^{-1}$ *and* $(I-\mathcal{F}^1)^{-1}$ *are right* ω*-coprime. This condition is necessary for* (i)–(iii) *to hold whenever* $\mathcal{D}(I-\mathcal{F}^2)^{-1}$ *and* $(I-\mathcal{F}^2)^{-1}$ *are right* ω*-coprime.*

Proof. Let us study the further extended system

$$\begin{bmatrix} \mathcal{A} & \mathcal{B} \\ \begin{bmatrix} \mathcal{C} \\ \mathcal{K}^1 \\ \mathcal{K}^2 \\ \mathcal{K}^2-\mathcal{K}^1 \end{bmatrix} & \begin{bmatrix} \mathcal{D} \\ \mathcal{F}^1 \\ \mathcal{F}^2 \\ \mathcal{F}^2-\mathcal{F}^1 \end{bmatrix} \end{bmatrix},$$

where the last line is the difference between the two previous lines. If we use here the output feedback operator $\begin{bmatrix} 0 & I & 0 & 0 \end{bmatrix}$, then we get an extended version of the closed loop system $\Psi_\flat$ in the statement of the lemma (that we still denote by $\Psi_\flat$), namely,

$$\begin{aligned} \Psi_\flat &= \begin{bmatrix} \mathcal{A}_\flat & \mathcal{B}_\flat \\ \begin{bmatrix} \mathcal{C}_\flat \\ \mathcal{K}^1_\flat \\ \mathcal{K}^2_\flat \\ \mathcal{K}^2_\flat-\mathcal{K}^1_\flat \end{bmatrix} & \begin{bmatrix} \mathcal{D}_\flat \\ \mathcal{F}^1_\flat \\ \mathcal{F}^2_\flat \\ \mathcal{F}^2_\flat-\mathcal{F}^1_\flat \end{bmatrix} \end{bmatrix} \\ &= \begin{bmatrix} \mathcal{A}+\mathcal{B}\tau\left(I-\mathcal{F}^1\right)^{-1}\mathcal{K}^1 & \mathcal{B}\left(I-\mathcal{F}^1\right)^{-1} \\ \begin{bmatrix} \mathcal{C}+\mathcal{D}\left(I-\mathcal{F}^1\right)^{-1}\mathcal{K}^1 \\ \left(I-\mathcal{F}^1\right)^{-1}\mathcal{K}^1 \\ \mathcal{K}^2+\mathcal{F}^2\left(I-\mathcal{F}^1\right)^{-1}\mathcal{K}^1 \\ \mathcal{K}^2-\left(I-\mathcal{F}^2\right)\left(I-\mathcal{F}^1\right)^{-1}\mathcal{K}^1 \end{bmatrix} & \begin{bmatrix} \mathcal{D}\left(I-\mathcal{F}^1\right)^{-1} \\ \left(I-\mathcal{F}^1\right)^{-1}-I \\ \mathcal{F}^2\left(I-\mathcal{F}^1\right)^{-1} \\ \left(I-\mathcal{F}^2\right)\left(I-\mathcal{F}^1\right)^{-1}-I \end{bmatrix} \end{bmatrix}. \end{aligned}$$

If we instead use the output feedback operator $\begin{bmatrix} 0 & 0 & I & 0 \end{bmatrix}$, then we get the system

$$\begin{aligned} \Psi_\natural &= \begin{bmatrix} \mathcal{A}_\natural & \mathcal{B}_\natural \\ \begin{bmatrix} \mathcal{C}_\natural \\ \mathcal{K}^1_\natural \\ \mathcal{K}^2_\natural \\ \mathcal{K}^2_\natural-\mathcal{K}^1_\natural \end{bmatrix} & \begin{bmatrix} \mathcal{D}_\natural \\ \mathcal{F}^1_\natural \\ \mathcal{F}^2_\natural \\ \mathcal{F}^2_\natural-\mathcal{F}^1_\natural \end{bmatrix} \end{bmatrix} \\ &= \begin{bmatrix} \mathcal{A}+\mathcal{B}\tau\left(I-\mathcal{F}^2\right)^{-1}\mathcal{K}^2 & \mathcal{B}\left(I-\mathcal{F}^2\right)^{-1} \\ \begin{bmatrix} \mathcal{C}+\mathcal{D}\left(I-\mathcal{F}^2\right)^{-1}\mathcal{K}^2 \\ \mathcal{K}^1+\mathcal{F}^1\left(I-\mathcal{F}^2\right)^{-1}\mathcal{K}^2 \\ \left(I-\mathcal{F}^2\right)^{-1}\mathcal{K}^2 \\ -\mathcal{K}^1+\left(I-\mathcal{F}^1\right)\left(I-\mathcal{F}^2\right)^{-1}\mathcal{K}^2 \end{bmatrix} & \begin{bmatrix} \mathcal{D}\left(I-\mathcal{F}^2\right)^{-1} \\ \mathcal{F}^1\left(I-\mathcal{F}^2\right)^{-1} \\ \left(I-\mathcal{F}^2\right)^{-1}-I \\ \left(I-\mathcal{F}^1\right)\left(I-\mathcal{F}^2\right)^{-1}-I \end{bmatrix} \end{bmatrix}. \end{aligned}$$

By Proposition 3.3, we get the same system by using $\begin{bmatrix} 0 & -I & I & 0 \end{bmatrix}$ as a feedback operator for $\Psi_\flat$ or, equivalently, by using the feedback operator $\begin{bmatrix} 0 & 0 & 0 & I \end{bmatrix}$. This proves the equivalence of (i) and (ii). We know that the operators on the first, second, and fourth row of $\Psi_\natural$ are stable (since $\begin{bmatrix} \mathcal{K}^2 & \mathcal{F}^2 \end{bmatrix}$ is stabilizing for Ψ), so the full system $\Psi_\natural$ is stable iff the two conditions listed in (iii) hold (recall that the last line is the difference between the two previous lines).

Suppose that $\mathcal{D}(I-\mathcal{F}^1)^{-1}$ and $(I-\mathcal{F}^1)^{-1}$ are right coprime. We claim that (i)–(iii) then hold. To prove this we choose operators $\widetilde{\mathcal{Y}}$ and $\widetilde{\mathcal{X}}$ in TIC that together

with $\mathcal{D}_\flat$ and $I+\mathcal{F}^1_\flat$ satisfy the Bezout identity

$$\widetilde{\mathcal{Y}}\mathcal{D}_\flat + \widetilde{\mathcal{X}}\left(I+\mathcal{F}^1_\flat\right) = \widetilde{\mathcal{Y}}\mathcal{D}\left(I-\mathcal{F}^1\right)^{-1} + \widetilde{\mathcal{X}}\left(I-\mathcal{F}^1\right)^{-1} = I.$$

Then

$$\begin{aligned}\widetilde{\mathcal{Y}}\mathcal{D}_\natural + \widetilde{\mathcal{X}}\left(I+\mathcal{F}^2_\natural\right) &= \widetilde{\mathcal{Y}}\mathcal{D}\left(I-\mathcal{F}^2\right)^{-1} + \widetilde{\mathcal{X}}\left(I-\mathcal{F}^2\right)^{-1}\\ &= \left(\widetilde{\mathcal{Y}}\mathcal{D}+\widetilde{\mathcal{X}}\right)\left(I-\mathcal{F}^2\right)^{-1}\\ &= \left(I-\mathcal{F}^1\right)\left(I-\mathcal{F}^2\right)^{-1},\end{aligned}$$

and this shows that $\left(I-\mathcal{F}^1\right)\left(I-\mathcal{F}^2\right)^{-1}$ is stable; hence, $\mathcal{F}^1\left(I-\mathcal{F}^2\right)^{-1}$ is stable. A similar computation

$$\begin{aligned}\widetilde{\mathcal{Y}}\left(\mathcal{C}_\natural-\mathcal{C}_\flat\right)+\widetilde{\mathcal{X}}\left(\mathcal{K}^2_\natural-\mathcal{K}^2_\flat\right) &= \left(\widetilde{\mathcal{Y}}\mathcal{D}+\widetilde{\mathcal{X}}\right)\left(\left(I-\mathcal{F}^2\right)^{-1}\mathcal{K}^2-\left(I-\mathcal{F}^1\right)^{-1}\mathcal{K}^1\right)\\ &= -\mathcal{K}^1+\left(I-\mathcal{F}^1\right)\left(I-\mathcal{F}^2\right)^{-1}\mathcal{K}^2\end{aligned}$$

shows that $-\mathcal{K}^1+\left(I-\mathcal{F}^1\right)\left(I-\mathcal{F}^2\right)^{-1}\mathcal{K}^2$ is stable; hence, $\mathcal{K}^1+\mathcal{F}^1\left(I-\mathcal{F}^2\right)^{-1}\mathcal{K}^2$ is stable.

Finally, let us assume that (iii) holds, and that $\mathcal{D}(I-\mathcal{F}^2)^{-1}$ and $(I-\mathcal{F}^2)^{-1}$ are right coprime. By interchanging the two feedback pairs with each other and using the statement that we have just proved, we find that $(I-\mathcal{F}^2)(I-\mathcal{F}^1)^{-1}$ is invertible in $TIC(U)$. This, combined with the coprimeness of $\mathcal{D}(I-\mathcal{F}^2)^{-1}$ and $(I-\mathcal{F}^2)^{-1}$ and Lemma 4.3, implies that $\mathcal{D}(I-\mathcal{F}^1)^{-1}$ and $(I-\mathcal{F}^1)^{-1}$ must be right coprime. ∎

5. Dynamic stabilization. As is well known, if the input/output map $\mathcal{D}$ has a right ω-coprime factorization $(\mathcal{N},\mathcal{M})$, and if $\widetilde{\mathcal{Y}}$ and $\widetilde{\mathcal{X}}$ together with $\mathcal{N}$ and $\mathcal{M}$ satisfy the Bezout identity

$$\widetilde{\mathcal{Y}}\mathcal{N}+\widetilde{\mathcal{X}}\mathcal{M}=I,$$

then $\mathcal{Q}=\widetilde{\mathcal{X}}^{-1}\widetilde{\mathcal{Y}}$ is an ω-stabilizing compensator for $\mathcal{D}$, provided it is possible to make sense out of $\widetilde{\mathcal{X}}^{-1}$. A similar statement is true in the case where $\mathcal{D}$ has a left ω-coprime factorization. If $\mathcal{D}$ has a doubly ω-coprime factorization, then, with the notations of Definition 4.2, the stabilizing compensator $\mathcal{Q}$ is given by $\mathcal{Q}=\widetilde{\mathcal{X}}^{-1}\widetilde{\mathcal{Y}}=\mathcal{Y}\mathcal{X}^{-1}$, still provided $\widetilde{\mathcal{X}}^{-1}$ and $\mathcal{X}^{-1}$ make sense. If $\widetilde{\mathcal{X}}$ and $\mathcal{X}$ do not have inverses in TIC_α for any $\alpha>0$, then $\mathcal{Q}$ does not belong to $TIC_\alpha(Y;U)$ for any $\alpha>0$, and $\mathcal{Q}$ cannot be realized as the input/output map of a well-posed linear system (see the discussion following Definition 4.2). Thus it is natural to impose this extra condition on a doubly coprime factorization as follows.

DEFINITION 5.1. *Let $\alpha>\omega$, and let $\mathcal{D}\in TIC_\alpha(U;Y)$ and $\mathcal{Q}\in TIC_\alpha(Y;U)$. A joint doubly ω-coprime factorization of $\mathcal{D}$ and $\mathcal{Q}$ consists of eight operators in TIC_ω (with the appropriate dimensions) satisfying* (4.1)*, and, in addition, we require that $(\mathcal{N},\mathcal{M})$ is a right and $(\widetilde{\mathcal{M}},\widetilde{\mathcal{N}})$ is a left ω-coprime factorization of $\mathcal{D}$, and that $(\mathcal{Y},\mathcal{X})$ is a right and $(\widetilde{\mathcal{X}},\widetilde{\mathcal{Y}})$ is a left ω-coprime factorization of $\mathcal{Q}$. (In particular, all the denominators $\mathcal{M}$, $\widetilde{\mathcal{M}}$, $\mathcal{X}$, and $\widetilde{\mathcal{X}}$ are invertible in TIC_α.)*

LEMMA 5.2. *Let $\Psi=\left[\begin{smallmatrix}\mathcal{A}&\mathcal{B}\\ \mathcal{C}&\mathcal{D}\end{smallmatrix}\right]$ be a jointly ω-stabilizable and detectable well-posed linear system, and let Ψ_{ext}, $\Psi_\flat$, and $\Psi^\sharp$ denote the systems in Definition* 3.11 *and Lemma* 3.13*. Then the following conditions are equivalent:*

(i) *the operator $\left[\begin{smallmatrix}I&0\\0&I\end{smallmatrix}\right]$ is an admissible output feedback operator for Ψ_{ext};*

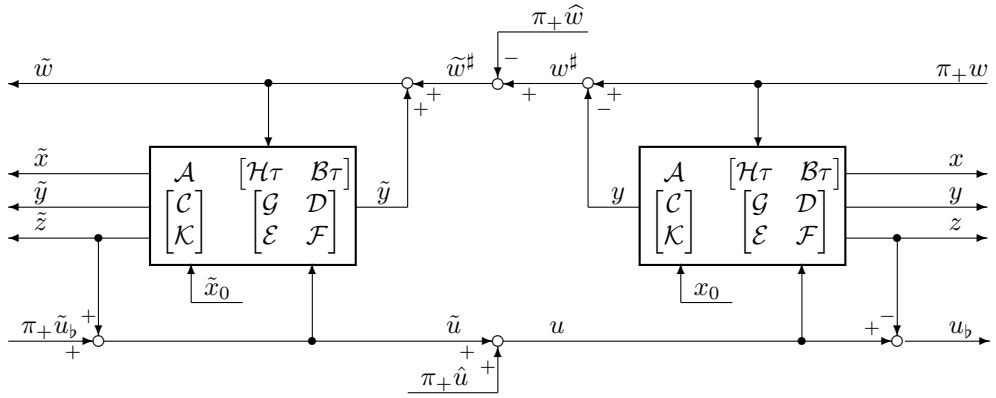


FIG. 5.1. *Dynamic stabilization.*

(ii) $I - \mathcal{G}_\flat$ *has an inverse in* $TIC_\alpha(Y)$ *for some* $\alpha \geq \omega$*;*
(iii) $I - \mathcal{F}^\sharp$ *has an inverse in* $TIC_\alpha(U)$ *for some* $\alpha \geq \omega$.

In these cases the system $\Psi_\flat^\sharp$ *drawn in the left half of Figure* 5.1 *with inputs* $\widetilde{w}^\sharp$ *and* $\tilde{u}_\flat$ *and outputs* $\tilde{y}$, $\tilde{z}$, $\tilde{w}$, *and* $\tilde{u}$ *(i.e., the system that we get by using* $\left[\begin{smallmatrix} I & 0\\ 0 & I\end{smallmatrix}\right]$ *as an output feedback operator for* Ψ_{ext}*) is a well-posed linear system, and the coprime factorization presented in Theorem* 4.4 *is a joint doubly* ω*-coprime factorization of* $\mathcal{D}$ *and* $\mathcal{Q}$*, where* $\mathcal{Q} = \mathcal{E}_\flat(I - \mathcal{G}_\flat)^{-1} = (I - \mathcal{F}^\sharp)^{-1}\mathcal{E}^\sharp$.

This follows from Proposition 3.3.

In the situation described above the input/output map of the closed loop system $\Psi_\flat^\sharp$ is equal to the stabilizing compensator $\mathcal{Q}$, and we can use the observer connection drawn in Figure 5.1 to stabilize the system as shown in Theorem 5.3.

THEOREM 5.3. *Let* $\Psi = \left[\begin{smallmatrix}\mathcal{A} & \mathcal{B}\\ \mathcal{C} & \mathcal{D}\end{smallmatrix}\right]$ *be a jointly* ω*-stabilizable and detectable well-posed linear system, and let* Ψ_{ext} *denote the system in Definition* 3.11. *Then the connection drawn in Figure* 5.1 *defines an* ω*-stable well-posed linear system. Moreover, the input/output maps for the two additional outputs* $\tilde{w}$ *and* $u_\flat$ *are given by*

$$\begin{bmatrix}\tilde{w}\\ u_\flat\end{bmatrix} = \begin{bmatrix}\pi_+ w\\ \pi_+\tilde{u}_\flat\end{bmatrix} + \begin{bmatrix}-I - \mathcal{G}^\sharp & -\mathcal{D}^\sharp\\ -\mathcal{E}^\sharp & I - \mathcal{F}^\sharp\end{bmatrix}\begin{bmatrix}\pi_+\widehat{w}\\ \pi_+\hat{u}\end{bmatrix}.$$

Proof. By Remark 3.14, we can regard Ψ_{ext} as a state feedback perturbed version of the closed loop system $\Psi_\flat$; see Figure 3.7. By substituting this system for Ψ_{ext} in Figure 5.1 we get the equivalent Figure 5.2, which can be interpreted as a feedback connection for an ω-stable system consisting of two copies of $\Psi_\flat$. By part (i) of Lemma 3.5, it suffices to show that the two internal inputs $u_\flat$ and $\tilde{w}$ depend continuously in L^2_ω on the four inputs. By using the equations describing the summation junctions in Figure 5.2 we can eliminate the variables u, $\tilde{u}$, $w^\sharp$, and $\widetilde{w}^\sharp$ to get

$$\begin{bmatrix}I - \mathcal{G}_\flat & -\mathcal{D}_\flat\\ \mathcal{E}_\flat & I + \mathcal{F}_\flat\end{bmatrix}\begin{bmatrix}\pi_+ w - \tilde{w}\\ u_\flat - \pi_+\tilde{u}_\flat\end{bmatrix} = \begin{bmatrix}\pi_+\widehat{w}\\ \pi_+\hat{u}\end{bmatrix}.$$

By Theorem 4.4, the operator on the left-hand side has an inverse in TIC_ω. Inverting this operator we get the formula given in Theorem 5.3. ∎

REMARK 5.4. *Observe that Theorem* 5.3 *is true even if the equivalent conditions listed in Lemma* 5.2 *are false. However, if this is the case, then although the system*

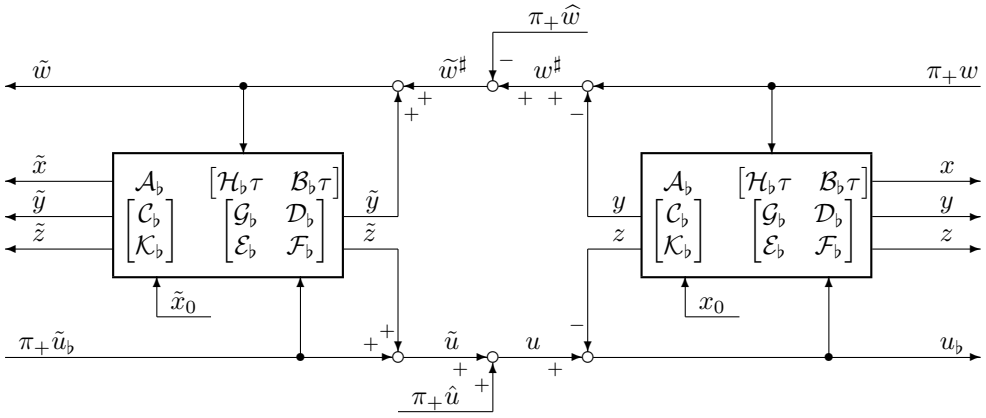


FIG. 5.2. *Alternative version of compensator connection.*

in Figure 5.1 *is an* ω*-stable well-posed linear system, the compensator (the left half of Figures* 5.1 *and* 5.2*) is not well-posed by itself, and the well posedness is lost if the feedback loop from the compensator to the original system is opened.*

REMARK 5.5. *By using Theorem* 5.3 *one can easily develop a Youla parameterization of the set of all stabilizing compensators for* Ψ_{ext}*. The key observation is that the input/output map from* $\left[\begin{smallmatrix}\pi_+ w\\ \pi_+\tilde{u}_\flat\end{smallmatrix}\right]$ *to* $\left[\begin{smallmatrix}\tilde{w}\\ u_\flat\end{smallmatrix}\right]$ *is the identity map. To get the Youla parameterization we simply connect the Youla parameter* Q *from* $\tilde{w}$ *to* $\tilde{u}_\flat$ *in the equivalent Figures* 5.1 *and* 5.2*. This does not affect the stability of the system since* $\tilde{w}$ *does not depend on* $\tilde{u}_\flat$*. The resulting input/output map from* w *to* $u_\flat$ *will be equal to* Q*. The proofs of these claims are essentially the same as the ones given in* [6].

REMARK 5.6. *All the main results of this paper remain valid if throughout we replace the algebra of time-invariant bounded linear operators from* $L^2(\mathbf{R};U)$ *into* $L^2(\mathbf{R};Y)$ *by various subalgebras, for example the algebra of convolution operators induced by measures with finite total variation. This is the algebra around which* [22] *and* [23] *were built.*

REFERENCES

[1] R. F. CURTAIN, *Representations of infinite-dimensional systems*, in Three Decades of Mathematics Systems Theory. A Collection of Surveys at the Occasion of the 50th Birthday of Jan C. Willems, Lecture Notes in Control and Information Sciences 135, H. Nijmeijer and J. M. Schumacher, eds., Springer-Verlag, Berlin, New York, 1989, pp. 101–128.

[2] R. F. CURTAIN, *A synthesis of time and frequency domain methods for the control of infinite-dimensional systems: A system theoretic approach*, in Control and Estimation in Distributed Parameter Systems, Frontiers in Applied Mathematics, H. T. Banks, ed., SIAM, Philadelphia, PA, 1992, pp. 171–224.

[3] R. F. CURTAIN AND G. WEISS, *Well posedness of triplets of operators (in the sense of linear systems theory)*, in Control and Optimization of Distributed Parameter Systems, Birkhäuser-Verlag, Basel, 1989, pp. 401–416.

[4] R. F. CURTAIN AND G. WEISS, *Dynamic stabilization of regular linear systems*, IEEE Trans. Automat. Control, 42 (1996), pp. 1–18.

[5] R. F. CURTAIN, G. WEISS, AND M. WEISS, *Coprime factorization for regular linear systems*, Automatica, 32 (1996), pp. 1519–1531.

[6] R. F. CURTAIN, G. WEISS, AND M. WEISS, *Stabilization of irrational transfer functions by controllers with internal loop*, Automatica, 1997, submitted.

[7] F. FLANDOLI, I. LASIECKA, AND R. TRIGGIANI, *Algebraic (Riccati) equations with non-*

smoothing observation arising in hyperbolic and (Euler–Bernuolli) boundary control problems, Ann. Mat. Pura Appl., 153 (1988), pp. 307–382.

[8] Y. Fourès and I. E. Segal, *Causality and analyticity*, Trans. Amer. Math. Soc., 78 (1955), pp. 385–405.

[9] T. T. Georgiou and M. C. Smith, *Graphs, causality, and stabilizability: Linear, shift-invariant systems on* $L^2[0,\infty)$, Math. Control Signals Systems, 6 (1993), pp. 195–223.

[10] S. Hansen and G. Weiss, *New results on the operator Carleson measure criterion*, IMA J. Math. Control Inform., 14 (1997), pp. 3–32.

[11] B. Jacob and H. Zwart, *Equivalent Conditions for Stabilizability of Infinite-dimensional Systems with Admissible Control Operators*, preprint, 1996.

[12] H. Logemann, *Stabilization and regulation of infinite-dimensional systems using coprime factorizations*, in Analysis and Optimization of Systems: State and Frequency Domain Approaches for Infinite-Dimensional Systems, Lecture Notes in Control and Information Sciences 185, Springer-Verlag, Berlin, New York, 1993, pp. 102–139.

[13] K. A. Morris, *State feedback and estimation of well-posed systems*, Math. Control Signals Systems, 7 (1994), pp. 351–388.

[14] R. Ober and S. Montgomery-Smith, *Bilinear transformation of infinite-dimensional state-space systems and balanced realizations of nonrational transfer functions*, SIAM J. Control Optim., 28 (1990), pp. 438–465.

[15] A. J. Pritchard and D. Salamon, *The linear-quadratic control problem for retarded systems with delays in control and observation*, IMA J. Math. Control Inform., 2 (1985), pp. 335–362.

[16] A. J. Pritchard and D. Salamon, *The linear quadratic control problem for infinite dimensional systems with unbounded input and output operators*, SIAM J. Control Optim., 25 (1987), pp. 121–144.

[17] R. Rebarber, *Conditions for the equivalence of internal and external stability for distributed parameter systems*, IEEE Trans. Automat. Control, 38 (1993), pp. 994–998.

[18] D. Salamon, *Control and Observation of Neutral Systems*, Pitman Publishing Ltd., London, 1984.

[19] D. Salamon, *Infinite dimensional linear systems with unbounded control and observation: A functional analytic approach*, Trans. Amer. Math. Soc., 300 (1987), pp. 383–431.

[20] D. Salamon, *Realization theory in (Hilbert) space*, Math. Systems Theory, 21 (1989), pp. 147–164.

[21] M. C. Smith, *On stabilization and the existence of coprime factorizations*, IEEE Trans. Automat. Control, 34 (1989), pp. 1005–1007.

[22] O. J. Staffans, *Quadratic optimal control of stable systems through spectral factorization*, Math. Control Signals Systems, 8 (1995), pp. 167–197.

[23] O. J. Staffans, *Quadratic Optimal Control Through Coprime and Spectral Factorizations*, Technical Reports on Computer Science & Mathematics, Series A 178, Åbo Akademi University, Åbo, Finland, 1996.

[24] O. J. Staffans, *Quadratic optimal control of stable well-posed linear systems*, Trans. Amer. Math. Soc., 349 (1997), pp. 3679–3715.

1997, to appear.

[25] O. J. Staffans, *Quadratic optimal control of well-posed linear systems*, SIAM J. Control Optim., to appear.

[26] B. van Keulen, H_∞*-Control for Distributed Parameter Systems: A State Space Approach*, Birkhäuser-Verlag, Basel, Boston, Berlin, 1993.

[27] G. Weiss, *Admissibility of unbounded control operators*, SIAM J. Control Optim., 27 (1989), pp. 527–545.

[28] G. Weiss, *Admissible observation operators for linear semigroups*, Israel J. Math., 65 (1989), pp. 17–43.

[29] G. Weiss, *Representations of shift-invariant operators on* L^2 *by* H^∞ *transfer functions: An elementary proof, a generalization to* L^p*, and a counterexample for* L^∞, Math. Control Signals Systems, 4 (1991), pp. 193–203.

[30] G. Weiss, *Transfer functions of regular linear systems. Part* I*: Characterizations of regularity*, Trans. Amer. Math. Soc., 342 (1994), pp. 827–854.

[31] G. Weiss, *Regular linear systems with feedback*, Math. Control Signals Systems, 7 (1994), pp. 23–57.

[32] G. Weiss and H. Zwart, *An example in linear quadratic optimal control*, Systems Control Lett., 1998.

[33] M. Weiss and G. Weiss, *Optimal control of stable weakly regular linear systems*, Math. Control Signals Systems, 1997.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 4, pp. 1293–1314, July 1998

010

SOME PATHOLOGICAL TRAPS FOR STOCHASTIC APPROXIMATION*

ODILE BRANDIÈRE†

Abstract. We consider different kinds of "pathological traps" for stochastic algorithms, thus extending a previous study on regular traps. An illustration is given by the complete proof of the convergence of a principal component analysis (PCA) algorithm when the eigenvalues are multiple.

Key words. stochastic approximation, ordinary differential equations, traps, neural networks

AMS subject classifications. 62L20, 60F99, 58F21, 92B20

PII. S036301299630759X

1. Introduction.

We consider the $\mathbb{R}^d$-valued stochastic algorithm, defined on a probability space $(\Omega, \mathcal{A}, \mathcal{P})$:

$$Z_{n+1} = Z_n + \gamma_n h(Z_n) + \eta_{n+1}, \tag{1}$$

where

- h is a continuous function from an open set $G \subseteq \mathbb{R}^d$ to $\mathbb{R}^d$,
- (γ_n) is a decreasing-to-zero deterministic real sequence satisfying

$$\sum_{n \geq 0} \gamma_n = \infty, \tag{2}$$

- (η_n) is a "small" stochastic disturbance.

The ordinary differential equation (ODE) method (see [3], [8], [14] and others) associates the possible limit sets of (1) with the properties of the associated ODE

$$\frac{dz}{dt} = h(z). \tag{3}$$

When the algorithm is bounded, these sets are compact connected invariant and "chain-recurrent" in the Benaïm sense (see [2]) for the ODE.

It is natural to think that, thanks to the random disturbance, the algorithm (1) avoids some of those limit sets which we shall call "traps."

The most simple limit sets of (1) (and (3)) are the "regular zeros of h": z^* is a "regular trap of h" if z^* is an isolated zero with a neighborhood where h is $\mathcal{C}^1$ with a Lipschitz differential Dh, having at z^* at least one eigenvalue with a positive real part. Such "regular traps" have been studied in [6], [16], and [25].

The aim of this paper is to study some other compact connected chain-recurrent sets L such as

- periodic cycles for the ODE,
- singular equilibria and other "repulsive regions,"
- connected sets of equilibria.

*Received by the editors July 31, 1996; accepted for publication (in revised form) June 16, 1997; published electronically May 15, 1998.
http://www.siam.org/journals/sicon/36-4/30759.html

†Université de Marne-la-Vallée, Equipe d'Analyse et de Mathématiques Appliquées, 2, rue de la Butte Verte, 93166 Noisy-le-Grand, France. Current address: Département de Mathématiques, Université Paris Sud, Bâtiment 425, 91405 Orsay, France (brandiere@jm.u-psud.fr).

Our question is: "does $\{\omega\ ;\ d(Z_n(\omega), L) \to 0\}$ have probability zero?"

Throughout this paper, we shall always make the following assumptions.

Assumptions A1. On the "small disturbance"

$$\eta_{n+1} = c_n(\varepsilon_{n+1} + r_{n+1}), \tag{4}$$

where (c_n) is a nonnegative deterministic sequence such that

$$\bullet \gamma_n = O(c_n),\ \Sigma c_n^2 < \infty, \text{ and } c_n \neq 0 \text{ infinitely often}, \tag{5}$$

• (ε_n) and (r_n) are $\mathbb{R}^d$-valued random vector sequences, defined on $(\Omega, \mathcal{A}, \mathcal{P})$ adapted with respect to an increasing sequence of σ-fields $(\mathcal{F}_n)_{n\geq 0}$ and satisfying almost surely (a.s.) on $\{\omega\ ;\ d(Z_n(\omega), L) \to 0\}$

$$E(\varepsilon_{n+1}|\mathcal{F}_n) = 0 \quad \text{and} \quad \Sigma\|r_n\|^2 < \infty. \tag{6}$$

Now we state our main results. The proofs will be given in the following sections. Then we will illustrate some results by the proof of the convergence of a PCA (principal component analysis) algorithm.

Symbols. We denote by

• $\lambda_{min}(A)$, the smallest eigenvalue of the symmetric matrix A,

• $\underline{\lambda}(A)$ (resp., $\overline{\lambda}(A)$), the smallest (resp., largest) real part of eigenvalues of the matrix A,

• $\Gamma(L) = \{\omega\ ;\ d(Z_n(\omega), L) \to 0\}$,

• $L_r = \{x \in \mathbb{R}^d\ ;\ d(x, L) < r\}$,

• $\overline{L_r} = \{x \in \mathbb{R}^d\ ;\ d(x, L) \leq r\}$,

• **C**, a generic positive constant whose value may change,

• i.i.d., independent and identically distributed.

The "ODE" will always be the ODE (3); a "solution" of the ODE, $t \to z(t)$, will be considered for $t \geq 0$, $z(0)$ being the initial condition.

1.1. Cyclic traps. Here we extend the framework of a result of Benaïm [1].

Set $L \subseteq G$, a closed and periodic orbit, solution to the ODE (3). We assume that h is $\mathcal{C}^2$ on a neighborhood of L and that L is a periodic and hyperbolic cycle having at least one characteristic exponent with a positive real part; we shall call such an L a *cyclic trap.*

Then we claim the following theorem.

THEOREM 1.1. *Let L be a cyclic trap of the stochastic algorithm* (1) *under the Assumptions* A1. *Assume that h is $\mathcal{C}^2$ on a neighborhood of L and that for some $a > 2$, a.s. on $\Gamma(L)$,*

$$\limsup_n E(\|\varepsilon_{n+1}\|^a|\mathcal{F}_n) < \infty \quad \textit{and} \quad \liminf_n E(\lambda_{min}(\varepsilon_{n+1}(\varepsilon_{n+1})^T)|\mathcal{F}_n) > 0. \tag{7}$$

Then $P(\Gamma(L)) = 0$; the cyclic trap L is avoided by (1).

The proof of Theorem 1.1 is given in section 2.1.

1.2. Repulsion and singular equilibria. In this section we introduce the notion of repulsive set.

DEFINITION 1. *A compact connected set L which is invariant and chain-recurrent for the ODE will be called* repulsive *if there exists an $r > 0$ such that any solution to the ODE, $(z(t))_{t\geq 0}$, starting from $x \in L_r \setminus L$ leaves L_r within a finite time.*

Some results will be given in section 2.2.3 for such general repulsive sets. The easiest case is the case of singular repulsive equilibria. z^* is a *singular equilibrium* of

h (or of the associated ODE) if it is an isolated zero of h ($L = \{z^*\}$) such that h is $\mathcal{C}^1$ on a neighborhood of z^*, with a Lipschitz differential Dh verifying $Dh(z^*) = 0$.

For those results we consider the algorithm (1) with

$$\gamma_n = c_n = \frac{g}{n}, \qquad g > 0.$$

THEOREM 1.2 (d-dimensional repulsive singular equilibrium). *Set* (1) *under the Assumptions* A1. *We consider an isolated zero z^* of h, repulsive for the ODE, h being $\mathcal{C}^1$ on a neighborhood of z^* with a differential Dh verifying $Dh(z^*) = 0$. Let us assume that a.s. on $\Gamma(z^*)$ and for $a > 4$,*

$$\limsup_n E(\|\varepsilon_{n+1}\|^a | \mathcal{F}_n) < \infty \quad \text{and} \quad \liminf_n E(\|\varepsilon_{n+1}\|^2 | \mathcal{F}_n) > 0. \tag{8}$$

Then $\Gamma(z^*)$ has probability zero.

We prove Theorem 1.2 in section 2.2.1. In one dimension, we obtain the following proposition about singular, but not necessarily repulsive, equilibria.

PROPOSITION 1.3 (one-dimensional general singular equilibrium). *Set* (1) *under the Assumptions* A1 *with $d = 1$. We consider an isolated zero z^* of h such that on a neighborhood of z^*,*

$$h(z) \simeq \alpha(z - z^*)^p \text{ with } \alpha \neq 0, \ p \text{ integer and } p > 1.$$

We assume that a.s. on $\Gamma(z^*)$, (8) is verified with $a > \frac{2p}{p-1}$.

Then, if $\alpha > 0$ and if p is odd, z^* is a.s. avoided by (1).

Otherwise, a.s. on $\Gamma(z^*)$, when $n \to \infty$,

$$|Z_n - z^*| \simeq [|\alpha|(p-1)g \log n]^{-\frac{1}{p-1}},$$

and $(Z_n - z^*)$ has, for n large enough, a constant sign which is necessarily the same as $(-\alpha)$ when p is even.

This proposition is proved in section 2.2.2.

For $p = 2$ and $\alpha > 0$, the almost sure convergence rate,

$$\log n \|Z_n - z^*\| \to \frac{1}{\alpha g},$$

has been obtained by Kersting [13] in a more restrictive framework and later by Wei [27] with a square integrable noise.

1.3. Connected sets of equilibria. Let L be a nonempty, compact, and connected part included in $\{h = 0\}$. It is a possible limit set of (1).

Clearly, when L has a nonempty interior, the algorithm might get "bogged" in L, thus converging to a random point of L. Hence we only consider L with empty interior.

The following framework looks somewhat restrictive but it will be helpful for the proof of Theorem 3.1 on the principal component analysis. We set the following definition.

DEFINITION 2. *A compact connected set of equilibria L is called* homogeneous *if:*

- *on a neighborhood of L, h is $\mathcal{C}^1$ and Dh is Lipschitz;*
- *for all $x \in L$, $\mathbb{R}^d$ is the direct sum of the repulsive subspace K_r, associated to the eigenvalues of $Dh(x)$ with a positive real part, and of the nonrepulsive subspace K_a*

associated to the eigenvalues of $Dh(x)$ with a nonpositive real part. And in a suitable basis $\mathcal{B}$ (the first vectors of $\mathcal{B}$ belong to K_r and the others belong to K_a):

$$\forall x \in L,\ Dh(x) = \begin{pmatrix} J_+ & 0 \\ 0 & J_-(x) \end{pmatrix},$$

where $J_-(x)$ may depend of x but not on J_+; J_+ is repulsive ($\underline{\lambda}(J_+) > 0$) and $\overline{\lambda}(J_-(x)) \leq 0$.

- *K_a contains L.*

Denote by $\varepsilon_{n+1}^{(r)}$ the projection of ε_{n+1} on K_r in the direction of K_a. We claim the following result.

THEOREM 1.4. *Assume that*

- *L is a compact connected set of equilibria, homogeneous and nonattractive ($K_r \neq \{0\}$),*
- *Assumptions* A1 *are satisfied,*
- *a.s. on $\Gamma(L)$,*

$$\limsup_n E(\|\varepsilon_{n+1}\|^2 | \mathcal{F}_n) < \infty \quad \textit{and} \quad \liminf_n E(\|\varepsilon_{n+1}^{(r)}\| | \mathcal{F}_n) > 0; \tag{9}$$

then $P(\Gamma(L)) = 0$.

The proof is given in section 2.3.

2. Proofs of results.

2.1. Proof of Theorem 1.1 about the cyclic traps. A smooth "distance" to the stable manifold of the cycle. Let L be a cyclic trap (see the definition in section 1.1) and denote by $W^s(L)$ the stable manifold of L and by $W^r(L)$ the repulsive (called unstable in [1]) manifold of L.

By an adaptation of the Pemantle method [25] owing to Benaïm [1], under the assumptions of Theorem 1.1, there exist a neighborhood $V(L)$ of L and an $\mathbb{R}_+$-valued function η, defined on $V(L)$, vanishing on $W^s(L) \supset L$ and satisfying the following properties.

- On $V(L) \setminus W^s(L)$, η is $\mathcal{C}^2$ and

$$\forall x \in V(L) \setminus W^s(L)\ \|\nabla \eta(x)\| \geq c_1 > 0, \tag{10}$$

where $\nabla \eta$ is the gradient of η. And for $v \in \mathbb{R}^d$,

$$D\eta(x)v = \langle \nabla \eta(x), v \rangle.$$

- On $V(L) \bigcap W^s(L)$, there exists a "right derivative" that we also denote $D\eta(x)$ and which associates to any vector v of $\mathbb{R}^d$:

$$D\eta(x)v = \lim_{\substack{t \to 0 \\ t>0}} \frac{\eta(x+tv) - \eta(x)}{t}.$$

This "derivative" is continuous, convex, and positively homogeneous.

For all $x \in V(L) \bigcap W^s(L)$ there exist a linear subspace $E_2(x)$ of the repulsive directions and a linear subspace $E_1(x)$ of the stable directions satisfying $\mathbb{R}^d = E_1(x) \bigoplus E_2(x)$ and

$$D\eta(x)v \geq u_x^T v_{E_2(x)}, \tag{11}$$

where $v_{E_2(x)}$ is the component of v on $E_2(x)$ and u_x a vector of $E_2(x)$ such that $\|u_x\| = c_1 > 0$ (c_1 is the same constant as in (10)). Moreover, $x \to u_x$ is continuous and $x \to E_2(x)$ is a $\mathcal{C}^1$ map from $V(L) \bigcap W^s(L)$ into the Grassmann manifold of linear subspaces of the appropriate dimension (see [1]).

• There exist a $k > 0$ and a neighborhood of 0, U_0, satisfying for all v of U_0 and for all x of $V(L)$:

$$\eta(x+v) \geq \eta(x) + D\eta(x)v - k\|v\|^2. \tag{12}$$

• There exists a $\lambda > 0$ such that for all x of $V(L)$ and for all v of $\mathbb{R}^d$,

$$D\eta(x)(h(x)+v) = D\eta(x)h(x) + D\eta(x)v, \tag{13}$$

$$D\eta(x)h(x) \geq \lambda\eta(x). \tag{14}$$

Proof of Theorem 1.1. On $\Gamma(L)$, for n large enough, $Z_n \in V(L)$ and $\gamma_n h(Z_n) + c_n(\varepsilon_{n+1} + r_{n+1}) \in U_0$. Denote

$$\Gamma_N(L) = \Gamma(L) \bigcap \{\forall n \geq N\ ,\ Z_n \in V(L) \text{ and } \gamma_n h(Z_n) + c_n(\varepsilon_{n+1} + r_{n+1}) \in U_0\}.$$

On $\Gamma_N(L)$, for $n \geq N$,

$$Z_{n+1} = Z_n + \gamma_n h(Z_n) + c_n(\varepsilon_{n+1} + r_{n+1}),$$

and by (12)

$$\begin{aligned}\eta(Z_{n+1}) \geq \eta(Z_n) &+ D\eta(Z_n)(\gamma_n h(Z_n) + c_n(\varepsilon_{n+1} + r_{n+1})) \\ &- k\|\gamma_n h(Z_n) + c_n(\varepsilon_{n+1} + r_{n+1})\|^2.\end{aligned}$$

By (13), (14), and (11),

$$\begin{aligned}\eta(Z_{n+1}) \geq \eta(Z_n)(1+\lambda\gamma_n) &+ c_n D\eta(Z_n)\ \mathbb{I}_{\{\eta(Z_n)>0\}}\varepsilon_{n+1} \\ &+ c_n D\eta(Z_n)\ \mathbb{I}_{\{\eta(Z_n)>0\}} r_{n+1} \\ &+ c_n u_{Z_n}^T(\varepsilon_{n+1} + r_{n+1})_{E_2(Z_n)}\ \mathbb{I}_{\{\eta(Z_n)=0\}} \\ &- k\|\gamma_n h(Z_n) + c_n(\varepsilon_{n+1} + r_{n+1})\|^2.\end{aligned}$$

We also obtain

$$\begin{aligned}\eta(Z_{n+1}) \geq \eta(Z_n)(1+\lambda\gamma_n) &+ c_n D\eta(Z_n)\varepsilon_{n+1}\ \mathbb{I}_{\{\eta(Z_n)>0\}} \\ &+ c_n u_{Z_n}^T(\varepsilon_{n+1})_{E_2(Z_n)}\ \mathbb{I}_{\{\eta(Z_n)=0\}} \\ &+ c_n(D\eta(Z_n)\ \mathbb{I}_{\{\eta(Z_n)>0\}} r_{n+1} + u_{Z_n}^T(r_{n+1})_{E_2(Z_n)}\ \mathbb{I}_{\{\eta(Z_n)=0\}}) \\ &- k\|\gamma_n h(Z_n) + c_n(\varepsilon_{n+1} + r_{n+1})\|^2.\end{aligned}$$

Then,

$$\eta(Z_{n+1}) \geq \eta(Z_n)(1+\lambda\gamma_n) + c_n(e_{n+1} + \rho_{n+1}), \tag{15}$$

where

$$e_{n+1} = D\eta(Z_n)\varepsilon_{n+1}\ \mathbb{I}_{\{\eta(Z_n)>0\}} + u_{Z_n}^T(\varepsilon_{n+1})_{E_2(Z_n)}\ \mathbb{I}_{\{\eta(Z_n)=0\}}$$

and

$$\rho_{n+1} = D\eta(Z_n)\ \mathbb{I}_{\{\eta(Z_n)>0\}} r_{n+1} + u_{Z_n}^T (r_{n+1})_{E_2(Z_n)}\ \mathbb{I}_{\{\eta(Z_n)=0\}}$$

$$-c_n k \left\| \frac{\gamma_n}{c_n} h(Z_n) + \varepsilon_{n+1} + r_{n+1} \right\|^2 .$$

On $V(L)$, $h(z)$ and $D\eta(z)$ are bounded. (e_n) and (ρ_n) are two random real sequences that are $(\mathcal{F}_n)$-measurable. By (6) and (7) a.s., on $\Gamma_N(L)$, (e_n) is a noise satisfying

$$E(e_{n+1}|\mathcal{F}_n) = 0 \quad \text{and} \quad \limsup_n E(|e_{n+1}|^a \mid \mathcal{F}_n) < \infty,$$

and

$$\liminf_n E(e_{n+1}^2 \mid \mathcal{F}_n) > 0.$$

By (5), (6), and (7), a.s. on $\Gamma_N(L)$, $\Sigma \rho_n^2 < \infty$.

Theorem 4.1 given in Appendix 1 (section 4.1) applies to $\eta(Z_n) = \zeta_n$ and $P(\Gamma_N(L)) = 0$. Hence $\Gamma(L) = \bigcup \Gamma_N(L)$ and $\Gamma(L)$ has probability zero. □

2.2. Proofs of results about singular traps and repulsive regions. The basic tool in this section is an accompanying result of Benaïm and Hirsch [1], [10] stated in Appendix 2 (section 4.2). Roughly speaking, this result states that under some conditions, (Z_n) and a given solution of the ODE $(z(t))_{t\geq 0}$ have the same asymptotic behavior.

2.2.1. Proof of Theorem 1.2 (*d*-dimensional repulsive singular equilibria). Let V_{z^*} be a neighborhood of z^* where Dh is Lipschitz and such that any solution of the ODE starting from $V_{z^*} \setminus z^*$ leaves V_{z^*} within a finite time. As $Dh(z^*) = 0$, for all $z \in V_{z^*}$, $\|h(z)\| \leq \mathbf{C}\|z - z^*\|^2$.

Under the assumptions of Theorem 1.2, by Appendix 2, there exist on $\Gamma(L)$ a random vector $Y \in V_{z^*}$ and a positive random variable T such that the solution of the ODE $t \to z(t)$ starting from Y satisfies

$$\limsup_n \frac{1}{\log n} \log(\|Z_n - z(s_{n-1} - T)\|) \leq -\left(\frac{1}{2} - \frac{1}{a}\right),$$

where $s_{n-1} = \sum_{j=0}^{n-1} \gamma_j \simeq g \log n$.

Thus we cannot simultaneously have

$$\lim_{n\to\infty} \|Z_n - z^*\| = 0\ ,\ \lim_{n\to\infty} \|Z_n - z(s_{n-1} - T)\| = 0, \text{and } \limsup_{t\to\infty} \|z(t) - z^*\| \neq 0.$$

So if $\lim_{n\to\infty} \|Z_n - z^*\| = 0$, then $Y = z^*$ and $P(\{Y \neq z^*\} \bigcap \Gamma(z^*)) = 0$.

On $\{Y = z^*\} \bigcap \Gamma(z^*)$, $z(0) = z^*$ and since $a > 4$,

$$\limsup \frac{1}{\log n} \log \|Z_n - z^*\| \leq -\left(\frac{1}{2} - \frac{1}{a}\right) < -\frac{1}{4} \tag{16}$$

and

$$0 = Z_n - z^* + \sum_{j=n}^{\infty} c_j \left(\frac{\gamma_j}{c_j} h(Z_j) + \varepsilon_{j+1} + r_{j+1} \right).$$

Set $\Gamma_p = \{Y = z^*\} \bigcap \Gamma(z^*) \bigcap \{\omega\ ;\ Z_n(\omega) \in V_{z^*} \text{ for } n \geq p\}$.
A.s. on Γ_p,

$$\|h(Z_n)\|^2 = O(\|Z_n - z^*\|^4);$$

by (16), $\sum_{j=n}^{\infty} \|h(Z_n)\|^2 < \infty$ and

$$\sum_{j=n}^{\infty} \left\| \frac{\gamma_j}{c_j} h(Z_j) + r_{j+1} \right\|^2 < \infty.$$

By Theorem A of [6] about the distribution of a regressive series, under properties (8) of the noise, $P(\Gamma_p) = 0$, and seeing that $\{Y = z^*\} \bigcap \Gamma(z^*) = \bigcup \Gamma_p$, $P(\{Y = z^*\} \bigcap \Gamma(z^*)) = 0$. So $P(\Gamma(z^*)) = 0$. □

2.2.2. Proof of Proposition 1.3 (one-dimensional general singular equilibria). For any $c \in]0, 1[$, there exists a neighborhood W of z^* such that on W, h is $\mathcal{C}^1$ and

$$(1-c)|\alpha(z-z^*)^p| \leq |h(z)| \leq (1+c)|\alpha(z-z^*)^p|,$$
$$\text{for } z \neq z^* \text{ and } p \text{ even } \alpha h(z) > 0,$$
$$\text{for } z \neq z^* \text{ and } p \text{ odd, } \alpha(z-z^*)h(z) > 0.$$

Set $t \to z(t)$, a solution of the ODE, defined for $t \geq 0$, with an orbit in W and the initial condition $z(0) = y \neq z^*$:

$$\int_y^{z(t)} \frac{dz}{h(z)} = t.$$

Therefore, $z(t) \neq z^*$, $(z(t) - z^*)$ keeps the sign of $(y - z^*)$, and $h(z(t))$ keeps the sign of $\alpha(y - z^*)^p$.

Thus, for $(y - z^*)\alpha > 0$ and p even or for $\alpha > 0$ and p odd, $z(t)$ doesn't converge to z^*.

In the other cases the convergence is slow because

$$(|\alpha|(1-c)(p-1)t + y^{1-p})^{\frac{1}{1-p}} \leq |z(t) - z^*| \leq (|\alpha|(1+c)(p-1)t + y^{1-p})^{\frac{1}{1-p}}.$$

By a similar argument as in section 2.2.1, we obtain on $\Gamma(z^*)$ two random variables $Y \in W$ and $T \geq 0$ such that

$$\limsup_n \frac{1}{\log n} \log(\|Z_n - z(s_{n-1} - T)\|) \leq -\left(\frac{1}{2} - \frac{1}{a}\right) < -\frac{1}{2p},$$

with $z(0) = Y$. On $\Gamma(z^*) \bigcap \{Y = z^*\}$, $\limsup \frac{1}{\log n} \log(\|Z_n - z^*\|) < -\frac{1}{2p}$,

$$0 = Z_n - z^* + \sum_{j=n}^{\infty} c_j \left(\frac{\gamma_j}{c_j} h(Z_j) + \varepsilon_{j+1} + r_{j+1} \right).$$

For $\Gamma_q = \{Y = z^*\} \bigcap \Gamma(z^*) \bigcap \{\omega\ ;\ Z_n(\omega) \in W \text{ for } n \geq q\}$, a.s. on Γ_q,

$$h(Z_n)^2 = O((Z_n - z^*)^{2p}),$$

$\sum_{j=q}^{\infty} h(Z_j)^2 < \infty$, and $P(\{Y = z^*\} \bigcap \Gamma(z^*)) = 0$.

• If p is odd and if $\alpha > 0$, there is no solution of the ODE starting from a state distinct from z^* in W and converging to z^*. So $P(\{Y \neq z^*\} \bigcap \Gamma(z^*)) = 0$. Then z^* is avoided.

• In the other cases the convergence is possible, but for p even and for n large enough, $\alpha(Z_n - z^*) < 0$ on $\Gamma(z^*)$. In addition, a.s. on $\Gamma(z^*)$, for all $\delta < \frac{1}{2p}$,

$$|Z_n - z^*| \geq [(1-c)|\alpha|(p-1)s_{n-1} + |Y - z^*|^{1-p}]^{-\frac{1}{p-1}} + O(n^{-\delta}),$$

$$|Z_n - z^*| \leq [(1+c)|\alpha|(p-1)s_{n-1} + |Y - z^*|^{1-p}]^{-\frac{1}{p-1}} + O(n^{-\delta}).$$

Since c is arbitrary in $]0,1[$ and $s_{n-1} \simeq g \log n$, we have a.s. on $\Gamma(z^*)$,

$$|Z_n - z^*| \simeq [|\alpha|(p-1)g \log n]^{-\frac{1}{p-1}}.$$

This almost sure convergence rate to z^* is very slow. □

2.2.3. Some more general repulsive regions. In this section we give some properties of repulsive regions. First, we obtain a sufficient condition to claim that a region L is repulsive.

PROPOSITION 2.1. *Let L be a compact set connected and invariant for the ODE. If h is $\mathcal{C}^1$ on a neighborhood of L and if for all $z \in L$,*

$$\lambda_{min}(Dh(z) + (Dh(z))^T) > 0, \tag{17}$$

then L is repulsive in the sense of Definition 1.

Proof. Let L_r be a neighborhood of L where h is $\mathcal{C}^1$, and by the continuity of the spectrum of $Dh(z)$, there exists a constant $\lambda > 0$ such that

$$\inf_{z \in L} [\lambda_{min}(Dh(z) + (Dh(z))^T)] > 2\lambda,$$

and for r small enough,

$$\lambda_{min}(Dh(z) + (Dh(z))^T) \geq \lambda \quad \text{on } \overline{L_r}. \tag{18}$$

Set $(\varphi_t(x))_{t \geq 0}$ as the solution to the ODE starting from $x \in L_r \setminus L$ and

$$\tau = \inf\{t \ ; \ \varphi_t(x) \notin L_r\}.$$

Denote by $D\varphi$ the differential of φ with respect to x. For $t < \tau$ and for all $v \in \mathbb{R}^d$,

$$\begin{aligned} \frac{d}{dt}\|D\varphi_t(x)v\|^2 &= 2v^T(D\varphi_t(x))^T Dh(\varphi_t(x))D\varphi_t(x)v \\ &\geq \|D\varphi_t(x)v\|^2 \lambda_{min}(Dh(D\varphi_t(x)) + Dh(D\varphi_t(x))^T). \end{aligned}$$

And by (18),

$$\begin{aligned} \frac{d}{dt}\|D\varphi_t(x)\|^2 &\geq \lambda \|D\varphi_t(x)\|^2, \\ \|D\varphi_t(x)\| &\geq \exp\left(\frac{\lambda t}{2}\right). \end{aligned} \tag{19}$$

Set $z \in L$ such that $d(x, L) = \|z - x\|$. By the smoothness of φ,

$$\varphi_t(x) - \varphi_t(z) = D\varphi_t(x)(x - z) + o(x - z).$$

Then for $0 < t < \tau$ and for $r > 0$ small enough, by (19),

$$\|\varphi_t(x) - \varphi_t(z)\| \geq a\|x - z\|,$$

where $a > 1$. And for all integers $k > 0$,

$$\|\varphi_{kt}(x) - \varphi_{kt}(z)\| \geq a^k d(x, L),$$

as long as $\varphi_t(x) \in L_r$. This implies that $\varphi_t(x)$, $t > 0$, eventually leaves L_r; τ is bounded and L is repulsive. □

Now we still use the accompanying result of Appendix 2, so we assume that $\gamma_n = c_n = \frac{g}{n}$ with $g > 0$, and we denote by L_1 the closure of alpha and omega limit sets of points of L (as L is a compact connected chain recurrent set $L_1 \subseteq L$).

Then we obtain the following proposition.

PROPOSITION 2.2. *Consider the stochastic algorithm* (1) *under Assumptions* A1 *and the following assumptions:*

- *h is $\mathcal{C}^1$ on a neighborhood of L, and for all $z \in L_1$,*

$$\lambda_{min}(Dh(z) + (Dh(z))^T) \geq 0;$$

- *a.s. on $\Gamma(L)$,* (8) *is verified.*

Then, if L is a repulsive region, there exist a random variable T and a solution to the ODE, $t \to z(t)$ $(t \geq 0)$ with values in the invariant set L such that, a.s. on $\Gamma(L)$,

$$\limsup_n \frac{1}{\log n} \log(\|Z_n - z(s_{n-1} - T)\|) \leq -\left(\frac{1}{2} - \frac{1}{a}\right).$$

Proof. By Appendix 2, for $r > 0$ small enough, there exist a random variable $Y \in L_r$ and a positive random variable T such that

$$\limsup_n \frac{1}{\log n} \log(\|Z_n - z(s_{n-1} - T)\|) \leq -\left(\frac{1}{2} - \frac{1}{a}\right)$$

with $z(0) = Y$. But, as in section 2.2.1, L being repulsive, $P(Y \notin L) = 0$. □

So, if all solutions of the ODE included in L have limit sets in $L_1 \subseteq L$, the study of $\Gamma(L)$ reduces to the study of $\Gamma(L_1)$. For example, if L contains a finite number of equilibria, and if the solutions to the ODE included in L converge to one of these equilibria (it is the case of "equilibria cycles"; see [2], [5], and [8]), we reduce the study of $\Gamma(L)$ to the study of $\Gamma(z^*)$ for z^* equilibria of the ODE contained in L. When z^* is regular this case was studied in [6]; the case when we are not in the framework of singular traps is treated in Theorem 1.2.

Example. Let (Z_n) be defined in polar components (ρ_n, θ_n) by

$$\begin{cases} \rho_{n+1} = \rho_n + \frac{1}{n}[(\cos^2\theta_n - \frac{1}{2})g(\rho_n) + \xi_{n+1}], \\ \theta_{n+1} = \theta_n + \frac{1}{n}(\sin^2\theta_n + \xi_{n+1}), \end{cases}$$

and let $L = \{\rho = 1\}$. We assume that g is $\mathcal{C}^1$ on a neighborhood of 1, $g(1) = 0$, $g'(1) > 0$; (ξ_n) is a sequence of i.i.d. random variables with a uniform distribution on $[-\frac{1}{2}, \frac{1}{2}]$. For the ODE, $z_1 = (1, 0)$ and $z_2 = (1, \pi)$ are two equilibria, and the circle L is a repulsive region. The proposition applies and a.s.

$$(d(Z_n, L) \to 0) \Longrightarrow (Z_n \to z_1 \text{ or } Z_n \to z_2).$$

We have reduced the study of $\Gamma(L)$ to the case of repulsive regular traps z_1 and z_2. Thus $\Gamma(L)$ has probability zero.

2.3. Proof of Theorem 1.4.

2.3.1. Linearization of the ODE. Let P be the change of basis matrix from the canonical basis to $\mathcal{B}$. On $\mathcal{B}$, we denote the decomposition of a vector v as

$$v = \begin{pmatrix} v^+ \\ v^- \end{pmatrix} \quad \text{and}$$

$$\begin{pmatrix} g_+(v) \\ g_-(v) \end{pmatrix} = g(v) = Ph(v) = P \begin{pmatrix} h_+(v) \\ h_-(v) \end{pmatrix}.$$

For y in a neighborhood of L, there exists $x \in L$ such that $\|y - x\| = d(y, L)$ and

$$g_+(y^+, y^-) = J_+ y^+ + q_+(y^+, y^-),$$

where the function q_+ vanishes on x and its differential with respect to y^+, D_+q_+ also vanishes on x.

On a suitable neighborhood of x, $\|D_+q_+(y)\| \leq \mathbf{C}\ d(y, L)$. Besides, $g_-(x) = 0$, $D_+g_-(x) = 0$, and $\|D_+g_-(y)\| \leq \mathbf{C}d(y, L)$.

So the solution of the ODE starting from y satisfies

$$\begin{cases} z^+(t) = \exp(tJ_+)y^+ + Q_+(t, y), \\ z^-(t) = \exp(tJ_-(x))(y^- - x) + x + Q_-(t, y), \end{cases}$$

where $Q_\pm(t, x) = 0$, $D_+Q_\pm(t, x) = 0$, and $\|D_+Q_\pm(t, y)\| \leq \mathbf{C}\ d(y, L)$. Set

$$\begin{cases} f_+(y) = \exp J_+ y^+ + Q_+(1, y), \\ f_-(y) = \exp(J_-(x))(y^- - x) + x + Q_-(1, y). \end{cases}$$

Then, $f_-(x) = x$, $D_+f_-(x) = 0$, and $\|D_+f_-(y)\| \leq \mathbf{C}\ d(y, L)$.

Now we describe the linearization method (see Hartman [9, Corollary 5.2, p. 240]). We build recursively a sequence of functions (G_n) from K_a to K_r, by the relations

$$\begin{cases} G_0(x + v^-) = 0, \\ G_n(x + v^-) = G_{n-1}(x + v^-) \\ \qquad\qquad +(\exp J_+)^{-1}[G_{n-1}(x + (f_-(G_{n-1}(x + v^-), x + v^-) - x)) \\ \qquad\qquad\qquad\qquad -f_+(G_{n-1}(x + v^-), x + v^-)]. \end{cases}$$

For all $x \in L$, there exists an $r(x) > 0$ such that for $v^- \in K_s$ and $\|v^-\| \leq r(x)$, $G_n(x + v^-) \to G(x + v^-)$; if $z(0)$ is close enough to x and if $z(0)^+ = G(z(0)^-)$, the relation $z(t)^+ = G(z(t)^-)$ remains true for $t \leq t_0$, t_0 small enough. On a neighborhood of x,

$$[y^+ = G(y^-)] \Longrightarrow [h_+(y) = DG(y^-)h_-(y)].$$

Then, we infer that

• G is $\mathcal{C}^1$ in a ball $\mathbf{B}$ with a center x and a radius $r(x)$, $G(x) = 0$, $DG(x) = 0$, and DG is Lipschitz on $\mathbf{B}$;

• if $\|y - x\| \leq r(x)$, we set $v = y - x$ and

$$\varphi(y) = h_+(y) - DG(y^-)h_-(y).$$

Then $\varphi(G(y^-), y^-) = 0$ and $\varphi(y) = (J^+ + \Delta(y))(y^+ - G(y^-))$, with

$$\begin{aligned} \|\Delta(y)\| &\leq \sup_{0 \leq t \leq 1} \|D_+q_+(ty^+ + (1-t)G(y^-), y^-) \\ &\qquad\qquad -DG(y^-)D_+q_-(ty^+ + (1-t)G(y^-), y^-)\| \\ &\leq \mathbf{C}d(y, L), \end{aligned}$$

because $D_+q_\pm$ is Lipschitz.

We can cover L by a finite number of open balls, centered on $x_1, \ldots, x_q$, belonging to L, and with radius $r(x_1), \ldots, r(x_q)$. Set $r > 0$ such that

$$r < \min\{r(x_1), \ldots, r(x_q)\}.$$

Consider $L_r = \{y \in \mathbb{R}^d \ ; \ d(y^-, L) + \|y^+\| < r\}$. If $y \in L_r$, there exists j, $1 \leq j \leq q$, such that y^- is in the ball with the center x_j and the radius $r(x_j)$.

$$G_n(y^-) = G_n(x_j + (y^- - x_j)) \to G(y^-),$$

and, setting on L_r,

$$\varphi(y) = h_+(y) - DG(y^-)h_-(y), \tag{20}$$

then $\varphi(G(y^-), y^-) = 0$ and

$$\varphi(y) = (J^+ + \Delta(y))(y^+ - G(y^-)), \tag{21}$$

$$\text{where } \Delta(y) = O(d(y, L)). \tag{22}$$

Consequently, if $t \to z(t)$ is a solution of the ODE (3) which converges to L, setting $u(t) = z^+(t) - G(z^-(t))$, we obtain the repulsive ODE

$$\frac{du(t)}{dt} = (J_+ + \delta(t))u(t),$$

where $\|\delta(t)\| = O(d(z(t), L)) \to 0$.

2.3.2. Transformation of the algorithm. Set

$$Y_{n+1} = PZ_{n+1} = Y_n + \gamma_n g(Y_n) + c_n(P(\varepsilon_{n+1} + r_{n+1})).$$

As in [6, p. 401], by using a result of Lai and Wei [15], it is enough to prove Theorem 1.4 with the more restrictive condition that there exist three constants, $K < \infty$, $A > 0$, and $B < \infty$ such that

$$\Sigma\|r_{n+1}\|^2 \leq K, \ E(\|\varepsilon_{n+1}\|^2|\mathcal{F}_n) \leq B, \quad \text{and} \quad E(\|\varepsilon_{n+1}^{(r)}\||\mathcal{F}_n) \geq A > 0.$$

Set

$$\Gamma_p = \{\omega \ ; \ PZ_n(\omega) \in L_r \text{ for } n \geq p\} \bigcap \Gamma(L)$$

and $U_n = Y_n^+ - G(Y_n^-)$. On Γ_p, for $n \geq p$, by (20), (21), and (22), we obtain

$$U_{n+1} = U_n + \gamma_n(J_+ + \Delta_n)U_n + c_n(e_{n+1} + \rho_{n+1}),$$

with $\lim_n \Delta_n = 0$, $e_{n+1} = (P\varepsilon_{n+1})^+ - DG(Y_n^-)(P\varepsilon_{n+1})^-$, and (ρ_n) a sequence adapted with respect to $\mathbb{F}$ and verifying $\Sigma\|\rho_n\|^2 < \infty$ a.s. on $\Gamma(L)$.

$$E(\|DG(Y_n^-)(P\varepsilon_{n+1})^-\|^2|\mathcal{F}_n) \leq \|DG(Y_n^-)\|^2 E(\|(P\varepsilon_{n+1})^-\|^2|\mathcal{F}_n).$$

On $\Gamma(L)$, by the properties of G, $DG(Y_n^-)$ vanishes to 0, and by the assumptions on the noise, a.s. on $\Gamma(L)$,

$$\limsup_n E(\|e_{n+1}\|^2|\mathcal{F}_n) < \infty$$

and

$$\liminf_n E(\|e_{n+1}\| \ |\mathcal{F}_n) = \liminf_n E(\|(P\varepsilon_{n+1})^+\||\mathcal{F}_n) > 0.$$

So Proposition 4 in [6] applies to (U_n) and $P(\Gamma(L)) = 0$. □

3. An application of Theorem 1.4.

3.1. About a principal component analysis algorithm. A scatter plot of N data points of $\mathbb{R}^d$ $(x_1, \dots, x_N)$ with empirical mean zero $(\frac{1}{N}\sum_{k=1}^N x_k = 0)$ and with empirical covariance matrix $C = \frac{1}{N}\sum_{k=1}^N x_k (x_k)^T$ is analyzed.

We search for the j-dimensional principal subspace $(1 \le j \le d)$, where the projection of this scatter plot is the best.

If $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_d > 0$ are the eigenvalues of C, such a subspace has an orthonormal basis $\{V_1, \dots, V_j\}$, where V_i is an eigenvector associated with λ_i $(1 \le i \le j)$.

For j unitary and orthogonal $\mathbb{R}^d$-vectors $a_1, \dots, a_j$, we denote by $[a_1 \dots a_j]$ the $j \times d$ matrix whose column vectors are these vectors; $\mathcal{M}$ is the set of such orthonormalized matrices.

A PCA algorithm is intended for converging to $[V_1 \dots V_j] \in \mathcal{M}$, where V_i is an eigenvector associated with λ_i $(1 \le i \le j)$.

In the framework of the study of linear neural network, Oja [12], [20] naturally built a recursive PCA algorithm by suggesting the following method.

Assume that (X_n) is an i.i.d. sequence of points picked in the scatter plot with a uniform distribution. Then, if $Z_n = [Z_n^1 \dots Z_n^j]$ is the approximation of $[V_1 \dots V_j]$ at the nth step, it is natural, according to the neuronal intuition, to set

$$\tilde{Z}_{n+1} = Z_n + \gamma_n X_{n+1} (X_{n+1})^T Z_n, \tag{23}$$

$$Z_{n+1} = S_{n+1} \tilde{Z}_{n+1}, \tag{24}$$

where S_{n+1} is a matrix which depends on $\tilde{Z}_{n+1}$ and performs the Gram–Schmidt orthonormalization on the columns of $\tilde{Z}_{n+1}$.

(γ_n) is a decreasing nonnegative deterministic sequence such as

$$\sum_{n \ge 0} \gamma_n = \infty \quad \text{and} \quad \sum_{n \ge 0} \gamma_n^2 < \infty.$$

We claim the following result.

THEOREM 3.1. *If C is nonsingular, for any j, $1 \le j \le d$, $[Z_n^1 \dots Z_n^j]$ converges a.s. to a stochastic orthonormalized matrix $[W^1 \dots W^j]$, where $W^k \in \mathcal{S}_k$ for $1 \le k \le j$, $\mathcal{S}_k$ being the unit sphere of the eigensubspace associated to the eigenvalue λ_k.*

3.2. Previous results.

• The same algorithm was proposed by Benzécri [4], and later by Lebart [17]. They use algebraic arguments. The case where $j = 1$ is entirely treated, but the generalization is just given roughly (see [4]). With similar arguments Monnez [19] considers some analogous algorithms. He obtains the almost sure convergence of (Z_n^1) to the subspace associated to λ_1 and, only in the case of the distinct eigenvalues, the almost sure convergence of (Z_n) to $[\pm V^1, \dots, \pm V^d]$.

• Independently, in the framework of the study of the neural behavior, Oja, Karhunen, Sanger, Hornick, Kuan, Becker, Williams, Ogawa, Wangviwattana, and others [11], [12], [20], [21], [22], [23], [24], [26], [28] consider similar algorithms.

For the sequence (Z_n^1), we can find a partial proof in [12] when the eigenvalues have a unit multiplicity, and Delyon [7] treats the case of the multiple eigenvalues but without proving that the traps are avoided.

In the general case, and for the distinct and positive eigenvalues, Oja studies the ODE associated to (23), (24) and determines its asymptotically stable zeros. Then,

thanks to many simulations, he claims that the algorithm (23), (24) converges to these zeros, but without a theoretical proof.

We have chosen the Oja algorithm which better agrees with our framework and seems quite natural. It is a very suitable illustration of our previous study. Our method applies to other similar algorithms aforementioned.

3.3. Proof of Theorem 3.1. By (23) we obtain an algorithm which satisfies

$$\begin{aligned} Z_n &= [Z_n^1 \dots Z_n^j] \in \mathcal{M}, \text{ and for } 1 \le k \le j, \\ Z_{n+1}^k &= Z_n^k + \gamma_n \Bigg[(C_{n+1} - ((Z_n^k)^T C_{n+1} Z_n^k)) Z_n^k \\ &\qquad -2 \sum_{i=1}^{k-1} ((Z_n^k)^T C_{n+1} Z_n^i) Z_n^i \Bigg] + O(\gamma_n^2), \end{aligned} \tag{25}$$

with $C_{n+1} = X_{n+1} X_{n+1}^T$. We denote by $\mathcal{F}_n$ the σ-fields generated by $X_1, \dots, X_n$. Let $\mathcal{B} = \{V^1, \dots, V^d\}$ be an orthonormalized basis of eigenvectors of C, such that V^k is a eigenvector associated to λ_k, and let $\mathcal{S}_k$ be the unit sphere of the eigensubspace F_k associated to λ_k. The algorithm (25) satisfies

$$Z_{n+1} = Z_n + \gamma_n (h(Z_n) + \varepsilon_{n+1} + r_{n+1}), \tag{26}$$

where

$$\begin{aligned} h(Z) &= [h_1(Z) \dots h_j(Z)] \text{ with} \\ h_1(Z) &= CZ^1 - ((Z^1)^T C Z^1) Z^1, \end{aligned} \tag{27}$$

and for $1 \le k \le j$,

$$h_k(Z) = CZ^k - ((Z^k)^T C Z^k) Z^k - 2 \sum_{i=1}^{k-1} ((Z^i)^T C Z^k) Z^i, \tag{28}$$

$$\begin{aligned} \varepsilon_{n+1} &= [\varepsilon_{n+1}^1 \dots \varepsilon_{n+1}^j] \text{ with} \\ \varepsilon_{n+1}^1 &= (C_{n+1} - C) Z_n^1 - ((Z_n^1)^T (C_{n+1} - C) Z_n^1) Z_n^1, \end{aligned} \tag{29}$$

and for $1 \le k \le j$,

$$\begin{aligned} \varepsilon_{n+1}^k &= (C_{n+1} - C) Z_n^k - ((Z_n^k)^T (C_{n+1} - C) Z_n^k) Z_n^k \\ &\qquad -2 \sum_{i=1}^{k-1} ((Z_n^i)^T (C_{n+1} - C) Z_n^k) Z_n^i, \end{aligned} \tag{30}$$

$$\|r_{n+1}\| = O(\gamma_n). \tag{31}$$

(ε_n) and (r_n) are two bounded random sequences that are $(\mathcal{F}_n)$-measurable and satisfy (6).

Step 1. *Possible limit sets.*

LEMMA 3.2. *The possible limit sets L of solutions of the algorithm* (26) *are subsets of $\mathcal{M}$ such that, i being a map from $\{1, \dots, j\}$ to $\{1, \dots, d\}$, $z = [z^1 \dots z^j] \in L$ is characterized by: for $1 \le k \le j$, $z^k \in G_{i(k)}$, where $G_{i(k)}$ is a compact connected set contained in $\mathcal{S}_{i(k)}$, the unit sphere of the eigensubspace $F_{i(k)}$ associated to $\lambda_{i(k)}$.*

Therefore, for $z \in L$, $z^T C z = \operatorname{diag}[\lambda_{i(1)} \dots \lambda_{i(j)}]$.

Proof of Lemma 1. Following [2], we know that L is a compact connected subset of $\mathcal{M}$, invariant and "chain-recurrent" for the ODE.

First, remark that $h_1(x) = 0$(h_1 is given in (27)) if and only if x is a unit eigenvector of C. So the connected components of $\{h_1 = 0\}$ are contained in a unit sphere $\mathcal{S}_i$ $(1 \leq i \leq d)$. If λ_i has unit multiplicity, $\mathcal{S}_i = \{V^i, -V^i\}$ and the corresponding connected components of $\{h_1 = 0\}$ are $\{V^i\}$ or $\{-V^i\}$.

For $x \in \mathbb{R}^d$, set $V(x) = \frac{\exp(\|x\|^2)}{x^T(I+C)x}$ (it is the Lyapounov function used by Delyon in [7]).

$$\nabla V(x) = \frac{\exp(\|x\|^2)}{(x^T(I+C)x)^2}[(x^T(I+C)x)2x - 2(I+C)x],$$
$$\nabla V(x) = -2\frac{V(x)}{x^T(I+C)x}[h_1(x) - (\|x\|^2 - 1)x].$$

(a) For $j = 1$, if $z_0 \in L$, $z(t) \in L$ for all t and

$$\frac{dV(z^1(t))}{dt} = -2\frac{V(z^1(t))}{z^1(t)^T(I+C)z^1(t)}\left\langle h_1(z^1(t)), \frac{dz^1(t)}{dt}\right\rangle$$
$$= -2\frac{V(z^1(t))}{z^1(t)^T(I+C)z^1(t)}\|h_1(z^1(t))\|^2.$$

$V(z^1(t))$ is nonnegative and decreasing. Then, $V(z^1(t))$ converges to V_∞ when $t \to \infty$ and $h_1(z^1(t)) \to 0$. $(z^1(t))$ has a limit set which is a connected component of $\{h_1 = 0\}$ contained in $\mathcal{S}_{i(1)}$, one of the unit spheres described above.

But since L is chain-recurrent, for all t, $z^1(t) \in \mathcal{S}_{i(1)}$ and $L \subseteq \mathcal{S}_{i(1)}$.

(b) For $j > 1$, let $z(t) = [z^1(t), \ldots, z^j(t)]$ be a solution of the ODE with values in L. Assume that there exist $(k-1)$ integers $i(1), \ldots, i(k-1)$ with $z^\ell(t) \in \mathcal{S}_{i(\ell)}$, $1 \leq \ell \leq k-1$. Then, since $z(t) \in \mathcal{M}$, $\|z^k(t)\| = 1$, $z^k(t)$ is orthogonal to $z^i(t)$ if $i \neq k$ and

$$\frac{dz^k(t)}{dt} = h_k(z(t)) = h_1(z^k(t)),$$
$$\frac{dV(z^k(t))}{dt} = -2\frac{V(z^k(t))}{z^k(t)^T(I+C)z^k(t)}\left\langle h_1(z^k(t)), \frac{dz^k(t)}{dt}\right\rangle$$
$$= -2\frac{V(z^k(t))}{z^k(t)^T(I+C)z^k(t)}\|h_1(z^k(t))\|^2.$$

As in (a), we infer that for a given integer $i(k)$, $z^k(t) \in \mathcal{S}_{i(k)}$ for all t.

We have proved by recurrence the characterization of L. □

Remark. By Lemma 3.2, the *traps* of the algorithm (26) correspond to $(i(1), \ldots, i(j)) \neq (1, \ldots, j)$. We shall prove that these traps are regular if any eigenvalue has unit multiplicity while they are homogeneous flat traps in the sense of Theorem 1.4 for multiple eigenvalues.

Step 2. *Proof of* $(i(1) \ldots, i(j)) = (1, \ldots, j)$.

Now we prove the following result.

PROPOSITION 3.3. *For C nonsingular with possibly multiple eigenvalues, the Oja algorithm* (26) *satisfies*

$$Z_n^k \to \mathcal{S}_k \quad for\ 1 \leq k \leq j,$$

where $\mathcal{S}_k$ is the unit sphere of the eigensubspace F_k associated to λ_k.

Proof. Let L be defined as in Lemma 3.2. We proceed by recurrence to prove that $(i(1)\dots,i(j)) = (1,\dots,j)$.

(1) First we consider $j = 1$ and

$$Z^1_{n+1} = Z^1_n + \gamma_n(h_1(Z^1_n) + \varepsilon^1_{n+1} + r^1_{n+1}). \tag{32}$$

Assume that $i(1) = p > 1$ and $\mathcal{S}_p \neq \mathcal{S}_1$.

Let $\mathcal{B}$ be the orthonormalized basis described at the beginning of this section. For $Z^1 \in \mathbb{R}^d$, set $Z^1 = \sum_{i=1}^d z^{1i}V^i$ and $h_1(Z^1) = \sum_{i=1}^d h_1^i(Z^1)V^i$; $h_1^i(Z^1) = \lambda_i z^{1i} - (\sum_{j=1}^d \lambda_j (z^{1j})^2)z^{1i}$.

Determination of Dh_1.

$$\frac{\partial h_1^i}{\partial z^{1i}} = \lambda_i - \sum_{n=i}^d \lambda_n (z^{1n})^2 - 2\lambda_i (z^{1i})^2, \tag{33}$$

$$\frac{\partial h_1^i}{\partial z^{1k}} = -2\lambda_k z^{1k} z^{1i} \quad \text{if } k \neq i. \tag{34}$$

If $\mathcal{S}_p$, with a π-dimension, is generated by $V^{j(1)},\dots,V^{j(\pi)}$, for $p \geq 1$:

$$\text{if } z^1 \in \mathcal{S}_p \ , \ z^1 = \sum_{k=1}^{\pi} \langle z^1, V^{j(k)}\rangle V^{j(k)}.$$

And for $1 \leq k \leq \pi$ and $1 \leq k' \leq \pi$,

$$\begin{cases} \frac{\partial h_1^{j(k)}}{\partial z^{1j(k)}}(z^1) = -2\lambda_p \langle z^1, V^{j(k)}\rangle^2, \\ \frac{\partial h_1^{j(k)}}{\partial z^{1j(k')}}(z^1) = -2\lambda_p \langle z^1, V^{j(k)}\rangle \langle z^1, V^{j(k')}\rangle. \end{cases}$$

By denoting

$$J(z^1) = \begin{bmatrix} \langle z^1, V^{j(1)}\rangle \\ \cdots \\ \cdots \\ \langle z^1, V^{j(\pi)}\rangle \end{bmatrix} [\langle z^1, V^{j(1)}\rangle \dots \langle z^1, V^{j(\pi)}\rangle],$$

$J(z^1)$ is a matrix $\pi \times \pi$ that is symmetric, semidefinite positive, and for $z^1 \in \mathcal{S}_p$,

$$Dh_1(z^1) = \begin{pmatrix} J_+ & 0 & 0 \\ 0 & -2\lambda_p J(z^1) & 0 \\ 0 & 0 & J_- \end{pmatrix},$$

with

$$J_+ = \begin{pmatrix} \lambda_1 - \lambda_p & 0 & . & . \\ 0 & \lambda_2 - \lambda_p & 0 & . \\ \cdots & . & . & . \\ \cdots & . & . & \lambda_{j(1)-1} - \lambda_p \end{pmatrix}$$

and

$$J_- = \begin{pmatrix} \lambda_{j(\pi)+1} - \lambda_p & 0 & . & . \\ 0 & \lambda_{j(\pi)+2} - \lambda_p & 0 & . \\ \cdots & . & . & . \\ \cdots & . & . & \lambda_d - \lambda_p \end{pmatrix}.$$

We are in the situation of Theorem 1.4 with $L \subseteq \mathcal{S}_p$, a nonattractive and homogeneous set of equilibria. Indeed, for all $z^1 \in \mathcal{S}_p$, K_r is generated by $\{V^1, \dots, V^{p-1}\}$, $K_a = (K_r)^\perp$, J_+ is the matrix as aforesaid, and

$$J_-(z^1) = \begin{pmatrix} -2\lambda_p J(z^1) & 0 \\ 0 & J_- \end{pmatrix}.$$

To apply Theorem 1.4, it is enough to prove that the noise excitation is sufficient in the repulsive direction, i.e., with the same notations as in section 1.3, that

$$\liminf_n E(\|(\varepsilon_{n+1})^{(r)}\||\mathcal{F}_n) > 0.$$

For $1 \le k \le p-1$, a.s. on $\Gamma(L)$,

$$E([\langle \varepsilon^1_{n+1}, V^k\rangle - \langle (C_{n+1} - C)Z^1_n, V^k\rangle]^2|\mathcal{F}_n) \le K\langle Z^1_n, V^k\rangle^2,$$

and a.s., $\lim_n \langle Z^1_n, V^k\rangle = 0$. Hence

$$\lim_n E(\langle \varepsilon^1_{n+1}, V^k\rangle^2|\mathcal{F}_n) = \lim_n E([(V^k)^T X_{n+1} X^T_{n+1} Z^1_n - V^k C Z^1_n]^2|\mathcal{F}_n)$$

$$\lim_n E(\langle \varepsilon^1_{n+1}, V^k\rangle^2|\mathcal{F}_n) = E[\langle V^k, X_1\rangle^2 \langle X_1, V^p\rangle^2] > 0 \tag{35}$$

and

$$E(\|(\varepsilon^1_{n+1})^{(r)}\|^2|\mathcal{F}_n) = \sum_{k=1}^{p-1} \lim_n E(\langle \varepsilon^1_{n+1}, V^k\rangle^2|\mathcal{F}_n) > 0.$$

And since (ε_n) has a conditional moment with an order larger than 2, the excitation conditions are checked. $P(\Gamma(L)) = 0$ and the proposition is proved for $j = 1$.

(2) Assume that for $1 \le k \le j-1$, $i(k) = k$, and that $i(j) = p > j$ and $\mathcal{S}_p \ne \mathcal{S}_j$. We have to prove that $P(\Gamma(L)) = 0$.

Consider $\mathcal{M}$ provided with $\mathcal{B}_1 = \{M^{11} \dots M^{jd}\}$, where

$$M^{11} = [V^1 0 \dots 0], \dots, M^{1d} = [V^d 0 \dots 0], \qquad M^{21} = [0 V^1 0 \dots 0], \dots, M^{jd} = [0 \dots V^d].$$

Setting $Z = [Z^1 \dots Z^j] \in \mathcal{M}$, we denote $Z = \sum_{k=1}^j \sum_{m=1}^d z^{km} M^{km}$ and $Z^k = \sum_{m=1}^d z^{km} V^m$. If h^m_k designates the m-component of h_k in $\mathcal{B}$, we have

$$h^m_k(Z) = \lambda_m - \left(\sum_{n=1}^d \lambda_n (z^{kn})^2\right) z^{km} - 2\sum_{i=1}^{k-1} \left(\sum_{n=1}^d \lambda_n z^{kn} z^{in}\right) z^{im}.$$

Hence

$$\frac{\partial h^m_k}{\partial z^{km}}(Z) = \lambda_m - \sum_{n=1}^d \lambda_n (z^{kn})^2 - 2\lambda_m (z^{km})^2 - 2\sum_{i=1}^{k-1} \lambda_m (z^{im})^2, \tag{36}$$

$$\frac{\partial h^m_k}{\partial z^{kn}}(Z) = -2\lambda_n z^{kn} z^{km} - 2\sum_{i=1}^{k-1} \lambda_n z^{in} z^{im} \quad \text{if } n \ne m, \tag{37}$$

$$\frac{\partial h^m_k}{\partial z^{im}}(Z) = -2\lambda_m z^{km} z^{im} - 2\sum_{n=1}^{d} \lambda_n z^{kn} z^{in} \quad \text{for } 1 \le i \le k-1, \tag{38}$$

$$\frac{\partial h^m_k}{\partial z^{in}}(Z) = -2\lambda_n z^{kn} z^{im} \quad \text{if } n \ne m \text{ and } 1 \le i \le k-1, \tag{39}$$

$$\frac{\partial h^m_k}{\partial z^{in}}(Z) = 0 \quad \text{if } i > k. \tag{40}$$

For $z = [z^1 \dots z^{j-1} z^j]$, with $z^k \in \mathcal{S}_k$ for $1 \le k < j$ and $z^j \in \mathcal{S}_p$,

$$Dh(z) = \begin{pmatrix} \Delta_1 & . & . & . & . \\ \cdots & \Delta_2 & . & . & . \\ \cdots & . & . & . & . \\ \cdots & . & . & . & . \\ \cdots & . & . & . & \Delta_j \end{pmatrix}.$$

Denote by $0 < \alpha_\rho < \cdots < \alpha_1$ the distinct eigenvalues of C. By using the previous calculations, for $1 \le k \le j$, if $\lambda_k = \alpha_r$ and if $\nu(r)$ is the multiplicity order of α_r,

$$\Delta_k = \begin{pmatrix} -\lambda_1 - \lambda_k & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & -\alpha_{r-1} - \lambda_k & . & . & . & . \\ . & . & . & -2\lambda_k J(z^k) & . & . & . \\ . & . & . & . & \alpha_{r+1} - \lambda_k & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & \lambda_d - \lambda_k \end{pmatrix},$$

where $J(z^k)$ is a symmetric, semidefinite positive matrix $\nu(r) \times \nu(r)$ as $J(z^1)$.

So, for $1 \le k \le j-1$, all the matrices Δ_k have nonpositive eigenvalues. For $k = j$, only the first block of Δ_j has some positive eigenvalues and it doesn't depend on z.

$\{[z^1 \dots z^j] \in \mathcal{M};\ z^1 \in \mathcal{S}_1, \dots, z^{j-1} \in \mathcal{S}_{j-1}, z^j \in \mathcal{S}_p\}$ is a compact connected set of equilibria, homogeneous and nonattractive. K_r is generated by $\{M^{jj}, M^{j(j+1)}, \dots, M^{j(p-1)}\}$ and has a $(p-j)$-dimension, and $K_a = (K_r)^\perp$. About the noise excitation, $(\varepsilon_{n+1})^{(r)} = \sum_{k=j+1}^{p-1} M^{jk} (\varepsilon_{n+1})^T M^{jk}$, and by a similar calculation to (35), we obtain

$$E(\|(\varepsilon_{n+1})^{(r)}\|^2 | \mathcal{F}_n) = \sum_{k=j}^{p-1} E(\langle V^k, X_1 \rangle^2 \langle X_1, V^p \rangle^2) > 0.$$

So, by Theorem 1.4, $P(\Gamma(L)) = 0$. □

Step 3. *Proof of the a.s. convergence to a solution of the PCA*. Knowing that $d(Z_n^k, S_k) \to 0$, we have to show that, when λ_k is multiple, Z_n^k converges to a random vector of $\mathcal{S}_k$.

(a) For a Q matrix $d \times d$ and $Qx \ne 0$, if we set $\varphi(x) = \frac{Qx}{\|Qx\|}$, then

$$D\varphi(x) = \frac{Q}{\|Qx\|} - \frac{(Qx)^T (Qx) Q}{\|Qx\|^3}. \tag{41}$$

(b) For $z = [z^1 \dots z^d]$ such that, for $1 \le j \le d$, $d(z^j, \mathcal{S}_j) \le r$ and for r small enough, if P_j is the orthogonal projection on F_j, $[P_1(z^1) \dots P_d(z^d)]$ is nonsingular. We can perform a Gram–Schmidt orthonormalization on $[P_1(z^1) \dots P_d(z^d)]$:

$$z = [z^1 \dots z^d] \to [\varphi_1(z^1) \dots \varphi_d(z^1, \dots, z^d)] = \varphi(z).$$

We shall prove that $D\varphi(z) h(z) = 0$.

$$\begin{aligned} P_1 h_1(z^1) &= P_1 C z^1 - (z^1)^T C z^1 P_1 z^1 \\ &= (\lambda_1 - (z^1)^T C z^1) P_1 z^1, \end{aligned}$$

and by (41), with $Q = P_1$, $\frac{\partial}{\partial z^1} \varphi_1(z^1) h_1(z^1) = 0$.

Assume that for $1 \leq k < j$, $(z^1, \ldots, z^k) \to \varphi_k(z^1, \ldots, z^k)$ satisfies, for all $i \leq k$,

$$\frac{\partial}{\partial z^i}\varphi_k(z^1, \ldots, z^k)h_i(z^1, \ldots, z^i) = 0.$$

Let I be the set of i such as $\lambda_i = \lambda_j$ and $i < j$. Denote $P = P_j$ and $\varphi_i = \varphi(z^1, \ldots, z^i)$; then

$$\varphi_j(z^1, \ldots, z^j) = \frac{Pz^j - \sum_{i\in I}\langle \varphi_i, Pz^j\rangle \varphi_i}{\|Pz^j - \sum_{i\in I}\langle \varphi_i, Pz^j\rangle \varphi_i\|}.$$

Setting $Qx = Px - \sum_{i\in I}\langle Px, \varphi_i\rangle\varphi_i$ and using (41), we have

$$\varphi_j(z^1, \ldots, z^j) = \frac{Qz^j}{\|Qz^j\|},$$
$$\frac{\partial}{\partial z^j}\varphi_j(z^1, \ldots, z^j) = \frac{Q}{\|Qz^j\|} - \frac{Qz^j(Qz^j)^T P}{\|Qz^j\|^2}.$$

Now $Qx = Px - \sum_{i\in I}\langle Px, \varphi_i\rangle\varphi_i$, and for $k \in I$,

$$\begin{cases} Pz^k = \displaystyle\sum_{i\in I}\langle Pz^k, \varphi_i\rangle\varphi_i, \\ Qz^k = 0, \end{cases}$$

and for j, $Qz^j = Pz^j - \sum_{i\in I}\langle Pz^j, \varphi_i\rangle\varphi_i \in F_j$. Hence

$$\begin{aligned} Qh_j(z^1, \ldots, z^j) &= Q\left(Cz^j - (z^j)^T Cz^j z^j - \sum_{i<j}(z^i)^T C(z^j)z^i\right) \\ &= (\lambda_j - (z^j)^T Cz^j)z^j \end{aligned}$$

and

$$\frac{\partial}{\partial z^j}\varphi_j(z^1, \ldots, z^j)h_j(z^1, \ldots, z^j) = 0.$$

For $i < j$, $\frac{\partial}{\partial z^i}\varphi_j(z^1, \ldots, z^j) = \sum_{k\in I,\ k\geq i} A_k(\frac{\partial}{\partial z^i}\varphi_k(z^1, \ldots, z^k))$, where A_k is a matrix dependent on $z^1, \ldots, z^j$, but it is unnecessary to make it explicit because the recurrence assumption sets

$$\frac{\partial}{\partial z^i}\varphi_j(z^1, \ldots, z^j)h_i(z^1, \ldots, z^i) = 0.$$

We also proved that $D\varphi(z)h(z) = 0$.

(c) For all j,

$$\varphi_j(Z^1_{n+1}, \ldots, Z^j_{n+1}) - \varphi_j(Z^1_n, \ldots, Z^j_n) = \gamma_n \sum_{k=1}^{j}\frac{\partial}{\partial z^k}\varphi_j(Z^1_n, \ldots, Z^j_n)\varepsilon^k_{n+1} + \rho_{n+1};$$

the regressive series, with the general term $\gamma_n \sum_{k=1}^{j}\frac{\partial}{\partial z^k}\varphi_j(Z^1_n, \ldots, Z^j_n)\varepsilon^k_{n+1}$, converges, as does $\Sigma\rho_{n+1}$. It implies the a.s. convergence of $\varphi(Z_n)$ to $(W^1, \ldots, W^j)$, and thus by Proposition 3.3, of (Z_n) to $(W^1, \ldots, W^j)$. □

4. Appendices.

4.1. Appendix 1. Stochastic iterative inequality.

THEOREM 4.1 (of repulsion). *Set (ζ_n), a positive sequence, defined on a probability space $(\Omega, \mathcal{A}, \mathcal{P})$ provided with an increasing sequence of σ-fields $(\mathcal{F}_n)_{n\geq 0}$ and satisfying a.s. on an $\mathcal{F}_\infty$-measurable set $\Omega_0 \subseteq \Omega$,*

$$\zeta_{n+1} \geq (1+\lambda\gamma_n)\zeta_n + c_n(\varepsilon_{n+1} + r_{n+1}), \tag{42}$$

with $\lambda > 0$. We assume that

• *(ε_n) and (r_n) are real random sequences, defined on $(\Omega, \mathcal{A}, \mathcal{P})$, $(\mathcal{F}_n)$-measurable and satisfying (6) a.s on Ω_0, and*

$$\limsup_n E(\varepsilon_{n+1}^2|\mathcal{F}_n) < \infty \quad \textit{and} \quad \liminf_n E(|\varepsilon_{n+1}||\mathcal{F}_n) > 0; \tag{43}$$

• *(γ_n) and (c_n) are real nonnegative deterministic sequences satisfying (2) and (5).*

Then the event $\Omega_0 \bigcap \{\omega\ ;\ \zeta_n(\omega) \to 0\}$ has probability zero.

Proof. Step 1. It is sufficient to prove Theorem 4.1, assuming that a.s. on Ω,

$$E(\varepsilon_{n+1}|\mathcal{F}_n) = 0 \quad \text{and} \quad \limsup_n E(\varepsilon_{n+1}^2|\ \mathcal{F}_n) < \mathbf{C} < \infty, \tag{44}$$

$$\liminf_n E(|\varepsilon_{n+1}||\mathcal{F}_n) > \mathbf{C} > 0 \quad \text{and} \quad \Sigma r_n^2 < \mathbf{C} < \infty. \tag{45}$$

See [15] and [6, p. 401] for the way to achieve this simplification.

Step 2. Set $G = \Omega_0 \bigcap \{\omega\ ;\ \zeta_n(\omega) \to 0\}$. On G,

$$\zeta_{n+1} = (1+\lambda\gamma_n)\zeta_n + c_n(\varepsilon_{n+1} + r_{n+1}) + c_n U_{n+1}$$

with (U_n) a random positive real sequence that is $(\mathcal{F}_n)$-measurable.

Set

$$G_N = P(G|\mathcal{F}_N);$$

G is $\mathcal{F}_\infty$-measurable and (G_N) converges to $\mathbb{I}_G$ a.s. and in L^p for all $p \geq 1$. Thus

$$E((G_N - \mathbb{I}_G)^2) \to 0. \tag{46}$$

Setting $\beta_n = \prod_{j=0}^n (1+\gamma_j\lambda)$, we have

$$\zeta_n = \beta_n \left[\zeta_0 + \sum_{j=0}^n \frac{c_j}{\beta_j}(\varepsilon_{j+1} + r_{j+1} + U_{j+1}) \right].$$

Step 3. Prove that

$$E\left(\mathbb{I}_G \sum_{n=N}^\infty \frac{c_n}{\beta_n}\|U_n\| \right) = o\left(\left(\sum_{n=N}^\infty \frac{c_n^2}{\beta_n^2} \right)^{\frac{1}{2}} \right). \tag{47}$$

We have

$$\zeta_{N+n} = \beta_{N+n} \left[\frac{\zeta_N}{\beta_N} + \sum_{j=N}^{n+N} \frac{c_j}{\beta_j}(\varepsilon_{j+1} + r_{j+1} + U_{j+1}) \right],$$

and on G, $\zeta_{n+N} \to 0$ and $\beta_{N+n} \to \infty$. Hence on G,

$$-\frac{\zeta_N}{\beta_N} = \sum_{j=N}^{\infty} \frac{c_j}{\beta_j}(\varepsilon_{j+1} + r_{j+1} + U_{j+1})$$

and

$$\sum_{j=N}^{\infty} \frac{c_j}{\beta_j} U_{j+1} = -\frac{\zeta_N}{\beta_N} - \sum_{j=N}^{\infty} \frac{c_j}{\beta_j}(\varepsilon_{j+1} + r_{j+1}).$$

Thus

$$E\left[\mathbb{I}_G\left(\sum_{n=N}^{\infty} \frac{c_n}{\beta_n} U_{n+1}\right)\right] \le -E\left[\mathbb{I}_G \sum_{n=N}^{\infty} \frac{c_n}{\beta_n}(\varepsilon_{n+1} + r_{n+1})\right]$$

$$\le E\left[(G_N - \mathbb{I}_G) \sum_{n=N}^{\infty} \frac{c_n}{\beta_n}(\varepsilon_{n+1} + r_{n+1})\right] - E\left[G_N \sum_{n=N}^{\infty} \frac{c_n}{\beta_n}(\varepsilon_{n+1} + r_{n+1})\right]$$

$$\le [E((G_N - \mathbb{I}_G)^2)]^{\frac{1}{2}} \left[E\left(\sum_{n=N}^{\infty} \frac{c_n}{\beta_n}(\varepsilon_{n+1} + r_{n+1})^2\right)\right]^{\frac{1}{2}} + E\left(\sum_{n=N}^{\infty} \frac{c_n}{\beta_n}|r_{n+1}|\right)$$

$$\le \mathbf{C}[E((G_N - \mathbb{I}_G)^2)]^{\frac{1}{2}} \left(\sum_{n=N}^{\infty} \frac{c_n^2}{\beta_n^2}\right)^{\frac{1}{2}} + \left(E\left(\sum_{n=N}^{\infty} r_{n+1}^2\right)\right)^{\frac{1}{2}} \left(\sum_{n=N}^{\infty} \frac{c_n^2}{\beta_n^2}\right)^{\frac{1}{2}}$$

$$= o\left(\left(\sum_{n=N}^{\infty} \frac{c_n^2}{\beta_n^2}\right)^{\frac{1}{2}}\right).$$

So, by (44) and (46), (47) is proved.

We now use a theorem, proved in [6] (Theorem A in the appendix), that extends the framework of a result of Levy [18], Lai and Wei [15], and others about the sum of a convergent regressive series.

Here Theorem A of [6] applies with $H = G$, $c_n = \frac{c_n}{\beta_n}$, and $R_n = U_n$. So $P(G) = 0$. □

4.2. Appendix 2. An accompanying result. We use a version of a theorem of Benaïm and Hirsch [2], [10] which specifies the conditions under which any trajectory solution to (1) is a.s. asymptotic to a forward trajectory solution to ODE (3). From the study of Benaïm and Hirsch, we can claim the following result.

THEOREM 4.2. *Set* (1) *under Assumption* A1 *and the following assumptions:*

- *h is $\mathcal{C}^1$ on a neighborhood of L and*

$$\inf_{x \in L_1} \frac{1}{2} \lambda_{min}(Dh(x) + (Dh(x))^T) = -\mu, \tag{48}$$

where $L_1 \subset L$ is the closure of alpha and omega limit sets of points in L;

- *a.s. on $\Gamma(L)$, for $a > 2$,*

$$\limsup_n E(\|\varepsilon_{n+1}\|^a | \mathcal{F}_n) < \infty; \tag{49}$$

- *$\gamma_n = c_n = \frac{g}{n}$ $(g > 0)$ and $\frac{1}{2g} - \frac{1}{ag} > \max(0, \mu)$.*

Then, for any neighborhood W of L, there exist a random variable T and a random vector $Y \in W$, defined on $\Gamma(L)$, such that, a.s. on $\Gamma(L)$,

$$\limsup_n \frac{1}{s_{n-1}} \log(\|Z_n - z(s_{n-1} - T)\|) \leq -\left(\frac{1}{2g} - \frac{1}{ag}\right), \tag{50}$$

where $s_n = \sum_{j=0}^{n} \gamma_j$ and the function $t \to z(t)$ is the solution to the ODE (3), defined on $\mathbb{R}_+$, such that $z(0) = Y$, and whose orbit is in W.

Acknowledgments. I wish to thank Marie Duflo for her personal encouragements. I am particularly grateful to her for many remarks and fruitful ideas on the topics in this paper. I also thank the referee and Michel Benaïm for their judicious comments to ameliorate this work.

REFERENCES

[1] M. Benaïm and M. W. Hirsch, *Dynamics of Morse-Smale urn processes*, Ergodic Theory Dynam. Systems, 15 (1995), pp. 1005–1030.
[2] M. Benaïm, *A dynamical system approach to stochastic approximation*, SIAM J. Control Optim., 34 (1996), pp. 437–472.
[3] A. Benveniste, M. Métivier, and P. Priouret, *Algorithmes adaptatifs et approximations stochastiques*, Masson, Paris, 1987.
[4] M. Benzécri, *Approximation stochastique dans une algèbre normée non commutative*, Bull. Soc. Math. France, 97 (1969), pp. 225–241.
[5] O. Brandière, *Autour des pièges des algorithmes stochastiques*, Thèse, Université de Marne-la-Vallée, Noisy-le-Grand, France, 1996.
[6] O. Brandière and M. Duflo, *Les algorithmes stochastiques contournent-ils les pièges?*, Ann. Inst. H. Poincaré, 32 (1996), pp. 395–427.
[7] B. Delyon, *General convergence result on stochastic approximation*, IEEE Trans. Automat. Control, 41 (1996), pp. 1245–1255.
[8] M. Duflo, *Algorithmes Stochastiques*, Math. Appl. 23, Springer-Verlag, Heidelberg, 1996.
[9] P. Hartman, *Ordinary Differential Equations*, 2nd ed, Wiley, New York, 1982.
[10] M. W. Hirsch, *Asymptotic phase, shadowing and reaction diffusion*, in Control Theory, Dynamical Systems and Geometry of Dynamics, K. D. Elworthy and W. N. Everitts, eds., Marcel Dekker, New York, 1993, pp. 87–99.
[11] K. Hornik and C. M. Kuan, *Convergence analysis of local feature extraction algorithms*, Neural Networks, 5 (1992), pp. 229–240.
[12] J. Karhunen and E. Oja, *On stochastic approximation of the eigenvectors and eigenvalues of expectation of a random matrix*, J. Math. Anal. Appl., 106 (1995), pp. 69–84.
[13] G. D. Kersting, *Some results on the asymptotic behavior of the Robbins-Monro procedure*, Bull. Int. Stat. Inst., 47 (1977), pp. 327–335.
[14] H. J. Kushner and D. S. Clark, *Stochastic Approximation for Constrained and Uncontrained Systems*, Appl. Math. Sci. 26, Springer-Verlag, New York, 1978.
[15] T. Z. Lai and C. Z. Wei, *A note on martingale difference sequences satisfying the local Marcinkiewicz-Zygmund condition*, Bull. Inst. Math. Acad. Sinica, 11 (1983), pp. 1–13.
[16] V. A. Lazarev, *Convergence of stochastic approximation procedures in case of regression equation with several roots*, Problemy Peredachi Informatsii, 28 (1992), pp. 75–88 (in Russian).
[17] L. Lebart, *Sur les calculs impliqués par la description de certains grands tableaux*, Annales de l'INSEE 22-23, 1931.
[18] P. Levy, *Sur les séries dont les termes sont des variables éventuellement indépendantes*, Stud. Math., 3 (1931), pp. 119–155.
[19] J. M. Monnez, *Convergence d'un processus d'approximation stochastique en analyse factorielle*, Internal Report, Institut de Statistique de l'Université de Paris, 38, 1994, pp. 37–56.
[20] E. Oja, *Simplified neuron model as a principal component analyser*, J. Math. Biol., 15 (1982), pp. 267–273.
[21] E. Oja, *Subspace Methods of Pattern Recognition*, Research Studies Press and Wiley, Letchworth, England, 1983.
[22] E. Oja, C. Y. Suen, and L. Xu, *Modified Hebbian learning for curve and surface fitting*, Neural Networks, 5 (1992), pp. 441–457.

[23] E. Oja, *Principal components, minor components, and linear neural networks*, Neural Networks, 5 (1992), pp. 927–935.
[24] E. Oja, H. Ogawa, and J. Wangviwattana, *Principal component analysis by homogeneous neural networks, Parts* I *and* II, IEICE Trans. Inf. Syst., E75-D, 3 (1992), pp. 316–382.
[25] R. Pemantle, *Nonconvergence to unstable points in urn models and stochastic approximations*, Ann. Probab., 18 (1990), pp. 698–712.
[26] T. D. Sanger, *Optimal unsupervised learning in a single-layer linear feedforward network*, Neural Networks, 2 (1989), pp. 459–473.
[27] C. Z. Wei, *Martingale transforms with non-atomic limits and stochastic approximation*, Probab. Theory Related Fields, 95 (1993), pp. 103–114.
[28] R. Williams, *Feature Discovery through Error-Correcting Learning*, Technical Report 8501, University of California, Institute of Cognitive Science, San Diego, CA, 1985.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 4, pp. 1315–1330, July 1998

011

PARAMETER ESTIMATION IN ACOUSTIC MEDIA USING THE ADJOINT METHOD*

ELENA M. FERNÁNDEZ-BERDAGUER†

Abstract. We present an algorithm based on the adjoint method to locate points that provide an approximate solution to the parameter estimation problem for the acoustic model. The parameter belongs to infinite-dimensional sets. We prove the existence of the directional derivative of the solution with respect to the parameter in some dense set of directions of the set of parameters. This derivative is the solution of a differential boundary value problem. The adjoint problem is presented. A result on the convergence of the iterations is proved.

Key words. inverse problems, direct algorithms, wave equations

AMS subject classifications. 86A22, 8608

PII. S0363012996299326

1. Introduction.

Algorithms to estimate coefficients in the acoustic problem have been discussed in detail in many works, but mostly in the case in which the coefficients are piecewise continuous; see, for example, [3, 4, 8, 9]. Here we consider more general sets of parameters: parameters in an infinite-dimensional space, of which the piecewise continuous case appears as a particular case.

Parameter estimation problems of the type considered here are typical of seismic exploration; they also appear in many different fields such as remote sensing, imaging, and nondestructive testing. The direct problem models the propagation of waves in acoustic media, and in the case of seismic exploration the parameter represents the bulk modulus of the media.

We will approach the estimation problem in the usual least squares sense, that is, as an optimization one. It consists of the minimization of a quadratic functional involving the observed data and a functional of the traces of the model.

Minimization problems where the model is a partial differential equation can be approached by discretizing the differential problem first and then applying optimization algorithms to the discretized version, or applying optimization algorithms to the continuous problem and discretizing as a final step. The first approach was the most commonly employed; among the many works using it we can mention as examples [2, 3, 4, 16, 17]. The second type of algorithm was used in applications in geophysics, mostly without rigorous proofs (see [19, 21, 20]), and more recently in several works, for example, [6, 15, 13, 8, 9].

We present an algorithm of the second type based on the adjoint method for a hyperbolic problem with absorbing boundary conditions. The method is used without a priori information on the parametrization of the function to be identified. Here we present the mathematical details of the differential equations involved, and the operators and conditions for the convergence of the method. The highlights of the method are the sensitivity equations (2.11)–(2.13) and the adjoint integral (3.10) that allow us to calculate the different steps by solving equations similar to the one for the direct model but with a different source function. Those features make the algorithm

*Received by the editors February 28, 1996; accepted for publication (in revised form) June 16, 1997; published electronically May 22, 1998.
http://www.siam.org/journals/sicon/36-4/29932.html
†CONICET, Instituto de Cálculo FCEyN y Facultad de Ingenieria de la Universidad de Buenos Aires, Ciudad Universitaria, Pabellón II, Buenos Aires, Argentina (efernan@ulises.ic.fcen.uba.ar).

a fast one as needed for the application that we have in mind, which even in the layered case has a large number of parameters.

A similar algorithm that employs the sensitivity equations was used previously by us in the case of piecewise constant coefficients (this amounts to parameters in a finite-dimensional space) with very good numerical results (see [7, 8, 9]).

Basic questions when dealing with parameter estimation are those of existence and uniqueness of solutions of the inverse problem. Conditions for the existence of solutions to the estimation problem are known only for particular sets of parameters, mainly in the case where the parameter belongs to a space of finite dimension, as, for example, sets of piecewise constant functions. In general, inverse problems are ill posed since the solutions are very sensitive to changes in input data. An excellent discussion of the aforementioned problems can be found in the book by Banks and Kunisch [5]. Also, the identifiability problem, for systems different from those treated here, is clearly presented in the most recent articles by Giudici [10, 11]. For the particular problem presented in this work we refer the reader to the early publication of Bamberger, Chavent, and Lailly [1].

Here we do not address those problems; instead we present an algorithm which under conditions on the intervening operators converges to a critical point of the cost functional.

There is a vast body of research in the field of algorithms to identify parameters in problems governed by partial differential equations. A good update in the case of elliptic problems is the paper by Kunisch [14] and the bibliography therein.

The algorithm presented here does not include a regularization term. Our proofs on the differentiability with respect to the parameter suggest the use of an L^m-regularization term (with $m > 2$) instead of the usual L^2-norm term.

We use the customary notation as follows. Let $\Omega = (0,1)^n, n = 1, 2,$ or $3, \Gamma = \partial\Omega$. For all nonnegative integers s, let $(H^s(\Omega), \|\cdot\|_s)$ denote the usual Sobolev space. In particular, $H^0(\Omega) = L^2(\Omega)$ and $\|\|_0$ is the usual L^2-norm, with inner product

$$(v, w) = \int_\Omega v\, w\, dx.$$

For notational convenience, let

$$[v, w] = \int_\Gamma v\, w\, d\sigma$$

denote the inner product on Γ, with the associated norm denoted by $|\cdot|_0 = ([\cdot,\cdot])^{1/2}$, $d\sigma$ being the surface measure on Γ.

The following lemma about a trace inequality, for the particular domain with which we are dealing, will be needed. Its proof is given in the Appendix.

LEMMA 1.1. *Let $f \in H^1(\Omega)$ and m an integer such that $m \geq 2$. Then*

$$\|f\|^m_{L^m(\Gamma)} \leq C(\|f\|^m_{L^m(\Omega)} + \|f\|^{(m-1)}_{L^{2(m-1)}(\Omega)} \|f'\|_{L^2(\Omega)}).$$

Above, and in what follows, C stands for generic constants which may be different at different places.

The paper is organized as follows. In section 2 we present the direct model and state a theorem on the regularity of its solution. We set forth the inverse problem using the output least squares criterion. Also, we prove the existence of a continuous first

Gâteaux derivative of the solution of the direct model with respect to the parameter. We present the sensitivity equations for the derivatives. In section 3 we find an expression for the adjoint of the observation operator derivative. Finally, in section 4, we present an algorithm that under suitable hypotheses allows us to locate critical points of the cost functional.

2. The problem, the direct model, and the Gâteaux derivatives. The problem is to estimate the parameter $K(x)$ in the usual model for wave propagation in acoustic media:

$$\frac{1}{K}p_{tt}(K,x,t) - \nabla\cdot\left(\frac{1}{\rho}\nabla p(K,x,t)\right) = \frac{1}{K}S(x,t), \quad x\in\Omega, \quad t\in[0,T], \tag{2.1}$$

with initial condition

$$p(K,x,t=0) = p_t(K,x,t=0) = 0, \quad x\in\Omega, \tag{2.2}$$

and boundary condition

$$-\frac{1}{\rho(x)}\frac{\partial p(K,x,t)}{\partial \nu} = \frac{1}{\alpha(K,x)}p_t(K,x,t), \quad x\in\Gamma, \quad t\in[0,T], \tag{2.3}$$

where $\alpha(K,x) = \sqrt{K(x)\rho(x)}$. In seismic exploration, $p(K,x,t)$ represents the pressure, $K(x)$ the bulk modulus, and $\rho(x)$ the density of the medium. The function $S(x,t)$ in the right-hand side of (2.1) is the external source function, and equation (2.3) is an absorbing boundary condition so that waves arriving at Γ normally are absorbed completely; see [12, 18, 22].

We assume that $\rho(x)$ and $K(x)$ are measurable functions satisfying the following constraints that must be imposed from physical considerations:

$$\begin{array}{lccccc} \text{(i)} & \rho_* & \le & \rho(x) & \le & \rho^*, \\ \text{(ii)} & K_* & \le & K(x) & \le & K^*. \end{array} \tag{2.4}$$

Also, for the boundary condition (2.3) to make sense, we have to restrict K in a neighborhood of the boundary. For $0 < a_i < b_i < 1$, $i = 1,\dots,n$, let $S_n = \prod_{i=1}^n (a_i, b_i)$, $\tilde{S}_n = \Omega \setminus S_n$, and consider the space $\mathcal{A}$ spanned by a fixed set of continuous functions $g_1,\dots,g_\ell$ defined on $\tilde{S}_n$. A weaker condition could be imposed but we want $\mathcal{P}$ to be contained in a complete space. A similar condition must be imposed on ρ. The set of admissible parameters, denoted by $\mathcal{P}$, is given by

$$\mathcal{P} = \{K \text{ is measurable in } \Omega,\ K|_{\tilde{S}_n}\in\mathcal{A},\ K_*\le K(x)\le K^*,\ \text{a.e. in } \Omega\}.$$

The weak form of the direct problem (2.1)–(2.3) is: find $p(K,x,t)\in H^1(\Omega)$ such that

$$\left(\frac{1}{K}p_{tt}, v\right) + \left(\frac{1}{\rho}\nabla p, \nabla v\right) + \left[\frac{1}{\alpha}p_t, v\right] = \left(\frac{1}{K}S, v\right), \qquad v\in H^1(\Omega), \quad t\in[0,T]. \tag{2.5}$$

Let $\mathcal{V} = W^{1,\infty}(0,T,L^2(\Omega)) \cap L^\infty(0,T,H^1(\Omega))$. The following result on the solutions of (2.1)–(2.3) is proved in [9], so its proof is omitted here.

THEOREM 2.1. *Assume that ρ and K are measurable functions satisfying* (2.4)*, and that for an integer $q\ge 1$,*

$$\frac{\partial^{q-1}S}{\partial t^{q-1}} \in L^2(0,T,L^2(\Omega)).$$

Then the solution $p(K,x,t)$ *of* (2.5) *is such that*

$$\frac{\partial^{q-1}p(K,\cdot,\cdot)}{\partial t^{q-1}} \in \mathcal{V} \tag{2.6}$$

and satisfies the estimates

$$\left\|\frac{\partial^q p(K,\cdot,\cdot)}{\partial t^q}\right\|_{L^\infty(0,T,L^2(\Omega))} + \left\|\frac{\partial^{q-1}p(K,\cdot,\cdot)}{\partial t^{q-1}}\right\|_{L^\infty(0,T,H^1(\Omega))} \leq C\left\|\frac{\partial^{q-1}S}{\partial t^{q-1}}\right\|_{L^2(0,T,L^2(\Omega))},$$

where the positive constant C *depends only on the total time* T *and the upper and lower bounds for* $\rho(x)$ *and* $K(x)$.

We endow the set of admissible parameters with the $L^m(\Omega)\cap L^m(\Gamma)$-topology with $m=6$. The reason for the choice of the topology becomes clear in the proof of the differentiability in Theorem 2.4. We could have chosen the $H^1(\Omega)$-topology; but even in the case $n=1$ that choice overrides the space of piecewise constant functions. The choice of $\mathcal{P}$ as a subset of $H^1(\Omega)$ would have simplified many proofs. Also, the L^∞-topology is possible, but most regularization theorems require the parameter to belong to a reflexive space.

In [7] and [9] we have considered as the set of admissible parameters

$$\mathcal{Q}^N = \left\{ K:\ K(x) = \sum_{i=1}^{N} k_i \chi_{\Omega_i}(x), \quad K_* \leq k_i \leq K^* \right\};$$

there we estimated the parameters k_i while N and Ω_i were assumed to be fixed; in those cases we used the $L^2(\Omega)$-topology in the set of parameters. That was possible because the functions on the basis of the set of parameters were finite in number, then uniformly bounded.

Thus we endow the set $\mathcal{P}$ with the topology given by the norm

$$|||K|||_m = \|K\|_{L^m(\Omega)} + \|K\|_{L^m(\Gamma)}.$$

We assume that the observations are recorded at points x_{r_i} inside Ω, $1\leq i\leq N_r$, and denote by $p^{obs}\in L^2(0,T;R^{N_r})$ the vector of observations $p_i(t)=p^{obs}(x_{r_i},t)$, $1\leq i\leq N_r$, $t\in[0,T]$. We define the model for the observations as follows. For a given measurable set $E\subset R^n$, $|E|$ denotes the usual Lebesgue measure of E. Let B_i be the ball with center x_{r_i} and radius a, where a is small enough so that $B_i\cap B_j=\emptyset$ for $i\neq j$ and $B_i\subset\Omega$, $1\leq i,j\leq N_r$. The observation map $\Phi:\mathcal{P}\mapsto L^2(0,T;R^{N_r})$ is defined as the following nonlinear map:

$$\Phi(K)(t) = (\Phi_i(K)(t))_{1\leq i\leq N_r} \tag{2.7}$$

with

$$\Phi_i(K)(t) = \frac{1}{|B_i|}\int_{B_i} p(K,x,t)\,dx. \tag{2.8}$$

In the one-dimensional case, if the source function S satisfies $\partial S/\partial t\in L^2(0,T;L^2(\Omega))$, we can use point evaluation $p(K,x_{r_i},t)$, $i=1,\dots,N_r$, as the model for the observations because (Theorem 2.1) $p(K,\cdot,t)\in H^1(\Omega)$; therefore, for $n=1$, the function $p(K,\cdot,t)$ is absolutely continuous.

The estimation problem will be solved using the output least squares criterion. The problem is to minimize the functional

$$J(K) = \frac{1}{2} \left\| \Phi(K) - p^{obs} \right\|^2_{L^2(0,T;\mathbb{R}^{N_r})} \tag{2.9}$$

over the set $\mathcal{P}$.

A version of the next theorem, about the continuity of the solution of the direct model with respect to the parameter, was proved in [9] for the set $\mathcal{P}$ endowed with the $L^2(\Omega)$-topology. The difference is that here we show under additional hypotheses that the function p is Lipschitz continuous in the parameter K. A brief proof is given in the Appendix.

THEOREM 2.2. *Let $K_1(x), K_2(x) \in \mathcal{P}$, and $p(K_1, x, t)$, and $p(K_2, x, t)$ be the corresponding solutions of the direct problem* (2.1)–(2.3) *for $K = K_1$ and $K = K_2$, respectively. Set*

$$d(K_1, K_2, x, t) = p(K_1, x, t) - p(K_2, x, t),$$

and assume that for an integer $q \geq 1$,

$$\frac{\partial^{q+1} S}{\partial t^{q+1}}, \frac{\partial^{q+2} S}{\partial t^{q+2}} \in L^2(0, T, L^2(\Omega)).$$

Assume also that

(i) $\operatorname{supp}(K_1 - K_2) \cap \operatorname{supp} S(., t) = \emptyset$

or

(ii) $S \in L^2(0, T; L^3(\Omega))$.

Then $d(K_1, K_2, \cdot, \cdot) \in \mathcal{V}$. Moreover, d satisfies the estimate

$$\begin{aligned} &\left\| \frac{\partial^{q+1} d(K_1, K_2)}{\partial t^{q+1}} \right\|_{L^\infty(0,T,L^2(\Omega))} + \left\| \frac{\partial^q d(K_1, K_2)}{\partial t^q} \right\|_{L^\infty(0,T,H^1(\Omega))} \\ &\qquad \leq C \; |||K_1 - K_2|||_m , \end{aligned} \tag{2.10}$$

where $C = C(\Omega, T, \rho_*, \rho^*, K_*, K^*) C(S)$ with

$$C(S) = \left\| \frac{\partial^{q+1} S}{\partial t^{q+1}} \right\|_{L^2(0,T,L^2(\Omega))} + \left\| \frac{\partial^{q+2} S}{\partial t^{q+2}} \right\|_{L^2(0,T,L^2(\Omega))}$$

in case (i), *and*

$$C(S) = \|S\|_{L^2(0,T,L^3(\Omega))} + \left\| \frac{\partial^{q+1} S}{\partial t^{q+1}} \right\|_{L^2(0,T,L^2(\Omega))} + \left\| \frac{\partial^{q+2} S}{\partial t^{q+2}} \right\|_{L^2(0,T,L^2(\Omega))}$$

in case (ii).

COROLLARY 2.3. *The mapping*

$$\begin{array}{ccc} K & \longrightarrow & p(K, \cdot, \cdot), \\ \mathcal{P} & \longrightarrow & \mathcal{V} \end{array}$$

from the set $\mathcal{P}$ of parameters equipped with the $|||\cdot|||_m$-topology into $\mathcal{V}$ is continuous.

The following theorem will allow us to establish, in some cases, the existence of solutions of the output least squares problem.

THEOREM 2.4. *Let $\mathcal{Q} \subset \mathcal{P}$ be a compact set on $L^m(\Omega) \cap L^m(\Gamma)$; then the problem*

$$\text{minimize } J(K) \quad \text{over } \mathcal{Q}$$

has a solution.

Now we will prove that the function p has a Gâteaux derivative with respect to the parameter K for every $K \in \mathcal{P}$ and that this derivative is the solution of a differential problem.

Let $\mathcal{E}_0$ be a small neighborhood of the surface; for example, for $\eta \in (0,1)$, η small, we can choose the set $\mathcal{E}_0$ to be $[0,\eta]$ for $n=1$, $[0,\eta]\times[0,1]$ for $n=2$, and $[0,\eta]\times[0,1]\times[0,1]$ for $n=3$.

We choose as the space of perturbations for the parameter K one of the following subspaces of $L^m(\Omega) \cap L^m(\Gamma)$:

$$\mathcal{W}_0 = \{\delta K \in L^m(\Omega)\cap L^m(\Gamma),\ \delta K|_{\tilde{S}_n} \in \mathcal{A} \text{ and } \delta K(x) = 0 \text{ for } x \in \mathcal{E}_0\},$$

or

$$\mathcal{W} = \{\delta K \in L^m(\Omega)\cap L^m(\Gamma),\ \delta K|_{\tilde{S}_n} \in \mathcal{A}\}.$$

Since we can assume that we know the parameter K close to the surface, the condition $\delta K = 0$ in $\mathcal{E}_0$ is not an important restriction from the point of view of the application. We denote by Λ the quantity

$$\Lambda = \sum_{i=0}^{2} \left\| \frac{\partial^i S}{\partial t^i} \right\|_{L^2(0,T;L^2(\Omega))}.$$

The following theorem about the Gâteaux derivative of p with respect to K assumes different hypotheses depending on whether the perturbations belong to the set $\mathcal{W}_0$.

THEOREM 2.5. *Assume that the source function $S(x,t)$ is such that the quantity Λ is finite and that the parameter $K(x)$ in (1.1)–(1.3) belongs to the set $\mathcal{P}$. Assume also that either* (i) $\operatorname{supp} S \subset \mathcal{E}_0$ *and* $\delta K \in \mathcal{W}_0$ *or* (ii) $S \in L^2(0,T;L^3(\Omega))$ *and* $\delta K \in \mathcal{W}$. *Then the weak form of the problem*

$$\begin{aligned} &\frac{1}{K}(D'(K)\delta K)_{tt}(x,t) - \nabla\cdot\left(\frac{1}{\rho}\nabla D'(K)\delta K\right)(x,t) \\ &\qquad = \frac{\delta K(x)}{K^2(x)}\Big(p_{tt}(K,x,t) - S(x,t)\Big), \quad x\in\Omega, \quad t\in[0,T], \end{aligned} \tag{2.11}$$

with initial conditions

$$D'(K)\delta K(x,t=0) = (D'(K)\delta K)_t(x,t=0) = 0, \quad x\in\Omega, \tag{2.12}$$

and boundary conditions

$$-\frac{1}{\rho(x)}\frac{\partial D'(K)\delta K(x,t)}{\partial \nu} = \frac{(D'(K)\delta K)_t(x,t)}{\alpha(K,x)} - \frac{\delta K(x) p_t(K,x,t)}{2\sqrt{\rho(x)}K^{3/2}(x)}, \quad x\in\Gamma,\ t\in[0,T], \tag{2.13}$$

has a solution which satisfies the estimate

$$\|(D'(K)\delta K)_t\|^2_{L^\infty(0,T,L^2(\Omega))} + \|D'(K)\delta K\|^2_{L^\infty(0,T,H^1(\Omega))} \le C(\|\delta K\|^2_{L^m(\Omega)} + \|\delta K\|^2_{L^m(\Gamma)}), \tag{2.14}$$

where $C = C\,(K_*, K^*, \rho_*, \rho^*, T, \Omega, \Lambda)\, C(S)$, *with* $C(S)$ *as in Theorem* 2.2.

Moreover, if $\delta K \in L^\infty(\Omega)$, *the function* $D'(K)\delta K$ *is the Gâteaux derivative of* p (*the solution of* (2.5)) *with respect to* K *in the direction of* δK; *also, the limit*

$$\lim_{\lambda \to 0} \frac{p(K + \lambda \delta K, \cdot, \cdot) - p(K, \cdot, \cdot)}{\lambda}$$

exists in $\mathcal{V}$ *and is equal to* $D'(K)\delta K$.

Proof. The proof of (2.14) is a simplified version of the proof of the differentiability of p; thus we omit the former and we prove that for $D'(K)\delta K$ solution of (2.11)–(2.13), the function

$$\phi(K + \lambda \delta K, K, x, t) = D'(K)\delta K(x,t) - \frac{p(K + \lambda \delta K, x, t) - p(K, x, t)}{\lambda} \tag{2.15}$$

tends to zero in $\mathcal{V}$ as λ tends to zero.

We prove the theorem in case (i); case (ii) has extra terms that are treated as in the proof of Theorem 2.2.

First we prove the theorem for $\delta K \in L^\infty(\Omega)$. Let $\lambda_o = \frac{1}{2} K_* / \|\delta K\|_\infty$ and $|\lambda| \le \lambda_0$.

Now we use (2.5) for $K + \lambda \delta K$ and K, and the weak form of (2.11)–(2.13), in order to write the weak form for the differential problem for the function ϕ in (2.15).

Find $\phi \in H^1(\Omega)$ such that

$$\begin{aligned}
&\left(\frac{1}{K}\phi_{tt}, v\right) + \left(\frac{1}{\rho}\nabla\phi, \nabla v\right) + \left[\frac{1}{\alpha}\phi_t, v\right] \\
&\quad = \left(\frac{\lambda \delta K^2 p_{tt}(K + \lambda \delta K)}{K^2(K + \lambda \delta K)}, v\right) - \left(\frac{\delta K d_{tt}(K + \lambda \delta K, K)}{K^2}, v\right) \\
&\quad\quad + \left[\frac{\delta K}{\sqrt{\rho K}}\left(\frac{1}{K + \lambda\, K\, \delta K + \sqrt{K^2 + \lambda\, \delta K}} - \frac{1}{2K}\right) p_t(K + \lambda \delta K), v\right] \\
&\quad\quad - \left[\frac{\delta K d_t(K + \lambda \delta K, K)}{2\sqrt{\rho} K^{3/2}}, v\right], \qquad v \in H^1(\Omega), \quad t \in [0, T].
\end{aligned} \tag{2.16}$$

In the above formula, d is the difference function of Theorem 2.2.

In (2.16), choose $v = \phi_t$ to obtain

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\left(\left\|\frac{\phi_t}{K^{1/2}}\right|_0^2 + \left\|\frac{\nabla\phi}{\rho^{1/2}}\right|_0^2\right) + \left[\frac{1}{\alpha(K)}\phi_t, \phi_t\right] \\
&\quad = \left(\frac{\lambda \delta K^2 p_{tt}(K + \lambda \delta K)^2}{K^2(K + \lambda \delta K)^2}, \phi_t\right) - \left(\frac{\delta K d_{tt}(K + \lambda \delta K, K)}{K^2}, \phi_t\right) \\
&\quad\quad + \left[\frac{\delta K}{\sqrt{\rho K}}\left(\frac{1}{K + \lambda\, K\, \delta K + \sqrt{K^2 + \lambda \delta K}} - \frac{1}{2K}\right) p_t(K + \lambda \delta K), \phi_t\right] \\
&\quad\quad - \left[\frac{\delta K d_t(K + \lambda \delta K, K)}{2\sqrt{\rho} K^{3/2}}, \phi_t\right], \qquad v \in H^1(\Omega), \quad t \in [0, T].
\end{aligned} \tag{2.17}$$

Adding the inequality

$$\frac{1}{2}\frac{d}{dt}\left(\left\|\frac{\phi}{\rho^{1/2}}\right\|_0^2\right) \le \left\|\frac{\phi}{\rho^{1/2}}\right\|_0^2 + C\left\|\frac{\phi_t}{K^{1/2}}\right\|_0^2$$

to (2.17), integrating the resulting inequality from 0 to t, and using the initial conditions, we have

$$\begin{aligned}
&\left\|\frac{\phi_t}{K^{1/2}}\right\|_0^2(t)+\left\|\frac{\phi}{\rho^{1/2}}\right\|_1^2(t)+2\int_0^t\left[\frac{1}{\alpha(K)}\phi_t,\phi_t\right](\tau)\,d\tau\\
&\le\left|\int_0^t\left(\frac{\lambda\delta K^2p_{tt}(K+\lambda\delta K)}{K^2(K+\lambda\delta K)^2},\phi_t\right)(\tau)\,d\tau\right|+\left|\int_0^t\left(\frac{\delta K\,d_{tt}(K+\lambda\delta K,K)}{K^2},\phi_t\right)(\tau)\,d\tau\right|\\
&\quad+C\int_0^t\left(\left\|\frac{\phi}{\rho^{1/2}}\right\|_0^2+\left\|\frac{\phi_t}{K^{1/2}}\right\|_0^2\right)(\tau)\,d\tau\\
&\quad+\left|\int_0^t\left[\frac{\delta K}{\sqrt{\rho K}}\left(\frac{1}{K+\lambda\delta K+\sqrt{K^2+\lambda\,\delta K}}-\frac{1}{2K}\right)p_t(K+\lambda\delta K),\phi_t\right](\tau)d\tau\right|\\
&\quad+\left|\int_0^t\left[\frac{\delta Kd_t(K+\lambda\delta K,K)}{2\sqrt{\rho}K^{3/2}},\phi_t\right](\tau)d\tau\right|\\
&=T_1+T_2+I+B_1+B_2.
\end{aligned}\tag{2.18}$$

We will bound each term T_1, T_2, B_1, B_2 on the right-hand side separately:

$$\begin{aligned}
|T_1|&=\left|\int_0^t\left(\frac{\lambda\delta K^2p_{tt}(K+\lambda\delta K)}{K^2(K+\lambda\delta K)^2},\phi_t\right)(\tau)\,d\tau\right|\\
&\le C\int_0^t\frac{\lambda^2}{K_*^6}\left\|p_{tt}(K+\lambda\delta K)\delta K^2\right\|_0^2(\tau)\,d\tau+\int_0^t\left\|\frac{\phi_t}{K^{1/2}}\right\|_0^2(\tau)\,d\tau\\
&\le C\lambda^2\int_0^t\left\|\delta K^4\right\|_{L^{\alpha'}(\Omega)}\left\|p_{tt}^2(K+\lambda\delta K)\right\|_{L^{\alpha}(\Omega)}(\tau)\,d\tau+\int_0^t\left\|\frac{\phi_t}{K^{1/2}}\right\|_0^2(\tau)\,d\tau\\
&=C\lambda^2\int_0^t\|\delta K\|_{L^{4\alpha'}(\Omega)}^4\|p_{tt}(K+\lambda\delta K)\|_{L^{2\alpha}(\Omega)}^2(\tau)\,d\tau+\int_0^t\left\|\frac{\phi_t}{K^{1/2}}\right\|_0^2(\tau)\,d\tau.
\end{aligned}$$

Sobolev's imbedding theorem implies that $L^{2\alpha}(\Omega)\subset H^1(\Omega)$ for $2\alpha\ge 2$ if $n=1$ or 2 and $2\le 2\alpha\le 6$ if $n=3$. Then using that and Theorem 2.1, we have

$$|T_1|\le C\lambda^2\left\|S_{tt}\right\|_{L^2(0,T;L^2(\Omega))}^2\|\delta K\|_{L^m(\Omega)}^4+\int_0^t\left\|\frac{\phi_t}{K^{1/2}}\right\|_0^2(\tau)\,d\tau.\tag{2.19}$$

Similarly, we have

$$\begin{aligned}
|T_2|&=\left|\int_0^t\left(\frac{\delta K\,d_{tt}(K+\lambda\delta K,K)}{K^2},\phi_t\right)(\tau)\,d\tau\right|\\
&\le C\int_0^t\|\delta K\|_{L^m(\Omega)}^2\,\|d_{tt}(K+\lambda\delta K,K)\|_1^2(\tau)\,d\tau+\int_0^t\left\|\frac{\phi_t}{K^{1/2}}\right\|_0^2(\tau)\,d\tau.
\end{aligned}\tag{2.20}$$

For the boundary terms we have

$$B_1=\int_0^t\left[\frac{\lambda\,\delta K^2\,(\lambda\delta K-3K)p_t(K+\lambda\delta K)}{\sqrt{\rho K}2K\left(\left(K+\sqrt{K^2+\lambda K\,\delta K}\right)^2-(\lambda\delta K)\right)},\phi_t\right](\tau)d\tau.$$

Integrating by parts in t and using that $K\in L^\infty(\Omega)$ we have

$$|B_1|\le C\lambda\left(\left[\delta K^2\,p_t(K+\lambda\delta K),\phi\right](t)+\int_0^t\left[\delta K^2p_{tt}(K+\lambda\delta K),\phi\right](\tau)d\tau\right).$$

By a similar argument to the one used for T_1, T_2 using Lemma 1.1, Sobolev's embedding theorem, and Theorem 2.1 we have

$$\begin{aligned}|B_1| &\le C\,\lambda\,\|\delta K\|^4_{L^m(\Gamma)}\left(\|p_t(K+\lambda\delta K)\|^2_1(t)+\int_0^t\|p_{tt}(K+\lambda\delta K)\|^2_1(\tau)d\tau\right)\\ &\quad+\frac{1}{4}\left\|\frac{\phi}{\rho^{1/2}}\right\|^2_1(t)+\int_0^t\left\|\frac{\phi}{\rho^{1/2}}\right\|^2_1(\tau)d\tau\\ &\le C\lambda\left(\|S_t\|_{L^2(0,T;L^2(\Omega))}+\|S_{tt}\|_{L^2(0,T;L^2(\Omega))}\right)+\frac{1}{4}\left\|\frac{\phi}{\rho^{1/2}}\right\|^2_1(t)\\ &\quad+\int_0^t\left\|\frac{\phi}{\rho^{1/2}}\right\|^2_1(\tau)d\tau.\end{aligned}\tag{2.21}$$

Finally,

$$\begin{aligned}|B_2| &\le C\|\delta K\|^4_{L^m(\Gamma)}\left(\|d_t(K+\lambda\delta K,K)\|^2_1(t)+\int_0^t\|d_{tt}(K+\lambda K,K)\|^2_1(\tau)d\tau\right)\\ &\quad+\frac{1}{4}\left\|\frac{\phi}{\rho^{1/2}}\right\|^2_1(t)+\int_0^t\left\|\frac{\phi}{\rho^{1/2}}\right\|^2_1(\tau)d\tau.\end{aligned}\tag{2.22}$$

Replacing (2.19)–(2.22) in (2.18) and using Gronwall's theorem, we obtain

$$\begin{aligned}\|\phi_t\|^2_{L^\infty(0,T;L^2(\Omega))}+\|\phi\|^2_{L^\infty(0,T;H^1(\Omega))} &\le C\lambda\sum_{i=0}^{2}\left\|\frac{\partial^i S}{\partial t^i}\right\|^2_{L^2(0,T,L^2(\Omega))}\\ &\quad+\|d_t(K+\lambda K,K)\|^2_{L^2(0,T,H^1(\Omega))}+\|d_{tt}(K+\lambda\delta K,K)\|^2_{L^2(0,T;H^1(\Omega))},\end{aligned}\tag{2.23}$$

where $C=C(K_*,K^*,\rho_*,\rho^*,T,\Omega)$. Now using Theorem 2.2 in (2.23), we have that for a given $\epsilon>0$ there is λ_0 such that if $\lambda<\lambda_0$

$$\|\phi_t\|_{L^\infty(0,T;L^2(\Omega))}+\|\phi\|_{L^\infty(0,T;H^1(\Omega))}\le\epsilon\Lambda.\tag{2.24}$$

Thus the theorem is proved for $\delta K\in L^\infty(\Omega)$. Now using Hahn–Banach's theorem and (2.14), the theorem holds for $\delta K\in\mathcal{W}_0$. ∎

As a consequence of Theorem 2.4, we have the following corollary.

COROLLARY 2.6. *Assume that the source function $S(x,t)$ satisfies the hypotheses of Theorem* 2.4. *Then the functional J has a Gâteaux derivative with respect to the parameter K and it is given by*

$$J'(K)\delta K=\int_0^T\left(\Phi'(K)\delta K,(\Phi(K)-p^{obs})(t)\right)dt,\tag{2.25}$$

where $\Phi'(K)\delta K(t)$ is the operator whose entries $(\Phi'(K)\delta K)_i(t)$, $1\le i\le N$, are given by

$$(\Phi'(K)\delta K)_i(t)=\frac{1}{|B_i|}\int_{B_i}D'(K)\delta K(x,t)\,dx,\tag{2.26}$$

and $D'(K)\delta K\in\mathcal{V}$ is the solution of (2.11)–(2.13).

Proof. Using (2.14) of Theorem 2.4 and Theorem 2.1, we have that

$$\begin{aligned}|J'(K)\delta K|^2 &\le C\,|||\delta K|||^2_m\int_0^T|\Phi(K)-p^{obs}|(t)dt\\ &\le C(\|S\|^2_{L^2(0,T;L^2(\Omega))}+\left\|p^{obs}\right\|_{L^2(0,T;\mathbb{R}^{N_r})})\,|||\delta K|||^2_m.\end{aligned}$$

∎

The following theorem about the continuity of $D'(K)\delta K$ with respect to K is necessary to prove the convergence of the algorithm.

THEOREM 2.7. *Let K_1, $K_2 \in \mathcal{P}$ and $\delta K \in \mathcal{W}_0$ $(\mathcal{W})$ and assume that the quantity Λ is finite; then*

$$\begin{aligned}\|(D'(K_1) - D'(K_2))\,\delta K\|_{L^\infty(0,T;L^2(\Omega))} + \|((D'(K_1) - D'(K_2))\,\delta K)_t\|_{L^\infty(0,T;H^1(\Omega))} \\ \leq C\Lambda\,|||K_1 - K_2|||_m\,.\end{aligned}$$

Proof. The technique of the proof is the same as the one of Theorem 2.4, thus we omit it here. ∎

3. The adjoint operator. To describe the algorithm we need the adjoint formulation of (2.25).

THEOREM 3.1. *Assume that the source function $S(x,t)$ satisfies the hypotheses of Theorem 2.4 and that the function f is such that*

$$f \in L^2(0,T,L^2(\Omega));$$

then

$$\begin{aligned}&\int_0^T \int_\Omega D'(K)\delta K(x,t) f(x,t)\,dx\,dt \\ &\quad = \int_\Omega \delta K(x) \int_0^T \frac{p_{tt}(K,x,t)}{K^2(x)} u(K,x,t)\,dt\,dx \\ &\qquad + \int_\Gamma \delta K(x) \int_0^T \frac{p_t(K,x,t)}{2\sqrt{\rho}K^{3/2}(x)} u(K,x,t)\,dt\,dx,\end{aligned} \tag{3.1}$$

where the function $u \in \mathcal{V}$ is defined as $u(K,x,t) = w(K,x,T-t)$ and w is the solution of: find $w(x,t) \in H^1(\Omega)$ such that

$$\left(\frac{1}{K} w_{tt}, v\right) + \left(\frac{1}{\rho}\nabla w, \nabla v\right) + \left[\frac{1}{\alpha} w_t, v\right] = \left(\check{f}, v\right), \qquad v \in H^1(\Omega), \quad t \in [0,T], \tag{3.2}$$

where $\check{f}(\cdot,t) = f(\cdot, T-t)$ and

$$w(K,x,t=0) = w_t(K,x,t=0) = 0. \tag{3.3}$$

Proof. Notice that under the hypotheses of the theorem, $w_{tt} \in \mathcal{V}$. Using $v = D'(K)\delta K(\cdot,t)$ in (3.2) and integrating in t from 0 to T,

$$\begin{aligned}&\int_0^T \left(\left(\frac{1}{K} w_{tt}, D'(K)\delta K\right)(t) + \left(\frac{1}{\rho}\nabla w, \nabla D'(K)\delta K\right)(t) + \left[\frac{1}{\alpha} w_t, D'(K)\delta K\right](t)\right) dt \\ &\quad = \int_0^T \left(f, D'(K)\delta K\right)(t)\,dt.\end{aligned} \tag{3.4}$$

Integrating the first and third integrals by parts in t and using that $w(K,x,T-t) = u(K,x,t)$, we have that

$$\begin{aligned}&\int_0^T \left(\left(\frac{1}{K}\left(D'(K)\delta K\right)_{tt}, u\right)(t) + \left(\frac{1}{\rho}\nabla D'(K)\delta K, \nabla u\right)(t) \right. \\ &\qquad \left. + \left[\frac{1}{\alpha}\left(D'(K)\delta K\right)_t, u\right](t)\right) dt = \int_0^T \left(f, D'(K)\delta K\right)(t)\,dt.\end{aligned} \tag{3.5}$$

Now integrating the weak form for $D'(K)\delta K$ (for $v = u$) from 0 to T we have

$$\begin{aligned}\int_0^T \left(\frac{1}{K}\left(D'(K)\delta K\right)_{tt}, u\right)(t)\,dt \\ + \int_0^T \left(\left(\frac{1}{\rho}\nabla D'(K)\delta K, \nabla u\right)(t) + \left[\frac{1}{\alpha}\left(D'(K)\delta K\right)_t, u\right](t)\right) dt \\ = \int_0^T \left(\frac{\delta K\, p_{tt}(K)}{K^2}, u\right)(t)\,dt + \int_0^T \left[\frac{\delta K\, p_t(K)}{2\sqrt{\rho}K^{3/2}}, u\right](t)\,dt.\end{aligned} \tag{3.6}$$

Finally, (3.1) follows from (3.5) and (3.6). ∎

COROLLARY 3.2. *Assume that the hypotheses on S of the theorem hold; then*

$$\begin{aligned}J'(K)\delta K = \int_\Omega \delta K(x) \int_0^T \frac{p_{tt}(K,x,t)\,u(K,x,t)}{K^2(x)}\,dt\,dx \\ + \int_\Gamma \delta K(x) \int_0^T \frac{p_t(K,x,t)\,u(K,x,t)}{2\sqrt{\rho(x)}K^{3/2}(x)}\,dt\,dx,\end{aligned} \tag{3.7}$$

where u is the solution of (3.2)–(3.3) *with*

$$f(x,t) = \sum_{i=1}^{N_r} \frac{\chi_{B_i}(x)}{|B_i|}(\Phi(K) - p^{\text{obs}})_i(t), \quad x \in \Omega, \quad t \in [0,T]. \tag{3.8}$$

Proof. First note that under the conditions on the source function S using Theorem 2.1, $f \in L^2(0,T,L^2(\Omega))$; then the result follows immediately from (3.1) and

$$\begin{aligned}J'(K)\delta K &= \int_0^T \Phi'(K)\delta K(t)(\Phi(K) - p^{\text{obs}})(t)\,dt \\ &= \int_0^T \sum_{i=1}^{N_r} \left(\frac{1}{|B_i|}\int_\Omega D'(K)\delta K(x,t)\chi_{B_i}(x)\,dx\right)(\Phi(K) - p^{\text{obs}})_i(t)\,dt. \quad ∎\end{aligned} \tag{3.9}$$

Remark. If for $n = 1$ we take point evaluations of p as the model for the measurements, $p(K, x_{r_i}, t)$, the function f in the right-hand side of (3.2), should be $\sum_{i=1}^{N_r} \delta(x - x_{r_i})(p_i - p_i^{\text{obs}})(x,t)$. In this particular case (which is not under the hypotheses of Theorem 3.1) the proof of the existence of solutions of (3.2)–(3.3) is done using energy estimates as in Theorem 2.1.

Consider the observation map Φ defined by (2.7)–(2.8); for $K \in \mathcal{P}$ the operator $\Phi'(K) \in \mathcal{B}(\mathcal{W}, L^2(0,T;\mathbb{R}^{Nr}))$ is given by

$$\Phi'(K)\delta K = \left(\frac{1}{|Bi|}\int_{Bi} D'(K)\,\delta K(x,t)dx\right)_{(1\le i\le Nr)}, \quad \delta K \in \mathcal{W}_o \quad (\mathcal{W}).$$

Thus the adjoint operator of $\Phi'(K)$ is a continuous operator from $L^2(0,T;\mathbb{R}^{Nr})$ into $\mathcal{W}_o$ $(\mathcal{W})$. Then we have the following theorem.

THEOREM 3.3. *Let $h \in L^2(0,T,R^{N_r})$; then the adjoint operator of $\Phi'(K)$ is given by*

$$\begin{aligned}\left([\Phi'(K)]^*\,h, \delta K\right) = \left(\int_0^T \frac{p_{tt}(K)u(K)}{K^2}(\cdot,t)\,dt, \delta K\right) \\ + \left[\int_0^T \frac{p_t(K)\,u(K)}{2\sqrt{\rho}K^{3/2}}(\cdot,t)\,dt, \delta K\right],\end{aligned} \tag{3.10}$$

where u is the solution of (3.2)–(3.3) *with*

$$f(x,t) = \sum_{i=1}^{N_r} \frac{\chi_{B_i}(x)}{|B_i|}\, h_i(t), \quad x \in \Omega, \quad t \in [0,T], \tag{3.11}$$

and the domain of $[\Phi'(K)]^*$ *is* $L^2(0,T;R^{N_r})$.

4. The inverse problem and the algorithm. We turn now to the iterative algorithm. Let $M(K) \in \mathcal{B}(\mathcal{W},\mathcal{W})$ be defined as

$$M(K)\, h(x) = [\Phi'(K)]^*\, \Phi'(K) h(x), \quad h \in \mathcal{W}. \tag{4.1}$$

For every $K \in \mathcal{W}$ for which the operator $M(K)$ is invertible let

$$f : \mathcal{P} \to \mathcal{W}$$

be defined as

$$f(K) = [M(K)]^{-1}\, [\Phi'(K)]^*\, (\Phi(K) - p^{\text{obs}}). \tag{4.2}$$

Note that in (4.1), $\Phi'(K)h \in L^2(0,T;\mathbb{R}^{N_r})$; also in (4.2), $\Phi(K) - p^{\text{obs}} \in L^2(0,T;\mathbb{R}^{N_r})$.

The algorithm is defined by the iteration

$$K^{n+1} = \begin{cases} K^n + f(K^n), & K_* \le K^n + f(K^n) \le K^*, \\ K^*, & K^* < K^n + f(K^n), \\ K_*, & K^n + f(K^n) < K_*. \end{cases} \tag{4.3}$$

Remarks. In practice a regularization term $N(K)$ with $N : \mathcal{Q} \to \mathbb{R}^+$ which is a weakly lower semicontinuous map satisfiying $\lim_{|x|\to\infty} N(x) = \infty$ is added to the cost functional $J(K)$. By choosing $N(K)$ adequately, the condition on the invertibility of the operator M is obtained.

The fact that we have to use in $\mathcal{P}$ the topology given by $|||\cdot|||_m$ with $m > 2$ suggests the use of a regularization term of the form $\left|||K - K^{ref}|||_m^m$ instead of the usual $\|K - K^{ref}\|_0^2$.

There are many different hypotheses under which algorithm (4.3) will converge. We prove the following classical result as an illustration of our particular problem.

THEOREM 4.1. *Assume that the hypotheses of Theorem* 2.5 *hold and that there is a* $K^c \in \mathcal{P}$ *such that* $J'(K^c) = 0$. *Also assume that* $M(K^c)$ *is invertible. Then* K^c *is a point of attraction of iteration* (4.6).

Proof. Since $\Phi'(K^c)$, $\Phi(K^c)$, and $[M(K^c)]^{-1}$ are bounded operators we have the inequality

$$\|[M(K^c)]^{-1}\|\, \|[\Phi'(K^c)]^*(\Phi(K^c) - p^{\text{obs}})\| \le C\|J'(K^c)\| = 0,$$

which implies that K^c is a fixed point for iteration (4.3).

The function $G(K)$,

$$G(K) = K - [M(K)]^{-1}[\Phi'(K)]^*(\Phi(K) - p^{\text{obs}}),$$

is well defined in a neighborhood of K^c as follows: if $M(K)$ satisfies

$$\|M(K) - M(K^c)\|\, \|M(K^c)\| < 1, \tag{4.4}$$

then $M(K)$ is one to one; moreover, $M(K)$ is one to one in a neighborhood of K^c.

Now let $\beta = \|[M(K^c)]^{-1}\|$ and let ϵ satisfy $0 < \epsilon < (2\beta)^{-1}$. Choose $\delta > 0$ so that the ball of center K^c and radius δ is contained in $\mathcal{P}$ and

$$\|M(K^c) - M(K)\| \le \epsilon \quad \text{for } K \in S.$$

Now for $K \in S$

$$\|[M(K)]^{-1}\| \le \beta + \epsilon\beta\|[M(K)]^{-1}\|.$$

Thus

$$\|[M(K)]^{-1}\| \le \beta/(1 - \epsilon\beta) \le 2\beta.$$

Since Φ is Gâteaux differentiable and Φ' is continuous as a function of K, Φ is Fréchet differentiable as follows:

$$\|\Phi(K + \delta K) - \Phi(K) - \Phi'(K)\,\delta K\| \le \sup_{\lambda} \|\Phi'(K + \lambda\,\delta K) - \Phi'(K)\|\,\|\delta K\| \le \epsilon\|\delta K\|.$$

Now we prove that the Fréchet derivative of $[\Phi'(K)]^*(\Phi(K) - p^{\text{obs}})$ at $K = K^c$ is $M(K^c)$.

$$\begin{aligned}\|[\Phi'(K)]^*(\Phi(K) - p^{\text{obs}}) - [\Phi'(K^c)]^*(\Phi(K^c) - p^{\text{obs}}) - M(K^c)(K - K^c)\| \\ \le \|[\Phi'(K)]^*\big((\Phi(K) - p^{\text{obs}}) - (\Phi(K^c) - p^{\text{obs}}) - \Phi'(K^c)(K - K^c)\big)\| \\ + \|[\Phi'(K)]^*\Phi'(K^c)(K - K^c) - M(K^c)(K - K^c)\| = T_1 + T_2.\end{aligned}$$

Using that Φ is Fréchet differentiable and the continuity of Φ with respect to K we have

$$T_1 \le \|[\Phi'(K)]^*\|\epsilon\|K - K^c\|.$$

Using the definition of M and the continuity of $[\Phi(K)]^*$ with respect to K,

$$T_2 \le \epsilon\|[\Phi'(K^c)]^*\|\,\|K - K^c\|.$$

Next we prove that the Fréchet derivative of G at $K = K^c$ is 0:

$$\begin{aligned}\|G(K) - G(K^c)\| &= \|K - [M(K)]^{-1}[\Phi'(K)]^*(\Phi(K) - p^{\text{obs}}) - K^c\| \\ &\le \|[M(K)]^{-1}(\Phi'(K)]^*(\Phi(K) - p^{\text{obs}}) - [\Phi'(K^c)]^*(\Phi(K^c) - p^{\text{obs}}) \\ &\quad - M(K^c)(K - K^c)\big)\| + \|[M(K)]^{-1}M(K^c)(K - K^c) + K - K^c\| \\ &\le 2\beta(\|[\Phi'(K)]^*\| + \|[\Phi'(K^c)]^*\|)\,\epsilon\,\|K - K^c\| \\ &\quad + \|[M(K)]^{-1}\|M(K^c) - M(K)\|\|K - K^c\| \le 2\beta C\,\epsilon\,\|K - K^c\|.\end{aligned}$$

Finally,

$$\|G(K) - G(K^c)\| = \|G(K) - K^c\| \le \epsilon\|K - K^c\|. \qquad \square$$

Appendix.

Proof of Lemma 1.1. Let $f \in C^1(\Omega)$. We distinguish two cases: $n = 1$ and $n = 2, 3$.

Case $n = 1$:

$$
\begin{aligned}
|f(0)|^m &= \left| f(x)^m - \int_0^x \frac{d}{ds}(f^m(s))\, ds \right| \\
&= \left| f(x)^m - m \int_0^x f^{m-1}(s)\, f'(s)\, ds \right| \\
&\le C(|f(x)|^m + \|f\|^{m-1}_{L^{2(m-1)}(\Omega)}\ \|f'\|_{L^2(\Omega)}).
\end{aligned} \tag{A1}
$$

A similar inequality holds for $|f(1)|^m$. Adding both and integrating from 0 to 1 we have

$$
\|f\|^m_{L^m(\Gamma)} \le C\left(\|f\|^m_{L^m(\Omega)} + \|f\|^{m-1}_{L^{2(m-1)}(\Omega)}\ \|f'\|_{L^2(\Omega)} \right). \tag{A2}
$$

Case $n = 2, 3$: Let $\tilde{x}_n = x_1$ if $n = 2$ and $\tilde{x}_n = (x_1, x_2)$ if $n = 3$:

$$
|f(\tilde{x}_n, 0)|^m = \left| f(\tilde{x}_n, x)^m - m \int_0^x f^{m-1}(\tilde{x}_n, s)\, f'(\tilde{x}_n, s)\, ds \right|. \tag{A3}
$$

Integrating in $\tilde{x}_n$ on $\widetilde{\Omega}_n = proj(\Omega)$ over $x_n = 0$ we have

$$
\begin{aligned}
\int_{\widetilde{\Omega}_n} |f(\tilde{x}_n, 0)|^m\ d\tilde{x}_n &\le C \int_{\widetilde{\Omega}_n} \left(|f(\tilde{x}_n, x)|^m + \int_0^x \left| f^{m-1}(\tilde{x}_n, s)\, f'(\tilde{x}_n, s) \right|\, ds \right) d\tilde{x}_n \\
&\le C \left(\int_{\widetilde{\Omega}_n} |f(\tilde{x}_n, x)|^m\ d\tilde{x}_n + \|f\|^{m-1}_{2(m-1)}(\Omega)\ \|f'\|_{L^2(\Omega)} \right).
\end{aligned}
$$

Integrating in x we have

$$
\int_{\widetilde{\Omega}_n} |f(\tilde{x}_n, 0)|^m\ d\tilde{x}_n \le C(\|f\|^m_{L^m(\Omega)} + \|f\|^{(m-1)}_{L^2(m-1)(\Omega)}\ \|f'\|_{L^2(\Omega)}).
$$

We have similar inequalities for each face of the cube. Adding them, inequality (A2) is obtained for $n = 2, 3$.

Proof of Theorem 2.2. We prove the theorem for $q = 0$. Let $d = d(K_1, K_2, x, t)$; d is the solution of the weak form: find $d \in H^1(\Omega)$ such that

$$
\begin{aligned}
&\left(\frac{1}{K_1} d_{tt}, v \right) + \left(\frac{1}{\rho} \nabla d, \nabla v \right) + \left[\left(\frac{1}{\alpha(K_1)} - \frac{1}{\alpha(K_2)} \right) d_t(K_2), v \right] \\
&\quad = \left(\frac{1}{K_1} - \frac{1}{K_2} S, v \right) + \left(\frac{1}{K_2} - \frac{1}{K_1} p_{tt}(K_2), v \right) - \left[\frac{\left(\sqrt{K_1} - \sqrt{K_2} \right) p_t(K_1)}{\rho \sqrt{K_1\, K_2}}, v \right], \\
&\qquad v \in H^1(\Omega), \quad t \in [0, T].
\end{aligned}
$$

Choosing $v = d_t$ and using the usual arguments we obtain the inequality

$$
\begin{aligned}
\left\| \frac{d_t}{K_1^{1/2}} \right\|_0^2 + \left\| \frac{d}{\rho^{1/2}} \right\|_1^2 &\le \int_0^t \left(\left\| \frac{d_t}{K_1^{1/2}} \right\|_0^2 + \left\| \frac{d}{\rho^{1/2}} \right\|_0^2 \right)(\tau)\, d\tau \\
&\quad + \int_0^t \left(\left(\frac{K_2 - K_1}{K_1\, K_2} p_{tt}(K_2), d_t \right)(\tau) + \left(\frac{K_2 - K_1}{K_1\, K_2} S, d_t \right)(\tau) \right) d\tau \\
&\quad + \int_0^t \left[\frac{\sqrt{K_2} - \sqrt{K_1}}{\sqrt{\rho\, K_1\, K_2}} p_t(K_2), d_t \right](\tau) d\tau \\
&= I + T_1 + T_2 + B.
\end{aligned} \tag{A4}
$$

We bound each term separately:

$$
\begin{aligned}
|T_1| &\le C \int_0^t \int_\Omega (K_1 - K_2)^2(x) \left| p_{tt}(K_2, x, \tau) \right|^2 d\tau + \int_0^t \left\| \frac{d_t}{K_1^{1/2}} \right\|_0^2 (\tau) d\tau \\
&\le C \left\| K_1 - K_2 \right\|_m^2 \int_0^T \left\| p_{tt} \right\|_{L^{2m/m-2}}^2 (\tau)\, d\tau + \int_0^t \left\| \frac{d_t}{K_1^{1/2}} \right\|_0^2 (\tau)\, d\tau \\
&\le C \left\| K_1 - K_2 \right\|_{L^m(\Omega)}^2 \left\| S_{tt} \right\|_{L^2(0,T;L^2(\Omega))} + \int_0^t \left\| \frac{d_t}{K_1^{1/2}} \right\|_0^2 (\tau)\, d\tau.
\end{aligned} \tag{A5}
$$

Similarly,

$$
\begin{aligned}
|T_2| &\le C \int_0^t \int_\Omega |K_1 - K_2|^2(x) |S|^2(x,\tau) dx\, d\tau + \int_0^t \left\| \frac{d_t}{K_1^{1/2}} \right\|_0^2 (\tau)\, d\tau \\
&\le C \left\| K_1 - K_2 \right\|_m^2 \left\| S \right\|_{L^2(0,T;L^{2m/m-2})}^2 + \int_0^t \left\| \frac{d_t}{K_1^{1/2}} \right\|_0^2 (\tau)\, d\tau.
\end{aligned} \tag{A6}
$$

Finally, for the boundary term, we have

$$
\begin{aligned}
|B| &\le C \left\| K_1 - K_2 \right\|_{L^m(\Gamma)}^2 \left(\left\| p_t \right\|_{L^\infty(0,T;H^1(\Omega))}^2 + \left\| p_{tt} \right\|_{L^\infty(0,T;H^1(\Omega))}^2 \right) \\
&\quad + \frac{1}{4} \left\| \frac{d}{\rho^{1/2}} \right\|_1^2 (t) + \int_0^T \left\| \frac{d}{\rho^1/2} \right\|_1^2 (\tau) d\tau \\
&\le C \left\| K_1 - K_2 \right\|_{L^m(\Gamma)}^2 \left(\left\| S_t \right\|_{L^2(0,T;L^2(\Omega))}^2 + \left\| S_{tt} \right\|_{L^2(0,T;L^2(\Omega))} \right) \\
&\quad + \int_0^T \left\| \frac{d}{\rho^1/2} \right\|_1^2 (\tau) d\tau + \frac{1}{4} \left\| \frac{d}{\rho^{1/2}} \right\|_1^2 (t) \left\| S_{tt} \right\|_{L^2(0,T;L^2(\Omega))}.
\end{aligned} \tag{A7}
$$

Replacing (A2)–(A4) in (A1) and using Gronwall's inequality we obtain the conclusion of the theorem.

Acknowledgment. I thank Prof. E. Lami Dozo for many useful discussions.

REFERENCES

[1] A. Bamberger, G. Chavent, and P. Lailly, *About the stability of the inverse problem in* 1-*D wave equations—application to the interpretation of seismic profiles*, Appl. Math. Optim., 5 (1979), pp. 1–47.

[2] H. T. Banks, J. A. Burns, and E. Cliff, *Parameter estimation and identification for systems with delays*, SIAM J. Control Optim., 19 (1981), pp. 791–828.

[3] H. T. Banks, J. M. Crowley, and K. Kunisch, *Cubic spline approximation techniques for parameter estimation in distributed systems*, IEEE Trans. Automat. Control, AC-28 (1983), pp. 773–785.

[4] H. T. Banks, S. S. Gates, I. G. Rosen, and Y. Wang, *The identification of a distributed parameter model for a flexible structure*, SIAM J. Control Optim., 26 (1988), pp. 743–762.

[5] H. T. Banks and K. Kunisch, *Estimation Techniques for Distributed Parameter Systems*, Birkhauser, Boston, 1989.

[6] D. W. Brewer, J. A. Burns, and E. Cliff, *Parameter identification for an abstract Cauchy problem by quasilinearization*, Quart. Appl. Math., 51 (1993), pp. 1–22.

[7] E. Fernández-Berdaguer, P. Gauzellino, and J. E. Santos, *An algorithm for parameter estimation in acoustic media, practical issues*, Latin American Applied Research, 25 (1995), pp. 161–168.

[8] E. Fernández-Berdaguer and J. E. Santos, *On the solution of an inverse scattering problem in one-dimensional acoustic media*, Comput. Methods Appl. Mech. Engrg., 129 (1996), pp. 91–105.

[9] E. Fernández-Berdaguer, J. E. Santos, and D. Sheen, *An optimization procedure for estimation of variable coefficients in a hyperbolic system*, Appl. Math. Comput., 76 (1996), pp. 213–250.

[10] M. Giudici, *Identifiability of distributed physical parameters in diffusive-like systems*, Inverse Problems, 7 (1991), pp. 231–245.

[11] M. Giudici, *Identifiability of Physical Parameters for Transport Phenomena in Geophysics*, Technical Report, Universita di Milano, Milano, Italy, 1994.

[12] H. L. Higdon, *Absorving boundary conditions for difference approximations to the multi–dimensional wave equation*, Math. Comp., 34 (1991), pp. 437–459.

[13] Y. Jarny, M. N. Ozisik, and J. P. Bardon, *A general optimization method using adjoint equation for solving multidimensional inverse heat conduction*, Int. J. Heat Mass Transfer, 47 (1986), pp. 2911–2919.

[14] K. Kunisch, *Numerical Methods for Parameter Estimation Problems*, Preprint Reike Mathematic No. 411/1994, Technical University of Berlin, 1994.

[15] P. K. Lamm, *Regularization and the Adjoint Method of Solving Inverse Problems*, Lectures given at 3rd Annual Inverse Problems in Eng. Seminar, Michigan State University, East Lansing, 1990.

[16] P. K. Lamm and K. A. Murphy, *Coefficients and boundary parameters for hyperbolic sysytems*, Quart. Appl. Math., 46 (1988), pp. 1–22.

[17] L. Lines and S. Treitel, *A review of least-squares inversion and its application to geophysical problems*, Geophysical Prospecting, 32 (1984), pp. 159–186.

[18] R. A. Stephen, F. Cardo Casas, and C. H. Cheng, *Finite-difference synthetic acoustic logs*, Geophysics, 50 (1985), pp. 1588–1609.

[19] A. Tarantola, *Inversion of seismic reflection data in the acoustic approximation*, Geophysics, 49 (1984), pp. 1259–126.

[20] A. Tarantola, *Linearized inversion of seismic reflection data*, Geophysical Prospecting, 32 (1984), pp. 998–1015.

[21] A. Tarantola, *Inverse Problem Theory — Methods for Data Fitting and Model Parameter Estimation*, Elsevier, New York, 1987.

[22] L. N. Trefethen and L. Halpern, *Well-posedness of one-way wave equations and absorbing boundary conditions*, Math. Comp., 47 (1986), pp. 421–435.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 4, pp. 1331–1347, July 1998

012

SIMPLIFYING OPTIMAL STRATEGIES IN STOCHASTIC GAMES*

J. FLESCH†, F. THUIJSMAN†, AND O. J. VRIEZE†

Abstract. We deal with zero-sum limiting average stochastic games. We show that the existence of arbitrary optimal strategies implies the existence of stationary ε-optimal strategies, for all $\varepsilon > 0$, and the existence of Markov optimal strategies. We present such a construction for which we do not even need to know these optimal strategies. Furthermore, an example demonstrates that the existence of stationary optimal strategies is not implied by the existence of optimal strategies, so the result is sharp.

More generally, one can evaluate a strategy π for the maximizing player, player 1, by the reward $\phi_s(\pi)$ that π guarantees to him when starting in state s. A strategy π is called nonimproving if $\phi_s(\pi) \geq \phi_s(\pi[h])$ for all s and for all finite histories h with final state s, where $\pi[h]$ is the strategy π conditional on the history h. Using the evaluation ϕ, we may define the relation "ε-better" between strategies. A strategy π^1 is called ε-better than π^2 if $\phi_s(\pi^1) \geq \phi_s(\pi^2) - \varepsilon$ for all s. We show that for any nonimproving strategy π, for all $\varepsilon > 0$, there exists an ε-better stationary strategy and a (0-)better Markov strategy as well. Since all optimal strategies are nonimproving, this result can be regarded as a generalization of the above result for optimal strategies.

Finally, we briefly discuss some other extensions. Among others, we indicate possible simplifications of strategies that are only optimal for particular initial states by "almost stationary" ε-optimal strategies, for all $\varepsilon > 0$, and by "almost Markov" optimal strategies. We also discuss the validity of the above results for other reward functions. Several examples clarify these issues.

Key words. stochastic games, limiting average rewards, optimality, Markov strategies, stationary strategies

AMS subject classifications. 90D15, 90D20, 90D05

PII. S0363012996311940

1. Introduction.

A zero-sum stochastic game Γ can be described by a state space $S := \{1, \ldots, z\}$ and a corresponding collection $\{M_1, \ldots, M_z\}$ of matrices, where matrix M_s has size $m_s^1 \times m_s^2$ and, for $i_s \in I_s := \{1, \ldots, m_s^1\}$ and $j_s \in J_s := \{1, \ldots, m_s^2\}$, entry (i_s, j_s) of M_s consists of a payoff $r(s, i_s, j_s) \in \mathbb{R}$ and a probability vector $p(s, i_s, j_s) = (p(1|s, i_s, j_s), \ldots, p(z|s, i_s, j_s))$. The elements of S are called states and for each state $s \in S$ the elements of I_s and J_s are called the actions of player 1 and player 2 in state s. The game is to be played at stages in $\mathbb{N}$ in the following way. The play starts at stage 1 in an initial state, say, in state $s^1 \in S$, where, simultaneously and independently, both players are to choose an action: player 1 chooses an $i_{s^1}^1 \in I_{s^1}$, while player 2 chooses a $j_{s^1}^1 \in J_{s^1}$. These choices induce an immediate payoff $r(s^1, i_{s^1}^1, j_{s^1}^1)$ from player 2 to player 1. Next, the play moves to a new state according to the probability vector $p(s^1, i_{s^1}^1, j_{s^1}^1)$, say, to state s^2. At stage 2 new actions $i_{s^2}^2 \in I_{s^2}$ and $j_{s^2}^2 \in J_{s^2}$ are to be chosen by the players in state s^2. Then player 1 receives payoff $r(s^2, i_{s^2}^2, j_{s^2}^2)$ from player 2 and the play moves to some state s^3 according to the probability vector $p(s^2, i_{s^2}^2, j_{s^2}^2)$, and so on.

The sequence $(s^1, i_{s^1}^1, j_{s^1}^1; \ldots; s^{n-1}, i_{s^{n-1}}^{n-1}, j_{s^{n-1}}^{n-1}; s^n)$ is called the history up to stage n. The players are assumed to have complete information and perfect recall.

A mixed action for a player in state s is a probability distribution on the set of his actions in state s. Mixed actions in state s will be denoted by x_s for player 1 and

*Received by the editors November 11, 1996; accepted for publication (in revised form) June 5, 1997; published electronically May 27, 1998.

http://www.siam.org/journals/sicon/36-4/31194.html

†Department of Mathematics, University of Maastricht, P.O. Box 616, 6200 MD Maastricht, the Netherlands (flesch@math.unimaas.nl, frank@math.unimaas.nl, oj.vrieze@math.unimaas.nl).

by y_s for player 2, and the sets of mixed actions in state s by X_s and Y_s, respectively. A strategy is a decision rule that prescribes a mixed action for any finite history of the play. Such general strategies, so-called behavior strategies, will be denoted by π for player 1 and by σ for player 2, and $\pi(h)$ and $\sigma(h)$ will denote the mixed actions for history h. We use the notations Π and Σ, respectively, for the behavior strategy spaces of players 1 and 2. If for all finite histories, the mixed actions prescribed by a strategy only depend on the current stage and state, then the strategy is called Markov, while if they only depend on the state then the strategy is called stationary. Thus the stationary strategy spaces are $X := \times_{s\in S} X_s$ for player 1 and $Y := \times_{s\in S} Y_s$ for player 2, while the Markov strategy spaces are $F := \times_{n\in\mathbb{N}} X$ for player 1 and $G := \times_{n\in\mathbb{N}} Y$ for player 2. We will use the respective notations x and y for stationary strategies and f and g for Markov strategies for players 1 and 2. A stationary strategy is called pure if, for each state, it specifies one "pure" action to be used. Hence the spaces of pure stationary strategies are $I := \times_{s\in S} I_s$ for player 1 and $J := \times_{s\in S} J_s$ for player 2. Pure stationary strategies will be denoted by i and j, respectively.

Let H denote the set of finite histories, $H(\alpha,\omega)$ the set of finite histories with initial state α and final state ω, $H(\alpha,\cdot)$ the set of finite histories with initial state α, and $H(\cdot,\omega)$ the set of finite histories with final state ω. For any strategy π and for any given history $h \in H(\cdot,\omega)$, we can define the strategy $\pi[h]$ which prescribes a mixed action $\pi[h](\bar{h})$ to each history $\bar{h} \in H(\omega,\cdot)$ as if h had happened before $\bar{h}$, i.e., $\pi[h](\bar{h}) = \pi(h\bar{h})$, where $h\bar{h}$ is the history consisting of h concatenated with $\bar{h}$.

Payoffs and transition probabilities can be naturally extended to mixed actions as well. For $x_s \in X_s$ and $y_s \in Y_s$ let

$$r(s,x_s,y_s) := \sum_{i_s\in I_s,\, j_s\in J_s} x_s(i_s)\, y_s(j_s)\cdot r(s,i_s,j_s),$$

$$p(t|s,x_s,y_s) := \sum_{i_s\in I_s,\, j_s\in J_s} x_s(i_s)\, y_s(j_s)\cdot p(t|s,i_s,j_s).$$

For $x \in X,\ y \in Y$ we will also use the vector notation

$$r(x,y) := (r(s,x_s,y_s))_{s\in S}.$$

A pair of strategies (π,σ) with an initial state $s \in S$ determines a stochastic process on the payoffs. The sequences of payoffs are evaluated by the limiting average reward and by the β-discounted reward, $\beta \in (0,1)$, given by

$$\gamma(s,\pi,\sigma) := \liminf_{N\to\infty} \mathbb{E}_{s\pi\sigma}\left(\frac{1}{N}\sum_{n=1}^{N} r_n\right) = \liminf_{N\to\infty} \mathbb{E}_{s\pi\sigma}\left(R_N\right),$$

$$\gamma^{\beta}(s,\pi,\sigma) := \mathbb{E}_{s\pi\sigma}\left((1-\beta)\sum_{n=1}^{\infty}\beta^{n-1} r_n\right),$$

where r_n is the random variable for the payoff at stage $n \in \mathbb{N}$, and R_N for the average payoff up to stage N. We also use the vector notations

$$\gamma(\pi,\sigma) := (\gamma(s,\pi,\sigma))_{s\in S}, \qquad \gamma^{\beta}(\pi,\sigma) := \left(\gamma^{\beta}(s,\pi,\sigma)\right)_{s\in S}.$$

A pair of stationary strategies (x, y) determines a Markov chain with transition matrix P_{xy} on S, where entry (s,t) of P_{xy} is $p(t|s, x_s, y_s)$. With respect to this Markov chain, we can speak of transient and recurrent states (a state is called recurrent if, when starting there, it will be visited infinitely often with probability 1; otherwise the state is called transient). We can group the recurrent states into minimal closed sets, and into so-called ergodic sets (an ergodic set is a collection E of recurrent states with the property that, when starting in one of the states in E, all states in E will be visited and the play will remain in E with probability 1). Let

$$Q_{xy} := \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} (P_{xy})^n;$$

the limit is known to exist (cf. Doob [1953, Theorem 2.1, p. 175]). Entry (s,t) of the stochastic matrix Q_{xy}, denoted by $q(t|s,x,y)$, is the expected average number of stages the process is in state t when starting in s. The matrix Q_{xy} has the well-known properties (cf. Doob [1953]) that

$$Q_{xy} = Q_{xy}\, P_{xy} = P_{xy}\, Q_{xy}, \qquad Q_{xy}^2 = Q_{xy}. \tag{1.1}$$

By its definition, for the limiting average reward we have

$$\gamma(x,y) = Q_{xy}\, r(x,y), \tag{1.2}$$

hence by (1.1) we also obtain

$$\gamma(x,y) = Q_{xy}\, r(x,y) = Q_{xy}^2\, r(x,y) = Q_{xy}\, \gamma(x,y). \tag{1.3}$$

Against a fixed stationary strategy y there always exists a pure stationary best reply i of player 1 (cf. Hordijk, Vrieze, and Wanrooij [1983]); i.e.,

$$\gamma(i,y) \geq \gamma(\pi, y) \qquad \forall \pi.$$

Obviously a similar statement holds for the best replies of player 2.

For the limiting average reward, Mertens and Neyman [1981] showed that

$$\sup_{\pi} \inf_{\sigma} \gamma(s,\pi,\sigma) = \inf_{\sigma} \sup_{\pi} \gamma(s,\pi,\sigma) =: v_s \qquad \forall s \in S. \tag{1.4}$$

Here $v := (v_s)_{s\in S}$ is called the limiting average value and v is known to satisfy the following equations:

$$v_s = \mathrm{Val}\,(A_s) \qquad \forall s \in S, \tag{1.5}$$

where

$$A_s := \left[\sum_{t\in S} p(t|s, i_s, j_s)\, v_t\right]_{i_s\in I_s, j_s\in J_s} \tag{1.6}$$

and Val stands for the matrix game value. The sets of optimal mixed actions in A_s, for any $s \in S$, are nonempty polytopes. A strategy π of player 1 is called optimal for initial state $s \in S$ if

$$\gamma(s,\pi,\sigma) \geq v_s \quad \forall \sigma \in \Sigma,$$

and ε-optimal for initial state $s \in S$, $\varepsilon > 0$ if

$$\gamma(s, \pi, \sigma) \geq v_s - \varepsilon \quad \forall \sigma \in \Sigma.$$

If a strategy of player 1 is optimal or ε-optimal for all initial states in S, then the strategy is called optimal or ε-optimal, respectively. Optimality for strategies of player 2 is analogously defined. Although for all $\varepsilon > 0$, by (1.4), there exist ε-optimal strategies for both players, the famous example of Gillette [1957], the Big Match, examined by Blackwell and Ferguson [1968], demonstrates that in general the players need not have optimal strategies, and for achieving ε-optimality, behavior strategies are indispensable.

For the β-discounted reward, $\beta \in (0, 1)$, using a fixed-point argument, Shapley [1953] showed that

$$\sup_{\pi} \inf_{\sigma} \gamma^{\beta}(s, \pi, \sigma) = \inf_{\sigma} \sup_{\pi} \gamma^{\beta}(s, \pi, \sigma) =: v_s^{\beta} \qquad \forall s \in S.$$

Here $v^{\beta} := (v_s^{\beta})_{s \in S}$ is called the β-discounted value. Optimality can be similarly defined as for the limiting average reward. Stationary β-discounted optimal strategies always exist, and x is β-discounted optimal if and only if

$$v_s^{\beta} \leq (1 - \beta)\, r(s, x_s, y_s) + \beta \sum_{t \in S} p(t|s, x_s, y_s)\, v_t^{\beta} \qquad \forall y_s \in Y_s,\ \forall s \in S.$$

We will also make use of the N-stage game Γ^N, $N \in \mathbb{N}$, which is played up to stage N and where the reward is defined by the expected average payoff up to stage N. The N-stage game Γ^N, $N \in \mathbb{N}$, is known to have a value v^N, and both players have N-stage Markov optimal strategies. Bewley and Kohlberg [1976] showed, using Puiseux series, that both $\lim_{\beta \uparrow 1} v^{\beta}$ and $\lim_{N \to \infty} v^N$ exist and

$$\lim_{\beta \uparrow 1} v^{\beta} = \lim_{N \to \infty} v^N,$$

while Mertens and Neyman [1981] proved that the limiting average value is equal to the limit of the β-discounted values, i.e.,

$$v = \lim_{\beta \uparrow 1} v^{\beta}.$$

Although both the β-discounted value and the limiting average value exist, they cannot usually be easily calculated. In general, only iterative algorithms are available. We refer to Raghavan and Filar [1991] for a survey on algorithms.

We will often deal with specific restricted games derived from Γ. Assume that $S' \subset S$ is a nonempty set of states and $X_s' \subset X_s$, $Y_s' \subset Y_s$ are nonempty polytopes for all $s \in S'$. If all pairs of mixed actions in $X_s' \times Y_s'$, for all $s \in S'$, only induce transitions to states in S', then we may define a restricted game Γ', derived from Γ, where the state space is S' and the players are restricted to use strategies that only prescribe mixed actions in X_s' and Y_s' if the play is in state $s \in S'$. Let $\Pi' \subset \Pi$ and $\Sigma' \subset \Sigma$ denote the sets of these strategies. Clearly, the stationary strategy spaces in Γ' are $X' := \times_{s \in S'} X_s'$ and $Y' := \times_{s \in S'} Y_s'$. For the restricted game Γ', with respect to the β-discounted reward, $\beta \in (0, 1)$, similar results can be shown by using a fixed-point argument as for the original game Γ. Thus

$$\sup_{\pi \in \Pi'} \inf_{\sigma \in \Sigma'} \gamma^{\beta}(s, \pi, \sigma) = \inf_{\sigma \in \Sigma'} \sup_{\pi \in \Pi'} \gamma^{\beta}(s, \pi, \sigma) =: v_s'^{\beta} \qquad \forall s \in S'.$$

Here $v'^{\beta} := (v'^{\beta}_s)_{s\in S'}$ is called the β-discounted value for Γ'. Stationary β-discounted optimal strategies in Γ' always exist and $x \in X'$ is β-discounted optimal if and only if

$$v'^{\beta}_s \le (1-\beta)\, r(s, x_s, y_s) + \beta \sum_{t\in S'} p(t|s, x_s, y_s)\, v'^{\beta}_t \qquad \forall y_s \in Y'_s,\ \forall s \in S'. \tag{1.7}$$

The results of Bewley and Kohlberg [1976] apply for Γ' as well, so $\lim_{\beta\uparrow 1} v'^{\beta}$ and $\lim_{N\to\infty} v'^{N}$ exist and

$$v'^{1} := \lim_{\beta\uparrow 1} v'^{\beta} = \lim_{N\to\infty} v'^{N}. \tag{1.8}$$

Note that we do not claim that v'^{1} is the limiting average value of Γ', for even though the players only observe pure actions, these do not correspond one-to-one to extreme points of the restricted spaces of mixed actions. However, one can show, by using an appropriate sequence of discount factors as in Mertens and Neyman [1981], that, against any fixed strategy in Π', for any $\varepsilon > 0$ player 2 can make sure that player 1's limiting average reward is at most $v'^{1} + \varepsilon$; i.e.,

$$\sup_{\pi\in\Pi'} \inf_{\sigma\in\Sigma'} \gamma(s, \pi, \sigma) \le v'^{1}_s \qquad \forall s \in S'. \tag{1.9}$$

From now on, when we speak of rewards, values, or optimal strategies, we will always have the limiting average reward in mind, unless mentioned otherwise.

The organization of the paper is as follows. In section 2 we will deal with optimal strategies. We show that the existence of arbitrary optimal strategies implies the existence of stationary ε-optimal strategies, for all $\varepsilon > 0$, and the existence of Markov optimal strategies. We give such a construction for which we do not even need to know any optimal strategy. This remarkable result should not only be regarded as a simplification of optimal strategies, but also as a sufficient condition for the existence of stationary ε-optimal strategies or Markov optimal strategies. For many classes of stochastic games, where on the payoff or transition structures special conditions are imposed, stationary ε-optimal strategies exist, for all $\varepsilon > 0$, while about sufficient conditions for the existence of Markov optimal strategies, comparatively little is known. Here, instead of providing such structural conditions, the existence of optimal strategies will be proven to be sufficient. Moreover, an example will be provided to show that the existence of stationary optimal strategies is not implied by the existence of optimal strategies, so the result is sharp.

In section 3 we show that simplification of strategies can also be employed for a class of strategies, containing the optimal ones, in view of the rewards they guarantee. For this purpose we will evaluate a strategy π by the reward $\phi_s(\pi)$ that π guarantees when starting in state $s \in S$. A strategy π is called "nonimproving" if $\phi_s(\pi) \ge \phi_s(\pi[h])$ for all s and for all finite histories h with final state s, where $\pi[h]$ is the strategy π conditional on the history h, as defined above. Intuitively, a nonimproving strategy, for any state, cannot guarantee a larger reward conditional on any past history than initially. Using the evaluation ϕ, we may naturally define the relation "ε-better" between strategies. A strategy π^1 is called ε-better than π^2 if $\phi_s(\pi^1) \ge \phi_s(\pi^2) - \varepsilon$ for all $s \in S$. We show that for any nonimproving strategy π, for all $\varepsilon > 0$, there exists an ε-better stationary strategy and a (0-)better Markov strategy as well. Optimal strategies are clearly nonimproving, since they guarantee the value and more cannot be guaranteed; hence this result implies the above result for optimal strategies.

In section 4 we briefly discuss some extensions of the above results. We indicate possible simplifications of strategies that are only optimal for particular initial

states by "almost stationary" ε-optimal strategies and by "almost Markov" optimal strategies. We also discuss the validity of the results when other rewards are used to evaluate the long-term average payoffs. Some remarks concerning the proofs and some consequences are mentioned.

2. Optimal strategies. In this section we show the following result.

THEOREM 2.1. *If player* 1 *has an optimal strategy then, for all* $\varepsilon > 0$*, player* 1 *has stationary* ε*-optimal strategies and Markov optimal strategies as well.*

The proof will be constructive. We present such a construction for which we do not even need to know the optimal strategy.

For $s \in S$ let

$$X_s^* := \left\{ x_s \in X_s | \sum_{t \in S} p(t|s, x_s, y_s)\, v_t \geq v_s \ \ \forall y_s \in Y_s \right\}, \qquad X^* := \times_{s \in S} X_s^*,$$

so X_s^* is the set of optimal mixed actions for player 1 in the matrix game A_s (cf. (1.6)). The sets X_s^* are nonempty polytopes. For $s \in S$ let

$$Y_s^* := \left\{ y_s \in Y_s | \sum_{t \in S} p(t|s, x_s, y_s)\, v_t = v_s \ \ \forall x_s \in X_s^* \right\}, \qquad Y^* := \times_{s \in S} Y_s^*;$$

the sets Y_s^*, called the equalizers in the corresponding matrix games, are nonempty polytopes (in fact, by (1.5) all optimal mixed actions of player 2 in A_s belong to Y_s^*). Note the asimilarity in the definitions of X_s^* and Y_s^*, $s \in S$. It is easy to verify that, for any $s \in S$, there exists a $J_s^* \subset J_s$ such that $Y_s^* = \text{conv}(J_s^*)$, where conv stands for the convex hull of a set. Let

$$J^* := \times_{s \in S} J_s^*.$$

As described in the Introduction, we may define a restricted game Γ^*, derived from Γ, where the state space is S and the players are restricted to use strategies that only prescribe mixed actions in X_s^* and Y_s^* if the play is in state $s \in S$. The sets of these strategies are denoted by Π^* and Σ^*. Let $v^{*\beta}$ denote the β-discounted value for Γ^*, and let $v^{*1} := \lim_{\beta \uparrow 1} v^{*\beta}$.

By the finiteness of the state and action spaces there exists a countable subset of discount factors $\mathcal{B} \subset (0,1)$ such that 1 is a limit point of $\mathcal{B}$ and there are stationary β-discounted optimal strategies $x^\beta \in X^*$ in the restricted game Γ^* such that the sets $\{i_s \in I_s | \, x_s^\beta(i_s) > 0\}$, $s \in S$, are independent of $\beta \in \mathcal{B}$. In the sequel each time that we are dealing with discount factors, discounted optimal strategies, or limits when the discount factors converge to 1, we will have such a subset of discount factors $\mathcal{B}$ in mind.

The following lemma clarifies why the sets X^* and Y^* play an important role when player 1 has an optimal strategy in the original game Γ. This lemma states that if π is an optimal strategy for player 1 in Γ then, for any history with a positive occurrence probability with respect to (π, σ) for some $\sigma \in \Sigma^*$, the strategy π prescribes a mixed action belonging to X^*. In other words, if player 2 uses a strategy $\sigma \in \Sigma^*$ then the optimal strategy π will behave as a strategy in Π^*.

LEMMA 2.2. *Let* $\pi \in \Pi$ *be an optimal strategy for player* 1 *in the game* Γ*. Then for all* $h \in H(\alpha, \omega)$*, for any* $\alpha, \omega \in S$*, we have* $\pi(h) \in X_\omega^*$ *if* $\mathbb{P}_{\alpha\pi\sigma}(h) > 0$ *for some* $\sigma \in \Sigma^*$*. Here* $\mathbb{P}_{\alpha\pi\sigma}(h)$ *denotes the probability that the finite history* h *occurs when the strategies* π *and* σ *are played and the initial state is* α*.*

Proof. Suppose the opposite. Then there exists a shortest history $\bar{h}^n \in H(\alpha,\omega)$, say, with length n, for some $\alpha,\omega \in S$, and a $\sigma \in \Sigma^*$ with $\mathbb{P}_{\alpha\pi\sigma}(\bar{h}^n) > 0$ such that $\pi(\bar{h}^n) \notin X^*_\omega$. Since $\pi(\bar{h}^n) \notin X^*_\omega$ there exists a $j_\omega \in J_\omega$ such that

$$\tau := v_\omega - \sum_{s\in S} p(s|\omega,\pi(\bar{h}^n),j_\omega)\, v_s > 0.$$

Let $s^1 := \alpha$, let s^k, $k \geq 2$, denote the random variable for the state at stage k, and let h^k denote the history up to stage $k \in \mathbb{N}$. Let

$$\delta \in \left(0, \mathbb{P}_{s^1\pi\sigma}(\bar{h}^n)\cdot\tau\right).$$

Let $\sigma^\delta \in \Sigma$ be the strategy that prescribes to play as follows: play σ for the first $n-1$ stages and then, if $h^n = \bar{h}^n$, play j_ω, while if $h^n \neq \bar{h}^n$ then play an optimal mixed action in the matrix game A_{s^n}; and finally, play a δ-optimal strategy afterwards. Note that

$$\mathbb{P}_{s^1\pi\sigma^\delta}(\bar{h}^n) = \mathbb{P}_{s^1\pi\sigma}(\bar{h}^n) > 0.$$

Since we have chosen a shortest history $\bar{h}^n$ with the above property, by the definitions of X^* and Y^* we have

$$\mathbb{E}_{s^1\pi\sigma^\delta}\left(v_{s^n}\right) = v_{s^1},$$

and by the used mixed actions at stage n

$$\mathbb{E}_{s^1\pi\sigma^\delta}\left(v_{s^{n+1}}\right) \leq \mathbb{E}_{s^1\pi\sigma^\delta}\left(v_{s^n}\right) - \mathbb{P}_{s^1\pi\sigma^\delta}(\bar{h}^n)\cdot\tau.$$

From stage $n+1$, player 2 plays a δ-optimal strategy, so the choice of δ yields

$$\begin{aligned}\gamma(s^1,\pi,\sigma^\delta) &\leq \mathbb{E}_{s^1\pi\sigma^\delta}\left(v_{s^{n+1}}\right) + \delta \leq \mathbb{E}_{s^1\pi\sigma^\delta}\left(v_{s^n}\right) - \mathbb{P}_{s^1\pi\sigma^\delta}(\bar{h}^n)\cdot\tau + \delta \\ &= v_{s^1} - \mathbb{P}_{s^1\pi\sigma}(\bar{h}^n)\cdot\tau + \delta < v_{s^1},\end{aligned}$$

which contradicts the optimality of π. □

Based on the fact that any optimal strategy of player 1 in Γ guarantees the value v and, in view of the previous lemma, it only prescribes mixed actions in X^*_s, if the play is in state s, against any strategy of player 2 in Σ^*, we show that player 1 can guarantee at least v in the restricted game Γ^*. On the other hand, as discussed in (1.9), player 1 cannot guarantee more than the limit of the β-discounted values in Γ^*.

LEMMA 2.3. *Suppose that player* 1 *has an optimal strategy* $\pi \in \Pi$. *Then*

$$v_s \leq \sup_{\pi^*\in\Pi^*}\inf_{\sigma^*\in\Sigma^*}\gamma(s,\pi^*,\sigma^*) \leq v^{*1}_s \qquad \forall s \in S.$$

Proof. The second inequality follows from (1.9), so we only have to show the first one. For $\alpha,\omega \in S$ let

$$\bar{H}(\alpha,\omega) := \{h \in H(\alpha,\omega) |\, \mathbb{P}_{\alpha\pi\sigma^*}(h) > 0 \text{ for some } \sigma^* \in \Sigma^*\}.$$

Take an arbitrary $x \in X^*$. Using Lemma 2.2 we may define a strategy $\pi^* \in \Pi^*$ as follows: for $h \in H(\alpha,\omega)$ let

$$\pi^*(h) := \begin{cases} \pi(h) & \text{if } h \in \bar{H}(\alpha,\omega), \\ x_\omega & \text{otherwise.}\end{cases}$$

Then, by the optimality of π and by the definition of π^*, we have

$$v_s \le \gamma(s,\pi,\sigma^*) = \gamma(s,\pi^*,\sigma^*) \qquad \forall \sigma^* \in \Sigma^*,\ \forall s \in S,$$

which implies the first inequality. □

The next result shows the effectiveness of the β-discounted optimal strategies in the restricted game Γ^*.

LEMMA 2.4. *Let $\varepsilon > 0$. For $\beta \in \mathcal{B}$, let $x^\beta \in X^*$ be a β-discounted optimal strategy of player 1 in Γ^*, and let $y \in Y^*$. Suppose that $E \subset S$ is a closed set of states with respect to (x^β, y) for all $\beta \in \mathcal{B}$. Then, for large $\beta \in \mathcal{B}$,*

$$\gamma(s,x^\beta,y) \ge \min_{t\in E} v_t^{*1} - \varepsilon \qquad \forall s \in E.$$

Proof. Using inequality (1.7) for Γ^* we have

$$(1-\beta)\, r(x^\beta,y) + \beta\, P_{x^\beta y}\, v^{*\beta} \ge v^{*\beta} \qquad \forall \beta \in \mathcal{B}.$$

By (1.1), multiplying this inequality with $Q_{x^\beta y}$ yields

$$Q_{x^\beta y}\, r(x^\beta,y) \ge Q_{x^\beta y}\, v^{*\beta} \qquad \forall \beta \in \mathcal{B}.$$

The closedness of E implies that, for any $s \in E$, if $q(t|s,x^\beta,y) > 0$ then $t \in E$. Hence for all $s \in E$ and for large $\beta \in \mathcal{B}$, using (1.2), we have

$$\begin{aligned}\gamma(s,x^\beta,y) &= \sum_{t\in E} q(t|s,x^\beta,y)\, r(t,x_t^\beta,y_t) \ge \sum_{t\in E} q(t|s,x^\beta,y)\, v_t^{*\beta} \\ &\ge \sum_{t\in E} q(t|s,x^\beta,y)(v_t^{*1}-\varepsilon) \ge \min_{t\in E} v_t^{*1} - \varepsilon,\end{aligned}$$

so the proof is complete. □

Next we discuss some properties of stationary strategies belonging to X^* or to $\mathrm{Relint}(X^*)$, where $\mathrm{Relint}(X^*)$ stands for the relative interior of the polytope X^* and is defined as the set of points in X^* which can be written as a convex combination of all the extreme points of X^* with only strictly positive coefficients.

LEMMA 2.5. *Let $x \in X^*$ and $y \in Y$. Suppose E is an ergodic set with respect to (x,y). Then $v_s = v_t$ for all $s,t \in E$. Furthermore, if $x \in \mathrm{Relint}(X^*)$, then necessarily $y_s \in Y_s^*$ for all $s \in E$.*

Proof. By $x \in X^*$ we obtain

$$\sum_{t\in E} p(t|s,x_s,y_s)\, v_t \ge v_s \qquad \forall s \in E.$$

Let $\bar{E} := \{s \in E |\, v_s = \max_{t\in E} v_t\}$. The above inequalities imply that $\bar{E}$ is a closed set of states for (x,y), so since E is an ergodic set for (x,y) (minimal closed set of states), we have $\bar{E} = E$. Therefore, $v_s = v_t =: v_E$ for all $s,t \in E$.

Now suppose that $x \in \mathrm{Relint}(X^*)$. Then $(\bar{x}_s, y_s)$ only induces transitions to states in E for any $\bar{x}_s \in X_s^*$, $s \in E$; hence

$$\sum_{t\in S} p(t|s,\bar{x}_s,y_s)\, v_t = \sum_{t\in E} p(t|s,\bar{x}_s,y_s)\, v_E = v_E = v_s \quad \forall \bar{x}_s \in X_s^*,\ \forall s \in E,$$

which implies that $y_s \in Y_s^*$ for all $s \in E$. □

An important property of convex combinations of stationary strategies is stated in the next lemma.

LEMMA 2.6. *For $\tau \in (0,1)$, $x^1, x^2 \in X$ let $x^\tau := \tau x^1 + (1-\tau)x^2$. Suppose that E is an ergodic set with respect to (x^τ, y) for some $y \in Y$. Let $\varepsilon > 0$ and $d \in \mathbb{R}$. If*

$$\gamma(s, x^1, y) \geq d \qquad \forall s \in E,$$

then for sufficiently large τ

$$\gamma(s, x^\tau, y) \geq d - \varepsilon \qquad \forall s \in E.$$

Proof. Let $\delta \in (0,1)$. Since

$$\gamma(s, x^1, y) \geq d \qquad \forall s \in E,$$

there exists a K^δ satisfying

$$\mathbb{E}_{sx^1y}(R_N) \geq d - \delta \qquad \forall N \geq K^\delta,\ \forall s \in E,$$

where R_N denotes the average payoff up to stage N. Choose $\tau \in (0,1)$ such that

$$\tau^{K^\delta} \geq 1 - \delta.$$

The strategy x^τ can be interpreted as playing x^1 with probability τ and x^2 with probability $1-\tau$ at each stage, so the last inequality means that x^1 will be played at each K^δ consecutive stages with probability at least $1-\delta$. Hence, with probability at least $1-\delta$, the expected average of the payoffs will be at least $d-\delta$ for any K^δ consecutive stages. Let r denote the smallest payoff in the game. Then if δ is small, by the law of large numbers we have

$$\gamma(s, x^\tau, y) \geq (1-\delta)(d-\delta) + \delta r \geq d - \varepsilon \qquad \forall s \in E,$$

so the proof is complete. □

The next lemma will enable us to construct Markov optimal strategies from stationary ε-optimal strategies which prescribe optimal mixed actions in the matrix games A_s, $s \in S$ (cf. (1.6)). Here we present a short proof, which uses some arguments of Bewley and Kohlberg [1978] on so-called irreducible games.

LEMMA 2.7. *Suppose that for all $\varepsilon > 0$ player 1 has a stationary ε-optimal strategy $x^\varepsilon \in X^*$ in Γ. Then player 1 also has a Markov optimal strategy f in Γ.*

Proof. Consider the restricted game $\Gamma^*(1)$, derived from Γ, where player 1 is restricted to use strategies that only prescribe mixed actions in X_s^*, if the play is in state $s \in S$. As before, Π^* will denote the set of these strategies for player 1. (Note that here only player 1 is restricted, in contrast with the game Γ^*, where both players have a restriction.) Let $v^{*\beta}(1)$ denote the β-discounted value in $\Gamma^*(1)$ and let $v^{*1}(1) := \lim_{\beta \uparrow 1} v^{*\beta}(1)$. Let $v^{*N}(1)$ denote the value of the N-stage game $\Gamma^{*N}(1)$, and let f^N be an N-stage Markov optimal strategy in $\Gamma^{*N}(1)$. Using the assumption that $x^\varepsilon \in X^*$ is ε-optimal in Γ for all $\varepsilon > 0$ and using (1.9) and (1.8), we obtain

$$v_s \leq \sup_{\pi \in \Pi^*} \inf_{\sigma \in \Sigma} \gamma(s, \pi, \sigma) \leq v_s^{*1}(1) = \lim_{N \to \infty} v_s^{*N}(1) \qquad \forall s \in S. \tag{2.1}$$

Let f be the Markov strategy of player 1 which prescribes to play as follows: at stage 1, play f^1; at the next two stages, play f^2; at the next three stages, play f^3; and so

on. We show that f is optimal. Let s^1 be the initial state and let s^N, $N \geq 2$, denote the state for the first stage when f^N is to be played. Take an arbitrary $\sigma \in \Sigma$. Notice that $f \in \Pi^*$, hence by the definition of X^*,

$$\mathbb{E}_{s^1 f\sigma}(v_{s^N}) \geq v_{s^1} \qquad \forall N \in \mathbb{N}.$$

Thus using the N-stage optimality of f^N and (2.1), for any $\delta > 0$ if N is large, then

$$\mathbb{E}_{s^1 f\sigma}(R^N) \geq \mathbb{E}_{s^1 f\sigma}\left(v^{*N}_{s^N}(1)\right) \geq \mathbb{E}_{s^1 f\sigma}\left(v_{s^N}\right) - \delta \geq v_{s^1} - \delta, \tag{2.2}$$

where R^N denotes the average payoff for those N consecutive stages when f^N is played. Let $\phi(K)$ be such that $f^{\phi(K)}$ is to be played at stage K. Observe that

$$\lim_{K\to\infty}\left[\frac{\sum_{N<\phi(K)} N}{K}\right] = 1, \qquad \lim_{K\to\infty}\left[\frac{K - \sum_{N<\phi(K)} N}{K}\right] = 0,$$

so if R_K denotes the average payoff up to stage K and r denotes the smallest payoff in the game, then (2.2) gives

$$\begin{aligned}
\gamma(s^1, f, \sigma) &= \liminf_{K\to\infty} \mathbb{E}_{s^1 f\sigma}\left(R_K\right) \\
&\geq \liminf_{K\to\infty} \mathbb{E}_{s^1 f\sigma}\left(\frac{\sum_{N<\phi(K)} N\cdot R^N + \left[K - \sum_{N<\phi(K)} N\right]\cdot r}{K}\right) \\
&= \liminf_{K\to\infty} \frac{\sum_{N<\phi(K)} N\cdot \mathbb{E}_{s^1 f\sigma}\left(R^N\right)}{K} \\
&\geq v_{s^1},
\end{aligned}$$

which implies that f is optimal in Γ. □

Now we are ready to prove Theorem 2.1.

Proof of Theorem 2.1. We show the existence of stationary ε-optimal strategies for all $\varepsilon > 0$, and then the existence of Markov optimal strategies follows from Lemma 2.7.

For $\beta \in \mathcal{B}$, let $x^\beta \in X^*$ be a β-discounted optimal strategy of player 1 in Γ^* and let $x \in \text{Relint}(X^*)$. For all $\tau \in (0,1)$ and $\beta \in \mathcal{B}$ let

$$x^{\tau\beta} := \tau x^\beta + (1-\tau)x.$$

By the convexity of X^* and by $x \in \text{Relint}(X^*)$ we have $x^{\tau\beta} \in \text{Relint}(X^*)$ for all $\tau \in (0,1)$ and $\beta \in \mathcal{B}$.

We show that, for any $\varepsilon > 0$, for large $\tau \in (0,1)$ and for large $\beta \in \mathcal{B}$ the strategy $x^{\tau\beta}$ is ε-optimal. Let $\varepsilon > 0$. Since against a stationary strategy there always exists a pure stationary best reply, and there are only finitely many pure stationary strategies, it suffices to show that, for all $j \in J$, if $\tau \in (0,1)$ and $\beta \in \mathcal{B}$ are large, then

$$\gamma(x^{\tau\beta}, j) \geq v - \varepsilon\, 1_z,$$

where $1_z = (1,\ldots,1) \in \mathbb{R}^z$. Take a $j \in J$ and let $E \subset S$ be an arbitrary ergodic set with respect to $(x^{\tau\beta}, j)$. We start with showing that for large $\tau \in (0,1), \beta \in \mathcal{B}$ we have

$$\gamma(s, x^{\tau\beta}, j) \geq v_s - \varepsilon \qquad \forall s \in E. \tag{2.3}$$

Since $x^{\tau\beta} \in \text{Relint}(X^*)$, by Lemma 2.5 we obtain $v_s = v_t := v_E$ for all $s, t \in E$ and $j_s \in J_s^*$ for all $s \in E$. Let $j_s^* := j_s$ for all $s \in E$ and let $j_s^* \in J_s^*$ for all $s \notin E$; so $j^* \in J^*$. By the definition of $x^{\tau\beta}$ and by the properties of $\mathcal{B}$, the set of states E is closed with respect to (x^β, j) for all $\beta \in \mathcal{B}$, so with respect to (x^β, j^*) for all $\beta \in \mathcal{B}$ as well. Thus, applying Lemma 2.4 for Γ^* and using Lemma 2.3 yields that for large $\beta \in \mathcal{B}$

$$\gamma(s, x^\beta, j) = \gamma(s, x^\beta, j^*) \geq \min_{t \in E} v_t^{*1} - \frac{1}{2}\varepsilon \geq \min_{t \in E} v_t - \frac{1}{2}\varepsilon = v_E - \frac{1}{2}\varepsilon \qquad \forall s \in E.$$

Now Lemma 2.6 yields that for large $\tau \in (0,1)$ and for large $\beta \in \mathcal{B}$

$$\gamma(s, x^{\tau\beta}, j) \geq v_E - \varepsilon = v_s - \varepsilon \qquad \forall s \in E,$$

which proves (2.3).

Using that $x^{\tau\beta} \in X^*$ we have

$$P_{x^{\tau\beta} j}\, v \geq v,$$

therefore

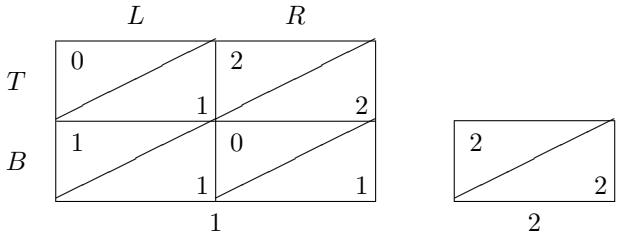
$$Q_{x^{\tau\beta} j}\, v \geq v.$$

For any $s \in S$, $q(t|s, x^{\tau\beta}, j) > 0$ implies that $t \in E$ for some ergodic set E with respect to $(x^{\tau\beta}, j)$; hence by (1.3) and (2.3), for large $\tau \in (0,1)$ and $\beta \in \mathcal{B}$, we obtain

$$\gamma(x^{\tau\beta}, j) = Q_{x^{\tau\beta} j}\, \gamma(x^{\tau\beta}, j) \geq Q_{x^{\tau\beta} j}\, (v - \varepsilon\, 1_z) = Q_{x^{\tau\beta} j}\, v - \varepsilon\, 1_z \geq v - \varepsilon\, 1_z,$$

which completes the proof. □

Example 1.



Here player 1 chooses rows and player 2 chooses columns. In each entry, the corresponding payoff is placed in the upper-left corner, while the transition is placed in the bottom-right corner. In this game each transition is represented by the number of the state to which transition should occur with probability 1. Notice that state 2 is absorbing. The value of this game is $v = (1, 2)$. It is not hard to show that there are optimal strategies for player 1 (later we will construct optimal Markov strategies). Following the construction for stationary ε-optimal strategies, we have $X^* = X$, $Y_1^* = \{(1,0)\}$, $Y_2^* = \{(1)\}$. Now the β-discounted optimal strategy of player 1 in Γ^* is $x^\beta = ((0,1),(1))$ for all $\beta \in (0,1)$. Take a strategy $x \in \text{Relint}(X^*)$, for example, $x = ((\frac{1}{2}, \frac{1}{2}), (1))$. Then for $\tau, \beta \in (0,1)$,

$$x^{\tau\beta} = \tau \cdot x^\beta + (1-\tau) \cdot x = \left(\left(\frac{1}{2} - \frac{1}{2}\tau, \frac{1}{2} + \frac{1}{2}\tau \right), (1) \right),$$

so $x^{\tau\beta}$ is ε-optimal for large τ and β (the strategies $((p,1-p),(1))$ are ε-optimal for $p \in (0,\varepsilon]$). Note that player 1 has no stationary optimal strategy in this game.

Also, a Markov optimal strategy can be constructed as in Lemma 2.7. In this game $X = X^*$, hence the restricted game $\Gamma^*(1)$ is just the original game Γ. The one-stage Markov optimal strategy and the one-stage value are

$$f^1 = \left(\left(\frac{1}{3},\frac{2}{3}\right),(1)\right), \qquad v^1 = v^{*1}(1) = \left(\frac{2}{3},2\right);$$

the two-stage Markov optimal strategy and the two-stage value are

$$f^2 = \left(\left(\left(\frac{3}{13},\frac{10}{13}\right),(1)\right);\left(\left(\frac{1}{3},\frac{2}{3}\right),(1)\right)\right), \qquad v^2 = v^{*2}(1) = \left(\frac{28}{39},2\right);$$

and so on. So, as shown before, the Markov strategy f which prescribes to play f^1 at the first stage, then f^2 at the next two stages, f^3 at the next three stages, and so on, is optimal.

3. Nonimproving strategies. It is in the spirit of zero-sum games to evaluate a strategy π of player 1 by the reward $\phi(\pi)$ it guarantees against any strategy of the opponent. For a strategy $\pi \in \Pi$ let

$$\phi_s(\pi) := \inf_{\sigma} \gamma(s,\pi,\sigma) \quad \forall s \in S, \qquad \phi(\pi) := (\phi_s(\pi))_{s\in S}\,.$$

Using this evaluation ϕ we may naturally define the relation "ε-better" between strategies of player 1. A strategy π^1 is called ε-better than π^2 if $\phi_s(\pi^1) \geq \phi_s(\pi^2) - \varepsilon$ holds for all $s \in S$. 0-better strategies will simply be called better. We will call a strategy π nonimproving if for any state $s \in S$ and for any history $h \in H(\cdot,s)$ we have

$$\phi_s(\pi) \geq \phi_s(\pi[h]).$$

Intuitively, a nonimproving strategy, for any state, cannot guarantee a larger reward conditional on any past history than initially. Obviously, all stationary strategies are nonimproving strategies. Also, optimal strategies are always nonimproving, since they guarantee the value, and no higher reward can be guaranteed.

In this section we will indicate how the following result, which is a generalization of Theorem 2.1, can be shown by using similar techniques as in section 2.

THEOREM 3.1. *For any nonimproving strategy, for any $\varepsilon > 0$, there exists an ε-better stationary strategy and a better Markov strategy as well.*

First we focus on the proof for the existence of ε-better stationary strategies, $\varepsilon > 0$, and afterwards we explain how the existence of a better Markov strategy will follow. Let π denote a fixed nonimproving strategy and let

$$w := \phi(\pi).$$

For $s \in S$ let

$$B_s := \left[\sum_{t\in S} p(t|s,i_s,j_s)\, w_t\right]_{i_s\in I_s,\, j_s\in J_s}, \qquad W_s := \mathrm{Val}\,(B_s),$$

where Val stands for the matrix game value. By using the nonimprovingness of π we obtain

$$\begin{aligned} w_s = \phi_s(\pi) &\leq \sum_{t\in S}\sum_{i_s\in I_s} \pi(s)(i_s)\, p(t|s,i_s,j_s)\cdot \phi_s(\pi[s,i_s,j_s,t]) \\ &\leq \sum_{t\in S}\sum_{i_s\in I_s} \pi(s)(i_s)\, p(t|s,i_s,j_s)\cdot w_t = \sum_{t\in S} p(t|s,\pi(s),j_s)\, w_t \qquad \forall j_s \in J_s,\ \forall s \in S, \end{aligned}$$

hence

$$w_s \le W_s = \mathrm{Val}\,(B_s) \qquad \forall s \in S. \tag{3.1}$$

This is the counterpart of (1.5), however, for w equality does not hold as for the value v, which causes some additional difficulties. We will define a restricted game here as well, but this restricted game will only be defined on a set of states s where $w_s = W_s$, so that we can use similar arguments as in section 2. Let

$$\tilde{X}_s := \left\{ x_s \in X_s | \sum_{t\in S} p(t|s,x_s,y_s)\, w_t \ge w_s \ \ \forall y_s \in Y_s \right\}, \qquad \tilde{X} := \times_{s\in S}\tilde{X}_s,$$

so the set $\tilde{X}_s$, which is a polytope, is the set of mixed actions of player 1 in state s which assure that after transition w will not decrease in expectation. The inequalities (3.1) imply that, for any state $s \in S$, all optimal mixed actions of player 1 in the matrix game B_s belong to $\tilde{X}_s$, which also means that the sets $\tilde{X}_s$ are nonempty.

Fix an arbitrary $x \in \mathrm{Relint}(\tilde{X})$. For a pure stationary strategy $j \in J$ let $R(j)$ denote the set of recurrent sets with respect to (x, j). Let

$$S' := \cup_{j\in J} R(j).$$

For $s \in S'$ let

$$J'_s := \cup_{j\in J,\, s\in R(j)}\{j_s\}, \quad Y'_s := \mathrm{conv}(J'_s), \quad Y' := \times_{s\in S'} Y'_s,$$

where conv stands for the convex hull of a set. Notice that the sets $R(j), S', J'_s, Y'_s, Y'$ are independent of the choice of $x \in \mathrm{Relint}(\tilde{X})$ and also that the sets Y'_s are nonempty polytopes. One can verify that all states $s \in S'$ are recurrent with respect to (x, y), if $y \in Y$ satisfies $y_s \in \mathrm{Relint}(Y'_s)$ for all $s \in S'$. If E is an ergodic set with respect to (x, y) with $y_s \in \mathrm{Relint}(Y'_s)$ for all $s \in S'$, then, as in Lemma 2.5, one can show that $w_s = w_t$ for all $s, t \in E$. Since $x \in \mathrm{Relint}(\tilde{X})$, this also yields that $w_s = W_s$ for all $s \in S'$, so w_s has a similar property as v_s in (1.5). The sets S' and Y' also have the property that, for any $y \in Y$, if $s \in S$ is recurrent with respect to (x, y), then $s \in S'$ and $y_s \in Y'_s$. Let

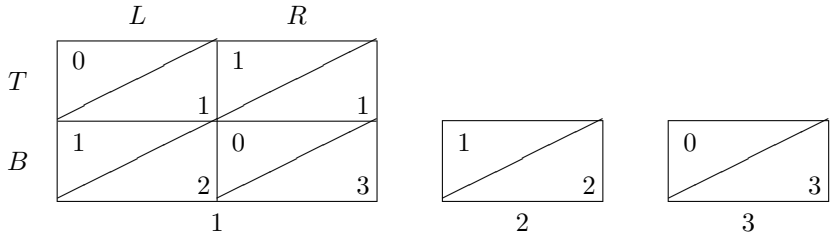
$$X' := \times_{s\in S'}\tilde{X}_s.$$

Let Γ' be the restricted game, derived from Γ, where the state space is S' and the players are restricted to using strategies that only prescribe mixed actions in X'_s and Y'_s, respectively, if the play is in state $s \in S'$. Note that, by the above property of S' and Y', if player 1 uses mixed actions in $\mathrm{Relint}(X'_s)$, $s \in S'$, then whatever stationary strategy y player 2 uses, the play will eventually reach an ergodic set $E \subset S'$ in such a way that w does not decrease in expectation, and $y_s \in Y'_s$ for all $s \in E$, so intuitively the play will eventually proceed in Γ'. Now, using $w_s = W_s$ for all $s \in S'$, for the restricted game Γ', similar results can be shown as for the restricted game in section 2, which completes the proof for the existence of ε-better stationary strategies.

Now the existence of better Markov strategies can be shown along similar lines as the proof of Lemma 2.7. One has to define a restricted game $\Gamma'(1)$, derived from Γ, where player 1 is restricted to use strategies that only prescribe mixed actions in X'_s if the play is in state $s \in S$. Notice that $\Gamma'(1)$ is the counterpart of $\Gamma^*(1)$ defined in the proof of Lemma 7 and also that the above constructed ε-better stationary strategies

belong to X', hence player 1 may use these strategies in the restricted game $\Gamma'(1)$ as well. Now in the game $\Gamma'(1)$, analogous equalities and inequalities can be derived as for $\Gamma^*(1)$ in the proof of Lemma 2.7, but w has to be used instead of v, which leads to the conclusion that better Markov strategies indeed exist.

Example 2.



This example, known as the Big Match (cf. Gillette [1957], Blackwell and Ferguson [1968]), clarifies that, although optimality implies nonimprovingness, improving strategies are indispensable for achieving ε-optimality. The notation is the same as in Example 1. Notice that states 2 and 3 are absorbing. For initial state 1, the limiting average value is $v_1 = \frac{1}{2}$ and player 1 has neither optimal strategies nor stationary ε-optimal strategies for small $\varepsilon > 0$, but for any $N \in \mathbb{N}$ player 1 can guarantee $\frac{1}{2} - \frac{1}{2(N+1)}$ by playing the following strategy π^N: for any history h without absorption, if $k(h)$ denotes the number of stages where player 2 has chosen action R minus the number of stages where player 2 has chosen action L, player 1 has to play the mixed action

$$\pi^N(h) := \left(1 - \frac{1}{(k(h)+N+1)^2}, \frac{1}{(k(h)+N+1)^2}\right).$$

This strategy π^N is clearly improving, since for the history $h = (1, T, R, 1)$ we have $\pi^N[h] = \pi^{N+1}$. Note that, in fact, all strategies that are ε-optimal for small $\varepsilon > 0$ must be improving; otherwise, by Theorem 3.1, player 1 would have stationary ε-optimal strategies (and Markov optimal strategies as well).

4. Concluding remarks. Finally we discuss some consequences. For the sake of simplicity, we only focus on the results of section 2 here.

Remarks on the restricted game Γ^*. In Lemma 2.3 we showed that $v_s^{*1} \geq v_s$ for all $s \in S$. In fact, this is the only statement for which we needed the condition that player 1 has an optimal strategy. Therefore, if in a zero-sum game $v_s^{*1} \geq v_s$ holds for all $s \in S$, then stationary ε-optimal strategies, $\varepsilon > 0$, and Markov optimal strategies can be constructed exactly as in section 2. It also means that $v_s^{*1} \geq v_s$ for all $s \in S$ holds if and only if player 1 has an optimal strategy.

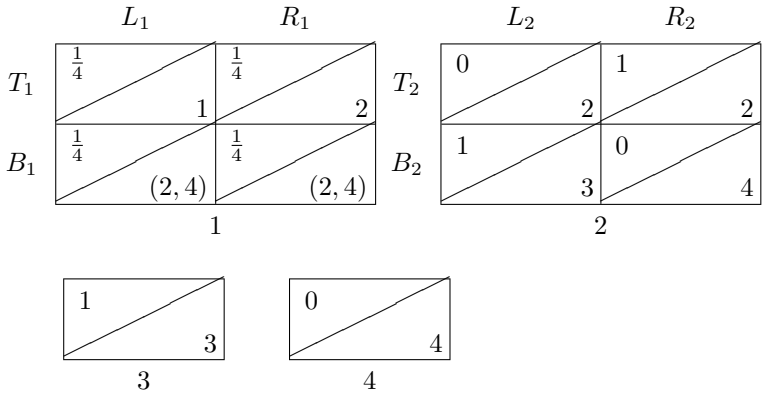
We also remark that, even if player 1 has an optimal strategy, one can find examples where $v_s^{*1} > v_s$ for some state s. However, if E is an ergodic set with respect to some $(x, y) \in \text{Relint}(X^*) \times \text{Relint}(Y^*)$, then there exists a state $s \in E$ such that $v_s^{*1} = v_E$ (recall that the value v is a constant on E by Lemma 2.5). To see this one can argue as follows. Suppose to the contrary that $v_s^{*1} \geq v_E + \mu$ for all $s \in E$, where $\mu > 0$. Let $x^{\tau\beta} \in \text{Relint}(X^*)$ be defined as in the proof of Theorem 2.1. Then Lemmas 2.4 and 2.6 imply that for large τ and β we have

$$\gamma(s, x^{\tau\beta}, j) \geq \min_{t \in E} v_t^{*1} - \frac{\mu}{2} \geq v_E + \frac{\mu}{2} \qquad \forall s \in E,\ \forall j \in J^*. \tag{4.1}$$

Here we used that there are only finitely many pure stationary strategies. Let player 1 play the strategy π^δ, $\delta > 0$, which prescribes to play as follows: play $x^{\tau\beta}$ as long as player 2 chooses actions in J_s^*, $s \in E$, and start playing a δ-optimal strategy as soon as player 2 chooses an action in $J_s \setminus J_s^*$ in some state $s \in E$. Note that if player 2 always chooses actions in J_s^*, $s \in E$, then (4.1) assures that the reward is at least $v_E + \frac{\mu}{2}$ (recall that against a stationary strategy there always exists a pure stationary best reply). On the other hand, if player 2 chooses an action in $J_s \setminus J_s^*$ in some state $s \in E$, then one can show that $x_s^{\tau\beta} \in \mathrm{Relint}(X_s^*)$ yields that the original value v increases in expectation by at least some $\nu > 0$, so if $\delta \in (0, \frac{\nu}{2})$, by the definition of π^δ, the reward is at least $v_E + \frac{\nu}{2}$ in this case. Therefore, π^δ, with $\delta \in (0, \frac{\nu}{2})$, guarantees a reward of at least $v_E + \frac{1}{2}\min(\mu, \nu) > v_E$, which contradicts the definition of the value. So we have shown that $v_s^{*1} = v_E$ holds for some state $s \in E$.

Optimal strategies for particular initial states. We briefly discuss a generalization of the results of section 2, which concerns strategies that are only optimal for particular initial states. Let $\tilde{S}$ denote the set of states for which player 1 has an optimal strategy. First note that in each stochastic game there always exists at least one initial state for which player 1 has optimal strategies (cf. Thuijsman and Vrieze [1991]), so the set $\tilde{S}$ is always nonempty. Using similar techniques as in section 2, one can show that, for any $\varepsilon > 0$, player 1 has a strategy ξ^ε which for all initial states $\alpha \in \tilde{S}$ satisfies the following criteria: (i) ξ^ε is ε-optimal, (ii) ξ^ε is stationary until leaving $\tilde{S}$, (iii) there exist stationary best replies of player 2 against ξ^ε, (iv) the probability of ever leaving $\tilde{S}$ is zero with respect to $(\alpha, \xi^\varepsilon, \sigma)$, if σ is a best reply. The difference between this result and the corresponding result of section 2 is mainly due to the fact that stationary strategies are not effective in states outside $\tilde{S}$, so player 1 may have to start playing a behavior δ-optimal strategy if the play leaves $\tilde{S}$, for some $\delta > 0$. Furthermore, one can also show that player 1 has a strategy χ which for all initial states $\alpha \in \tilde{S}$ satisfies the following criteria: (v) χ is optimal, (vi) χ is Markov until leaving $\tilde{S}$, (vii) there exist Markov best replies of player 2 against χ, (viii) the probability of ever leaving $\tilde{S}$ is zero with respect to (α, χ, σ), if σ is a best reply. We remark here that Markov best replies do not necessarily exist against a Markov strategy, but a Markov strategy χ can be constructed so that (vii) holds.

Example 3.



This example clarifies the existence of such "almost stationary" ε-optimal strategies and "almost Markov" optimal strategies for initial states in $\tilde{S}$. The notation is the same as in Example 1 except for two "mixed" transition vectors in entries (B_1, L_1) and (B_1, R_1), which lead to state 2 with probability $\frac{1}{2}$ and to state 4 with probability

$\frac{1}{2}$. For the sake of simplicity, we only focus on the possible simplifications by "almost stationary" ε-optimal strategies. Notice that if the initial state is state 2, then this game reduces to Example 2. So here the value is $v = (\frac{1}{4}, \frac{1}{2}, 1, 0)$. As mentioned, for initial state 2, player 1 has no optimal strategy, so $\tilde{S} = \{1, 3, 4\}$. Since initial states $3, 4 \in \tilde{S}$ are trivial, we assume the initial state to be $1 \in \tilde{S}$. Consider the strategy ξ for player 1 which prescribes playing action T_1 as long as the play is in state 1, and as soon as the play visits state 2 then prescribes starting to play a behavior $\frac{1}{8}$-optimal strategy. This strategy ξ is optimal and clearly satisfies properties (i), (ii), (iii), and (iv). Note that switching to a behavior strategy when entering state 2 is crucial, because by stationary strategies player 1 could only guarantee 0 for initial state 2. Note also that the use of action B_1 would violate property (iv).

An alternative proof for Lemma 2.7. We wish to remark that, under the condition of Lemma 2.7, other Markov optimal strategies exist as well . Let ε_n be a positive sequence converging to zero. One can show that the Markov strategy which prescribes x^{ε_1} for the first N_1 stages, x^{ε_2} for the next N_2 stages, and so on, is optimal for a well-chosen increasing sequence N_n.

Subgame optimality. Note that the Markov strategy f, constructed in section 2, is "subgame optimal"; namely, the strategy $f[h]$ is optimal for any finite history h.

Alternative rewards and optimality. It is worthwhile to mention that sometimes other rewards are used to evaluate the long-term average payoffs. The most common rewards are the following ones:

$$\gamma^1(s,\pi,\sigma) = \mathbb{E}_{s\pi\sigma}\left(\liminf_{N\to\infty} R_N\right), \qquad \gamma^2(s,\pi,\sigma) = \liminf_{N\to\infty} \mathbb{E}_{s\pi\sigma}\left(R_N\right),$$

$$\gamma^3(s,\pi,\sigma) = \limsup_{N\to\infty} \mathbb{E}_{s\pi\sigma}\left(R_N\right), \qquad \gamma^4(s,\pi,\sigma) = \mathbb{E}_{s\pi\sigma}\left(\limsup_{N\to\infty} R_N\right),$$

where R_N is the random variable for the average payoff up to stage $N \in \mathbb{N}$. It holds that $\gamma^1 \le \gamma^2 \le \gamma^3 \le \gamma^4$. Notice that we have used $\gamma = \gamma^2$ so far. Mertens and Neyman [1981] showed that the value is the same for all these rewards. Optimality and ε-optimality can be defined with respect to any of these four rewards. Sometimes a fifth alternative is to require uniformity from the optimal strategy; i.e., π is uniformly optimal for state $s \in S$ if

$$\forall \delta > 0\ \exists N^\delta\colon\ \mathbb{E}_{s\pi\sigma}\left(R_N\right) \ge v_s - \delta \quad \forall N \ge N^\delta,\ \forall \sigma \in \Sigma.$$

The definition of uniform ε-optimality is similar.

Focussing only on section 2 again, we briefly examine the validity of the results for all these criteria. First notice that it makes no difference in our results in which way the strategy of player 1 is optimal. Furthermore, for stationary strategy pairs, all the above optimality criteria are known to be equivalent (for example, cf. Bewley and Kohlberg [1978]), so the simplifications by stationary strategies remain valid with respect to all these alternatives. For Markov strategies, however, it is somewhat different. Notice first that the Markov strategy constructed in section 2 is uniformly optimal (see the proof of Lemma 2.7). Since $\gamma^2 \le \gamma^3 \le \gamma^4$ we have that this Markov strategy is also optimal for rewards γ^3, γ^4. However, when player 1 has an optimal strategy, the existence of Markov optimal strategies for reward γ^1 is not straightforward, not even by using an approach as in the alternative proof for Lemma 2.7.

REFERENCES

T. Bewley and E. Kohlberg [1976], *The asymptotic theory of stochastic games*, Math. Oper. Res., 1, pp. 197–208.

T. Bewley and E. Kohlberg [1978], *On stochastic games with stationary optimal strategies*, Math. Oper. Res., 3, pp. 104–125.

D. Blackwell and T. S. Ferguson [1968], *The big match*, Ann. Math. Stat., 33, pp. 159–163.

J. L. Doob [1953], *Stochastic Processes*, Wiley, New York.

D. Gillette [1957], *Stochastic games with zero stop probabilities*, in Contributions to the Theory of Games III, M. Dresher, A. W. Tucker, and P. Wolfe, eds., Ann. of Math. Stud. 39, Princeton University Press, Princeton, NJ, pp. 179–187.

A. Hordijk, O. J. Vrieze, and G. L. Wanrooij [1983], *Semi-Markov strategies in stochastic games*, Internat. J. Game Theory, 12, pp. 81–89.

J. F. Mertens and A. Neyman [1981], *Stochastic games*, Internat. J. Game Theory, 10, pp. 53–66.

T. E. S. Raghavan and J. A. Filar [1991], *Algorithms for stochastic games*, Z. Oper. Res., 35, pp. 437–472.

L. S. Shapley [1953], *Stochastic games*, Proc. Nat. Acad. Sci. U.S.A, 39, pp. 1095–1100.

F. Thuijsman and O. J. Vrieze [1991], *Easy initial states in stochastic games*, in Stochastic Games and Related Topics, T. E. S. Raghavan, T. S. Ferguson, O. J. Vrieze, and T. Parthasarathy, eds., Kluwer, Dordrecht, the Netherlands, pp. 85–100.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 4, pp. 1348–1375, July 1998

013

A VISCOSITY APPROACH TO INFINITE-DIMENSIONAL HAMILTON–JACOBI EQUATIONS ARISING IN OPTIMAL CONTROL WITH STATE CONSTRAINTS*

MACIEJ KOCAN† AND PIERPAOLO SORAVIA‡

Abstract. We consider nonlinear optimal control problems with state constraints and nonnegative cost in infinite dimensions, where the constraint is a closed set possibly with empty interior for a class of systems with a maximal monotone operator and satisfying certain stability properties of the set of trajectories that allow the value function to be lower semicontinuous. We prove that the value function is a viscosity solution of the Bellman equation and is in fact the minimal nonnegative supersolution.

Key words. viscosity solutions, nonlinear semigroups, accretive operators, dynamic programming, optimal control with state constraints, optimality principles

AMS subject classifications. 49L20, 49L25, 49J24

PII. S0363012996301622

1. Introduction. We study optimal control infinite horizon problems with state constraints in infinite dimensions. Let H be a real Hilbert space and let A be a maximal monotone operator in H, possibly nonlinear and multivalued. Let U be a Hilbert space and let $f\colon \overline{D(A)} \times U \to H$; the precise assumptions on f are stated in section 2. For $x \in \overline{D(A)}$ (where $D(A)$ denotes the domain of A) and $u \in L^2 \equiv L^2(0,\infty;U)$ let $y(\cdot) = y(\cdot,x,u) \in C([0,\infty);H)$ be the mild solution of the state equation

$$\begin{cases} y'(t) + Ay(t) \ni f(y(t),u(t)) & \text{for} \quad t \geq 0, \\ y(0) = x. \end{cases} \tag{1.1}$$

Given a lower semicontinuous function $g\colon \overline{D(A)} \to [0,+\infty]$, we seek to minimize the cost functional

$$J(x,u) = \int_0^\infty \left(g(y(t)) + \tfrac{1}{2}\|u(t)\|^2 \right) dt \tag{1.2}$$

over all controls $u \in L^2$, where $y(\cdot) = y(\cdot,x,u)$ is a trajectory of the system (1.1).

The value function $V\colon \overline{D(A)} \to [0,+\infty]$ associated with this problem is

$$V(x) = \inf \left\{ J(x,u)\colon\ u \in L^2(0,\infty;U) \right\}. \tag{1.3}$$

Our choice of the running cost g forces a state constraint on the trajectories of the system (1.1). Namely, let K denote the closure of $\mathrm{dom}(g) = \{x\colon\ g(x) \text{ is finite}\}$; we

*Received by the editors April 8, 1996; accepted for publication (in revised form) July 21, 1997; published electronically May 27, 1998.

http://www.siam.org/journals/sicon/36-4/30162.html

†Centre for Mathematics and Its Applications, Australian National University, Canberra, ACT 0200, Australia (kocan@maths.anu.edu.au). The research of this author was supported by National Science Foundation grant DMS93-02995 and an Australian Research Council grant.

‡Dipartimento di Matematica Pura e Applicata, Universita di Padova, via Belzoni 7, 35131 Padova, Italy (soravia@math.unipd.it). Part of this work was completed while the author was visiting the University of California at Santa Barbara. The research of this author was partially supported by Consiglio Nazionale delle Ricerche of Italy.

remark that K may have an empty interior (with respect to $\overline{D(A)}$). Then clearly $V(x) = +\infty$ for $x \in \overline{D(A)} \setminus K$, so that $\mathrm{dom}(V) \subset K$. Observe that for $x \in \mathrm{dom}(V)$ in (1.3) it is enough to take the infimum over all controls $u \in \mathcal{U}(x)$, where

$$\mathcal{U}(x) = \left\{u \in L^2(0,\infty;U): \quad J(x,u) < +\infty\right\}, \tag{1.4}$$

and we can write

$$V(x) = \inf\left\{J(x,u):\ u \in \mathcal{U}(x)\right\} \quad \text{for} \quad x \in \mathrm{dom}(V) \subset K. \tag{1.5}$$

In particular, for every $x \in \mathrm{dom}(V)$ there exists a control $u \in L^2$ such that $g(y(t))$ is finite for almost all $t \geq 0$; that is, u is admissible for the state constraint K and satisfies

$$y(t,x,u) \in K \quad \text{for all} \quad t \geq 0.$$

We refer to M. Soner [20] for sufficient conditions for existence of admissible controls at all points in the finite-dimensional case and to P. Cannarsa, F. Gozzi, and M. Soner [6] for infinite-dimensional systems. We also recall I. Capuzzo-Dolcetta and P. L. Lions [7] for a general study of constrained viscosity solutions in finite dimensions. In our study we do not require the condition $\mathcal{U}(x) \neq \emptyset$ for $x \in K$ and thus $\mathrm{dom}(V)$ may be strictly contained in K.

As usual, following the dynamic programming approach, the value function V is expected to solve in some weak sense the associated Bellman equation

$$\langle Ax, DV\rangle + \sup_{u\in U}\left\{-\langle f(x,u), DV\rangle - \tfrac{1}{2}\|u\|^2\right\} = g(x) \quad \text{in} \quad \overline{D(A)}. \tag{1.6}$$

In particular, note that for $f(x,u) = Bu$, where $B \in \mathcal{L}(U,H)$ is bounded linear, (1.6) becomes

$$\langle Ax, DV\rangle + \tfrac{1}{2}\|B^*DV\|^2 = g(x) \quad \text{in} \quad \overline{D(A)}. \tag{1.7}$$

The main technical problems in proving (1.6) are due to the fact that K, and, consequently, $\mathrm{dom}(V)$, may have an empty interior. A suitable concept of solution, taking into account this and the singularity of the Hamiltonian outside $\mathrm{dom}(g)$, is introduced in section 3. It is clear that even if $\mathrm{dom}(V) = \overline{D(A)}$, solutions of (1.6) are never unique, and this ill-posedness of the problem may not be fixed by prescribing solutions to vanish at a particular point; see the paper by the second author [22] for some examples.

A finite horizon version of the control problem we study, for a linear system, linear operator A and convex cost g, and the time-dependent version of (1.7) associated with it, were studied by P. Cannarsa and G. Di Blasio in [5]. In [5] the authors assume that $\mathcal{U}(x) \neq \emptyset$ for $x \in K$ and define weak solutions of the evolution equation corresponding to (1.7) as pointwise limits of increasing sequences of strong solutions of approximate equations that are associated with unconstrained convex control problems. Strong solutions are meant in the sense of convex control theory; see V. Barbu and G. Da Prato [3].

Our approach is different. We directly define solutions of (1.6) as viscosity solutions in the spirit of D. Tataru [24], [25] and M. G. Crandall and P. L. Lions [15], modifying their definition to take into account the failure of comparison between super- and subsolutions of (1.6) and to cope with extended real-valued solutions. This

idea allows us to eliminate the convexity assumption and to extend the framework of [5] to include certain nonlinear systems; see sections 2 and 6 for precise assumptions and examples. On the other hand we do not prove general explicit formulas for the optimal feedback law, as in [5]. This kind of result remains for now a peculiarity of convex control, where more regular (e.g., convex) value functions are available. We refer the reader also to [3] for these results in the case of unconstrained control problems. We prove that the value function in (1.5) is the minimal nonnegative viscosity supersolution of (1.6). To this end we extend to infinite-dimensional systems certain optimality principles for viscosity supersolutions and subsolutions proved by the second author in [21] and [22] in finite dimensions.

Our techniques have a broader scope. We decided to treat the infinite horizon problem without discount factor and the stationary problem (1.6) here because of the additional technical difficulties it poses, namely, the lack of comparison for viscosity solutions of the Bellman equation. In the case of the finite horizon problem and the infinite horizon problem with a positive discount factor, appropriate analogues of our main result hold and are in fact easier to obtain. When g is Lipschitz, the corresponding Hamilton–Jacobi equations do satisfy comparison and have unique solutions, as is known from Crandall and Lions [15], which can be used to simplify our proofs. However, if g is merely extended real-valued and lower semicontinuous, one only expects uniqueness results in the spirit of our Proposition 2.5.

The plan of the paper is as follows. Our assumptions and statements are discussed in section 2. The definitions and references for the theory of viscosity solutions, as well as some preliminary lemmas, are given in section 3. In section 4 we study the relationship between Hamilton–Jacobi–Bellman equations and value functions of control problems. In section 5 we complete the proof of our main result. Finally, in section 6 we present some examples of systems and constraints satisfying our assumptions.

2. Preliminaries: Assumptions and statements. H is a fixed real Hilbert space and A a maximal monotone (equivalently, m-accretive) operator in H. Then $-A$ generates a strongly continuous semigroup $S(t)$ of contractions on $\overline{D(A)}$; we refer the reader to V. Barbu [1] or H. Brézis [4] for the theory of nonlinear semigroups.

The function f in the state equation (1.1) will always satisfy

$$\text{(2.1)}\qquad \begin{cases} f\colon \overline{D(A)} \times U \to H \quad \text{is continuous and} \\ \text{there exist } L>0, q\in[1,2) \text{ such that for all } x,z\in\overline{D(A)}, u\in U, \\ \|f(x,u)-f(z,u)\| \le L\|x-z\| \quad \text{and} \quad \|f(x,u)\| \le L\left(1+\|x\|+\|u\|^q\right). \end{cases}$$

Note that from (2.1) and Hölder's inequality, for every $x,z\in\overline{D(A)}$, $u\in L^2$, and $t>0$,

$$\begin{aligned} \|y(t,z,u)-x\| &\le \|y(t,z,u)-S(t)z\| + \|S(t)z-S(t)x\| + \|S(t)x-x\| \\ &\le \int_0^t \|f(y(s,z,u),u(s))\|ds + \|z-x\| + \|S(t)x-x\| \\ &\le L\int_0^t \left(1+\|y(s,z,u)\|+\|u(s)\|^q\right)ds + \|z-x\| + \|S(t)x-x\| \\ &\le Lt + L\|u\|^q_{L^2(0,t;U)}t^{1-\frac{q}{2}} + \|z-x\| + \|S(t)x-x\| \\ &\quad + L\int_0^t \|y(s,z,u)\|ds, \end{aligned} \qquad \text{(2.2)}$$

and then by Gronwall's lemma $\|y(t,z,u)\|$ stays bounded for t bounded, uniformly in u bounded in L^2 and z bounded in $\overline{D(A)}$. Using (2.2) again we deduce

$$y(t,z,u) \to x \ \text{ as } \ t \downarrow 0 \ \text{ and } \ z \to x, \ \text{uniformly for } u \text{ bounded in } L^2(0,T;U), \ T>0. \tag{2.3}$$

Our most restrictive, although natural, assumption on the system is one of the following two stability conditions:

$$\text{(W)} \qquad \begin{cases} \text{if } u_n \rightharpoonup u \text{ in } L^2(0,T;U) \text{ for some } T>0 \text{ and } x_n \rightharpoonup x \text{ in } \overline{D(A)}, \\ \text{then for every } t \in (0,T),\ y(t,x_n,u_n) \rightharpoonup y(t,x,u) \text{ in } H, \end{cases}$$

or

$$\text{(S)} \qquad \begin{cases} \text{if } u_n \rightharpoonup u \text{ in } L^2(0,T;U) \text{ for some } T>0 \text{ and } x_n \to x \text{ in } \overline{D(A)}, \\ \text{then for every } t \in (0,T),\ y(t,x_n,u_n) \to y(t,x,u) \text{ in } H. \end{cases}$$

Conditions (W) and (S) are not directly comparable. Note however that by (2.1) and Gronwall's inequality we have that

$$\|y(t,z,u) - y(t,x,u)\| \le \|z-x\| e^{Lt},$$

and therefore (S) is equivalent to its version with $x_n \equiv x \in \overline{D(A)}$. Moreover, the condition that $y(\cdot,x,u_n)$ converges pointwise to $y(\cdot,x,u)$ in (S) is equivalent to uniform convergence in $C([0,T];U)$; see Theorem 2.3.1 in Vrabie [26].

It is well known (and can be easily deduced from the Duhamel principle) that the weak condition (W) holds if the operator A is linear and $f(x,u) = Bu$, where $B \in \mathcal{L}(U,H)$. In particular, our framework extends that in [5]. If instead $-A$ generates a compact semigroup, then the strong condition (S) is satisfied when $f(x,u) = f_1(x) + f_2(x)u$ and f_i, $i = 1,2$ are Lipschitz; see Proposition 2.7 and Remark 2.8 below. The condition (S) is also satisfied if $f(x,u) = Bu$, $B \in \mathcal{L}(U,H)$ is compact, and A generates a weakly equicontinuous semigroup; see Remark 2.10 at the end of this section.

Remark 2.1. All the results we present in this paper hold with trivial changes even if we require the controls to satisfy the condition $u(t) \in C \subset U$ for a.e. $t \ge 0$, where $C \ni 0$ is such that $L^2(0,T;C)$ is weakly closed in $L^2(0,T;U)$ for $T>0$, for example, if C is closed and convex. For previous results on convex control problems with control constraints see also Di Blasio [17].

In what follows, $LSC(\Omega)$, $USC(\Omega)$, $w\text{-}LSC(\Omega)$, and $w\text{-}USC(\Omega)$ will stand for the spaces of all lower semicontinuous, upper semicontinuous, sequentially weakly lower semicontinuous, sequentially weakly upper semicontinuous (possibly extended real-valued) functions on Ω, respectively.

We will always assume that

$$g\colon \overline{D(A)} \to [0,+\infty] \tag{2.4}$$

is at least lower semicontinuous and denote

$$K = \overline{\operatorname{dom}(g)} \subseteq \overline{D(A)}.$$

One can impose state constraints on the system (1.1) on given closed set K by starting off with any $g\colon \overline{D(A)} \to [0,+\infty)$ and setting $g \equiv +\infty$ off K, so that the value

function V associated with such a running cost satisfies $\mathrm{dom}(V) \subset K$. Obviously $g \in LSC(\overline{D(A)})$ if and only if $g_{|K} \in LSC(K)$, and asking for $g \in w\text{-}LSC(\overline{D(A)})$ roughly amounts to $g_{|K} \in w\text{-}LSC(K)$ and the set K to be weakly closed.

We state the main result of the paper, which says that the value function V given by (1.3) can be uniquely characterized as the minimal nonnegative lower semicontinuous supersolution of (1.6). For the definition of viscosity solutions we refer to the next section. Recall that $V\colon \overline{D(A)} \to [0, +\infty]$ satisfies $\mathrm{dom}(V) \subset K$ and is given by (1.3).

THEOREM 2.2. *Assume* (2.1) *and* (2.4). *Suppose that* $g \in w\text{-}LSC(\overline{D(A)})$ ($g \in LSC(\overline{D(A)})$, *respectively*). *If* (W) ((S), *respectively*) *holds, then* V *is a viscosity solution of* (1.6) *and it is the minimal nonnegative, sequentially weakly lower semicontinuous (strongly lower semicontinuous, respectively) extended real-valued viscosity supersolution of* (1.6).

Theorem 2.2 will follow from the following series of propositions. The first statement is about the regularity of the value function. Note that the proof we give also shows existence of optimal controls for our problem; see section 5.

PROPOSITION 2.3. *Assume* (2.4) *and suppose that* (1.1) *has a unique mild solution for any* $u \in L^2(0, \infty; U)$ *and* $x \in \overline{D(A)}$. *Suppose that* $g \in w\text{-}LSC(\overline{D(A)})$ (*respectively,* $g \in LSC(\overline{D(A)})$). *If* (W) ((S), *respectively*) *holds then* $V \in w\text{-}LSC(\overline{D(A)})$ ($V \in LSC(\overline{D(A)})$, *respectively*).

The second statement relates the value function to the Hamilton–Jacobi equation.

PROPOSITION 2.4. *Assume* (2.1) *and* (2.4) *and suppose that* $g \in LSC(\overline{D(A)})$. *Then the value function* V *is a viscosity solution of* (1.6).

The third statement characterizes the value function through the equation.

PROPOSITION 2.5. *Assume* (2.1) *and* (2.4). *Suppose that* $g \in w\text{-}LSC(\overline{D(A)})$ (*respectively,* $g \in LSC(\overline{D(A)})$) *and* (W) ((S), *respectively*) *holds. Suppose that* $w \in w\text{-}LSC(\overline{D(A)})$ ($w \in LSC(\overline{D(A)})$, *respectively*) *is a nonnegative, extended real-valued viscosity supersolution of* (1.6). *Then* $w \geq V$ *on* $\overline{D(A)}$, *where* V *is the value function given by* (1.5).

We will prove more than the statement of Proposition 2.5, namely, an optimality principle for supersolutions of (1.6). We refer to section 5 for the actual statement of this result, see Lemma 5.5, and to [21] and [22] for similar results in the finite-dimensional case. Note in particular that from Proposition 2.5 it follows that there exists a viscosity supersolution of (1.6) which is finite at x if and only if $V(x)$ is finite and, therefore, if and only if there is at least one control providing a finite cost at x, i.e., $\mathcal{U}(x) \neq \emptyset$.

Remark 2.6. If A is linear, $-A$ generates a compact semigroup, $f(x, u) = Bu$, where $B \in \mathcal{L}(U, H)$, and (2.1) and (2.4) hold, then the following holds true (note that it is stronger than both (W) and (S)):

$$\begin{cases} \text{if } u_n \rightharpoonup u \text{ in } L^2(0, T; U) \text{ for some } T > 0 \text{ and } x_n \rightharpoonup x \text{ in } \overline{D(A)}, \\ \text{then for every } t \in (0, T),\ y(t, x_n, u_n) \to y(t, x, u) \text{ in } H. \end{cases}$$

Then statements a little bit stronger than the ones above hold; e.g., $g \in LSC(\overline{D(A)})$ implies that $V \in w\text{-}LSC(\overline{D(A)})$, etc.

We continue this section by proving the sufficient condition for (S) that we mentioned above.

PROPOSITION 2.7. *Suppose that* $f(y, u) = f_1(y) + f_2(y)u$, *where* $f_1\colon \overline{D(A)} \to H$ *and* $f_2\colon \overline{D(A)} \to \mathcal{L}(U, H)$ *are Lipschitz. If* $-A$ *generates a compact semigroup then* (S) *holds.*

Proof. Fix $T > 0$ and $\hat{x} \in \overline{D(A)}$ and suppose that $u_n \rightharpoonup u_\infty$ in $L^2(0,T;U)$. Then $\{u_n\}$ is bounded in $L^2(0,T;U)$. For $n = 1, 2, \dots, \infty$ and $w \in C([0,T];H)$ denote by $E_n[w](\cdot)$ the mild solution $y(\cdot)$ of

$$\begin{cases} y' + Ay \ni f_1(w) + f_2(w)u_n & \text{in} \quad [0,T], \\ y(0) = \hat{x}. \end{cases}$$

Then for every $0 < t \le T$, $E_n : C([0,t];H) \to C([0,t];H)$ and we claim that with respect to the sup-norm in $C([0,t];H)$ the map

$$E_n^k \quad \text{is} \quad \frac{L^k(\sqrt{T}+C)^k t^{k/2}}{\sqrt{k!}} - \text{Lipschitz} \quad \text{for every} \quad k = 1, 2, \dots, \tag{2.5}$$

where L denotes a Lipschitz constant for f_1 and f_2, and $C = \max_n\{\|u_n\|_{L^2(0,T;U)}\}$.

If $w_1, w_2 \in C([0,t];H)$ then for every $t \in [0,T]$,

$$\begin{aligned} &\|E_n[w_1] - E_n[w_2]\|_{L^\infty(0,t;H)} \\ &\quad \le \int_0^t (\|f_1(w_1(s)) - f_1(w_2(s))\| + \|f_2(w_1(s)) - f_2(w_2(s))\|\|u_n(s)\|)\, ds \\ &\quad \le L(t + C\sqrt{t})\|w_1 - w_2\|_{L^\infty(0,t;H)} \le L\sqrt{t}(\sqrt{T}+C)\|w_1 - w_2\|_{L^\infty(0,t;H)}, \end{aligned}$$

and (2.5) holds if $k = 1$. Suppose that (2.5) holds for $k \ge 1$. Then

$$\begin{aligned} \|E_n^{k+1}[w_1] - E_n^{k+1}[w_2]\|_{L^\infty(0,t;H)} &\le \int_0^t \|f_1(E_n^k[w_1](s)) - f_1(E_n^k[w_2](s))\| ds \\ &\quad + \int_0^t \|f_2(E_n^k[w_1](s)) - f_2(E_n^k[w_2](s))\|\|u_n(s)\| ds \\ &\le \int_0^t L\frac{L^k(\sqrt{T}+C)^k s^{k/2}}{\sqrt{k!}}(1 + \|u_n(s)\|) \\ &\quad \cdot \|w_1 - w_2\|_{L^\infty(0,s;H)} ds \\ &\le \frac{L^{k+1}(\sqrt{T}+C)^{k+1} t^{(k+1)/2}}{\sqrt{(k+1)!}}\|w_1 - w_2\|_{L^\infty(0,t;H)}, \end{aligned}$$

since

$$\int_0^t s^{k/2}(1 + \|u_n(s)\|) ds \le \tfrac{2}{k+2} t^{k/2+1} + \tfrac{1}{\sqrt{k+1}} t^{(k+1)/2}\|u_n\|_{L^2(0,T;U)}.$$

Then (2.5) follows by induction.

In particular, E_n^k is a contraction for big k. Note that $y' + Ay \ni f_1(y) + f_2(y)u_n$ if and only if y is a fixed point of E_n: $y = E_n[y]$. For $n = 1, 2, \dots, \infty$ denote $y_n(\cdot) = y(\cdot, \hat{x}, u_n)$. Then by (2.5) y_n can be obtained as a limit of iterations, starting from the constant function 0

$$y_n = \lim_{k\to\infty} E_n^k[0].$$

We claim that

$$\|y_n - E_n^k[0]\|_{L^\infty(0,T;H)} \le \Big(T\|f_1(0)\| + C\sqrt{T}\|f_2(0)\| + \sup_{s\in[0,T]}\|S(s)\hat{x}\|\Big)\sum_{i=k}^{\infty}\tfrac{(L(T+C\sqrt{T}))^i}{\sqrt{i!}}. \tag{2.6}$$

From (2.5) for any $m > k$,

$$\begin{aligned}\|E_n^m[0] - E_n^k[0]\|_{L^\infty(0,T;H)} &\le \sum_{i=k}^{m-1} \|E_n^{i+1}[0] - E_n^i[0]\|_{L^\infty(0,T;H)} \\ &\le \|E_n[0]\|_{L^\infty(0,T;H)} \sum_{i=k}^{m-1} \tfrac{(L(T+C\sqrt{T}))^i}{\sqrt{i!}}.\end{aligned} \tag{2.7}$$

Since

$$\begin{aligned}\|E_n[0]\|_{L^\infty(0,T;H)} &\le \|E_n[0] - S(\cdot)\hat{x}\|_{L^\infty(0,T;H)} + \sup_{s\in[0,T]} \|S(s)\hat{x}\| \\ &\le \int_0^T \|f_1(0) + f_2(0)u_n(s)\|ds + \sup_{s\in[0,T]} \|S(s)\hat{x}\|,\end{aligned}$$

letting $m \to \infty$ in (2.7) and Hölder's inequality give (2.6).

We will prove that

$$\lim_{n\to\infty} \|y_n - y_\infty\|_{L^\infty(0,T;H)} = 0. \tag{2.8}$$

Since for every $k \ge 1$,

$$y_\infty - y_n = \big(y_\infty - E_\infty^k[0]\big) + \big(E_\infty^k[0] - E_n^k[0]\big) + \big(E_n^k[0] - y_n\big)$$

and from (2.6) $y_\infty - E_\infty^k[0]$ and $E_n^k[0] - y_n$ go to 0 in $L^\infty(0,T;H)$ as $k \to \infty$, uniformly in n, to prove (2.8) it is enough to show that

$$\lim_{n\to\infty} \|E_\infty^k[0] - E_n^k[0]\|_{L^\infty(0,T;H)} = 0 \quad \text{for every} \quad k \ge 1. \tag{2.9}$$

Since $u_n \to u_\infty$ weakly in $L^2(0,T;H)$, it follows that $f_1(0) + f_2(0)u_n \to f_1(0) + f_2(0)u_\infty$ weakly in $L^1(0,T;H)$, and then $E_n[0] \to E_\infty[0]$ uniformly on $[0,T]$ as $n \to \infty$ by the Baras theorem; see, e.g., Corollary 2.3.1 in [26]. Thus (2.9) holds for $k = 1$. Suppose that (2.9) is proved for $k \ge 1$. Then $f_i(E_n^k[0])$, $i = 1, 2$, uniformly converge to $f_i(E_\infty^k[0])$ and then $f_1(E_n^k[0]) + f_2(E_n^k[0])u_n \to f_1(E_\infty^k[0]) + f_2(E_\infty^k[0])u_\infty$ weakly in $L^1(0,T;H)$, and using the Baras theorem again gives (2.9) for $k+1$. Therefore, (2.9) is proved by induction, and consequently (2.8) is satisfied, and the proof of the proposition is complete, since, as observed above, (S) is equivalent to its version with $x_n = \hat{x}$. □

Remark 2.8. The result of Proposition 2.7 has independent interest. However, since the proof of the main result of this paper requires condition (2.1) on the vector field f, there are really two cases where we can apply Proposition 2.7. Either $f_2 \equiv B \in \mathcal{L}(U,H)$ is constant, or we choose our controls in the restricted set $L^2(0,+\infty;C)$, where C is convex, closed, and bounded in U; see Remark 2.1. Note also that the proof of Proposition 2.7 shows existence and uniqueness of mild solutions of $y' + Ay \ni f_1(y) + f_2(y)u$ for any maximal monotone operator A.

Remark 2.9. Let $\Phi\colon H \to \mathbb{R} \cup \{+\infty\}$ be a lower semicontinuous and convex function $\mathrm{dom}(\Phi) \ne \emptyset$. Then its subgradient $\partial\Phi$ is an example of a multivalued, maximal monotone operator. Note that for our purposes it is not restrictive to assume that $\Phi \ge 0$. In fact, if $(x_0, p_0) \in \partial\Phi$, then defining

$$\tilde{\Phi}(x) = \Phi(x) - \Phi(x_0) - \langle p_0, x - x_0\rangle,$$

we get $\tilde{\Phi}(x_0) = 0$ and $(x_0, 0) \in \partial\tilde{\Phi}$. In particular $\partial\tilde{\Phi} = \partial\Phi - p_0$ and $\tilde{\Phi} \geq 0$, since $\tilde{\Phi}$ is convex. Therefore, for $A = \partial\tilde{\Phi}$ the properties of the system (1.1) are unaffected.

Assume now that Φ satisfies the following coercivity condition, namely, $\mathrm{dom}(\Phi) \subset V \subset H$, where V is a Hilbert space compactly and algebraically imbedded into H, and there are two functions $\omega, \rho\colon [0, +\infty) \to [0, +\infty)$, ρ nondecreasing, such that

$$\omega(\|y\|_V) \leq \rho(\|y\|) + \Phi(y) \quad \text{and} \quad \lim_{r\to+\infty} \omega(r) = +\infty.$$

Then for all $\lambda \geq 0$ the sublevel set $\{y \in H\colon \|y\|^2 + \Phi(y) \leq \lambda\}$ is bounded in V and is hence compact in H. This means that Φ is of compact type, which is equivalent to saying that $-\partial\Phi$ generates a compact semigroup; see [26, Proposition 2.2.2].

The situation we just described appears in the so-called abstract parabolic variational inequalities, where we are given an operator in H of the form $A = (\tilde{A} + \partial\varphi) \cap H \times H$. The operator $\tilde{A}$ is supposed linear, bounded, and symmetric; $\tilde{A}\colon V \to V'$; $V \subset H \subset V'$; V is a Hilbert space compactly and algebraically imbedded into H; V dense in H. Moreover, $\tilde{A}$ satisfies for $a > 0$, $b \in \mathbb{R}$,

$$\langle \tilde{A}y, y\rangle \geq a\|y\|_V^2 + b\|y\|^2, \quad y \in V.$$

The function $\varphi\colon H \to \mathbb{R} \cup \{+\infty\}$ is supposed to be nonnegative, lower semicontinuous and convex, $\mathrm{dom}(\varphi) \neq \emptyset$. For simplicity of notation we suppose $\tilde{A}$ monotone (otherwise $b < 0$ and one has to consider $\tilde{A} - bI$ in the following). Then by defining $\Phi\colon H \to \mathbb{R} \cup \{+\infty\}$,

$$\Phi(y) = \begin{cases} \frac{1}{2}\langle \tilde{A}y, y\rangle + \varphi(y), & y \in V, \\ +\infty & y \in H\backslash V, \end{cases}$$

we get $\partial\Phi = \tilde{A} + \partial\varphi$ and we fall into the situation above, namely, the operator $-A = -\partial\Phi$ generates a compact semigroup, and Proposition 2.7 applies.

Remark 2.10. Other situations are known in the literature where our condition (S) is satisfied. We can consider systems of the form

$$y'(t) + Ay(t) \ni Bu(t),$$

where $-A$ generates a weakly equicontinuous semigroup, $B \in \mathcal{L}(U, H)$ is compact, and our set of admissible controls is restricted to $L^2(0, +\infty; B_R^U)$ for some $R > 0$. For the proof of this statement, see [26, Theorem 2.9.2]. See also Remark 2.1 and section 6 for an explicit example.

3. Viscosity solutions. Over the last decade substantial progress in the theory of Hamilton–Jacobi–Bellman–Isaacs equations in infinite-dimensional spaces has been made due to the introduction of the notion of viscosity solution. In particular, very general results have been obtained for equations with unbounded terms; see, e.g. [18], [24], [15], and [25].

We seek to solve the partial differential equation

$$\langle Ax, Du\rangle + F(x, u(x), Du) = 0 \quad \text{in} \quad \Omega, \tag{3.1}$$

where $\Omega \subseteq \overline{D(A)}$. Whenever the function u is regular enough, the derivative Du is understood in the sense of Fréchet and is identified with an element of H. Thus $F\colon \Omega \times \mathbb{R} \times H \to \mathbb{R}$ is appropriate. However, in general our solutions will not be smooth in the above sense and we need to relax the concept of solution. In case of a

bounded A (see [10], [11], [12]) the theory of viscosity solutions for (3.1) is not much different from the finite-dimensional one for which we refer the reader to [8]. For an unbounded A one additional difficulty is in the interpretation of the term $\langle A\hat{x}, Du(\hat{x})\rangle$ when $\hat{x} \notin D(A)$. The classical approach (see, e.g., work of Barbu and Da Prato [3]) is restricted to equations with convex and Lipschitz ingredients and linear A. Crandall and Lions [13], [14], [16] introduced a notion of viscosity solution suitable for a linear A and more general nonlinearities F. Ishii [18] deals with the case of $A = \partial h$, where h is convex and lower semicontinuous. The definition of solution suitable for (3.1) with no additional assumptions on A besides maximal monotonicity has been introduced by Tataru [24] and later refined by Crandall and Lions [15] and Tataru [25]. Here we will follow Crandall–Lions's approach [15].

As explained in [24], for $\varphi \in C^1(H)$ it is natural to interpret the unbounded term $\langle Ax, D\varphi\rangle$ in terms of the derivatives of φ along the trajectories of $S(t)$. This motivates the following definition.

DEFINITION 3.1. *For a function* $\Phi\colon \overline{D(A)} \to \mathbb{R}$ *and* $\hat{x} \in \overline{D(A)}$ *define*

$$D_A^-\Phi(\hat{x}) = \liminf_{\substack{x\to\hat{x}\\ h\downarrow 0}} \frac{\Phi(x) - \Phi\left(S(h)x\right)}{h} \quad \textit{and} \quad D_A^+\Phi(\hat{x}) = \limsup_{\substack{x\to\hat{x}\\ h\downarrow 0}} \frac{\Phi(x) - \Phi\left(S(h)x\right)}{h}.$$

We refer to [15] and [19] for basic properties of these operators. Next we introduce "test functions." $Lip(\Omega)$ will denote the space of all Lipschitz continuous functions on Ω. Given $\psi \in Lip(\Omega)$, $L(\psi)$ will denote its best Lipschitz constant. Hereafter we denote with P the projection of H onto $\overline{D(A)}$.

DEFINITION 3.2. *We will say that* $\Phi = \varphi + \psi \in C^1(H) + Lip(H)$ *is a subtest (supertest, respectively) function if*

$$\varphi(Px) \le \varphi(x) \quad \textit{and} \quad \psi(Px) \le \psi(x) \quad \textit{for} \quad x \in H, \tag{3.2}$$

$$\left(\varphi(Px) \ge \varphi(x) \quad \textit{and} \quad \psi(Px) \ge \psi(x) \quad \textit{for} \quad x \in H, \quad \textit{respectively}\right). \tag{3.3}$$

Remark 3.3. As explained in [15], the restrictions on ψ in (3.2) and (3.3) are made only for notational convenience. Since only the values of ψ on $\overline{D(A)}$ matter, one can always extend ψ from $\overline{D(A)}$ to all of H via $\psi(x) = \psi(Px)$ without increasing its Lipschitz constant to guarantee that (3.2) and (3.3) hold.

The notion of viscosity solution we are about to introduce is specific for equations of the form (1.6) but trivially adapts to include more general equations arising in optimal control and differential games. The technical reasons for introducing a new concept instead of following [15] exactly are presented in Remark 3.7 below along with comments explaining the relationship between this notion and the one due to Tataru and Crandall–Lions.

Given an extended real-valued function w on $\overline{D(A)}$, we denote by w^* and w_* its upper and lower semicontinuous envelopes, respectively; i.e., for $x \in \overline{D(A)}$,

$$w^*(x) = \limsup_{\overline{D(A)}\ni y\to x} w(y), \quad w_*(x) = \liminf_{\overline{D(A)}\ni y\to x} w(y).$$

DEFINITION 3.4. *Suppose that* (2.1) *holds and let* g *as in* (2.4) *be lower semicontinuous. Then* $w \in USC(\overline{D(A)})$ *(*$w \in LSC(\overline{D(A)})$*, respectively) is a viscosity subsolution (respectively, supersolution) of* (1.6) *if for every subtest (respectively, supertest) function* $\Phi = \varphi + \psi \in C^1(H) + Lip(H)$ *and a local maximum (respectively,*

minimum) $\hat{x} \in \mathrm{dom}(w)$ *of* $w - \Phi$ *relative to* $\overline{D(A)}$ *we have*

$$D_A^- \Phi(\hat{x}) + \sup_{u \in U} \left\{ -\langle f(\hat{x}, u), D\varphi(\hat{x}) \rangle - L(\psi) \| f(\hat{x}, u) \| - \tfrac{1}{2} \|u\|^2 \right\} \leq g^*(\hat{x}), \tag{3.4}$$

$$\left(D_A^+ \Phi(\hat{x}) + \sup_{u \in U} \left\{ -\langle f(\hat{x}, u), D\varphi(\hat{x}) \rangle + L(\psi) \| f(\hat{x}, u) \| - \tfrac{1}{2} \|u\|^2 \right\} \geq g(\hat{x}) \right). \tag{3.5}$$

A function w *(not necessarily continuous) defined on* $\overline{D(A)}$ *is a solution of* (1.6) *if* w^* *is a subsolution and* w_* *is a supersolution of* (1.6).

In the following we write $B_r(x)$ for the closed ball in H of radius r centered at x. Note that in both (3.4) and (3.5) one can replace $L(\psi)$ by a local Lipschitz constant for ψ. Indeed, suppose that ψ is $L-$Lipschitz on $B_r(\hat{x})$. Let Q denote the (orthogonal) projection onto $B_r(\hat{x}) \cap \overline{D(A)}$ and put $\tilde{\psi}(x) = \psi(QPx)$. Then $\tilde{\psi}(Px) = \tilde{\psi}(x)$ for all x, $\tilde{\psi}$ is $L-$Lipschitz on H and coincides with ψ near $\hat{x}$ on $\overline{D(A)}$, and therefore can be used in place of ψ. Also by modifying φ away from $\hat{x}$ using an appropriate cut-off technique one can achieve that it is not restrictive to suppose $\varphi \in Lip(H)$; this can be done without destroying the subtest (or supertest) property.

Remark 3.5. Note that if u is a solution of (1.6) then the supersolution condition (3.5) may have to be checked at all points $\hat{x} \in \mathrm{dom}(w_*)$, even for $\hat{x} \in \mathrm{dom}(w_*) \setminus \mathrm{dom}(g)$, while the subsolution condition (3.4) is meaningful only at points $\hat{x} \in \overline{D(A)}$ such that both w and g are locally bounded from above near $\hat{x}$ on $\overline{D(A)}$, since otherwise either $\hat{x} \notin \mathrm{dom}(w^*)$ or $g^*(\hat{x}) = +\infty$ and (3.4) is trivially satisfied. In particular, it follows that (3.4) trivializes unless $\hat{x} \in \mathrm{dom}(g^*) \subset \mathrm{int}(K)$ (the interior is taken with respect to $\overline{D(A)}$). This shows that in our problem, at least when K has an empty interior, the role of subsolutions is neither particularly meaningful nor helpful. We considered the Hamilton–Jacobi–Bellman equation (1.6) for our problem in the whole of $\overline{D(A)}$ only for notational convenience. As a matter of fact the supersolution condition really plays a role only in $K = \overline{\mathrm{dom}(g)}$, as the following result shows.

PROPOSITION 3.6. *Let* $\Omega \subset \overline{D(A)}$ *be relatively open and* $u \in LSC(\overline{D(A)})$ *be bounded from below. Then* u *is a supersolution of*

$$\langle Ax, Du \rangle = +\infty \quad \text{in} \quad \Omega \tag{3.6}$$

if and only if $u \equiv +\infty$ *in* Ω.

In the proof of Proposition 3.6 (and also in the proof of Lemma 5.5 below) we will use a perturbed optimization technique due to Tataru (see [25] and also Crandall–Lions [15]). For $x, y \in \overline{D(A)}$ define the Tataru distance d by

$$d(x, y) = \inf_{t \geq 0} \{ t + \|x - S(t)y\| \};$$

d is almost a metric (it lacks symmetry). In [25] Tataru proved a version of the classical Ekeland ε-variational principle with d in place of the norm. Subsequently, this optimization technique was successfully employed in proofs of comparison and uniqueness of viscosity solutions by Tataru and then Crandall and Lions; see [24], [15], [25], and the proof of Lemma 5.5.

Proof of Proposition 3.6. Of course we only need to show the necessary part. Suppose that $u(z) < +\infty$, $z \in \Omega$. Choose $r > 0$ such that $\overline{B_{2r}(z)} \cap \overline{D(A)} \subset \Omega$ and pick any $y \in B_r(z) \cap D(A)$. Since u is bounded from below, by choosing sufficiently large $M > 0$ we can guarantee that

$$g(z) + \sigma < \inf \{ g(x) \colon\ x \in \overline{D(A)},\ \|x - z\| = 2r \}, \tag{3.7}$$

where $\sigma > 0$ and $g(x) = u(x) + M\|x-y\|^2$ ($\inf \emptyset = +\infty$). For $\epsilon > 0$ use Tataru's perturbed optimization (see [25]) to find $\hat{x} \in \overline{B_{2r}(z)} \cap \overline{D(A)}$ such that the mapping $x \mapsto g(x) + \epsilon\, d(x, \hat{x})$ has at $\hat{x}$ finite minimum over $\overline{B_{2r}(z)} \cap \overline{D(A)}$. From (3.7) it follows that if ϵ is sufficiently small then $\|\hat{x} - z\| < 2r$ and, consequently, $u(x) + M\|x-y\|^2 + \epsilon\, d(x, \hat{x})$ has at $\hat{x}$ local minimum relative to Ω. Denoting $\Phi(x) = -M\|x-y\|^2 - \epsilon\, d(x, \hat{x})$, Φ is a supertest function and

$$D_A^+\Phi(\hat{x}) \le -2M\langle A^\circ y, \hat{x} - y\rangle + \epsilon < +\infty$$

(see Lemma 2.2 in [15]), which contradicts (3.6). □

Remark 3.7. We recall here the definition of solution employed by Crandall and Lions in [15] and compare it with ours. In order to interpret the term "$D\psi$" for merely Lipschitz ψ in the general case of (3.1), for $(x, r, p) \in \Omega \times \mathbb{R} \times H$ and $\lambda > 0$, they introduce the notation

$$F_\lambda(x, r, p) = \inf\,\{F(x, r, p + q):\ q \in H,\ \|q\| \le \lambda\}$$

and

$$F^\lambda(x, r, p) = \sup\,\{F(x, r, p + q):\ q \in H,\ \|q\| \le \lambda\}\,.$$

A function $u \in USC(\Omega)$ ($u \in LSC(\Omega)$, respectively) is a CL-viscosity subsolution (respectively, CL-supersolution) of (3.1) with a continuous F if for every subtest (respectively, supertest) function $\Phi = \varphi + \psi \in C^1(H) + Lip(H)$ and a local maximum (respectively, minimum) $\hat{x} \in \mathrm{dom}(u)$ of $u - \Phi$ relative to Ω, we have

$$D_A^-\Phi(\hat{x}) + F_{L(\psi)}(\hat{x}, u(\hat{x}), D\varphi(\hat{x})) \le 0,$$
$$\Big(D_A^+\Phi(\hat{x}) + F^{L(\psi)}(\hat{x}, u(\hat{x}), D\varphi(\hat{x})) \ge 0, \quad \text{respectively}\Big).$$

These definitions still make sense for extended real-valued F. In this case, as well as in the case of discontinuous F, the definitions of sub- and supersolutions have to be modified in a usual way by inserting appropriate semicontinuous envelopes (e.g., F_* for the subsolution condition and F^* for supersolutions).

Note that if $F(x, r, p) = \sup_{u \in U}\left\{-\langle f(x, u), p\rangle - \frac{1}{2}\|u\|^2\right\} - g(x)$ is as in (1.6) then both notions of supersolutions clearly coincide, while the notion of Crandall–Lions' subsolution is stronger than the one given in Definition 3.4. The reason for introducing a new concept of solution as in Definition 3.4 is to eliminate "error terms" which appear while proving that the value function of a control problem or a differential game solves the associated Hamilton–Jacobi equation; see, e.g., the proof in the next section and also [12], [24], and [19]. Such error terms can be eliminated by means of comparison principle if it holds for the equation under consideration; see [12] and [19] for the appropriate argument. Problems of type (1.6) that we study here fail to have unique solutions and we dispense with error terms by relaxing the notion of solution. We observe however that our notion applied to Hamiltonians satisfying typical conditions required to carry out the proof of uniqueness of viscosity solutions yields the same unique solution as the one in [15].

Next two auxiliary lemmas on change of variables will be used in the proof of the main result in section 5. For convenience, Lemmas 3.8 and 3.9 are stated for general Hamiltonians and Crandall–Lions's notion of supersolution. We will apply them to Hamiltonians as in (1.6), recalling that in this case, by Remark 3.7, the two notions of supersolutions as given in [15] and in Definition 3.4 are equivalent.

LEMMA 3.8. *Suppose that* $w:\Omega \to [0,+\infty]$ *is a CL-supersolution of*

$$\langle Ax, Dw\rangle + F(x,Dw) = 0 \quad \text{in} \quad \Omega, \tag{3.8}$$

where $\Omega \subseteq \overline{D(A)}$ *and* $F:\Omega \times H \to \mathbb{R}$ *is upper semicontinuous and locally uniformly continuous in its second variable. Then* $W(x) = 1 - e^{-w(x)}$ *is a CL-supersolution of*

$$\langle Ax, DW\rangle + \overline{F}(x,W(x),DW) = 0 \quad \text{in} \quad \Omega \cap \{x: W(x) < 1\} = \Omega \cap \mathrm{dom}(w), \tag{3.9}$$

where $\overline{F}(x,r,p) = (1-r)F\left(x, p/(1-r)\right)$.

Proof. For $r \in \mathbb{R} \cup \{+\infty\}$ put $\rho(r) = 1 - e^{-r}$, where $\rho(+\infty) = 1$; then $W = \rho(w)$ and $0 \le W \le 1$, but $W < 1$ on $\mathrm{dom}(w)$. Let $\Phi = \varphi + \psi$ be a supertest function and suppose that $W - \Phi$ has a local minimum (relative to $\Omega \cap \mathrm{dom}(w)$) at $\hat{x} \in \Omega \cap \mathrm{dom}(w)$, so $W(\hat{x}) < 1$. Without loss of generality we may assume that

$$\varphi(\hat{x}) = W(\hat{x}), \quad \psi(\hat{x}) = 0, \quad \varphi + \psi < 1, \quad \varphi < 1 \quad \text{on} \quad H. \tag{3.10}$$

Locally near $\hat{x}$ on $\Omega \cap \mathrm{dom}(w)$,

$$W \ge \varphi + \psi = \rho(\tilde{\varphi}) + e^{-\tilde{\varphi}} e^{\tilde{\varphi}} \psi = \rho\left(\tilde{\varphi} + \rho^{-1}\left(e^{\tilde{\varphi}}\psi\right)\right),$$

with equality at $\hat{x}$, where $\tilde{\varphi} = \rho^{-1}(\varphi) = -\ln(1-\varphi)$. It follows that $w - \tilde{\varphi} - \tilde{\psi}$ has a local minimum relative to Ω at $\hat{x}$, where $\tilde{\psi} = \rho^{-1}(e^{\tilde{\varphi}}\psi) = -\ln(1 - \frac{\psi}{1-\varphi})$. We will show that $\tilde{\varphi} + \tilde{\psi}$ is a supertest function. Clearly $\tilde{\varphi}(Px) \ge \tilde{\varphi}(x)$ for all x, $\tilde{\varphi} \in C^1(H)$ and $D\tilde{\varphi}(\hat{x}) = D\varphi(\hat{x})/(1 - W(\hat{x}))$. Formally (but strictly a.e. in every direction)

$$D\tilde{\psi}(x) = \frac{(1-\varphi(x))D\psi(x) + \psi(x)D\varphi(x)}{(1-\varphi(x))(1-\varphi(x)-\psi(x))} \to \frac{D\psi(\hat{x})}{1 - W(\hat{x})} \quad \text{as} \quad x \to \hat{x}.$$

Therefore,

$$L(\tilde{\psi}_{|B_r(\hat{x})}) \le \frac{L(\psi)}{1 - W(\hat{x})} + o(1) \quad \text{as} \quad r \downarrow 0,$$

and from (3.8) and the locally uniform continuity of H

$$D_A^+(\tilde{\varphi} + \tilde{\psi})(\hat{x}) + F^{\frac{L(\psi)}{1-W(\hat{x})}}\left(\hat{x}, D\varphi(\hat{x})/(1 - W(\hat{x}))\right) \ge 0. \tag{3.11}$$

Now note that for any $y \in \overline{D(A)}$ and $h > 0$, from the mean value theorem

$$\begin{aligned}
&\tilde{\varphi}(y) + \tilde{\psi}(y) - \tilde{\varphi}\left(S(h)y\right) - \tilde{\psi}\left(S(h)y\right) \\
&\quad = \ln\left(1 - \varphi\left(S(h)y\right) - \psi\left(S(h)y\right)\right) - \ln(1 - \varphi(y) - \psi(y)) \\
&\quad = \tfrac{1}{\alpha}\left(\varphi(y) + \psi(y) - \varphi\left(S(h)y\right) - \psi\left(S(h)y\right)\right)
\end{aligned}$$

for some number α between $1 - \varphi\left(S(h)y\right) - \psi\left(S(h)y\right)$ and $1 - \varphi(y) - \psi(y)$. From (3.10) $\alpha \to 1 - W(\hat{x})$ as $y \to \hat{x}$ and $h \downarrow 0$ and consequently

$$D_A^+(\tilde{\varphi} + \tilde{\psi})(\hat{x}) \le \frac{1}{1 - W(\hat{x})} D_A^+(\varphi + \psi)(\hat{x}).$$

Combining this with (3.11) gives

$$D_A^+(\varphi + \psi)(\hat{x}) + \overline{F}^{L(\psi)}(\hat{x}, W(\hat{x}), D\varphi(\hat{x})) \ge 0$$

and (3.9) follows. ∎

LEMMA 3.9. *Let $\Omega \subseteq \overline{D(A)}$ and $F\colon \overline{D(A)} \times [0,1] \times H \to \mathbb{R} \cup \{+\infty\}$ be upper semicontinuous. Suppose that $W\colon \Omega \to [0,1)$ is a CL-supersolution of*

$$\langle Ax, DW\rangle + F(x, W(x), DW) = 0 \quad in \quad \Omega.$$

Define

$$\widetilde{W}(x) = \begin{cases} W(x) & if \quad x \in \Omega, \\ 1 & if \quad x \in \overline{D(A)}\backslash\Omega. \end{cases}$$

If $F(x,1,0) \geq 0$ for $x \in \overline{D(A)}\backslash\Omega$ and $\widetilde{W} \in LSC(\overline{D(A)})$, then it is a CL-supersolution of

$$\langle Ax, D\widetilde{W}\rangle + F(x, \widetilde{W}(x), D\widetilde{W}) \geq 0 \quad in \quad \overline{D(A)}.$$

Proof. Suppose that $\widetilde{W} - \Phi$ has a local minimum relative to $\overline{D(A)}$ at $\hat{x} \in \overline{D(A)}$. If $\hat{x} \in \Omega$ there is nothing to show. Otherwise, since Φ is a supertest function, it follows that

$$\Phi(\hat{x}) - \Phi(x) \geq \Phi(\hat{x}) - \Phi(Px) \geq \widetilde{W}(\hat{x}) - \widetilde{W}(Px) = 1 - \widetilde{W}(Px) \geq 0,$$

provided $x \in H$ is sufficiently close to $\hat{x}$, so Φ has a local maximum relative to H at $\hat{x}$ and therefore $\|D\varphi(\hat{x})\| \leq L(\psi)$. Moreover,

$$D_A^+\Phi(\hat{x}) \geq \limsup_{h\downarrow 0} \frac{\Phi(\hat{x}) - \Phi\left(S(h)\hat{x}\right)}{h} \geq 0$$

and then

$$D_A^+\Phi(\hat{x}) + F^{L(\psi)}\left(\hat{x}, \widetilde{W}(\hat{x}), D\varphi(\hat{x})\right) \geq F(\hat{x}, 1, 0) \geq 0. \qquad \square$$

4. Value functions, dynamic programming principle and Hamilton–Jacobi equations. In this section we develop the dynamic programming approach for the value function in (1.3) and an auxiliary value function which will be helpful in the proof of the main result. Let f and g be as in (2.1) and (2.4), respectively. In the following for $t, \lambda > 0$, $x \in \overline{D(A)}$, $u \in L^2$, and $w\colon \overline{D(A)} \to \mathbb{R}$, we denote

$$J(t,x,u) = \int_0^t \left(g(y(s)) + \tfrac{1}{2}\|u(s)\|^2\right) ds$$

and

$$J^\lambda(t,x,u,w) = \int_0^t e^{-\lambda s - J(s,x,u)} \left(g(y(s)) + \tfrac{1}{2}\|u(s)\|^2\right) ds + e^{-\lambda t - J(t,x,u)} w(y(t)), \tag{4.1}$$

where $y(\cdot, x, u)$ is the trajectory of the system (1.1) corresponding to the choice of control $u(\cdot)$ and initial point $x \in \overline{D(A)}$.

We start considering the value function V given by (1.3),

$$V(x) = \inf_{u \in L^2(0,+\infty;U)} \int_0^\infty \left(g(y(t)) + \tfrac{1}{2}\|u(t)\|^2\right) dt$$

and improve the representation formula in (1.5).

LEMMA 4.1. *Assume that $V(x)$ is finite. Then there is $M_x > 0$ such that*

$$V(x) = \inf\{J(x,u):\ u \in \mathcal{U}(x),\ \|u\|_2 \le M_x\}.$$

Proof. The proof is immediate if we observe that by the assumption and the fact that g is nonnegative for all ϵ-optimal controls u for $V(x)$ with $\epsilon \le 1$, we have $\frac{1}{2}\|u\|_2^2 < V(x) + \epsilon$, and then we can take, e.g., $M_x = \sqrt{2(V(x)+1)}$. □

As a consequence the following dynamic programming principle holds.

LEMMA 4.2. *Assume that $V(x)$ is finite. Then there is a constant $M_x > 0$ such that the constrained value function V satisfies, for all $t \ge 0$,*

$$V(x) = \inf\{J(t,x,u) + V(y(t,x,u)):\ u \in \mathcal{U}(x),\ \|u\|_2 \le M_x\} \quad \textit{for} \quad x \in \mathrm{dom}(V).$$

Proof. If $x \in \mathrm{dom}(V)$, the proof of the fact that

$$V(x) = \inf_{u \in \mathcal{U}(x)} \{J(t,x,u) + V(y(t,x,u))\}$$

is completely standard and we skip it. Since V is nonnegative by definition, the rest of the statement follows by the same argument as in the proof of Lemma 4.1. □

We now proceed with the proof of Proposition 2.4. We start recalling a preliminary technical lemma, which is based on ideas of Tataru, see [24] and [25], but for this particular version of it we refer the reader to the paper by the authors and Święch [19].

LEMMA 4.3. *Let Φ be a subtest function, $\hat{x} \in \overline{D(A)}$, and $D_A^-\Phi(\hat{x}) < +\infty$. Then there exists a modulus ρ (i.e., $\rho\colon [0,+\infty) \to [0,+\infty)$ is continuous, nondecreasing and $\rho(0) = 0$) such that if $v \in L^1(0,t;H)$ and $y(\cdot)$ solves*

$$\begin{cases} y'(s) + Ay(s) \ni v(s) & \textit{for} \quad 0 \le s \le t, \\ y(0) = \hat{y}, \end{cases} \tag{4.2}$$

then

$$\begin{aligned} \Phi(y(t)) - \Phi(\hat{y}) \le -tD_A^-\Phi(\hat{x}) + \int_0^t \langle D\varphi(y(s)), v(s)\rangle ds + L(\psi)\int_0^t \|v(s)\|ds \\ + t\rho\Big(\int_0^t \|v(s)\|ds + \sup_{s\in[0,t]} \|S(t)\hat{x} - \hat{x}\|\Big) \end{aligned} \tag{4.3}$$

for all $0 \le t < \bar{t}$, uniform for v bounded in $L^1(0,\bar{t};H)$ and $\hat{y}$ sufficiently close to $\hat{x}$.

Observe that a supertest version of the previous lemma holds as well by replacing Φ with $-\Phi$. To make the proof of Proposition 2.4 below self-contained we don't assume the lower semicontinuity of V (Proposition 2.3), and this is being taken care of in the course of the proof by introducing V_*. The proof simplifies somewhat if Proposition 2.3 is proved first.

Proof of Proposition 2.4. 1. We start with the supersolution case. Suppose that $V_* - \Phi$ has a local minimum at $\hat{x} \in \mathrm{dom}(V_*) \subset K$ and $V_*(\hat{x}) = \Phi(\hat{x})$. We start proving that

$$D_A^+\Phi(\hat{x}) > -\infty. \tag{4.4}$$

As we mentioned above, it is not restrictive to assume that Φ is Lipschitz. Let $\varepsilon \in (0,1]$ be fixed and $h > 0$ be sufficiently small. Let $x_n \to \hat{x}$ be such that $V(x_n) \to V_*(\hat{x})$.

From the proof of the dynamic programming principle Lemma 4.2 we can find a control u_n^h with $\|u_n^h\|_{L^2(0,h)} \le M_{\hat{x}}$ depending only on $\hat{x}$ such that

$$\varepsilon h + V(x_n) \ge \int_0^h \left(g(y(s,x_n,u_n^h)) + \tfrac{1}{2}\|u_n^h(s)\|^2\right) ds + V(y(h,x_n,u_n^h)),$$

and then as in (2.2), since g is nonnegative,

$$\begin{aligned}
\tfrac{1}{2}\int_0^h \|u_n^h(s)\|^2 ds &\le V(x_n) - V_*(\hat{x}) + V_*(\hat{x}) - V_*(y(h,x_n,u_n^h)) + \varepsilon h \\
&\le V(x_n) - V_*(\hat{x}) + \Phi(\hat{x}) - \Phi(y(h,x_n,u_n^h)) + \varepsilon h \\
&\le V(x_n) - V_*(\hat{x}) + \Phi(\hat{x}) - \Phi(S(h)\hat{x}) + \epsilon h \\
&\quad + L(\Phi)\|S(h)\hat{x} - y(h,x_n,u_n^h)\| \\
&\le V(x_n) - V_*(\hat{x}) + \Phi(\hat{x}) - \Phi(S(h)\hat{x}) + C_{\hat{x}}h + L(\Phi)Lh^{1-q/2}\|u_n^h\|^q_{L^2(0,h)} \\
&\quad + L(\Phi)\|x_n - \hat{x}\|,
\end{aligned}$$

where we put $C_{\hat{x}}$ to emphasize the dependence on $\hat{x}$. Rearranging the terms we can find $M > 0$ depending only on L, $L(\Phi)$, and q such that

$$V(x_n) - V_*(\hat{x}) + L(\Phi)\|x_n - \hat{x}\| + \Phi(\hat{x}) - \Phi(S(h)\hat{x}) + C_{\hat{x}}h \ge h\left(\tfrac{1}{2}r_n^2 - L(\Phi)Lr_n^q\right) \ge -Mh,$$

where $r_n = \|u_n^h\|_{L^2(0,h)}/\sqrt{h}$, and then, letting $n \to +\infty$ first, (4.4) follows.

We now argue by contradiction and suppose that

$$D_A^+\Phi(\hat{x}) + \sup_{u\in U}\{-\langle f(\hat{x},u), D\varphi(\hat{x})\rangle + L(\psi)\|f(\hat{x},u)\| - \tfrac{1}{2}\|u\|^2\} - g(\hat{x}) < -2\theta < 0.$$

That is, for all $u \in L^2(0,+\infty;U)$ and $t \ge 0$ we have

$$D_A^+\Phi(\hat{x}) - \langle f(\hat{x},u(t)), D\varphi(\hat{x})\rangle + L(\psi)\|f(\hat{x},u(t))\| - \tfrac{1}{2}\|u(t)\|^2 - g(\hat{x}) < -2\theta < 0.$$

Observe that for $x_n \to \hat{x}$ and $V(x_n) \to V_*(\hat{x})$, by the proof of Lemma 4.2, the constants $M_{x_n}, M_{\hat{x}}$ can be chosen uniformly bounded in n, say, by M. From (2.2), (2.3), and the assumptions on g, for every sufficiently small $t > 0$ and all $u \in \mathcal{U}(x_n)$, $\|u\|_{L^2} \le M$, we then have

$$D_A^+\Phi(\hat{x}) - \langle f(y_n(s),u(s)), D\varphi(y_n(s))\rangle + L(\psi)\|f(y_n(s),u(s))\| - \tfrac{1}{2}\|u(s)\|^2 - g(y_n(s)) < -\theta$$

for all $s \in [0,t]$, where $y_n(\cdot) = y(\cdot,x_n,u)$. Integrating from 0 to t gives

$$\begin{aligned}
t\left(D_A^+\Phi(\hat{x}) + \theta\right) &\le \int_0^t \left(\langle f\left(y_n(s),u(s)\right), D\varphi(y_n(s))\rangle - L(\psi)\|f(y_n(s),u(s))\|\right) ds \\
&\quad + \int_0^t \left(g\left(y_n(s)\right) + \tfrac{1}{2}\|u(s)\|^2\right) ds
\end{aligned}$$

for every $u \in \mathcal{U}(x_n)$, $\|u\|_2 \le M$, and thus

$$\begin{aligned}
t\left(D_A^+\Phi(\hat{x}) + \theta\right) \le \inf_{u\in\mathcal{U}(x_n),\|u\|_2\le M} \Big\{ &\int_0^t \left(g\left(y_n(s)\right) + \tfrac{1}{2}\|u(s)\|^2\right) ds \\
&+ \int_0^t \left(\langle f\left(y_n(s),u(s)\right), D\varphi(y_n(s))\rangle - L(\psi)\|f(y_n(s),u(s))\|\right) ds \Big\}.
\end{aligned}$$

We now use the dynamic programming principle Lemma 4.2, Lemma 4.3 (its supertest version), and (2.3) and get, for $t > 0$ sufficiently small,

$$\begin{aligned} t\theta &\le \inf_{u\in\mathcal{U}(x_n),\|u\|_2\le M}\Big\{\int_0^t (g(y_n(s))+\tfrac{1}{2}\|u(s)\|^2)ds + \Phi(y_n(s)) - \Phi(x_n)\Big\} \\ &\quad + t\rho(o(1)+\|x_n-\hat{x}\|) \\ &\le \inf_{u\in\mathcal{U}(x_n),\|u\|_2\le M}\Big\{\int_0^t (g(y_n(s))+\tfrac{1}{2}\|u(s)\|^2)ds + V(y_n(s)) - V_*(\hat{x})\Big\} \\ &\quad + \Phi(\hat{x}) - \Phi(x_n) + t\rho(o(1)+\|x_n-\hat{x}\|) \\ &= V(x_n) - V_*(\hat{x}) + \Phi(\hat{x}) - \Phi(x_n) + t\rho(o(1)+\|x_n-\hat{x}\|), \end{aligned}$$

where $o(1)$ is independent of n. Hence a contradiction as $n \to +\infty$ first and $t \downarrow 0$ next.

2. We now turn to the proof that V is a subsolution. Let Φ be a Lipschitz continuous subtest function and assume that $V^* - \Phi$ attains a maximum at $\hat{x} \in \mathrm{dom}(V^*)$; of course this implies that $\hat{x} \in \mathrm{int}(\mathrm{dom}(V^*)) \subset \mathrm{int}(K)$ as V is nonnegative. Moreover, it is not restrictive to assume $V^*(\hat{x}) = \Phi(\hat{x})$.

First we will prove that $D_A^-\Phi(\hat{x}) < +\infty$. Choose $x_n \to \hat{x}$ such that $V(x_n) \to V^*(\hat{x})$. We may also suppose that $\hat{x} \in \mathrm{dom}(g^*)$, as otherwise (3.4) is automatically satisfied. Then for sufficiently small $t > 0$ and large n, the control $u(s) \equiv 0$ for $s \in [0,t]$ and suitably defined afterward is an admissible element of $\mathcal{U}(x_n)$. Note to this end that V is locally bounded at $\hat{x}$; hence $\mathcal{U}(x) \neq \emptyset$ for x close to $\hat{x}$. From $\hat{x} \in \mathrm{dom}(g^*)$ and (2.3) it follows that $g(y(s,x_n,u))$ stay bounded uniformly in t small and n big. Denote this upper bound by $C_{\hat{x}}$; then by Lemma 4.2 we have

$$V(x_n) - V(y(t,x_n,u)) \le \int_0^t g(y(s,x_n,u))ds \le C_{\hat{x}}t.$$

Letting $n \to \infty$ and proceeding as in (2.2) gives, for t sufficiently small,

$$\begin{aligned} C_{\hat{x}}t &\ge V^*(\hat{x}) - V^*(y(t,\hat{x},u)) \ge \Phi(\hat{x}) - \Phi(y(t,\hat{x},u)) \\ &\ge \Phi(\hat{x}) - \Phi(S(t)\hat{x}) - L(\Phi)b\|y(t,\hat{x},u) - S(t)\hat{x}\| \ge \Phi(\hat{x}) - \Phi(S(t)\hat{x}) - L(\Phi)Ct, \end{aligned}$$

with some constant $C > 0$, which implies that $D_A^-\Phi(\hat{x}) < +\infty$.

We now again argue by contradiction and suppose that

$$D_A^-\Phi(\hat{x}) + \sup_{u\in U}\big\{-\langle f(\hat{x},u), D\varphi(\hat{x})\rangle - L(\psi)\|f(\hat{x},u)\| - \tfrac{1}{2}\|u\|^2\big\} - g^*(\hat{x}) > 2\theta > 0.$$

Then there is $u^* \in U$ such that

$$D_A^-\Phi(\hat{x}) - \langle f(\hat{x},u^*), D\varphi(\hat{x})\rangle - L(\psi)\|f(\hat{x},u^*)\| - g^*(\hat{x}) - \tfrac{1}{2}\|u^*\|^2 > 2\theta.$$

Let $x_n \to \hat{x}$ be such that $V(x_n) \to V^*(\hat{x})$. By the proof of Lemma 4.2 we can choose the constants $M_{x_n}, M_{\hat{x}}$ uniformly bounded, say, by M. Note that from $\hat{x} \in \mathrm{int}(\mathrm{dom}(V^*))$ and by (2.3), for a sufficiently small $t > 0$ and then for n large, we can find an admissible control $u^t \in \mathcal{U}(x_n)$ satisfying $u^t(s) = u^*$ for $s \in [0,t]$ (we can suitably extend it outside $[0,t]$). Denoting $y_n(\cdot) = y(\cdot,x_n,u^t)$ from (2.1) and (2.3) there exists $t > 0$ such that, for $s \in [0,t]$ and large n,

$$\begin{aligned} D_A^-\Phi(\hat{x}) &- \langle f(y_n(s),u^t(s)), D\varphi(y_n(s))\rangle \\ &- L(\psi)\|f(y_n(s),u^t(s))\| - g(y_n(s)) - \tfrac{1}{2}\|u^t(s)\|^2 > \theta. \end{aligned} \tag{4.5}$$

Integrating from 0 to t gives

$$t\left(D_A^-\Phi(\hat{x})-\theta\right)\geq\int_0^t\left(\langle f(y_n(s),u^t(s)),D\varphi(y_n(s))\rangle+L(\psi)\|f(y_n(s),u^t(s))\|\right)ds + \int_0^t\left(g(y_n(s))+\tfrac{1}{2}\|u^t(s)\|^2\right)ds$$

and consequently, denoting $y(\cdot)=y(\cdot,x_n,u)$ for a general $u\in\mathcal{U}(x_n)$,

$$t\left(D_A^-\Phi(\hat{x})-\theta\right)\geq\inf_{u\in\mathcal{U}(x_n),\|u\|_2\leq M}\Big\{\int_0^t\left(\langle f(y(s),u(s)),D\varphi(y(s))\rangle + L(\psi)\|f(y(s),u(s))\|\right)ds+\int_0^t\left(g(y(s))+\tfrac{1}{2}\|u(s)\|^2\right)ds\Big\}. \tag{4.6}$$

From this, using Lemma 4.3, we get

$$\begin{aligned}-t\theta&\geq\inf_{u\in\mathcal{U}(x_n),\|u\|_2\leq M}\Big\{\int_0^t\left(g(y(s))+\tfrac{1}{2}\|u(s)\|^2\right)ds+\Phi(y(t))-\Phi(x_n)\Big\}\\&\quad-t\rho_{\hat{x}}(o(1)+\|x_n-\hat{x}\|)\\&\geq\inf_{u\in\mathcal{U}(x_n),\|u\|_2\leq M}\Big\{\int_0^t\left(g(y(s))+\tfrac{1}{2}\|u(s)\|^2\right)ds+V^*(y(t))\Big\}-V^*(\hat{x})\\&\quad+\Phi(\hat{x})-\Phi(x_n)-t\rho_{\hat{x}}(o(1)+\|x_n-\hat{x}\|)\\&\geq V(x_n)-V^*(\hat{x})+\Phi(\hat{x})-\Phi(x_n)-t\rho_{\hat{x}}(o(1)+\|x_n-\hat{x}\|),\end{aligned}$$

where $o(1)$ does not depend on n as $t\downarrow 0$. Letting $n\to+\infty$ we then obtain $-t\theta\geq o(t)$, and this finally leads to a contradiction. ∎

In the rest of this section, we proceed with some results concerning an auxiliary problem we will need in the proof of Proposition 2.5, and for $\lambda>0$ and $w\colon\overline{D(A)}\to\mathbb{R}$ consider the value function

$$v^\lambda(x)=\inf_{u\in L^2}\sup_{t\geq 0}J^\lambda(t,x,u,w)\quad\text{for}\quad x\in\overline{D(A)}. \tag{4.7}$$

The function v^λ is known as the value of a stopping time control problem with stopping cost w.

LEMMA 4.4. *Let $g,w\colon\overline{D(A)}\to\mathbb{R}$ satisfy $0\leq g\leq C$ and $0\leq w\leq 1-\epsilon$ for some $\epsilon,C>0$. Then for every $\lambda>0$ we have $0\leq w\leq v^\lambda\leq(C/(\lambda+C))\wedge(1-\epsilon)<1$.*

Proof. Taking $t=0$ shows that $v^\lambda\geq w$. On the other hand, with the choice of $u(\cdot)\equiv 0$ we get

$$\begin{aligned}v^\lambda(x)&\leq\sup_{t\geq 0}\Big\{1-e^{-\lambda t-\int_0^t g(y(s))ds}(1-w(y(t)))-\lambda\int_0^t e^{-\lambda s-\int_0^s g(y(\sigma))d\sigma}ds\Big\}\\&\leq\sup_{t\geq 0}\Big\{1-\epsilon e^{-\lambda t-\int_0^t g(y(s))ds}-\lambda\int_0^t e^{-(\lambda+C)s}ds\Big\}\\&\leq\sup_{t\geq 0}\Big\{1-\epsilon e^{-(\lambda+C)t}+\tfrac{\lambda}{\lambda+C}\left(e^{-(\lambda+C)t}-1\right)\Big\},\end{aligned}$$

and then we easily reach the conclusion. ∎

LEMMA 4.5. *Under the assumptions and using the notation of Lemma* 4.4*, for a fixed* $\lambda > 0$ *there are* $T, K > 0$ *independent of* $x \in \overline{D(A)}$ *such that*

$$v^\lambda(x) = \inf_{\|u\|_{L^2(0,T)} \le K} \sup_{t \ge 0} J^\lambda(t, x, u, w).$$

Proof. By definition of v^λ, since g and w are nonnegative, we immediately obtain

$$v^\lambda(x) \ge \inf_{u \in L^2} \sup_{t \ge 0} e^{-\lambda t} \rho\Big(\int_0^t \tfrac{1}{2} \|u(s)\|^2 ds \Big),$$

where we recall that $\rho(r) = 1 - e^{-r}$. Therefore, for any ϵ-optimal control u for $v^\lambda(x)$ with $\epsilon \le \bar{\epsilon}$ independent of x, from Lemma 4.4 we have that

$$1 > M \ge v^\lambda(x) + \epsilon \ge e^{-\lambda t} \rho\Big(\int_0^t \tfrac{1}{2} \|u(s)\|^2 ds \Big) \quad \text{for all} \quad t \ge 0.$$

Therefore, if $T > 0$ is sufficiently small we get

$$\rho\Big(\int_0^t \tfrac{1}{2} \|u(s)\|^2 ds \Big) \le M e^{\lambda t} \le M e^{\lambda T} < 1 \quad \text{for all} \quad t \le T$$

and the conclusion by definition of ρ. ∎

As a consequence of the previous two lemmas, the following dynamic programming principle holds for the auxiliary value function v^λ defined by (4.7).

LEMMA 4.6. *Under the assumptions and using the notation of Lemma* 4.4*, for all* $x \in \overline{D(A)}$ *and* $t \ge 0$ *we have that*

$$v^\lambda(x) \ge \inf_{\|u\|_{L^2(0,T) \le K}} J^\lambda(t, x, u, v^\lambda). \tag{4.8}$$

If moreover $(v^\lambda)^*(x) > w^*(x)$*, there is* $\epsilon > 0$ *such that for* $z \in \overline{D(A)}$*,* $\|z - x\| < \epsilon$*, and* $|v^\lambda(z) - (v^\lambda)^*(x)| < \epsilon$*, we have*

$$v^\lambda(z) = \inf_{\|u\|_{L^2(0,T) \le K}} J^\lambda(t, z, u, v^\lambda) \quad \textit{for every } 0 \le t \le \epsilon.$$

Proof. The first part of the statement follows easily from the definition of $v^\lambda(x)$, Lemma 4.5, and the usual dynamic programming principle arguments, and we skip its proof.

We now assume that $x \in \overline{D(A)}$ and $(v^\lambda)^*(x) > w^*(x)$. If the statement was false, we could find sequences $x_n \in \overline{D(A)}$ and $0 < \delta_n, t_n < 1/n$ such that $\|x_n - x\| < 1/n$, $|v^\lambda(x_n) - (v^\lambda)^*(x)| < 1/n$, and

$$v^\lambda(x_n) > \inf_{\|u\|_{L^2(0,T) \le K}} J^\lambda(t_n, x_n, u, v^\lambda) + \delta_n. \tag{4.9}$$

By definition of $v^\lambda(x_n)$, for large n, $1/n \le T$, we can choose a control $u_n \in L^2$, $\|u_n\|_{L^2(0,T)} \le K$, such that

$$J^\lambda(t_n, x_n, u_n, v^\lambda) < v^\lambda(x_n) - \delta_n < \sup_{t \ge 0} J^\lambda(t, x_n, u_n, w),$$

and then using the definition of v^λ and J^λ, we can modify the control u_n off $[0,t_n]$ (we still use the same notation however) in such a way that by a change of variables in the integrals on the left-hand side we get

$$\sup_{t\geq t_n} J^\lambda(t,x_n,u_n,w) < v^\lambda(x_n)-\delta_n < \sup_{t\geq 0} J^\lambda(t,x_n,u_n,w).$$

Hence there is a sequence $s_n\in[0,t_n]$ such that

$$J^\lambda(s_n,x_n,u_n,w)\geq v^\lambda(x_n)-\delta_n. \tag{4.10}$$

By the uniform L^2 estimate on the controls u_n, i.e., $\|u_n\|_{L^2(0,s_n)}\leq K$, the definition of J^λ and by (2.3), we know that

$$\sup_{t\in[0,s_n]}\|y(t,x_n,u_n)-x\|\leq o(1)\quad\text{as}\quad n\to+\infty.$$

Therefore, as $n\to\infty$ in (4.10) we obtain

$$w^*(x)\geq (v^\lambda)^*(x),$$

and then we have a contradiction. □

Given Lemma 4.6, with a proof similar to the one of Proposition 2.4 we can show the following result.

PROPOSITION 4.7. *Under the assumptions of Lemma* 4.4*, if* $w,g\in LSC(\overline{D(A)})$*, for any* $\lambda>0$ *the value function* v^λ *is a viscosity solution of*

$$\begin{aligned}\lambda v^\lambda(x)+\min\Big\{&\langle Ax,Dv^\lambda\rangle+\sup_{u\in U}\left\{-\langle f(x,u),Dv^\lambda\rangle+(v^\lambda(x)-1)(\tfrac12\|u\|^2+g(x))\right\},\\ &v^\lambda(x)-(1+\lambda)w\Big\}=0\quad in\quad \overline{D(A)},\end{aligned} \tag{4.11}$$

where solutions of (4.11) *are defined by adapting Definition* 3.4 *in an obvious way.*

It is well known that stopping time control problems give rise to "quasi-variational inequalities" of the form (4.11); see [23] and the references therein.

5. Regularity and optimality principle. In the course of the proof of our main result we will use inf-convolutions to regularize various functions, including g. For $h\colon H\to\mathbb{R}\cup\{+\infty\}$ and $\epsilon>0$ put

$$h_\epsilon(x)=\inf_{y\in H}\left\{h(y)+\tfrac{1}{2\epsilon}\|x-y\|^2\right\}.$$

If h is lower semicontinuous and bounded from below then h_ϵ converge to h pointwise from below as $\epsilon\downarrow 0$. It is also known that if h is bounded or uniformly continuous then h_ϵ is globally Lipschitz. Moreover, if h is weakly lower semicontinuous then so is h_ϵ; see [9]. Therefore, combining inf-convolutions with an appropriate cut-off technique, for any bounded from below, (weakly) lower semicontinuous function $h\colon\overline{D(A)}\to\mathbb{R}\cup\{+\infty\}$ one can construct a sequence $h_1\leq h_2\leq\cdots\leq h_n\leq\cdots$ of bounded, globally Lipschitz (weakly) lower semicontinuous functions on H such that $h=\sup_n h_n$ on $\overline{D(A)}$.

In what follows we will frequently rely on the following well-known simple fact. X is going to be H equipped with either weak or strong topology.

LEMMA 5.1. *Suppose that X is a topological space and let $\phi, \phi_n : X \to \mathbb{R} \cup \{+\infty\}$ be sequentially lower semicontinuous. If $\phi_1 \le \phi_2 \le \cdots \le \phi_n \le \cdots$ and $\phi = \sup_n \phi_n$ then*

$$\liminf_{n\to\infty} \phi_n(x_n) \ge \phi(x) \quad \text{whenever} \quad x_n \to x \quad \text{in} \quad X. \tag{5.1}$$

Proof. For every $N \ge 1$,

$$\liminf_{n\to\infty} \phi_n(x_n) \ge \liminf_{n\to\infty} \phi_N(x_n) \ge \phi_N(x),$$

and taking a supremum over N gives (5.1). ∎

The following lemma guarantees the existence of optimal controls in the problems we consider and relies on one of the crucial assumptions (W) or (S) (see also Theorem 3.4 in [17]).

LEMMA 5.2. *Assume that* (1.1) *has a unique mild solution for any $u \in L^2(0,\infty;U)$ and $x \in \overline{D(A)}$. Suppose that $g, g_n, \phi, \phi_n \in$ w-LSC($\overline{D(A)}$) (LSC($\overline{D(A)}$), respectively), $0 \le \phi_1 \le \phi_2 \le \cdots \le \phi_n \le \cdots$, $0 \le g_1 \le g_2 \le \cdots \le g_n \le \cdots$, $\sup_n \phi_n = \phi$ and $\sup_n g_n = g$. If* (W) ((S), *respectively*) *holds and*

$$C \equiv \sup_{n\ge 1} \inf_{u\in L^2} \sup_{t\ge 0} \left\{ \int_0^t \left(g_n(y(s)) + \tfrac{1}{2}\|u(s)\|^2 \right) ds + \phi_n(y(t)) \right\},$$

where $y(\cdot) = y(\cdot, x, u)$ is the mild solution of (1.1) *and $x \in \overline{D(A)}$ is fixed, then there exists $u^\# \in L^2$ such that*

$$C = \sup_{t\ge 0} \left\{ \int_0^t \left(g(y(s,x,u^\#)) + \tfrac{1}{2}\|u^\#(s)\|^2 \right) ds + \phi(y(t,x,u^\#)) \right\}. \tag{5.2}$$

Proof. It is clear by definition that C is smaller than the right-hand side of (5.2) for any choice of $u^\# \in L^2$. To prove the opposite inequality, we may assume that $C < +\infty$. For every n there exists $u_n \in L^2$ such that for every $t \ge 0$

$$C + \tfrac{1}{n} \ge \int_0^t \left(g_n(y_n(s)) + \tfrac{1}{2}\|u_n(s)\|^2 \right) ds + \phi_n(y_n(t)) \ge \int_0^t \tfrac{1}{2}\|u_n(s)\|^2 ds.$$

Hence u_n are uniformly bounded in L^2, and one can find $u^\# \in L^2$ such that $u_n \rightharpoonup u^\#$ weakly in $L^2(0,T;U)$ for every $T > 0$ (passing to a subsequence if necessary). By the stability assumption $y_n(s) \equiv y(s,x,u_n)$ converges to $y^\#(s) \equiv y(s,x,u^\#)$ weakly (respectively, strongly) in H for every $s > 0$. Using the lower semicontinuity of ϕ_n's, g_n's, and the norm, from Lemma 5.1 and Fatou's lemma we deduce

$$\begin{aligned} \liminf_{n\to\infty} &\int_0^t \left(g_n(y_n(s)) + \tfrac{1}{2}\|u_n(s)\|^2 \right) ds + \phi_n(y_n(t)) \\ &\ge \int_0^t \left(g(y^\#(s)) + \tfrac{1}{2}\|u^\#(s)\|^2 \right) ds + \phi(y^\#(t)) \end{aligned}$$

for every $t \ge 0$ and the result follows. ∎

COROLLARY 5.3. *Under the assumptions of Proposition* 2.3 *the value function V in* (1.5), *i.e.,*

$$V(x) = \inf_{u\in L^2} \int_0^\infty \left(g(y(s)) + \tfrac{1}{2}\|u(s)\|^2 \right) ds = \inf_{u\in L^2} \sup_{t\ge 0} \left\{ \int_0^t \left(g(s) + \tfrac{1}{2}\|u(s)\|^2 \right) ds \right\}$$

has optimal controls.

Proof. Apply Lemma 5.2 with $g_n \equiv g$ and $\phi_n \equiv 0$. ∎

Given existence of optimal controls, we show that V is lower semicontinuous.

Proof of Proposition 2.3. Suppose for instance that $g \in w\text{-}LSC(\overline{D(A)})$ and that (W) holds (the other case follows similarly). Let $x_n \rightharpoonup x$, $x_n, x \in \overline{D(A)}$. We may assume that $\liminf_{n\to\infty} V(x_n) < +\infty$, and, passing to a subsequence, that $V(x_n)$ converges to $\liminf_{n\to\infty} V(x_n)$. From Corollary 5.3 for every n there exists $u_n \in L^2$ an optimal control for $V(x_n)$; that is,

$$V(x_n) \geq \int_0^t \left(g(y(s, x_n, u_n)) + \tfrac{1}{2}\|u_n(s)\|^2\right) ds \quad \text{for every} \quad t \geq 0. \tag{5.3}$$

Then $\|u_n\|_2$ are uniformly bounded and there is $u^\# \in L^2$ such that $u_n \rightharpoonup u^\#$ weakly in $L^2(0,T;U)$ for all $T > 0$ (passing to a subsequence if necessary), and then by (W) $y(s, x_n, u_n) \rightharpoonup y^\#(s) \equiv y(s, x, u^\#)$ weakly in H for every $s > 0$. Taking $\liminf$ as $n \to \infty$ in (5.3) yields, as in the proof of Lemma 5.2,

$$\liminf_{n\to\infty} V(x_n) \geq \int_0^t \left(g(y^\#(s)) + \tfrac{1}{2}\|u^\#(s)\|^2\right) ds$$

for every $t > 0$ and therefore $\liminf_{n\to\infty} V(x_n) \geq V(x)$. ∎

Proposition 2.5 is contained in the statement of Lemma 5.5 below, which is an optimality principle for viscosity supersolutions of equation (1.6). First, however, we prove the following elementary statement.

LEMMA 5.4. *Let $F\colon \overline{D(A)} \times H \times \mathbb{R} \times [0,1) \to \mathbb{R}$ be defined by*

$$F(x,p,r,s) = \sup_{u\in U} \left\{-\langle f(x,u), p\rangle + \|f(x,u)\| r - \tfrac{1}{2}\|u\|^2(1-s)\right\},$$

where f satisfies (2.1). Then for all $R > 0$ and $k \in (0,1)$ there is $C = C(R,k) > 0$ such that if $x \in \overline{D(A)}$, $\|p\|, |r| \leq R$ and $s \in [0, 1-k]$ then $\|u^*\| \leq C(1 + \sqrt{\|x\|})$ for any $u^* \in U$ satisfying

$$F(x,p,r,s) - 1 \leq -\langle f(x,u^*), p\rangle + \|f(x,u^*)\| r - \tfrac{1}{2}\|u^*\|^2(1-s).$$

Proof. Observe that by the assumption (2.1), $F(x,p,r,s)$ is finite for all $(x,p,r,s) \in \overline{D(A)} \times H \times \mathbb{R} \times [0,1)$. From one side, by choosing $u = 0$ we have by (2.1)

$$F(x,p,r,s) - 1 \geq -1 - L(1 + \|x\|)(\|p\| + |r|) \geq -1 - 2LR(1 + \|x\|),$$

for $x \in \overline{D(A)}$, $\|p\|, |r| \leq R$, and $s \in [0,1)$. On the other hand, if $x \in \overline{D(A)}$, $\|p\|, |r| \leq R$, and $s \in [0, 1-k]$ then, again by (2.1), we have for all $u \in U$,

$$\begin{aligned} -\langle f(x,u), p\rangle + \|f(x,u)\| r - \tfrac{1}{2}\|u\|^2(1-s) &\leq L(1 + \|x\| + \|u\|^q)(\|p\| + |r|) - \tfrac{1}{2}k\|u\|^2 \\ &\leq 2LR(1 + \|x\| + \|u\|^q) - \tfrac{1}{2}k\|u\|^2. \end{aligned}$$

Therefore, to reach the conclusion it is enough to consider only u's such that

$$\|u\|^q - \tfrac{k}{4LR}\|u\|^2 \geq -2(1 + \|x\|) - \tfrac{1}{2LR}. \tag{5.4}$$

We will use the elementary fact that for every $q \in [1,2)$ there is a constant C depending only on q (namely, $C = (\frac{q}{2})^{q/(2-q)} - (\frac{q}{2})^{2/(2-q)}$) such that

$$s^q \leq \epsilon s^2 + C\left(\tfrac{1}{\epsilon}\right)^{q/(2-q)} \quad \text{for all} \quad \epsilon > 0,\ s \geq 0.$$

(One quick way to prove this is to realize that it is equivalent to proving $\psi(s\epsilon^{1/(2-q)}) \geq -C$ for $\psi(r) = r^2 - r^q$.) Taking $\epsilon = \frac{k}{8LR}$ from (5.4) it follows that only u's satisfying

$$-\tfrac{1}{2LR} - 2(1 + \|x\|) \leq -\tfrac{k}{8LR}\|u\|^2 + C(\tfrac{8LR}{k})^{q/(2-q)}$$

or equivalently

$$\|u\|^2 \leq \tfrac{16LR}{k}(1 + \|x\|) + C(\tfrac{8LR}{k})^{2/(2-q)} + \tfrac{4}{k}$$

are of interest, from which the conclusion follows easily. □

We are now left with the most delicate step of the proof.

LEMMA 5.5. *Assume* (2.1) *and* (2.4). *Suppose that* $g \in w\text{-}LSC(\overline{D(A)})$ ($g \in LSC(\overline{D(A)})$) *and* (W) ((S), *respectively*) *holds. Suppose that* $w \in w\text{-}LSC(\overline{D(A)})$ ($w \in LSC(\overline{D(A)})$, *respectively*) *is a nonnegative extended real-valued viscosity supersolution of* (1.6). *Then*

$$w(x) = \inf_{u\in L^2} \sup_{t\geq 0} \left\{ \int_0^t \left(g(y(s)) + \tfrac{1}{2}\|u(s)\|^2\right) ds + w(y(t)) \right\} \quad \text{for all} \quad x \in \overline{D(A)}. \tag{5.5}$$

In particular $w \geq V$.

Proof. We treat the weakly lower semicontinuous case, the other case being similar.

1. Suppose that $w \in w\text{-}LSC(\overline{D(A)})$ is a nonnegative supersolution of (1.6). Construct two increasing sequences $0 \leq g_1 \leq g_2 \leq \cdots$ and $0 \leq w_1 \leq w_2 \leq \cdots$ of bounded, globally Lipschitz and weakly lower semicontinuous functions defined on H such that on $\overline{D(A)}$ $g = \sup_n g_n$ and $w = \sup_n w_n$, as at the beginning of this section. For every n put $W_n(x) = 1 - e^{-w_n(x)} \equiv \rho(w_n(x))$; $0 \leq W_n < 1$. Note that from (2.1) the Hamiltonian $H_n(x,p) = \sup_{u\in U}\{-\langle f(x,u),p\rangle - \frac{1}{2}\|u\|^2\} - g_n(x)$ is uniformly continuous on bounded subsets of $\overline{D(A)} \times H$ and therefore from Lemma 3.8 for every $\lambda > 0$ and $n \in \mathbb{N}$, $W = \rho(w)$ is a CL-supersolution of

$$\lambda W(x) + \langle Ax, DW\rangle + \sup_{u\in U}\left\{-\langle f(x,u), DW\rangle + (W(x)-1)(g_n(x) + \tfrac{1}{2}\|u\|^2)\right\} = 0$$

on $\{x \in \overline{D(A)}\colon W(x) < 1\} = \mathrm{dom}(w)$. We extend W by 1 off $\mathrm{dom}(w)$ (we still call this extended function W). Since $W \in w\text{-}LSC(\overline{D(A)})$, by Lemma 3.9 W is a CL-supersolution of

$$\lambda W(x) + \min\left\{\langle Ax, DW\rangle + \overline{H}_n(x, W(x), DW),\ W(x) - (1+\lambda)W_n(x)\right\} = 0 \quad \text{in} \quad \overline{D(A)}, \tag{5.6}$$

where for $(x,r,p) \in \overline{D(A)} \times [0,1] \times H$ we denote

$$\overline{H}_n(x,r,p) = \sup_{u\in U}\left\{-\langle f(x,u),p\rangle + \tfrac{1}{2}(r-1)\|u\|^2\right\} + (r-1)g_n(x).$$

Observe that $\overline{H}_n(x,1,p)$ may be infinite, but $\overline{H}_n(x,1,0) = 0$ so that $\overline{H}_n^*(x,1,0) \geq 0$ and the proof of Lemma 3.9 applies. Also note that from Lemma 5.4 one can easily show that $\overline{H}_n$ is uniformly continuous on bounded closed subsets of $\overline{D(A)} \times [0,1) \times H$. By Remark 3.7, W is a supersolution of (5.6) in the sense of Definition 3.4.

2. For $n \in \mathbb{N}$ and $x \in \overline{D(A)}$ let $U(x) = \inf_{u\in L^2} \sup_{t\geq 0} J_n^\lambda(t,x,u,W_n)$ be as in (4.7) and (4.1) with $g \equiv g_n$. By Lemma 4.4 there is $\kappa > 0$ such that $0 \leq U \leq 1-2\kappa$. Moreover, U^* is a subsolution of (5.6) by Proposition 4.7. We will show that

$$U^* \leq W \quad \text{on} \quad \overline{D(A)}. \tag{5.7}$$

The proof of (5.7) follows along the lines of the standard comparison theorem despite the Hamiltonian in (5.6) being possibly discontinuous and extended real valued, and we will only highlight the main points. To show (5.7) we argue by contradiction and suppose that $U^*(\hat z) - W(\hat z) \equiv 2\gamma > 0$ for some $\hat z \in \overline{D(A)}$.

We first make the following general remark. Let $\Psi\colon H\times H \to \mathbb{R}$, $\Psi = \varphi + \psi$, be a nonnegative subtest function for the operator $A \times A$ on $\overline{D(A)} \times \overline{D(A)}$. Assume that $U^*(x) - W(y) - \Psi(x,y)$ attains a maximum point at $(\hat x, \hat y)$ and $U^*(\hat x) - W(\hat y) - \Psi(\hat x, \hat y) > 0$. In particular we get

$$0 \leq W(\hat y) \leq U^*(\hat x) \leq 1 - 2\kappa.$$

Hence $(\hat y, W(\hat y)) \in \operatorname{dom}(w) \times [0, 1-2\kappa]$ and at $\hat x$, $\hat y$ we can use the equations for U^*, W, respectively, and the uniform continuity of $\overline{H}_n$ on the bounded subsets of $\overline{D(A)} \times [0,1-2\kappa] \times H$. It follows that with such test functions the proof of the doubling Theorem 3.1 in [15] can be applied (no matter if we use the notion of solution introduced in Definition 3.4 instead of the one employed in [15]).

3. For the sake of simplicity we start assuming that $(0,0) \in A$, so that $\varphi(x) = \|x\|^2$ is a subtest function and $D_A^- \varphi \geq 0$, as easily checked. For $\alpha, \delta > 0$ and $x, y \in \overline{D(A)}$ let

$$\Phi(x,y) = U^*(x) - W(y) - \tfrac{\alpha}{2}\|x-y\|^2 - \delta\|x\|^2 - \delta\|y\|^2;$$

note that $\Phi \leq 1 - 2\kappa$ and, for sufficiently small δ,

$$\sup \Phi \geq U^*(\hat z) - W(\hat z) - 2\delta\|\hat z\|^2 \geq \gamma.$$

For every $\epsilon > 0$ use perturbed optimization with Tataru's distance to find $\hat x, \hat y \in \overline{D(A)}$ such that $\Phi(\hat x, \hat y) \geq \sup\Phi - \epsilon$ and the map $\Phi(x,y) - \epsilon d(x,\hat x) - \epsilon d(y,\hat y)$ has a strict global maximum at $(\hat x, \hat y)$. Note that

$$\sup\{\Phi(x,y) - \epsilon d(x,\hat x) - \epsilon d(y,\hat y)\} = \Phi(\hat x, \hat y) \geq \gamma - \epsilon \geq \frac{\gamma}{2} > 0 \tag{5.8}$$

for small ϵ.

Consider two cases. If $U^*(\hat x) \leq W_n(\hat x)$ then

$$U^*(\hat x) - W(\hat y) \leq W_n(\hat x) - W_n(\hat y). \tag{5.9}$$

Otherwise, as we mentioned above, we can apply the doubling theorem in [15] to obtain

$$\begin{aligned}
&\lambda(U^*(\hat x) - W(\hat y)) - 2\epsilon \\
&\quad\leq \sup_{u\in U}\Big\{ -\langle f(\hat y,u), \alpha(\hat x - \hat y) - 2\delta\hat y\rangle + \epsilon\|f(\hat y,u)\| + (W(\hat y)-1)(\tfrac12\|u\|^2 + g_n(\hat y))\Big\} \\
&\qquad - \sup_{u\in U}\Big\{ -\langle f(\hat x,u), \alpha(\hat x - \hat y) + 2\delta\hat x\rangle - \epsilon\|f(\hat x,u)\| + (U^*(\hat x)-1)(\tfrac12\|u\|^2 + g_n(\hat x))\Big\}.
\end{aligned}$$

Note that from (5.8) $W(\hat{y}) \le U^*(\hat{x})$. By standard arguments (see, e.g., Lemma 3.2 in [15]), we can show that

$$\limsup_{\alpha\to\infty} \limsup_{\delta\downarrow 0} \limsup_{\epsilon\downarrow 0} \left(\alpha\|\hat{x}-\hat{y}\|^2 + \delta\|\hat{x}\|^2 + \delta\|\hat{y}\|^2\right) = 0. \tag{5.10}$$

Moreover, since Φ decays quadratically at infinity, for α and δ fixed, $\hat{x}$ and $\hat{y}$ remain bounded uniformly in $\epsilon \downarrow 0$. Therefore, letting $\epsilon \downarrow 0$ and using Lemma 5.4 gives

$$\begin{aligned} &\limsup_{\epsilon\downarrow 0} \lambda(U^*(\hat{x}) - W(\hat{y})) \\ &\le \limsup_{\epsilon\downarrow 0} \left(\overline{H}_n(\hat{y}, U^*(\hat{x}), \alpha(\hat{x}-\hat{y}) - 2\delta\hat{y}) - \overline{H}_n(\hat{x}, U^*(\hat{x}), \alpha(\hat{x}-\hat{y}) + 2\delta\hat{x})\right). \end{aligned} \tag{5.11}$$

Let $R > 0$ and suppose that $x \in \overline{D(A)}$, $\|p\| + \delta\|x\| \le R$, and $0 \le r \le 1 - \frac{1}{R}$. By Lemma 5.4 we can find $C = C(R)$ such that

$$\overline{H}_n(x,r,p) = \sup_{\|u\| \le C(1+\sqrt{\|x\|})} \left\{-\langle f(x,u), p\rangle + \tfrac{1}{2}(r-1)\|u\|^2\right\} + (r-1)g_n(x).$$

From (2.1) we then get

$$\begin{aligned} |\overline{H}_n(x,r,p+\delta x) - \overline{H}_n(x,r,p)| &\le \delta\|x\| L\left(1 + \|x\| + C^q(1+\sqrt{\|x\|})^q\right) \\ &\le \delta\|x\| L\left(1 + \|x\| + (2C)^q\left(1 + \|x\|^{\frac{q}{2}}\right)\right) \to 0, \quad \text{if} \quad \delta + \delta\|x\|^2 \to 0, \end{aligned} \tag{5.12}$$

uniformly for $\delta\|x\| + \|p\| \le R$, $0 \le r \le 1 - \frac{1}{R}$.

From (5.11), (5.12), and (2.1), by letting $\delta \to 0$,

$$\begin{aligned} &\limsup_{\delta\downarrow 0} \limsup_{\epsilon\downarrow 0} \lambda(U^*(\hat{x}) - W(\hat{y})) \\ &\le \limsup_{\delta\downarrow 0} \limsup_{\epsilon\downarrow 0} \left(\overline{H}_n(\hat{y}, U^*(\hat{x}), \alpha(\hat{x}-\hat{y})) - \overline{H}_n(\hat{x}, U^*(\hat{x}), \alpha(\hat{x}-\hat{y}))\right) \\ &\le \limsup_{\delta\downarrow 0} \limsup_{\epsilon\downarrow 0} \left(L\alpha\|\hat{x}-\hat{y}\|^2 + |g_n(\hat{y}) - g_n(\hat{x})|\right), \end{aligned} \tag{5.13}$$

and finally taking lim sup as $\alpha \to \infty$ and using (5.10) and the uniform continuity of g_n yields

$$\limsup_{\alpha\to\infty} \limsup_{\delta\downarrow 0} \limsup_{\epsilon\downarrow 0} (U^*(\hat{x}) - W(\hat{y})) \le 0.$$

The same inequality also holds in the first case because of (5.9) and $\|\hat{x}-\hat{y}\| \to 0$, and then (5.8) yields a contradiction.

If $(0,0) \notin A$ then one replaces $\delta\|x\|^2$ by $\delta\|x-\bar{x}\|^2$ for any fixed $\bar{x} \in D(A)$. As $D_A^-\|x-\bar{x}\|^2 \ge 2\langle A^\circ\bar{x}, x-\bar{x}\rangle$, additional terms of the form $2\delta\langle A^\circ\bar{x}, \hat{x}-\bar{x}\rangle$ appear, but they will vanish when $\delta \downarrow 0$.

4. Thus (5.7) is proved and for every $\lambda > 0$ and $n \ge 1$,

$$W(x) \ge \inf_{u\in L^2} \sup_{t\ge 0} J_n^\lambda(t,x,u,W_n) \quad \text{for} \quad x \in \overline{D(A)}. \tag{5.14}$$

Fix $x \in \overline{D(A)}$. Letting $\lambda \downarrow 0$ in (5.14) we obtain for every fixed $T > 0$,

$$\begin{aligned} W(x) &\ge \inf_{u\in L^2} \sup_{t\in[0,T]} \left\{1 - e^{-\int_0^t (g_n(y(s)) + \frac{1}{2}\|u(s)\|^2)ds}(1 - W_n(y(t)))\right\} \\ &= \inf_{u\in L^2} \sup_{t\in[0,T]} \left\{1 - e^{-\int_0^t (g_n(y(s)) + \frac{1}{2}\|u(s)\|^2)ds - w_n(y(t))}\right\} \\ &= \rho\left(\inf_{u\in L^2} \sup_{t\in[0,T]} \left\{\int_0^t (g_n(y(s)) + \tfrac{1}{2}\|u(s)\|^2)ds + w_n(y(t))\right\}\right), \end{aligned}$$

which implies, for all $n \geq 1$ and $T > 0$,

$$w(x) \geq \inf_{u \in L^2} \sup_{t \in [0,T]} \left\{ \int_0^t (g_n(y(s)) + \tfrac{1}{2}\|u(s)\|^2) ds + w_n(y(t)) \right\}. \tag{5.15}$$

Sending $n \to \infty$ in (5.15) as in Lemma 5.2 (hence here we finally use the assumption (W)), we get, for every $x \in \overline{D(A)}$,

$$w(x) \geq \inf_{u \in L^2} \sup_{t \in [0,T]} \left\{ \int_0^t \left(g(y(s)) + \tfrac{1}{2}\|u(s)\|^2\right) ds + w(y(t)) \right\}. \tag{5.16}$$

In order to pass to the limit as $T \to \infty$ in (5.16) we proceed as follows. For given $\varepsilon > 0$, we apply (5.16) with $T = 1$ and find $u_1 \in L^2$ such that

$$w(x) + \tfrac{\varepsilon}{2} \geq \int_0^t \left(g(y(s, x, u_1)) + \tfrac{1}{2}\|u_1(s)\|^2\right) ds + w(y(t, x, u_1))$$

for $t \in [0, 1]$. Then we apply (5.16) again at $y(1, x, u_1)$ with $T = 1$ and find $u_2 \in L^2$ such that

$$w(y(1, x, u_1)) + \tfrac{\varepsilon}{2^2} \geq \int_0^t \left(g(y(s, y(1, x, u_1), u_2)) + \tfrac{1}{2}\|u_2(s)\|^2\right) ds + w(y(t, y(1, x, u_1), u_2))$$

for $t \in [0, 1]$, and so forth. We proceed recursively and define $u(t) = u_{[t]+1}(t - [t])$ for $t \geq 0$, where $[t]$ denotes the largest integer in $[0, t]$. It then follows that

$$w(x) + \varepsilon \geq \int_0^t \left(g(y(s, x, u)) + \tfrac{1}{2}\|u(s)\|^2\right) ds + w(y(t, x, u)) \quad \text{for all} \quad t \geq 0;$$

therefore, in particular $u \in L^2$ since g and w are nonnegative and $u \in \mathcal{U}(x)$. Since ϵ was arbitrary we obtain for all $x \in \overline{D(A)}$,

$$w(x) \geq \inf_{u \in L^2} \sup_{t \geq 0} \left\{ \int_0^t (g(y(s)) + \tfrac{1}{2}\|u(s)\|^2) ds + w(y(t)) \right\},$$

which concludes the proof since the other inequality follows immediately by choosing $t = 0$ on the right-hand side. ∎

6. Examples. In this section we will quickly present some examples of nonlinear systems satisfying the strong stability condition (S) to which the results of this paper can be applied. These examples are meant to show that the condition is quite natural and is known to be satisfied in many interesting instances. More examples of nonlinear partial differential equations leading to systems of the form (1.1) with $-A$ generating a compact semigroup on a Hilbert space, e.g., reaction-diffusion systems, can be found in [26] and [2]. We do not present examples with a linear operator A that, as we mentioned above, satisfy the weak stability condition (W), but instead we refer the reader to [5].

Example 6.1 (p-Laplace operator). Let Ω be a bounded domain in $\mathbb{R}^n$, $n \geq 1$, with smooth boundary Γ, and let $p \geq 2$. For $\lambda \geq 0$ consider the nonlinear parabolic

equation with distributed parameters

$$\text{(6.1)}\qquad \begin{cases} \dfrac{\partial y}{\partial t} - \displaystyle\sum_{i=1}^{n} \frac{\partial}{\partial x_i}\left(\left| \frac{\partial y}{\partial x_i} \right|^{p-2} \frac{\partial y}{\partial x_i} \right) + \lambda y|y|^{p-2} = f(y,u) \\ \qquad \text{for} \quad (t,x) \in (0,T)\times\Omega, \\ -\displaystyle\sum_{i=1}^{n} \left| \frac{\partial y}{\partial x_i} \right|^{p-2} \frac{\partial y}{\partial x_i} \cdot n_i \in \partial\phi(y) \quad \text{for} \quad (t,x) \in (0,T)\times\Gamma, \\ y(0,x) = y_0(x) \quad \text{for} \quad x \in \Omega. \end{cases}$$

Here $n = (n_1, \dots, n_n)$ denotes the outward normal to Γ and $\phi\colon \mathbb{R} \to [0, +\infty]$ is lower semicontinuous and convex, $\phi(0) = 0$. Then (6.1) gives rise to a system of the form (1.1) in $H = L^2(\Omega)$ with a maximal monotone A, which is in fact the subgradient of a lower semicontinuous and convex function on H and such that $-A$ generates a compact semigroup in H; see [26, Remark 2.2.5]. If f in (6.1) satisfies the assumptions of Proposition 2.7, then (S) holds.

In the previous example, by choosing specific functions ϕ we obtain a number of physical models interesting for the applications; see [2, section 4.3].

Example 6.2. Let Ω be a bounded domain in $\mathbb{R}^n$, $n \geq 2$, with smooth boundary Γ. Consider the boundary value problem

$$\text{(6.2)}\qquad \begin{cases} \dfrac{\partial y}{\partial t} - \Delta y + y^3 = f(y,u) \quad \text{for} \quad (t,x) \in (0,T)\times\Omega, \\ \alpha y(t,x) + \beta \dfrac{\partial y}{\partial n}(t,x) = 0 \quad \text{for} \quad (t,x) \in (0,T)\times\Gamma, \\ y(0,x) = y_0(x) \quad \text{for} \quad x \in \Omega, \end{cases}$$

where $\alpha, \beta \geq 0$ and $\alpha + \beta > 0$. Then again (6.2) can be written in the form (1.1) in $H = L^2(\Omega)$ with a nonlinear maximal monotone A, $D(A) \subset H^2(\Omega)$; see, e.g., [2, p. 256]. From the compact embedding theorem it follows that $-A$ generates a compact semigroup and Proposition 2.7 applies.

In the previous example, the nonlinear perturbation y^3 of the Laplacian is the derivative of a convex function. Such example therefore falls into the class of problems that can be described by means of abstract parabolic variational inequalities; see Remark 2.9. As a matter of fact, the term y^3 can be replaced by any subgradient of a lower semicontinuous and convex function $\phi\colon \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$. When ϕ has nontrivial domain then the corresponding operator is apparently multivalued and our generality is motivated. We will now describe a more explicit and specific example which is a special case of the one-phase Stefan problem; see [2, p. 279].

Example 6.3 (parabolic variational inequalities). Let Ω be a bounded domain in $\mathbb{R}^n$, $n \geq 2$, with smooth boundary Γ. Consider the boundary value problem

$$\text{(6.3)}\qquad \begin{cases} \dfrac{\partial y}{\partial t} - \Delta y \geq f(y,u) \quad \text{for} \quad (t,x) \in (0,T)\times\Omega, \\ \dfrac{\partial y}{\partial t} - \Delta y = f(y,u) \quad \text{if} \quad y(t,x) > 0, \\ y \geq 0, \\ \alpha y(t,x) + \beta \dfrac{\partial y}{\partial n}(t,x) = 0 \quad \text{for} \quad (t,x) \in (0,T)\times\Gamma, \\ y(0,x) = y_0(x) \quad \text{for} \quad x \in \Omega, \end{cases}$$

where $\alpha, \beta \geq 0$ and $\alpha + \beta > 0$. Here again $H = L^2(\Omega)$. In this example, we are in the situation described in Remark 2.9, where $V = H^1(\Omega)$ (or $V = H_0^1(\Omega)$ is $\beta = 0$) and

the operator $\tilde{A} \in \mathcal{L}(V, V')$ (for $\beta \neq 0$) is defined by the position

$$\langle \tilde{A}y, z\rangle = \int_\Omega Dy \cdot Dz \, dx + \tfrac{\alpha}{\beta} \int_{\partial\Omega} yz \, d\sigma.$$

Moreover, the function $\varphi\colon H \to \mathbb{R} \cup \{+\infty\}$ is the indicator function of the set $\{y \in V\colon y \geq 0\}$. Note that a state constraint ($y \geq 0$) appears in the formulation of the problem (6.3), but it is included here in the abstract definition of the operator rather than in the cost g.

We end this section by presenting an example that can be described in abstract form with a noncompact semigroup but still satisfies our stability condition (S); see Remark 2.10. For other examples of this sort, again we refer to [26].

Example 6.4 (nonlinear hyperbolic equations). Let Ω be a bounded domain in $\mathbb{R}^n$, $n \geq 2$, with smooth boundary Γ. Consider the boundary value problem

$$\text{(6.4)} \qquad \begin{cases} \dfrac{\partial^2 y}{\partial t^2} - \Delta y + \beta\left(\dfrac{\partial y}{\partial t}\right) = \displaystyle\sum_{i=1}^{k} w_i(x) u_i(t) \quad \text{for} \quad (t,x) \in (0,T) \times \Omega, \\ y(t,x) = 0 \quad \text{for} \quad (t,x) \in (0,T) \times \Gamma, \\ y(0,x) = y_0(x) \quad \text{for} \quad x \in \Omega, \\ \dfrac{\partial y}{\partial t}(0,x) = y_1(x) \quad \text{for} \quad x \in \Omega. \end{cases}$$

Here $w_i \in L^2(\Omega)$, the control set $B_R^U \subset U = \mathbb{R}^k$ is finite dimensional, so the linear operator defined by the right-hand side of (6.4) has finite-dimensional range, and $\beta\colon \mathbb{R} \to \mathbb{R}$ is continuous, nondecreasing, and of linear growth. One has to check that (6.4) can be described in an abstract form with an operator $-A$ in $H = L^2(\Omega) \times H_0^1(\Omega)$ generating a weakly equicontinuous semigroup, and for this purpose we refer to [26, Theorem 2.9.4]; see also Remark 2.10.

In all the examples above, there are many standard state constraints K that are of interest for the applications. Some instances are described by one of the conditions $\|y\|_{L^2(\Omega)} \leq R$, $y \geq 0$, $\|y\|_{L^\infty(\Omega)} \leq R$. Note that the last two define subsets of $L^2(\Omega)$ with empty interior. Any lower semicontinuous function $l\colon \overline{D(A)} \to [0, +\infty)$ could be combined with the indicator function of K,

$$I_K(y) = \begin{cases} 0, & y \in K, \\ +\infty, & y \notin K, \end{cases}$$

to provide a running cost acceptable for our statements, namely, $g = l + I_K$. An example, for $H = L^2(\Omega)$, which leads to a nonconvex cost is the integral of a w-shaped potential, e.g.,

$$g(y) = \int_\Omega (|y(x)|^2 - 1)^2 \, dx + I_{B_1^{L^\infty(\Omega)}}(y), \quad y \in H.$$

Note that again $\mathrm{dom}(g) = B_1^{L^\infty(\Omega)}$ has an empty interior. In some cases, if $y \equiv 1$ and $y \equiv -1$ are solutions of the parabolic equation with a given control $u(\cdot)$, such control can be proven to produce trajectories fulfilling the constraint for any initial condition in $\mathrm{dom}(g)$ by the maximum principle.

Acknowledgment. The authors wish to thank one of the referees for suggesting some examples of systems that satisfy the abstract assumptions of the paper.

REFERENCES

[1] V. Barbu, *Nonlinear Semigroups and Differential Equations in Banach Spaces*, Nordhoff, Leyden, 1976.

[2] V. Barbu, *Analysis and Control of Infinite Dimensional Systems*, Academic Press, Boston, 1993.

[3] V. Barbu and G. Da Prato, *Hamilton-Jacobi Equations in Hilbert Spaces*, Pitman, London, 1983.

[4] H. Brézis, *Operateurs Maximaux Monotones et Semi-groupes de Contractions dans les Espaces de Hilbert*, North–Holland, Amsterdam, 1973.

[5] P. Cannarsa and G. Di Blasio, *A direct approach to infinite-dimensional Hamilton-Jacobi equations and applications to convex control with state constraints*, Differential Integral Equations, 8 (1995), pp. 225–246.

[6] P. Cannarsa, F. Gozzi, and M. Soner, *A boundary value problem for Hamilton-Jacobi equations in Hilbert spaces*, Appl. Math. Optim., 24 (1990), pp. 197–200.

[7] I. Capuzzo-Dolcetta and P. L. Lions, *Hamilton-Jacobi equations with state constraints*, Trans. Amer. Math. Soc., 318 (1990), pp. 643–683.

[8] M. G. Crandall, H. Ishii, and P. L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc., 27 (1992), pp. 1–67.

[9] M. G. Crandall, M. Kocan, and A. Święch, *On partial sup-convolutions, a lemma of P. L. Lions and viscosity solutions in Hilbert spaces*, Adv. Math. Sci. Appl., 3 (1993/4), pp. 1–15.

[10] M. G. Crandall and P. L. Lions, *Hamilton-Jacobi equations in infinite dimensions, Part* I. *Uniqueness of viscosity solutions*, J. Funct. Anal., 62 (1985), pp. 379–396.

[11] M. G. Crandall and P. L. Lions, *Hamilton-Jacobi equations in infinite dimensions, Part* II. *Existence of viscosity solutions*, J. Funct. Anal., 65 (1986), pp. 368–405.

[12] M. G. Crandall and P. L. Lions, *Hamilton-Jacobi equations in infinite dimensions, Part* III., J. Funct. Anal., 68 (1988), pp. 214–247.

[13] M. G. Crandall and P. L. Lions, *Hamilton-Jacobi equations in infinite dimensions, Part* IV. *Unbounded linear terms*, J. Funct. Anal., 90 (1990), pp. 237–283.

[14] M. G. Crandall and P. L. Lions, *Hamilton-Jacobi equations in infinite dimensions, Part* V. *B-continuous solutions*, J. Funct. Anal., 97 (1991), pp. 417–465.

[15] M. G. Crandall and P. L. Lions, *Hamilton-Jacobi equations in infinite dimensions, Part* VI. *Nonlinear A and Tataru's method refined*, in Evolution Equations, Control Theory, and Biomathematics, P. Clément and G. Lumer, eds., Lecture Notes in Pure and Appl. Math. 155, Marcel Dekker, New York, 1994, pp. 51–89

[16] M. G. Crandall and P. L. Lions, *Viscosity solutions of Hamilton-Jacobi equations in infinite dimensions, Part* VII. *The HJB equation is not always satisfied*, J. Funct. Anal., 125 (1994), pp. 111–148.

[17] G. Di Blasio, *Optimal control with infinite horizon for distributed parameter systems with constrained controls*, SIAM J. Control Optim., 29 (1991), pp. 909–925.

[18] H. Ishii, *Viscosity solutions for a class of Hamilton-Jacobi equations in Hilbert spaces*, J. Funct. Anal., 105 (1992), pp. 301–341.

[19] M. Kocan, P. Soravia, and A. Święch, *On differential games for infinite dimensional systems with nonlinear, unbounded operators*, J. Math. Anal. Appl., 211 (1997), pp. 395–423.

[20] M. Soner, *Optimal control with state-space constraint* I, SIAM J. Control Optim., 24 (1986), pp. 552–561.

[21] P. Soravia, *Optimality principles and representation formulas for viscosity solutions of Hamilton-Jacobi equations.* I. *Equations of unbounded and degenerate control problems without uniqueness*, Adv. Differential Equations, to appear.

[22] P. Soravia, *Optimality principles and representation formulas for viscosity solutions of Hamilton-Jacobi equations.* II. *Equations of control problems with state constraints*, Adv. Differential Equations, to appear.

[23] P. Soravia, *Stability of dynamical systems with competitive controls: The degenerate case*, J. Math. Anal. Appl., 191 (1995), pp. 428–449.

[24] D. Tataru, *Viscosity solutions for Hamilton-Jacobi equations with unbounded nonlinear terms*, J. Math. Anal. Appl., 163 (1992), pp. 345–392.

[25] D. Tataru, *Viscosity solutions for Hamilton-Jacobi equations with unbounded nonlinear term: A simplified approach*, J. Differential Equations, 111 (1994), pp. 123–146.

[26] I. I. Vrabie, *Compactness Methods for Nonlinear Evolution*, 2nd ed., Longman, London, 1995.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 4, pp. 1376–1422, July 1998

014

UNIFORM STABILIZABILITY OF A FULL VON KARMAN SYSTEM WITH NONLINEAR BOUNDARY FEEDBACK*

IRENA LASIECKA†

Abstract. Full von Karman system accounting for in-plane accelerations and describing the transient deformations of a thin, elastic plate subject to edge loading is considered. The energy dissipation is introduced via the nonlinear velocity feedback acting on a part of the edge of the plate. It is known [J. Puel and M. Tucsnak, *SIAM J. Control Optim.*, 33 (1995), pp. 255–273] that in the case of *linear* dissipation and "star-shaped" domains, boundary velocity feedback with the tangential derivatives of horizontal displacements leads to the exponential decay rates for the energy of the resulting closed loop system. The main goal of the paper is to derive the *uniform* energy decay rates valid for the model without the above-mentioned restrictions. In particular, it is shown that simple, monotone *nonlinear* feedback (without the tangential derivatives of the horizontal displacements) provides the uniform decay rates for the energy in the absence of *geometric hypotheses* imposed on the controlled part of the boundary. This is accomplished by establishing, among other things, "sharp" regularity results valid for the boundary traces of solutions corresponding to this nonlinear model and by employing a Holmgren-type uniqueness result proved recently in [V. Isakov, *J. Differential Equations*, 97 (1997), pp. 134–147] for the dynamical systems of elasticity which are overdetermined on the boundary.

Key words. full von Karman system, uniform stabilization, boundary control

AMS subject classifications. 35, 93

PII. S0363012996301907

1. Introduction.

We consider a model of dynamic nonlinear plate that is referred to as the *full von Karman* system and is introduced in [11]. We associate with this model a nonlinear damping represented by moments and shears applied to the edge of the plate. Here the variables w and $u = (u_1, u_2)$ represent, respectively, the vertical and in-plane displacement of a thin plate occupying a two-dimensional domain Ω with sufficiently smooth boundary $\Gamma = \Gamma_0 \cup \Gamma_1$. We shall assume that $\Gamma_0 \cap \Gamma_1 = \emptyset$. The governing equations are given by

$$\begin{aligned} u_{tt} + b_1 u_t - \operatorname{div}[\mathcal{C}[\boldsymbol{\epsilon}(u) + f(\nabla w)]] &= 0 \text{ in } \Omega \times (0,\infty), \\ [I - \gamma\Delta]w_{tt} + b_2 w_t + D\Delta^2 w - \operatorname{div}[\mathcal{C}[\boldsymbol{\epsilon}(u) + f(\nabla w)]\nabla w] &= 0 \text{ in } \Omega \times (0,\infty) \end{aligned} \tag{1.1}$$

with Dirichlet boundary conditions on the "uncontrolled part" of the boundary Γ_0,

$$u = w = \nabla w = 0 \text{ on } \Gamma_0 \times (0,\infty).$$

The dissipative boundary conditions on the "controlled part" of the boundary Γ_1 are given by

$$\begin{aligned} \mathcal{C}[\boldsymbol{\epsilon}(u) + f(\nabla w)]\nu &= -g(u_t), \\ D[\Delta w + (1-\mu)B_1 w] &= -h_1(D_n w_t), \\ D[D_n \Delta w + (1-\mu)B_2 w] - \gamma D_n w_{tt} - [\mathcal{C}[\boldsymbol{\epsilon}(u) + f(\nabla w)]\nu \cdot \nabla w] & \\ = -D_\tau h_2(D_\tau w_t). & \end{aligned} \tag{1.2}$$

*Received by the editors April 12, 1996; accepted for publication (in revised form) June 17, 1997; published electronically May 28, 1998. This research was partially supported by NSF grant DMS-9504822.

http://www.siam.org/journals/sicon/36-4/30190.html

†Department of Applied Mathematics, University of Virginia, Charlottesville, VA 22901 (il2v@virginia.edu).

With (1.1) and (1.2) we associate the initial conditions

$$u(0) = u_0, u_t(0) = u_1, w(0) = w_0, w_t(0) = w_1 \text{ in } \Omega. \tag{1.3}$$

The vector ν represents an outward normal and τ represents the tangential direction. D_n (resp., D_τ) stand for the normal and tangential derivatives. D represents the flexural rigidity, the constant $0 < \mu < 1/2$ is Poisson's modulus, and a positive constant γ is proportional to the thickness of the plate.

The fourth-order tensor $\mathcal{C}$ is defined by

$$\mathcal{C}(\boldsymbol{\epsilon}) \equiv \frac{E}{(1-2\mu)(1+\mu)}[\mu \; trace \; \boldsymbol{\epsilon} \; I + (1-2\mu)\boldsymbol{\epsilon}],$$

where $\boldsymbol{\epsilon}(u) \equiv 1/2(\nabla u + \nabla^T u)$. It can be easily verified that the tensor $\mathcal{C}$ is symmetric and strictly positive. The function f is given by

$$f(s) \equiv (1/2)s \times s, \; s \in R^2,$$

and the boundary operators are defined by

$$B_1 \equiv 2\nu_1\nu_2 D^2_{x,y} - \nu_1^2 D^2_{y,y} - \nu_2^2 D^2_{x,x},$$

$$B_2 \equiv D_\tau[(\nu_1^2 - \nu_2^2)D^2_{x,y} + \nu_1\nu_2(D^2_{y,y} - D^2_{x,x})] + lI.$$

The dissipation in the system is represented by a nonlinear vector function g and scalar functions h_i, which are assumed continuous, monotone increasing, zero at the origin, and of linear growth at infinity.

The following well-posedness/regularity results are proved for this model (see the Appendix).

PROPOSITION 1.1.

(1) Regular solutions. *We assume that* $h_1, h_2 \in C^1(R), g \in C^1(R^2)$ *are monotone, increasing functions with* $h_i{}'$ *(resp.,* g'*)* $\in L_\infty(R)$ *(resp.,* $L_\infty(R^2)$*). For any initial data*

$$u_0, u_1 \in [H^2(\Omega)]^2 \times [H^1(\Omega)]^2, \quad w_0, w_1 \in H^3(\Omega) \times H^2(\Omega)$$

subject to the compatibility conditions satisfied on the boundary Γ*,*

$$\begin{aligned} u_0 = w_0 = \nabla w_0 = w_1 = \nabla w_1 = 0 \quad & on \; \Gamma_0 \times (0,\infty), \\ \mathcal{C}[\boldsymbol{\epsilon}(u_0) + f(\nabla w_0)]\nu = -g(u_1) \quad & on \; \Gamma_1 \times (0,\infty), \\ D[\Delta w_0 + (1-\mu)B_1 w_0] = -h_1(D_n w_1) \quad & on \; \Gamma_1 \times (0,\infty), \end{aligned} \tag{1.4}$$

there exists a unique, global solution

$$(u,w) \in C(0,T;[H^2(\Omega)]^2 \times H^3(\Omega)), \quad (u_t, w_t) \in C(0,T;[H^1(\Omega)]^2 \times H^2(\Omega)),$$

where $T > 0$ *is arbitrary.*

(2) Weak solutions. *In the case of linear damping (i.e.,* $g'(s) = g_0 > 0, h_i{}'(s) = h_{i,0} > 0$*), there exist a* unique*, global solution of finite energy. This is to say that for any initial data*

$$u_0, u_1 \in [H^1(\Omega)]^2 \times [L_2(\Omega)]^2, \quad w_0, w_1 \in H^2(\Omega) \times H^1(\Omega),$$

subject to the boundary conditions satisfied on the boundary Γ_0*,*

$$u_0 = w_0 = \nabla w_0 = w_1 = 0 \quad on \; \Gamma_0 \times (0,\infty), \tag{1.5}$$

there exists a unique solution

$$(u,w)\in C(0,T;[H^1(\Omega)]^2\times H^2(\Omega)),\quad (u_t,w_t)\in C(0,T;[L_2(\Omega)]^2\times H^1(\Omega)),$$

where $T>0$ is arbitrary.

REMARK 1.1. *The existence of regular solutions, the result stated in the first part of Proposition* 1.1*, is proved in the appendix by using the nonlinear Galerkin method. It can also be proved by the same arguments (based on an application of Shaffer's theorem) as those used in* [3] *for the modified von Karman system with* nonlinear *dissipation.*

The uniqueness of regular solutions follows from a rather standard energy type argument which is given in section 7.2.

For the case of linear dissipation, the existence and uniqueness of strong solutions was proved in [23]*. (However, the techniques of* [23] *are not readily extendible for treating nonlinear boundary feedback.)*

In the case of linear damping the existence of weak *solutions (i.e.,* $(u,w)\in C(0,\infty;[H^1(\Omega)]^2\times H^2(\Omega));\ (u_t,w_t)\in C(0,\infty;[L_2(\Omega)]^2\times H^1(\Omega))$ *follows from the usual Galerkin-type argument. The uniqueness of weak solutions was recently proved in* [24] *for the case of Dirichlet boundary conditions and with* $\gamma=0$*. In section* 7.3 *we adopt Sedenko's method to prove the uniqueness of weak solutions for the model of interest (i.e., Proposition* 1.1(2)*). We also note that the uniqueness of weak solutions for the modified von Karman equations was proved recently in* [3]*. However, the arguments employed in* [3] *and based on sharp regularity of the Airy stress function are not applicable here.*

The main goal of this paper is to show that the solutions decay to zero at the uniform rate. To state this result, we recall that the energy functional associated with the plate model (1.1) is given by

$$E(t)=E_k(t)+E_p(t) \tag{1.6}$$

with the kinetic energy

$$E_k(t)=\int_\Omega |u_t|^2+w_t^2+\gamma|\nabla w_t|^2 d\Omega \tag{1.7}$$

and the potential energy

$$E_p(t)=a(w,w)+\int_\Omega [\mathcal{C}N(u,w)\cdot N(u,w)]\,d\Omega, \tag{1.8}$$

where the bilinear form $a(w,z)$ is defined by

$$\begin{aligned}a(w,z)\equiv D\int_\Omega [&w_{x,x}z_{x,x}+w_{y,y}z_{y,y}+\mu w_{x,x}z_{y,y}+\mu w_{y,y}z_{x,x}\\ &+2(1-\mu)w_{x,y}z_{x,y}]d\Omega+l\int_{\Gamma_1} wz d\Gamma_1,\end{aligned}$$

and the stress resultants $N(u,w)$ are given by

$$N(u,w)\equiv \boldsymbol{\epsilon}(u)+f(\nabla w).$$

It is well known that $E_p(t)$ is topologically equivalent to $H^2(\Omega)\times[H^1(\Omega)]^2$ topology.

To formulate our result we introduce the following functions:

$$\mathcal{H}_0(s) \equiv \mathcal{G}(s) + \mathcal{H}1(s) + \mathcal{H}2(s),$$

where the functions $\mathcal{G}, \mathcal{H}1, \mathcal{H}2$ are concave, strictly increasing functions, zero at the origin and such that the following inequalities are satisfied for $|s| \leq 1$

$$\mathcal{G}(\vec{s}g(\vec{s})) \geq |\vec{s}|^2 + |g(\vec{s})|^2, \quad \vec{s} \in R^2,$$

$$\mathcal{H}1(sh_1(s)) \geq s^2 + h_1^2(s), \quad s \in R,$$

$$\mathcal{H}2(sh_2(s)) \geq s^2 + h_2^2(s), \quad s \in R.$$

Due to the assumed monotonicity of the nonlinear functions g, h_i, one can easily construct functions $\mathcal{G}, \mathcal{H}1, \mathcal{H}2$ with the properties listed above (see [19]).

We are ready to state the main result of this paper. To this end we introduce the following hypothesis which will be assumed throughout the paper.

ASSUMPTION 1.

(1) *Let* $\boldsymbol{h}(x)$ *be a vector field defined by*

$$\boldsymbol{h}(x) \equiv x - x_0, \textit{where} \quad x_0 \in R^2.$$

We assume that

$$\boldsymbol{h}\nu \leq 0 \textit{ on } \Gamma_0. \tag{1.9}$$

(2) *There exist positive constants* $0 < m \leq M$ *such that for* $|s| \geq R$ *with a constant* R *sufficiently large, we have*

$$m|s|^2 \leq (g(s), s)_{R^2} \leq M|s|^2, \quad ms^2 \leq h_i(s)s \leq Ms^2, \quad i = 1, 2. \tag{1.10}$$

(3) *The "coefficients"* $b_1, b_2 \in \mathcal{L}(L_2(\Omega))$ *representing a potential "light damping" are assumed to satisfy* $(b_1 u, u)_{[L_2(\Omega)]^2} \geq 0$, $(b_2 w, w)_{L_2(\Omega)} \geq 0$ *for all* $u \in [L_2(\Omega)]^2$ *and* $w \in L_2(\Omega)$.

THEOREM 1.2. *Let* u, w *be a strong solution to the original system* (1.1) *and let us assume that either* $l > 0$ *or* $\Gamma_0 \neq \emptyset$.

Part I. *In addition to Assumption* 1 *we make Assumption* 2 *as follows.*

ASSUMPTION 2. *Either* b_1 *or* b_2 *are injective or* Ω *is star-shaped.*

Then there exists a constant $T_0 > 0$ *such that the following estimate holds*

$$E(t) \leq C(E(0))s(t/T_0 - 1)), \quad t \geq T_0, \tag{1.11}$$

where a real variable function $s(t)$ *converges to zero as* $t \to \infty$ *and it obeys the ordinary differential equation*

$$s_t(t) + q(s(t)) = 0, \quad s(0) = E(0). \tag{1.12}$$

The (nonlinear), monotone increasing function $q(s)$ *is determined entirely from the behavior at the origin of the nonlinear functions* g, h_i, *and it is given by the following algorithm:*

$$q \equiv I - (I + p)^{-1}, \tag{1.13}$$

$$p \equiv (kI + \mathcal{H})^{-1}, \tag{1.14}$$

$$\mathcal{H} \equiv \mathcal{H}_0(\cdot/mes\Sigma_1), \tag{1.15}$$

where the constant k is proportional to $\frac{1}{mes\Sigma_1}(m^{-1}+M)$ with $\Sigma_1 \equiv \Gamma_1 \times (0,T)$.

Part II. *If Assumption* 2 *does not hold, then the conclusion of Part* I *holds for all initial data incrementally more regular, i.e.,*

$$u_0 \in H^{1+\epsilon}(\Omega), \quad w_0 \in H^{2+\epsilon}(\Omega),$$

where $\epsilon > 0$ is arbitrary and the constant C in (1.11) *may depend on the norms of these data, i.e.,*

$$C = C(|u_0|_{H^{1+\epsilon}(\Omega)}, \quad |w_0|_{H^{2+\epsilon}(\Omega)}, E(0)).$$

REMARK 1.2. *If the nonlinear functions g, h_1, h_2 are bounded from below by a linear function, then it can be shown that the decay rates predicted by Theorem* 1.2 *are exponential. This is to say that there exist positive constants C, ω possibly depending on $E(0)$ and such that*

$$E(t) \leq Ce^{-\omega t} \quad for\ t > T_0.$$

If instead these functions have a polynomial growth at the origin, then the decay rates are algebraic (see [19]*).*

REMARK 1.3. *The light damping terms represented by the coefficients b_1, b_2 correspond, typically, to a possibility of having small viscous damping. We note that this damping alone (even if it is fully active, i.e., b_1 and b_2 are uniformly positive on Ω) will not cause a uniform decay for the energy (it may, at most, cause the strong stability of the solutions). For the former, the presence of the boundary dissipation or of a much stronger viscous damping is necessary. From the mathematical point of view, the presence of light damping is beneficial at the level of eliminating lower-order terms from the appropriate inequalities (see section* 5*) which is done by using the new uniqueness result due to Isakov* [9]*. Indeed, the application of this uniqueness result requires sufficient regularity of solutions to the linearized equations. This, in turn, can be established if one of the b_i's is injective or Ω is star-shaped. In the general case, however, the needed regularity can be shown provided we start off with minimally more regular initial data (see Part* II *of Theorem* 1.2*). At this point, it is not known whether this regularity requirement is necessary for the result to hold.*

REMARK 1.4. *The decay rates guaranteed in Part* II *of Theorem* 1.2 *can be extended to hold for all $H^{1+\epsilon}(\Omega) \times H^{2+\epsilon}(\Omega)$ solutions with ϵ arbitrary small. Indeed, this can be done by applying the usual density argument* [11] *combined with the uniqueness result stated in the Appendix. Similarly, in view of the uniqueness result valid for weak solutions in the case of linear dissipation, the result of Part* I *of Theorem* 1.2 *could be extended to all such weak solutions.*

Literature relevant to the problem. Problems related to boundary stabilization of von Karman equations have attracted considerable attention in recent years. Indeed, starting with [11] and followed by papers [6], [7] uniform decay properties for the energy of the *modified* von Karman system with boundary dissipation were established. As it is well known, the *modified* von Karman system does not account for in-plane accelerations and, therefore, it can be "almost" decoupled via the Airy stress function. This is in contrast to the *full* von Karman system, where the nonlinear coupling is strong. Moreover, the additional difficulty (in the case of the two-dimensional model) results from the unboundedness of nonlinear terms in the topology induced by the energy functional.

Energy decay rates for the one-dimensional full von Karman model were first derived in [14]. The two-dimensional version of this model has been subsequently treated

in [13], where the uniform decay rates were proved for a combination of static/dynamic models with a nonlinear boundary dissipation of a linear growth at the infinity and of a polynomial growth at the origin. A fully dynamic von Karman system accounting for in-plane accelerations (as considered in this paper) was treated in [22]. The model considered in [22], and inspired by [13], accounts for the *linear* boundary dissipation of the form

$$\begin{aligned} \mathcal{C}[\epsilon(u) + f(\nabla w)]\nu &= -a(\boldsymbol{h}\cdot\nu)u_t - b(D_\tau u_2, -D_\tau u_1), \\ D[\Delta w + (1-\mu)B_1 w] &= -a(\boldsymbol{h}\cdot\nu)D_n w_t, \\ D[D_n\Delta w + (1-\mu)B_2 w] - \gamma D_n w_{tt} - [\mathcal{C}[\epsilon(u) + f(\nabla w)]\nu\cdot\nabla w] & \\ = a(h\cdot\nu)w_t + a(D_\tau(h\cdot\nu)D_\tau w_t), & \end{aligned} \tag{1.16}$$

where the constants a, b are *strictly positive*. The exponential decay rates obtained in [22] for this model above hold under the following conditions: (i) *geometric star-shaped conditions* are assumed also on the *controlled* portion of the boundary Γ_1; (ii) the constant "b" in (1.16) is *strictly positive* but suitably *small*. This "smallness" requirement is dictated by the existence theory presented in [23].

Putting aside, for a moment, questions related to unjustified, on physical grounds, geometric conditions imposed on the *controlled* portion of the boundary and the linear nature of boundary dissipation, the major somewhat disappointing feature of this result is the fact that while the decay rates derived in [22] depend critically on the positivity of the constant b in (1.16) (in fact, they go to zero when b tends to zero), the very same constant needs to be assumed small in order to guarantee the existence of the solutions (see [23]). Moreover, since the estimates break down when $b = 0$, the authors in [22] suggest the necessity of these tangential components of vector u in the structure of the stabilizing feedback. This leads to an unpleasant dichotomy where the "factotum/savior" for the stabilization result is "knocked down" by the existence theory. It was precisely this aspect of the problem that provided the main motivation for gaining a better understanding and searching for more powerful techniques adequate for studying this nonlinear problem. Thus, the main question being asked is, can we obtain the uniform decay rates for the energy *without* the additional tangential components of the horizontal displacement u present in the structure of the feedback? (For example, can the constant b in (1.16) be equal to zero?)

This paper provides an affirmative answer to the above question. In fact, the results stated in Theorem 1.2 yield the uniform decay rates for the model *without* the additional component of the boundary feedback corresponding to the b term in (1.16) (contrary to the authors' conjecture in [22]). Moreover, our feedback is *nonlinear* and without any assumptions on the growth at the origin. Finally, the result of Theorem 1.2 does not require any geometric hypotheses on the controlled portion of the boundary Γ_1, which is in agreement with the physical understanding of propagations.

Let us state at the outset that the technique employed in [22], and based on a combination of multipliers and the Lyapunov method (used earlier also in [13], [12]) cannot be extended to treat the problem at hand. Instead, our main strategy is to obtain a certain nonlinear algebraic relation for the energy function which then leads, via suitable comparison argument, to the ordinary differential equation describing the decay rates for the solutions. This approach, very different from the Lyapunov function approach used by many authors (including [22]), has been introduced in [19] in a context of the wave equation. The main advantage of this technique is its flexibility in handling various "unstructured" terms of the equation (in contrast to the Lyapunov function method which is very sensitive to the structure of the

problem). However, the main new ingredients critical to the proof of Theorem 2.1 are (i) appropriate "sharp" trace estimates for the solutions to the nonlinear system (1.1) and (ii) a Holmgren-type uniqueness result valid for the nonlinear system which is overdetermined on the boundary.

Indeed, as to the point in (i), these "trace" estimates are responsible for handling the geometrical conditions as well as for showing the decay rates *without* the additional tangential boundary terms present in (1.16). We note that a version of sharp trace estimates was used before in [21], [6] for the purpose of eliminating geometric conditions in the context of wave and plate equations. However, in the present case, the situation is more complicated due to strong and nonlinear coupling of the equations.

A unique continuation result (point (ii) above) is needed at the level of absorbing lower-order terms. We show the validity of this unique continuation property by applying a new uniqueness result due to Isakov [9]. However, in order to do this, we need to establish a priori the regularity of the "overdetermined" solutions. This, in turn, can be proved for all finite energy solutions, provided, however, there is some light damping in the system (see Part I of Theorem 1.2). In the general case, however, we need to assume an incremental a priori smoothness of the initial data. (See Part II.) We note that this is in contrast with a modified von Karman system where one can show that the solutions overdetermined on the boundary display (without any interior damping) an arbitrary level of regularity. Proof of this "smoothing" property, carried out in [17], is based on sharp regularity of the Airy stress function established in [3]. Unfortunately, we do not have an analogue of this property valid for the present model.

2. Preliminary results and trace regularity. In this section we shall formulate and prove several preliminary estimates which deal with the trace regularity of solutions to the nonlinear equations given by (1.1). These results, while important in proving the main theorem, are also of independent interest in their own right.

2.1. Dissipativity equality. A starting point is, as usual, the dissipativity equality which states that the energy of the entire system is nonincreasing. This fact alone does not prove, of course, that the energy is decaying, but it is a necessary preliminary step of stability analysis.

LEMMA 2.1. *Let u, w be a finite energy solution of system* (1.1). *Then, for any $s \le t$,*

$$E(t) + 2\int_s^t \int_{\Gamma_1} [g(u_t)\cdot u_t + h_1(D_n w_t) D_n w_t + h_2(D_\tau w_t) D_\tau w_t] d\Gamma_1 dt + 2\int_s^t \int_\Omega [b_1 u_t \cdot u_t + b_2 w_t w_t] d\Omega dt = E(s). \tag{2.1}$$

Proof. The proof is standard and it follows by the classical energy type of argument (we multiply (1.1) by u_t, w_t, integrate over $\Omega \times (s,t)$, and apply the divergence theorem first to smooth solutions and then we extend it by density to all weak solutions). □

the corresponding trace estimates valid for (i) the linear model of dynamic elasticity and (ii) the linear Kirchhoff model obtained by methods of microlocal analysis in [5] and [21], respectively. Here the main idea is to obtain the estimates for the tangential derivatives on the boundary in terms of the velocity traces and lower-order terms. To formulate these results we introduce some notation. Let $T > 0$ be fixed. In fact, from now on we shall assume that T is sufficiently large and greater than the finite speed of propagation corresponding to equation (1.1). We denote $Q \equiv [0,T]\times\Omega$, $\Sigma_\alpha \equiv [\alpha, T-\alpha]\times\Gamma_1$, where $\alpha < T/2$. We also have that $\Sigma_1 \equiv [0,T]\times\Gamma_1$, $\Sigma_0 \equiv [0,T]\times\Gamma_0$, $\Sigma \equiv [0,T]\times\Gamma$.

We shall also use the following notation for Sobolev norms:

$$|u|_{\alpha,\Omega} \equiv |u|_{H^\alpha(\Omega)},\ |u|_{\alpha,\Gamma} \equiv |u|_{H^\alpha(\Gamma)}$$

and for the inner products

$$(u,v)_\Omega \equiv (u,v)_{L_2(\Omega)};\ \langle u,v\rangle_\Gamma \equiv (u,v)_{L_2(\Gamma)}.$$

Using the same symbol we shall also denote norms/inner products of two copies of L_2 or H^α spaces. This should not create any confusion, since the meaning will be clear from the context.

The constant C is a generic constant, different in various occurrences. $C(E(0))$ denotes the quantities bounded in terms of $E(0)$.

LEMMA 2.2. *Let u, w be a finite energy solution corresponding to the system (1.1). Then, for any $\epsilon < 1/4$, there exists a constant $C(E(0))$ such that the following trace regularity takes place:*

$$\begin{aligned}\int_{\Sigma\alpha} |\nabla u|^2 d\Sigma_\alpha \le C \int_{\Sigma_1} [|u_t|^2 + |g(u_t)|^2] dxdt \\ +C(E(0)) \int_0^T [|w|^2_{2-\epsilon,\Omega} + |u|^2_{1-\epsilon,\Omega}] dt. \end{aligned} \tag{2.2}$$

REMARK 2.1. *Notice that the regularity of the trace of ∇u proclaimed by Lemma 2.2 (see also Lemma 2.5 below) does not follow from the standard interior regularity of finite energy solutions via the trace theory. These are independent regularity results that rely heavily on microlocal arguments applied to both the dynamic system of elasticity and the dynamic Kirchhoff plate.*

Proof.

Step 1. We shall begin with the following trace regularity result valid [5] for the linear model of dynamic elasticity (see also [20] where the analogous result was proved for the wave equation). Define

$$F(x,y,t) \equiv \mathrm{div}[\mathcal{C}f(\nabla w(x,y,t))], \tag{2.3}$$

where w is a finite energy solution corresponding to the system (1.1). Then the solution u satisfies the following "linear" system of dynamic elasticity

$$u_{tt} + b_1 u_t - \mathrm{div}\mathcal{C}[\boldsymbol{\epsilon}(u)] = F \text{ in } Q. \tag{2.4}$$

According to [5], [28] for all $\epsilon < 1/2$, we have the estimate

$$\int_{\Sigma_\alpha} |\nabla u \cdot \tau|^2 d\Sigma_\alpha \le C \int_0^T [|u_t|^2_{0,\Gamma_1} + |\boldsymbol{\epsilon}(u)\cdot\nu|^2_{0,\Gamma_1} + |F|^2_{-1/2,\Omega} + |u|^2_{1-\epsilon,\Omega}] dt \tag{2.5}$$

and using the boundary conditions satisfied on Γ_1 we have

$$\int_{\Sigma_\alpha} |\nabla u \cdot \tau|^2 d\Sigma_\alpha \le C \int_0^T [|u_t|^2_{0,\Gamma_1} + |g(u_t)|^2_{0,\Gamma_1} + |f(\nabla w)|^2_{0,\Gamma_1} + |F|^2_{-1/2,\Omega} + |u|^2_{1-\epsilon,\Omega}]dt. \tag{2.6}$$

REMARK 2.2. *The estimate in inequality* (2.6), *when applied to the homogeneous system of dynamic elasticity, states that the traces of the tangential derivatives of* u *are bounded by the traces of velocity modulo lower-order terms. A result of similar nature was obtained first for the classical wave equation in* [20].

Step 2. We shall estimate the fourth term on the right-hand side of the inequality in (2.6).

PROPOSITION 2.3. *Let* $\epsilon < 1/2$. *Then the function* F *defined in* (2.3) *satisfies, for all* $t \ge 0$,

$$|F(t)|_{-1/2,\Omega} \le C|w(t)|_{2,\Omega}|w(t)|_{2-\epsilon,\Omega}. \tag{2.7}$$

Proof. Let $\phi \in H^{1/2}(\Omega)$. Direct computations give

$$(F,\phi)_{0,\Omega} = (\mathrm{div}[\mathcal{C}f(\nabla w)],\phi)_{0,\Omega} \le C|w|_{2,\Omega}|Dw\phi|_{0,\Omega}, \tag{2.8}$$

where D stands for a first-order differential operator. But

$$|Dw\phi|_{0,\Omega} \le C|w|_{W^1_{2p}(\Omega)}|\phi|_{L_{2q}(\Omega)},$$

where $1/p + 1/q = 1$. By Sobolev's embedding

$$H^{2-\epsilon}(\Omega) \subset W^1_4(\Omega), \quad H^{1/2}(\Omega) \subset L_4(\Omega)\ \epsilon \le 1/2 \tag{2.9}$$

we obtain

$$|Dw\phi|_{0,\Omega} \le C|w|_{2-\epsilon,\Omega}|\phi|_{1/2,\Omega}, \tag{2.10}$$

and going back to (2.8) we obtain

$$|(F,\phi)_{0,\Omega}| \le C|w|_{2,\Omega}|w|_{2-\epsilon,\Omega}|\phi|_{1/2,\Omega}, \tag{2.11}$$

which, via duality, proves the assertion in the proposition. □

Step 3. We shall next estimate the normal derivatives of the vector u.

PROPOSITION 2.4. *For all* $\epsilon < 1/4$ *we have*

$$\int_{\Sigma_\alpha} |\nabla u \cdot \nu|^2 d\Sigma_1 \le C \int_{\Sigma_\alpha} [|g(u_t)|^2 + |\nabla u \cdot \tau|^2] d\Sigma_\alpha + C(E(0)) \int_\alpha^{T-\alpha} |w|^2_{2-\epsilon,\Omega} dt. \tag{2.12}$$

Proof. Reading off the boundary conditions for the variable u we obtain the relation

$$\boldsymbol{\epsilon}(u) \cdot \nu = \vec{g}, \tag{2.13}$$

where we introduce the variable

$$\vec{g} \equiv -\mathcal{C}^{-1} g(u_t) - f(\nabla w) \cdot \nu,$$

where $\vec{g}$ satisfies the estimate

$$|\vec{g}|^2_{L_2(\Sigma_\alpha)} \le C[|g(u_t)|^2_{L_2(\Sigma_\alpha)} + |(\nabla w)^2|^2_{L_2(\Sigma_\alpha)}]$$
$$\le C\left[|g(u_t)|^2_{L_2(\Sigma_\alpha)} + |w|^2_{L_\infty(0,T;W_4^1(\Gamma_1))}\int_\alpha^{T-\alpha} |w(t)|^2_{W_4^1(\Gamma_1)}\right]$$
$$\le C|g(u_t)|^2_{L_2(\Sigma_\alpha)} + C(E(0))\int_\alpha^{T-\alpha} |w|^2_{2-\epsilon,\Omega} \tag{2.14}$$

and where we have used the estimate

$$|(\nabla w)^2|^2_{L_2(\Sigma_\alpha)} \le C\int_\alpha^{T-\alpha} |\nabla w(t)|^4_{L_4(\Gamma)}dt \le CE(0)\int_\alpha^{T-\alpha} |w(t)|^2_{2-\epsilon,\Omega}dt. \tag{2.15}$$

This last estimate follows, in turn, from the trace theorem, the dissipativity equality (2.1), and the following Sobolev embedding:

$$H^{3/2-\epsilon}(\Gamma) \subset W_4^1(\Gamma), \quad \epsilon \le 1/4. \tag{2.16}$$

On the other hand, denoting

$$\vec{d} \equiv \nabla u \cdot \tau$$

and writing

$$\boldsymbol{\epsilon}(u) \cdot \nu = \vec{g}, \tag{2.17}$$

$$\nabla u \cdot \tau = \vec{d}, \tag{2.18}$$

leads to the algebraic linear system of the form

$$A\vec{u} = [\vec{g}, \vec{d}]^T,$$

where $\vec{u} \equiv [u_{1,x}, u_{1,y}, u_{2,x}, u_{2,y}]$ and the determinant of the matrix A is equal to $-1/2$. Solving the above system pointwise and integrating the result over Σ_α yields the inequality

$$\int_{\Sigma_\alpha} [|D_x u|^2 + |D_y u|^2]d\Sigma_\alpha \le C\int_{\Sigma_\alpha} [|\vec{g}|^2 + |\vec{d}|^2]d\Sigma_\alpha. \tag{2.19}$$

The above estimate together with (2.14) leads to the result in (2.12). □

Step 4. Collecting the results of the estimates (2.6), (2.12), (2.7) we obtain

$$\int_{\Sigma\alpha} |\nabla u|^2 d\Sigma_\alpha \le C\int_{\Sigma_1} [|u_t|^2 + |g(u_t)|^2 + |f(\nabla w)|^2]dxdt$$
$$+C(E(0))\int_0^T \left[|w(t)|_{2-\epsilon,\Omega}dt + \int_0^T |u(t)|_{1-\epsilon}\right]dt. \tag{2.20}$$

Estimating the term $|f(\nabla w)|^2_{0,\Sigma_1}$ once more and using the inequality in (2.15) leads us to the desired result in Lemma 2.2. □

Our next result deals with the improved trace regularity for the vertical displacement w.

LEMMA 2.5. *Let u, w be a finite energy solution to* (1.1) *with the boundary conditions* (1.2). *Then*

$$\int_{\Sigma\alpha} [|D_n^2 w|^2 + |D_\tau^2 w|^2 + |D_n D_\tau w|^2]d\Sigma_\alpha \le C_T\int_{\Sigma_1} [|\nabla w_t|^2$$
$$+|h_1(D_n w_t)|^2 + |h_2(D_\tau w_t)|^2 + C(E(0))|u_t|^2]d\Sigma + C_T(E(0))\int_0^T |w|^2_{2-\epsilon,\Omega}dt. \tag{2.21}$$

Proof.
Step 1. Define

$$F \equiv \operatorname{div}[\mathcal{C}[\boldsymbol{\epsilon}(u) + f(\nabla w)]\nabla w];$$

then w satisfies the following equation of linear Kirchhoff plate

$$[I - \gamma\Delta]w_{tt} + b_2 w_t + D\Delta^2 w = F \text{ in } Q. \tag{2.22}$$

According to the estimates in Theorem 2.1 in [21] the following improved trace regularity is valid for (2.22)

$$\begin{aligned}\int_{\Sigma\alpha}[|D_n^2 w|^2 + |D_\tau^2 w|^2 + |D_n D_\tau w|^2]d\Sigma_\alpha \\ \leq C_T \int_{\Sigma_1}[|\nabla w_t|^2 \\ +|h_1(D_n w_t)|^2 + |h_2(D_\tau w_t)|^2]d\Sigma_1 \\ + \int_0^T [|\mathcal{C}[\boldsymbol{\epsilon}(u) + f(\nabla w)]\nu \cdot \nabla w|^2_{-1,\Gamma_1} \\ +|F(t)|^2_{(H^{3/2}(\Omega))'} + C|w(t)|^2_{2-\epsilon,\Omega}]dt \\ \leq C_T \int_{\Sigma_1}[|\nabla w_t|^2 + |h_1(D_n w_t)|^2 \\ +|h_2(D_\tau w_t)|^2]d\Sigma_1 \\ +C_T \int_0^T [|g(u_t)\cdot\nabla w|^2_{-1,\Gamma_1} + |F(t)|^2_{(H^{3/2}(\Omega))'} + |w(t)|^2_{2-\epsilon,\Omega}]dt,\end{aligned} \tag{2.23}$$

where we used the boundary conditions for the "u" equation.

Step 2. We shall estimate first the contribution of the term F in (2.23). This is accomplished by the proposition below.

PROPOSITION 2.6. *For all $\epsilon < 1/2$ the following estimates hold:*

$$|F(t)|_{(H^{3/2}(\Omega))'} \leq C[|N(u(t), w(t))|_{0,\Omega}|w(t)|_{2-\epsilon,\Omega} + |u_t(t)|_{0,\Gamma_1}|w(t)|_{2-\epsilon,\Omega}]. \tag{2.24}$$

Hence

$$\int_0^T |F(t)|^2_{(H^{3/2}(\Omega))'}dt \leq C(E(0))\int_0^T [|w(t)|^2_{2-\epsilon,\Omega} + |u_t(t)|^2_{0,\Gamma_1}]dt. \tag{2.25}$$

Proof. Let $\phi \in H^{3/2}(\Omega)$. Applying the divergence theorem and accounting for the boundary conditions we obtain

$$\begin{aligned}(F,\phi)_{0,\Omega} = (\operatorname{div}[\mathcal{C}[\boldsymbol{\epsilon}(u) + f(\nabla w)]\nabla w], \phi)_{0,\Omega} = -(\mathcal{C}[\boldsymbol{\epsilon}(u) + f(\nabla w)]\nabla w, \nabla\phi)_{0,\Omega} \\ +(\mathcal{C}[\boldsymbol{\epsilon}(u) + f(\nabla w)]\nu\cdot\nabla w, \phi)_{0,\Gamma_1} \\ = (\mathcal{C}[\boldsymbol{\epsilon}(u) + f(\nabla w)], \nabla\phi\times\nabla w)_{0,\Omega} + (g(u_t)\nabla w, \phi)_{0,\Gamma_1}.\end{aligned} \tag{2.26}$$

The first interior term in (2.26) is estimated as follows:

$$\begin{aligned}(\mathcal{C}[\boldsymbol{\epsilon}(u) + f(\nabla w)], \nabla\phi\times\nabla w)_{0,\Omega} \leq C|N(u,w)|_{0,\Omega}|\nabla w\times\nabla\phi|_{0,\Omega} \\ \leq C|N(u,w)|_{0,\Omega}|\nabla w|_{L_4(\Omega)}|\nabla\phi|_{L_4(\Omega)} \leq C|N(u,w)|_{0,\Omega}|w|_{2-\epsilon,\Omega}|\phi|_{3/2,\Omega},\end{aligned} \tag{2.27}$$

where we have used Sobolev's embedding (2.9).

To estimate the boundary term that appeared in (2.26) we proceed as follows:

$$(g(u_t)\nabla w, \phi)_{0,\Gamma_1} \le |g(u_t)|_{0,\Gamma_1} |\nabla w \phi|_{0,\Gamma_1}$$
$$\le C|u_t|_{0,\Gamma_1}|w|_{1,\Gamma_1}|\phi|_{1,\Gamma_1} \le C|u_t|_{0,\Gamma_1}|w|_{3/2,\Omega}|\phi|_{3/2,\Omega}. \tag{2.28}$$

Combining inequalities in (2.26)–(2.28) leads to the result stated in the first part of the proposition. The second inequality follows simply by integrating the first part and taking into account the dissipativity property (2.1). □

Step 3.

PROPOSITION 2.7. *We have*

$$\int_0^T |g(u_t)\nabla w|^2_{-1,\Gamma_1} dt \le CE(0)|u_t|^2_{0,\Sigma_T}. \tag{2.29}$$

Proof. As a bypass of the computations in (2.28) we obtain

$$|g(u_t)\nabla w|_{-1,\Gamma_1} \le C|u_t|_{0,\Gamma_1}|w|_{1,\Gamma_1} \le C|u_t|_{0,\Gamma_1}|w|_{2-\epsilon,\Omega}.$$

Integrating the above inequality with respect to time and recalling the dissipativity property we obtain the result claimed in Proposition 2.7. □

Step 4. Complete the proof of Lemma 2.5.

Combining the results of both propositions together with inequality (2.23) leads to the final conclusion of Lemma 2.5. □

3. Stabilizability estimate. The main aim in this section is to prove the following stabilizability estimate in Lemma 3.1.

LEMMA 3.1. *Let u, w be a regular solution to* (1.1). *Assume the geometric condition on* Γ_0 (1.9). *Then there exists T large enough such that for any constant $\epsilon < 1/4$ the following estimate takes place:*

$$E(0) + E(T) + \int_0^T E(t)dt \le C_T(E(0)) \int_{\Sigma_1} [|u_t|^2 + |\nabla w_t|^2 + |g(u_t)|^2 + |h_1(D_n w_t)|^2$$
$$+|h_2(D_\tau w_t)|^2]d_{\Sigma_1} + C\int_0^T [(b_1 u_t, u_t)_\Omega + (b_2 w_t, w_t)_\Omega]dt + C_T(E(0))lot(u,w), \tag{3.1}$$

where we have used the notation for the lower-order terms (lot(u,w))

$$lot(u,w) \equiv \int_0^T [|u(t)|^2_{1-\epsilon,\Omega} + |w(t)|^2_{2-\epsilon,\Omega}]dt. \tag{3.2}$$

The estimate of Lemma 3.1, critical to the proof of the main stabilizability result, is an inverse type of estimate. Indeed, it allows us to reconstruct the energy of the system, modulo lower-order terms, from the measurements of velocities on the boundary. The remainder of this section is devoted to the proof of Lemma 3.1. Here, the strategy used for the proof is to first apply the usual " multipliers" method (for an exposition of this method see the books [15], [11], [10] and references therein), which leads to the estimate for the energy in terms of *all* boundary traces and lower-order terms. The next crucial step is to eliminate the "unwanted" boundary traces by using sharp trace regularity results presented in section 2.

REMARK 3.1. *The result of Lemma* 3.1 *can be extended to hold for all finite energy (weak) solutions. Indeed, this can be accomplished by applying a special "regularization" argument as in* [19]. *In order to avoid the additional technical complications and for the sake of clarity of exposition we shall not do this here, and for details we refer the reader to* [19], [6], [7].

3.1. Variational formulation and preliminary identities. We shall begin by writing system (1.1) in a variational form. To this end let us have two test functions $\phi \in H^1(\Omega) \times H^1(\Omega)$ and $\psi \in H^2(\Omega)$. The von Karman system admits the following variational form:

$$(u_{tt}, \phi)_\Omega + (b_1 u_t, \phi)_\Omega + (\mathcal{C}[\boldsymbol{\epsilon}(u) + f(\nabla w)], \boldsymbol{\epsilon}(\phi))_\Omega + \langle g(u_t), \phi \rangle_{\Gamma_1} - \langle \mathcal{C}\boldsymbol{\epsilon}(u)\nu, \phi \rangle_{\Gamma_0} = 0, \tag{3.3}$$

$$\begin{aligned}(w_{tt}, \psi)_\Omega + \gamma(\nabla w_{tt}, \nabla\psi)_\Omega + a(w, \psi) + (b_2 w_t, \psi)_\Omega + (\mathcal{C}[\boldsymbol{\epsilon}(u) + f(\nabla w)], \nabla\psi \times \nabla w)_\Omega \\ + \langle h_1(D_n w_t), D_n\psi \rangle_{\Gamma_1} + \langle h_2(D_\tau w_t), D_\tau \psi \rangle_{\Gamma_1} - \langle \Delta w, D_n \psi \rangle_{\Gamma_0} = 0.\end{aligned} \tag{3.4}$$

Note that we have used the boundary conditions satisfied on Γ.

We shall apply this variational form with various choices of test functions ϕ and ψ. In order to facilitate verification of rather tedious computations below, we will provide a few elementary tensor identities.

In what follows the vector field $\boldsymbol{h}$ always denotes the radial vector field

$$\boldsymbol{\epsilon}(\boldsymbol{h}\nabla u) = \boldsymbol{\epsilon}(u) + M, \tag{3.5}$$

where the tensor M is given by

$$M \equiv \begin{bmatrix} D^2_{x_1,x_i} u_1 \boldsymbol{h}_i & 1/2[D^2_{x_2,x_i} u_1 \boldsymbol{h}_i + D^2_{x_1,x_i} u_2 \boldsymbol{h}_i] \\ 1/2[D^2_{x_2,x_i} u_1 \boldsymbol{h}_i + D^2_{x_1,x_i} u_2 \boldsymbol{h}_i] & D^2_{x_2,x_i} u_2 \boldsymbol{h}_i \end{bmatrix}, \tag{3.6}$$

and we have adopted double index notation to indicate the summation of the terms. If A is any symmetric fourth-order tensor identified by its coefficients $a_{i,j}$,

$$A \equiv \{a_{i,j}\},$$

then it is straightforward to show that

$$A \cdot M = a_{k,j} D^2_{x_k,x_i} u_j \boldsymbol{h}_i, \tag{3.7}$$

$$\nabla(\nabla w \boldsymbol{h}) = \nabla w + [w_{x_1,x_i} \boldsymbol{h}_i, w_{x_2,x_i} \boldsymbol{h}_i], \tag{3.8}$$

and

$$\nabla(\nabla w \boldsymbol{h}) \times \nabla w = \nabla w \times \nabla w + \nabla w \times [w_{x_1,x_i} \boldsymbol{h}_i, w_{x_2,x_i} \boldsymbol{h}_i], \tag{3.9}$$

$$A \cdot (\nabla(\nabla w \boldsymbol{h}) \times \nabla w) = A \cdot \nabla w \times \nabla w + a_{k,j} w_{x_j} w_{x_k,x_i} \boldsymbol{h}_i. \tag{3.10}$$

Let B be another symmetric tensor such that

$$a_{j,i} = c_{j,l} b_{l,i}$$

with constant and symmetric coefficients $c_{j,i}$. Then

$$\operatorname{div}[A \cdot B\boldsymbol{h}] = 2A \cdot B + c_{i,l} D_{x_k} b_{l,j} b_{i,j} \boldsymbol{h}_k = 2A \cdot B + 2a_{j,i} D_{x_k} b_{j,i} \boldsymbol{h}_k. \tag{3.11}$$

In the particular case when the tensors A and B are given by

$$A = \mathcal{C}[\boldsymbol{\epsilon}(u) + f(\nabla w)],$$

$$B = \boldsymbol{\epsilon}(u) + f(\nabla w),$$

the formula above reads

$$\begin{aligned}\operatorname{div}[\mathcal{C}[\boldsymbol{\epsilon}(u) + f(\nabla w)] \cdot [\boldsymbol{\epsilon}(u) + f(\nabla w)]\boldsymbol{h}] = 2\mathcal{C}[\boldsymbol{\epsilon}(u) + f(\nabla w)] \cdot [\boldsymbol{\epsilon}(u) + f(\nabla w)] \\ + 2a_{i,j}[D^2_{x_k,x_j} u_i + w_{x_j} D^2_{x_i,x_k} w]\boldsymbol{h}_k.\end{aligned} \tag{3.12}$$

3.2. First estimate. In this subsection we shall prove a preliminary estimate which shows that the energy of the system is bounded by the boundary traces modulo the lower-order terms. Computations carried below, based on the "multiplier's method," are reminiscent of those performed earlier in [13], [12], and later in [22].

LEMMA 3.2. *Let u, w be a regular solution to* (1.1). *Assume the geometric condition* (1.9) *holds on Γ_0. Then there exists T large enough such that for any constant $\epsilon < 1/4$ the following estimate takes place:*

$$E(T) + \int_0^T E(t)dt \le C_T(E(0)) \int_{\Sigma_1} [|u_t|^2 + |\nabla w_t|^2 + |g(u_t)|^2 + |h_1(D_n w_t)|^2 + |h_2(D_\tau w_t)|^2]d\Sigma_1 + C \int_{\Sigma_1} [|\nabla u|^2 + |D_n^2 w|^2 + |D_n D_\tau w|^2 + |D_\tau^2 w|^2]d\Sigma_1 + C \int_0^T [(b_1 u_t, u_t)_\Omega + (b_2 w_t, w_t)_\Omega]dt + C_T(E(0)) lot(u, w). \tag{3.13}$$

Proof.

Step 1. We use $\phi \equiv \nabla u \boldsymbol{h}$ as a test function in the first variational equality. By the virtue of (3.5), (3.7) applied with

$$A \equiv \mathcal{C}[\boldsymbol{\epsilon}(u) + f(\nabla w)],$$

we obtain

$$(\mathcal{C}[\boldsymbol{\epsilon}(u) + f(\nabla w)], \epsilon(\nabla u \boldsymbol{h})) = (\mathcal{C}[\boldsymbol{\epsilon}(u) + f(\nabla w)], \epsilon(u)) + (a_{i,j}, D^2_{x_k,x_j} u_i \boldsymbol{h}_k) \tag{3.14}$$

and integrating over Q

$$(u_t, \boldsymbol{h}\nabla u)_\Omega|_0^T + \int_0^T (b_1 u_t, \boldsymbol{h}\nabla u)_\Omega dt - 1/2 \int_\Sigma |u_t|^2 \boldsymbol{h} \cdot \nu d\Sigma + \int_{Q_T} |u_t|^2 dQ + \int_0^T [(\mathcal{C}[\boldsymbol{\epsilon}(u) + f(\nabla w)], \epsilon(u))_\Omega + (a_{i,j}, D^2_{x_k,x_j} u_i \boldsymbol{h}_k)_\Omega + \langle g(u_t), \boldsymbol{h}\nabla u\rangle_{\Gamma_1} - \langle \mathcal{C}\boldsymbol{\epsilon}(u), \epsilon(u)\nu h\rangle_{\Gamma_0}]dt = 0, \tag{3.15}$$

where we have used the fact that u vanishes on Γ_0 and, therefore,

$$\mathcal{C}\boldsymbol{\epsilon}(u)\nu\nabla u\boldsymbol{h} = \mathcal{C}\boldsymbol{\epsilon}(u) \cdot \boldsymbol{\epsilon}(u)\nu\boldsymbol{h} \text{ on } \Gamma_0. \tag{3.16}$$

To see (3.16), it suffices to notice the following identities taking place on Γ_0

$$\nabla u\boldsymbol{h} = D_n u \nu \boldsymbol{h}, \ \textit{trace}\ \boldsymbol{\epsilon}(u) = D_n u \nu,$$
$$\boldsymbol{\epsilon}(u)\nu = [D_n u_1 \nu_1^2, D_n u_2 \nu_2^2] + (1/2)\nu^T (D_n u \nu^T),$$
$$\boldsymbol{\epsilon}(u) \cdot \boldsymbol{\epsilon}(u) = D_n u_i^2 \nu_i^2 + (1/2)(D_n u \nu^T)^2.$$

Hence

$$\boldsymbol{\epsilon}(u)\nu\nabla u\boldsymbol{h} = \{[D_n u_1 \nu_1^2, D_n u_2 \nu_2^2] + (1/2)\nu^T (D_n u \nu^T)\} D_n u \nu \boldsymbol{h} = (D_n u_i^2 \nu_i^2 + (1/2)(D_n u \nu^T)^2)\nu\boldsymbol{h} = \boldsymbol{\epsilon}(u) \cdot \boldsymbol{\epsilon}(u))\nu\boldsymbol{h}. \tag{3.17}$$

Similarly

$$\textit{trace}\ (\boldsymbol{\epsilon}(u))I\nu\nabla u\boldsymbol{h} = (D_n u\nu)^2 \nu\boldsymbol{h} = \textit{trace}\ \boldsymbol{\epsilon}(u) I \cdot \boldsymbol{\epsilon}(u)\nu\boldsymbol{h}. \tag{3.18}$$

The identity in (3.16) follows now from (3.17) and (3.18) and the definition of the tensor $\mathcal{C}$.

To obtain appropriate estimates for the second variable we apply the variational form with the test function $\psi \equiv \nabla w \boldsymbol{h}$. From (3.10) we have

$$(\mathcal{C}[\boldsymbol{\epsilon}(u) + f(\nabla w)], \nabla(\nabla w \boldsymbol{h}) \times \nabla w) = 2(\mathcal{C}\boldsymbol{\epsilon}(u) + f(\nabla w), f(\nabla w)) + (a_{i,j} w_{x_j}, D^2_{x_i,x_k} w \boldsymbol{h}_k). \tag{3.19}$$

Integrating the result over Q we have

$$\begin{aligned}(w_t, \boldsymbol{h}\nabla w)_\Omega|_0^T + \gamma(\nabla w_t, \nabla(\boldsymbol{h}\nabla w))_\Omega|_0^T + \int_0^T [|w_t|^2_{0,\Omega} + (b_2 w_t, \boldsymbol{h}\nabla w)_\Omega + a(w,w) \\ +2(\mathcal{C}[\epsilon(u) + f(\nabla w)], f(\nabla w))_\Omega + (a_{i,j} w_{x_j}, D^2_{x_i,x_k} w \boldsymbol{h}_k)_\Omega] dt \\ = 1/2 \int_{\Sigma_1} [|w_t|^2 + \gamma|\nabla w_t|^2 - D/2[w^2_{x,x} + w^2_{y,y} + 2\nu w_{x,x} w_{y,y} + 2(1-\mu) w^2_{x,y}] \boldsymbol{h}\nu \\ +h_1(D_n w_t) D_n(\boldsymbol{h}\nabla w) + h_2(D_\tau w_t, D_\tau \boldsymbol{h}\nabla w)] d\Sigma_1 - D/2 \int_{\Sigma_0} |\Delta w|^2 \boldsymbol{h}\nu d\Sigma_0.\end{aligned} \tag{3.20}$$

Adding inequalities in (3.15) and (3.20) yields

$$\begin{aligned}\int_0^T [|u_t|^2_{0,\Omega} + |w_t|^2_{0,\Omega} + a(w,w) + (\mathcal{C}N(u,w), N(u,w))_\Omega + (\mathcal{C}N(u,w), f(\nabla w))_\Omega \\ +(a_{i,j}, [D^2 x_k, x_j u_i + w_{x_j} D^2_{x_j,x_k} w] \boldsymbol{h}_k)_\Omega] dt \le C[E(0) + E(T)] \\ +1/2 \int_{\Sigma_1} [|u_t|^2 + g(u_t) \boldsymbol{h}\nabla u + |w_t|^2 + \gamma|\nabla w_t|^2 + h_1(D_n w_t) D_n(\boldsymbol{h}\nabla w) \\ +h_2(D_\tau w_t) D_\tau(\boldsymbol{h}\nabla w) - D/2[w^2_{x,x} + w^2_{y,y} + 2\nu w_{x,x} w_{y,y} + 2(1-\nu) w^2_{x,y}] \boldsymbol{h}\nu] d\Sigma_1 \\ + \int_0^T [(b_1 u_t, \boldsymbol{h}\nabla u)_\Omega + (b_2 w_t, \boldsymbol{h}\nabla w)_\Omega] dt + \int_{\Sigma_0} [\mathcal{C}(\boldsymbol{\epsilon}(u)) \nu \boldsymbol{h}\nabla u + |\Delta w|^2 \boldsymbol{h}\nu] d\Sigma_0.\end{aligned} \tag{3.21}$$

Using the relations (3.12) in (3.21), applying the divergence theorem, and recalling (3.16) yields

$$\begin{aligned}\int_0^T [|u_t|^2_{0,\Omega} + |w_t|^2_{0,\Omega} + a(w,w) + (\mathcal{C}N(u,w), f(\nabla w))_\Omega] dt \le C[E(0) + E(T)] \\ +1/2 \int_{\Sigma_1} [|u_t|^2 + g(u_t) \boldsymbol{h}\nabla u + |w_t|^2 + \gamma|\nabla w_t|^2 + h_1(D_n w_t) D_n(\boldsymbol{h}\nabla w) \\ +h_2(D_\tau w_t) D_\tau(\boldsymbol{h}\nabla w) - D/2[w^2_{x,x} + w^2_{y,y} + 2\nu w_{x,x} w_{y,y} + 2(1-\nu) w^2_{x,y}] \boldsymbol{h}\nu] d\Sigma_1 \\ + \int_{\Sigma_0} [1/2 \mathcal{C}(\boldsymbol{\epsilon}(u)) \nu \boldsymbol{h}\nabla u + |\Delta w|^2 \boldsymbol{h}\nu] d\Sigma_0 - 1/2 \int_{\Sigma_1} \mathcal{C}N(u,w) \cdot N(u,w) \boldsymbol{h}\nu d\Sigma_1 \\ + \int_0^T [(b_1 u_t, \boldsymbol{h}\nabla u)_\Omega + (b_2 w_t, \boldsymbol{h}\nabla w)_\Omega] dt.\end{aligned} \tag{3.22}$$

On the other hand,

$$\int_{\Sigma_1} |N(u,w)|^2 d\Sigma_1 \le C \int_{\Sigma_1} |\nabla u|^2_{0,\Gamma_1} d\Sigma_1 + C(E(0)) \int_0^T |w|_{2-\epsilon,\Omega} dt, \tag{3.23}$$

$$(b_1 u_t, \boldsymbol{h}\nabla u)_\Omega \le \epsilon_1 |b_1^{*1/2} (\boldsymbol{h}\nabla u)|^2_{0,\Omega} + C_{\epsilon_1}(b_1 u_t, u_t)_\Omega \le \epsilon_0 |u|^2_{1,\Omega} + C_{\epsilon_0}(b_1 u_t, u_t)_\Omega, \tag{3.24}$$

$$(b_2 w_t, \boldsymbol{h}\nabla w)_\Omega \le \epsilon_1 |b_2^{*1/2} (\boldsymbol{h}\nabla w)|^2_{0,\Omega} + C_{\epsilon_1}(b_2 w_t, w_t)_\Omega \le \epsilon_0 |w|^2_{1,\Omega} + C_{\epsilon_0}(b_2 w_t, w_t)_\Omega. \tag{3.25}$$

Combining with (3.22) we obtain

$$\int_0^T [|u_t|^2_{0,\Omega} + |w_t|^2_{0,\Omega} + a(w,w) + (\mathcal{C}N(u,w), f(\nabla w))_\Omega]dt \le C[E(0) + E(T)$$
$$+1/2 \int_{\Sigma_1} [|u_t|^2 + g(u_t)\boldsymbol{h}\nabla u + |w_t|^2 + \gamma|\nabla w_t|^2 + h_1(D_n w_t)D_n(\boldsymbol{h}\nabla w) + h_2(D_\tau w_t)$$
$$\cdot D_\tau(\boldsymbol{h}\nabla w)]d\Sigma_1 - \int_{\Sigma_1} D/2[w^2_{x,x} + w^2_{y,y} + 2\nu w_{x,x}w_{y,y} + 2(1-\nu)w^2_{x,y}]\boldsymbol{h}\nu d\Sigma_1$$
$$+ \int_{\Sigma_0} [1/2\mathcal{C}(\boldsymbol{\epsilon}(u))\nu\boldsymbol{h}\nabla u + |\Delta w|^2\boldsymbol{h}\nu]d\Sigma_0 + \int_{\Sigma_1} |\nabla u|^2_{0,\Gamma_1} d\Sigma_1 + C(E(0)) \int_0^T |w|_{2-\epsilon,\Omega}dt$$
$$+C_{\epsilon_0} \int_0^T [(b_1 u_t, u_t)_\Omega + (b_2 w_t, w_t)_\Omega]dt + \epsilon_0 \int_0^T |u|^2_{1,\Omega}dt. \tag{3.26}$$

REMARK 3.2. *Note that in a special case when the geometric conditions are assumed also on* Γ_1*, then the last three boundary integrals in* (3.26) *can be discarded.*

Step 2. We apply the variational equality with the following test functions: $\phi = u$, $\psi = w$. We have

$$(u_t, u)_\Omega|_0^T + \int_0^T [(\mathcal{C}N(u,w), \boldsymbol{\epsilon}(u))_\Omega - |u_t|^2_{0,\Omega} + (b_1 u_t, u)_\Omega + \langle g(u_t), u\rangle_{\Gamma_1}$$
$$-\langle \mathcal{C}\boldsymbol{\epsilon}(u)\nu, u\rangle_{\Gamma_0}]dt = 0, \tag{3.27}$$

$$(w_t, w)_\Omega|_0^T + \gamma(\nabla w_t, \nabla w)_\Omega|_0^T + \int_0^T [a(w,w) + (b_2 w_t, w)_\Omega + (\mathcal{C}N(u,w), \nabla w \times \nabla w)_\Omega$$
$$-|w_t|^2_{0,\Omega} - \gamma|\nabla w_t|^2_{0,\Omega} - \langle h_1(D_n w_t), D_n w\rangle_{\Gamma_1} + \langle h_2(D_\tau w_t), D_\tau w\rangle_{\Gamma_1}]dt = 0. \tag{3.28}$$

Here we took into account the boundary conditions on Γ_0 for the variable w.

Multiplying equality (3.27) by a positive constant A, multiplying equality (3.28) by a negative constant B, adding the result to the inequality (3.26), and accounting for a correct sign of the boundary terms on Γ_0 yields

$$\int_0^T [|u_t|^2_{0,\Omega} + |w_t|^2_{0,\Omega} + a(w,w) + (\mathcal{C}N(u,w), f(\nabla w))_\Omega$$
$$+A\{(\mathcal{C}N(u,w), \boldsymbol{\epsilon}(u))_\Omega - |u_t|^2_{0,\Omega}\}$$
$$+B\{a(w,w) + 2(\mathcal{C}N(u,w), f(\nabla w))_\Omega - |w_t|^2_{0,\Omega} - \gamma|\nabla w_t|^2_{0,\Omega}\}]dt$$
$$\le C[E(0) + E(T)] + C \int_{\Sigma_1} [|u_t|^2 + g(u_t)\boldsymbol{h}\nabla u + |w_t|^2 + \gamma|\nabla w_t|^2 + h_1(D_n w_t)D_n(\boldsymbol{h}\nabla w)$$
$$+h_2(D_\tau w_t)D_\tau(\boldsymbol{h}\nabla w) - D/2[w^2_{x,x} + w^2_{y,y} + 2\nu w_{x,x}w_{y,y} + 2(1-\nu)w^2_{x,y}]\boldsymbol{h}\nu$$
$$-|\nabla u|^2]d\Sigma_1 + C \int_0^T [\langle g(u_t), u\rangle_{\Gamma_1} + \langle h_1(D_n w_t), D_n w\rangle_{\Gamma_1} + \langle h_2(D_\tau w_t), D_\tau w\rangle_{\Gamma_1}$$
$$+C_{\epsilon_0}((b_1 u_t, u_t)_\Omega + (b_2 w_t, w_t)_\Omega) + (b_1 u_t, u)_\Omega + (b_2 w_t, w)_\Omega + \epsilon_0|u|^2_{1,\Omega}$$
$$+C_\epsilon(E(0))|w|^2_{2-\epsilon,\Omega}]dt. \tag{3.29}$$

Selecting constants $A = 1/2, B = -1/4$ and upper bounding the boundary terms by

the trace theorem yields

$$\int_0^T [|u_t|^2_{0,\Omega} + |w_t|^2_{0,\Omega} + \gamma|\nabla w_t|^2_{0,\Omega} + a(w,w) + (\mathcal{C}N(u,w), N(u,w))_\Omega]dt$$
$$\leq C[E(0)+E(T)] + C\int_{\Sigma_1} [|u_t|^2 + |g(u_t)|^2 + |w_t|^2 + \gamma|\nabla w_t|^2 + |h_1(D_n w_t)|^2$$
$$+|h_2(D_\tau w_t)|^2 + |\nabla u|^2 + |D_n^2 w|^2 + |D_n D_\tau w|^2 + |D_\tau^2 w|^2]d\Sigma_1$$
$$+\int_0^T [C_{\epsilon_0}((b_1 u_t, u_t)_\Omega + (b_2 w_t, w_t)_\Omega) + \epsilon_0 |u|^2_{1,\Omega}]dt + C(E(0))\int_0^T [|w|^2_{2-\epsilon,\Omega} + |u|^2_{1-\epsilon,\Omega}]dt. \tag{3.30}$$

Recalling the definition of the energy together with the coercivity of E_p in $H^2(\Omega) \times H^1(\Omega)$ and taking ϵ_0 small enough yields

$$1/2TE(T) + 1/2\int_0^T E(t)dt \leq CE(T)$$
$$+C\int_{\Sigma_1} [|u_t|^2 + |g(u_t)|^2 + |w_t|^2 + \gamma|\nabla w_t|^2 + |h_1(D_n w_t)|^2$$
$$+|h_2(D_\tau w_t)|^2 + |\nabla u|^2 + |D_n^2 w|^2 + |D_n D_\tau w|^2 + |D_\tau^2 w|^2]d\Sigma_1$$
$$+C\int_0^T [(b_1 u_t, u_t)_\Omega + (b_2 w_t, w_t)_\Omega]dt + C(E(0))\int_0^T [|w|^2_{2-\epsilon,\Omega} + |u|^2_{1-\epsilon,\Omega}]dt. \tag{3.31}$$

Taking T large enough (note that the constant C in front of the term $E(T)$ is independent on T) yields the conclusion of Lemma 3.2 □

REMARK 3.3. *We note that the proof carried above provides a (new?) uniqueness result for the problem which is overdetermined on the boundary and defined on the star-shaped domain. Indeed, in the special case when the geometric conditions are satisfied also on* Γ_1,$b_1 = b_2 = 0$, *and solutions are required to have zero velocity traces on the boundary for* $t \in [0,T]$, *with* T *sufficiently large, the argument above gives*

$$E(t) \equiv 0.$$

Indeed, applying inequalities (3.22), (3.27), (3.28) *to the case considered above gives*

$$\int_0^T [|u_t|^2_{0,\Omega} + |w_t|^2_{0,\Omega} + a(w,w) + (\mathcal{C}N(u,w), f(\nabla w))_\Omega]dt \leq C[E(0)+E(T)], \tag{3.32}$$

$$(u_t, u)_\Omega|_0^T + \int_0^T [(\mathcal{C}N(u,w), \boldsymbol{\epsilon}(u))_\Omega - |u_t|^2_{0,\Omega}]dt = 0, \tag{3.33}$$

$$(w_t, w)_\Omega|_0^T + \gamma(\nabla w_t, \nabla w)_\Omega|_0^T + \int_0^T [a(w,w) + (\mathcal{C}N(u,w), \nabla w \times \nabla w)_\Omega$$
$$-|w_t|^2_{0,\Omega} - \gamma|\nabla w_t|^2_{0,\Omega}]dt = 0. \tag{3.34}$$

Multiplying equality (3.33) *by a positive constant* A, *multiplying equality* (3.34) *by a negative constant* B, *and adding the result to the inequality* (3.32) *yields*

$$\int_0^T [|u_t|^2_{0,\Omega} + |w_t|^2_{0,\Omega} + a(w,w) + (\mathcal{C}N(u,w), f(\nabla w))_\Omega$$
$$+A\{(\mathcal{C}N(u,w), \boldsymbol{\epsilon}(u))_\Omega - |u_t|^2_{0,\Omega}\}$$
$$+B\{a(w,w) + 2(\mathcal{C}N(u,w), f(\nabla w))_\Omega - |w_t|^2_{0,\Omega} - \gamma|\nabla w_t|^2_{0,\Omega}\}]dt$$
$$\leq C[E(0)+E(T)]. \tag{3.35}$$

Selecting constants $A = 1/2, B = -1/4$ *yields*

$$\int_0^T [|u_t|^2_{0,\Omega} + |w_t|^2_{0,\Omega} + \gamma|\nabla w_t|^2_{0,\Omega} + a(w,w) + (\mathcal{C}N(u,w), N(u,w))_\Omega]dt \leq C[E(0) + E(T)]. \tag{3.36}$$

Since, in our case, $E(t) = E(T)$, (3.36) *gives*

$$TE(T) = \int_0^T E(t)dt \leq CE(T) \tag{3.37}$$

and taking T *large enough we conclude* $E(t) \equiv 0$. *Hence* $u = w \equiv 0$, *which is the desired uniqueness result for the nonlinear problem at hand.*

3.3. Absorption of boundary traces and completion of the proof of Lemma 3.1. In this subsection we shall show that the boundary traces involving the second-order derivatives of w and the first-order derivatives of u are redundant. This will be done with the help of trace regularity results formulated in section 2.

LEMMA 3.3. *Under the assumptions of Lemma* 3.2 *we have*

$$E(T) + \int_0^T E(t)dt \leq C_T(E(0)) \int_{\Sigma_1} [|u_t|^2 + |\nabla w_t|^2 + |g(u_t)|^2 + |h_1(D_n w_t)|^2 + |h_2(D_\tau w_t)|^2]d_{\Sigma_1} + C\int_0^T [(b_1 u_t, u_t)_\Omega + (b_2 w_t, w_t)_\Omega]dt + C_{T,\epsilon}(E(0))lot(u,w). \tag{3.38}$$

Proof. From the result of Lemma 3.2 applied to the interval $[\alpha, T-\alpha]$ we obtain

$$\begin{aligned}\int_\alpha^{T-\alpha} E(t)dt &\leq C_T(E(0)) \int_{\Sigma_\alpha} [|u_t|^2 + |\nabla w_t|^2 + |g(u_t)|^2 + |h_1(D_n w_t)|^2 \\ &+|h_2(D_\tau w_t)|^2]d_{\Sigma_\alpha} + C\int_{\Sigma_\alpha} [|\nabla u|^2 + |D_n^2 w|^2 + |D_n D_\tau w|^2 + |D_\tau^2 w|^2]d\Sigma_\alpha \\ &+C\int_0^T [(b_1 u_t, u_t)_\Omega + (b_2 w_t, w_t)_\Omega]dt + C_T(E(0))lot(u,w).\end{aligned} \tag{3.39}$$

Here we took advantage of the dissipativity property in Lemma 2.1 which allows us to upper bound $E(\alpha)$ by $E(0)$. On the other hand, from regularity results stated in Lemmas 2.2 and 2.5 we infer the estimates

$$\begin{aligned}&\int_{\Sigma\alpha} [|\nabla u|^2 + |D_n^2 w|^2 + |D_\tau^2 w|^2 + |D_n D_\tau w|^2]d\Sigma_\alpha \\ &\leq C\int_{\Sigma_1} [|u_t|^2 + |g(u_t)|^2 + |\nabla w_t|^2 + h_1(D_n w_t)|^2 + |h_2(D_\tau w_t|^2 \\ &+C(E(0))|u_t|^2]d\Sigma_1 + C(E(0))\int_0^T [|w|^2_{2-\epsilon,\Omega} + |u|^2_{1-\epsilon,\Omega}]dt.\end{aligned} \tag{3.40}$$

Combining (3.39) and (3.40) and recalling, again, the dissipativity equality in Lemma

2.1 gives

$$E(T-\alpha)+\int_{\alpha}^{T-\alpha}E(t)dt\le C_T(E(0))\int_{\Sigma_1}[|u_t|^2+|\nabla w_t|^2+|g(u_t)|^2$$
$$+|h_1(D_nw_t)|^2+|h_2(D_\tau w_t)|^2]d_{\Sigma_1}+C\int_0^T[(b_1u_t,u_t)_\Omega+(b_2w_t,w_t)_\Omega]dt$$
$$+C(E(0))\int_0^T[|w|^2_{2-\epsilon,\Omega}+|u|^2_{1-\epsilon,\Omega}]dt. \tag{3.41}$$

To complete the proof of the lemma we need to estimate the contribution of the energy on the subintervals $[0,\alpha]$ and $[T-\alpha,T]$. To accomplish this we denote the right-hand side of (3.41) by $\mathcal{F}$. From dissipativity relation (2.1) and (3.41) we have, for all $t\in[0,T]$,

$$E(t)\le E(T-\alpha)+2\int_{\Sigma_1}[|u_t|^2+|\nabla w_t|^2+|g(u_t)|^2+|h_1(D_nw_t)|^2$$
$$+|h_2(D_\tau w_t)|^2]d\Sigma_1\le C\mathcal{F} \tag{3.42}$$

and by (3.41)

$$\int_0^{\alpha}E(t)dt+\int_{T-\alpha}^{T}E(t)dt\le C\mathcal{F}, \tag{3.43}$$

which completes the proof of the Lemma 3.3. □

Lemma 3.1 follows from Lemma 3.3 and from the dissipativity equality in Lemma 2.1. □.

4. Absorption of lower-order terms. Our next step is to eliminate lower-order terms from the inequality in Lemma 3.1. This is done by applying an appropriate compactness/uniqueness argument where a critical role is played by a recent uniqueness result due to Isakov which applies to domains with C^4 boundaries and with star-shaped unobserved portions of the boundaries (see [9, Remark 1.2], and [8, p. 750], where the geometry of the domains of unique continuation is explicitly described).

LEMMA 4.1. *Let u,w be a solution to* (1.1). *Then, there exist $T>0$ large enough so that the following parts hold.*

Part I. *Under Assumption* 2 (*in Theorem* 1.2) *we have*

$$lot(u,w)\equiv\int_0^T[|u(t)|^2_{1-\epsilon,\Omega}+|w(t)|^2_{2-\epsilon,\Omega}]dt\le C_T(E(0))\left[\int_{\Sigma_1}[|u_t|^2+|\nabla w_t|^2+|g(u_t)|^2\right.$$
$$\left.+|h_1(D_nw_t)|^2+|h_2(D_\tau w_t)|^2\right]d_{\Sigma_1}+C\int_0^T[(b_1u_t,u_t)_\Omega+(b_2w_t,w_t)_\Omega]dt. \tag{4.1}$$

Part II. *If the above assumption fails, then the constant C_T in* (4.1) *depends on*

$$E_\alpha(0)\equiv E(0)+|u_0|^2_{1+\alpha,\Omega}+|w_0|^2_{2+\alpha,\Omega},$$

where α can be taken to be arbitrarily small.

Proof. We argue by contradiction. Let u_n,w_n be a pair of solutions to (1.1) such that

$$\frac{lot(u_n,w_n)}{P(u_n,w_n)}\to\infty \text{ when } n\to\infty, \tag{4.2}$$

where

$$P(u,w) \equiv \int_{\Sigma_1} [|u_t|^2 + |\nabla w_t|^2 + |g(u_t)|^2 + |h_1(D_n w_t)|^2$$
$$+|h_2(D_\tau w_t)|^2]d_{\Sigma_1} + \int_0^T [(b_1 u_t, u_t)_\Omega + (b_2 w_t, w_t)_\Omega]dt. \tag{4.3}$$

From the boundedness of the initial energy ($E(0) \leq M$) we conclude that the sequence u_n, w_n satisfies

$$E_n(t) \leq M,$$

where E_n denotes $E(u_n, w_n)$.

Hence, on a subsequence,

$$\begin{aligned} u_n &\to u \text{ in } L_\infty(0,T;[H^1(\Omega)]^2) \text{ weakly}^\star, \\ u_{n,t} &\to u_t \text{ in } L_\infty(0,T;[L_2(\Omega)]^2) \text{ weakly}^\star, \\ w_n &\to w \text{ in } L_\infty(0,T;H^2(\Omega)) \text{ weakly}^\star, \\ w_{n,t} &\to w_t \text{ in } L_\infty(0,T;H^1(\Omega)) \text{ weakly}^\star. \end{aligned} \tag{4.4}$$

Thus, by the compactness of the lower-order terms (with respect to the topology induced by the energy) we conclude that

$$lot(u_n, w_n) \to lot(u, w). \tag{4.5}$$

We shall consider two separate cases.

Case 1. We have that

$$lot(u,w) \neq 0.$$

Then, $P(u_n, w_n) \to 0$ and

$$\begin{aligned} u_{n,t} &\to 0 \text{ in } [L_2(\Sigma)]^2, \\ \nabla w_{n,t} &\to 0 \text{ in } L_2(\Sigma), \\ g(u_{n,t}) &\to 0 \text{ in } [L_2(\Sigma)]^2, \\ h_1(D_n w_{n,t}) &\to 0 \text{ in } L_2(\Sigma), \\ h_2(D_\tau w_{n,t}) &\to 0 \text{ in } L_2(\Sigma), \\ b_1 u_{n,t} &\to 0 \text{ in } [L_2(\Omega)]^2, \\ b_2 w_{n,t} &\to 0 \text{ in } L_2(\Omega). \end{aligned} \tag{4.6}$$

Passing with the limit as $n \to \infty$ on the original equation (this is straightforward due to weak continuity of the nonlinear terms) we deduce that the limit functions u, w satisfy the original equations

$$\begin{aligned} u_{tt} + b_1 u_t - \operatorname{div}[\mathcal{C}[\boldsymbol{\epsilon}(u) + f(\nabla w)]] &= 0 \text{ in } \ Q, \\ [I - \gamma\Delta]w_{tt} + b_2 w_t + D\Delta^2 w - \operatorname{div}[\mathcal{C}[\boldsymbol{\epsilon}(u) + f(\nabla w)]\nabla w] &= 0 \ \text{ in } \ Q, \end{aligned} \tag{4.7}$$

$$\begin{aligned} \mathcal{C}[\boldsymbol{\epsilon}(u) + f(\nabla w)]\nu &= 0 \text{ in } \ \Sigma_1, \\ D[\Delta w + (1-\mu)B_1 w] &= 0 \text{ in } \ \Sigma_1, \\ D[D_n \Delta w + (1-\mu)B_2 w] &= 0 \text{ in } \ \Sigma_1, \\ u = w = \nabla w &= 0 \text{ in } \ \Sigma_0 \end{aligned} \tag{4.8}$$

with the overdetermined boundary conditions

$$u_t = 0, \nabla w_t = 0 \text{ on } \Sigma, \tag{4.9}$$

and

$$b_1 u_t = 0, b_2 w_t = 0 \text{ in } Q.$$

Denoting

$$\tilde{u} \equiv u_t, \quad \tilde{w} \equiv w_t,$$

we obtain the following system statisfied by the new variables

$$\begin{aligned} \tilde{u}_{tt} + b_1 \tilde{u}_t - \text{div}\mathcal{C}[\boldsymbol{\epsilon}(\tilde{u})] = L_1(\tilde{w}, w) \text{ in } \ Q, \\ [I - \gamma\Delta]\tilde{w}_{tt} + b_2\tilde{w}_t + D\Delta^2\tilde{w} = L_2(\tilde{u}, u, \tilde{w}, w) = 0 \ \text{ in } \ Q \end{aligned} \tag{4.10}$$

with the overdetermined boundary conditions on the boundary Γ

$$\tilde{u} = \nabla\tilde{w} = 0 \text{ on } \Sigma, \ \ \boldsymbol{\epsilon}(\tilde{u})\nu = 0 \text{ on } \Sigma_1, \ \ \Delta\tilde{w} = 0, \ D_n\Delta\tilde{w} = 0 \text{ on } \Sigma_1,$$

where we have used the notation

$$L_1(w, \tilde{w}) \equiv 1/2 \ \text{div}\mathcal{C}[\nabla\tilde{w} \times \nabla w] + 1/2 \ \text{div}\mathcal{C}[\nabla w \times \nabla\tilde{w}], \tag{4.11}$$

$$\begin{aligned} L_2(\tilde{u}, u, \tilde{w}, w) \equiv \text{div}[\mathcal{C}(\boldsymbol{\epsilon}(\tilde{u}))\nabla w + \mathcal{C}(\epsilon(u))\nabla\tilde{w} \\ + 1/2\mathcal{C}(\nabla\tilde{w} \times \nabla w)\nabla w + 1/2\mathcal{C}(\nabla w \times \nabla\tilde{w})\nabla w + \mathcal{C}f(\nabla w)\nabla\tilde{w}]. \end{aligned} \tag{4.12}$$

Our goal is to show that

$$u \equiv 0, \quad w \equiv 0 \text{ in } Q. \tag{4.13}$$

We shall treat separately cases considered in Parts I and II of Lemma 4.1. Let us notice first that in the case when the geometric conditions on the controlled portion of the boundary Γ_1 are satisfied (i.e., Ω is star-shaped), the conclusion in (4.13) follows at once from the uniqueness of solutions to the original equations (4.7), (4.8), (4.9) (see Remark 3.3). Thus, under the assumptions of Part I it suffices to consider the cases when either b_1 or b_2 are injective. Let us assume first that b_1 is injective. Then, we have $u_t \equiv 0$ and consequently

$$\begin{aligned} \text{div}\mathcal{C}N(u, w) = 0 \text{ in } \Omega, \\ N(u, w)\nu = 0 \text{ on } \Gamma_1, \ \ u = w = \nabla w = 0 \text{ on } \Gamma_0, \end{aligned} \tag{4.14}$$

which, in turn, implies

$$\text{div}\frac{d}{dt}\mathcal{C}(f(\nabla w)) = 0 \text{ in } Q. \tag{4.15}$$

Since a priori $f(\nabla w)\epsilon H^{1-\epsilon}(\Omega), \text{div} f(\nabla w) \in C(0, T; H^{-\epsilon}(\Omega))$ we obtain $\text{div}\mathcal{C}\epsilon(u) \in C(0, T; H^{-\epsilon}(\Omega))$ and by (4.14) $\epsilon(u) \cdot \nu|_{\Gamma_1} \in C(0, T; H^{1/2-\epsilon}(\Gamma_1))$. From elliptic regularity, we conclude the improved regularity for the variable u, i.e.,

$$u \in C(0, T; H^{2-\epsilon}(\Omega)). \tag{4.16}$$

Going back to the equation satisfied by w, denoting by $H(w)$ the matrix of the second derivatives of w (i.e., Hessian), and taking into account (4.14) we infer

$$\begin{aligned} [I - \gamma\Delta]w_{tt} + b_2 w_t + D\Delta^2 w = \mathcal{C}N(u, w) \cdot H(w) \text{ in } Q_T, \\ D[\Delta w + (1 - \mu)B_1 w] = 0 \text{ in } \ \Sigma_1, \\ D[D_n\Delta w + (1 - \mu)B_2 w] = 0 \text{ in } \ \Sigma_1, \end{aligned} \tag{4.17}$$

and the zero clamped boundary conditions on Γ_0. Differentiating (in the sense of distributions) the above equation in time yields

$$\begin{aligned}[I-\gamma\Delta]\tilde{w}_{t,t}+b_2\tilde{w}_t+D\Delta^2\tilde{w}=\mathcal{C}N(u,w)H(\tilde{w})+\mathcal{C}[1/2(\nabla\tilde{w}\times\nabla w)\\ +1/2(\nabla w\times\nabla\tilde{w})]H(w) \text{ in } Q,\\ D[\Delta\tilde{w}+(1-\mu)B_1\tilde{w}]=0 \text{ in } \Sigma_1,\\ D[D_n\Delta\tilde{w}+(1-\mu)B_2\tilde{w}]=0 \text{ in } \Sigma_1\end{aligned} \tag{4.18}$$

with the overdetermined boundary conditions on Γ_1

$$w=\nabla w=0 \text{ on } \Sigma.$$

Equation (4.18) is a linear equation in the variable $\tilde{w}$ with the coefficients depending on u, w. Therefore, provided that these coefficients are smooth enough and the solution itself is regular enough (and this is established in section 6, Theorem 6.2), we are in a position to apply the uniqueness result due to Isakov [9, Theorem 1.2], valid for the Kirchhoff plate, which gives

$$\tilde{w}=w_t=0 \text{ in } Q.$$

Thus, we have $u_t=0, w_t=0$ in Q. With this information we go back to the original equation (4.7), which now reads

$$\begin{aligned}\operatorname{div}\mathcal{C}N(u,w)=0 \text{ in } \Omega,\\ \mathcal{C}N(u,w)\nu=0 \text{ on } \Gamma_1, \quad u=0 \text{ on } \Gamma_0,\\ D\Delta^2 w-\operatorname{div}\mathcal{C}[N(u,w)\nabla w]=0 \text{ in } \Omega,\\ D[\Delta\tilde{w}+(1-\mu)B_1\tilde{w}]=0 \text{ on } \Gamma_1,\\ D[D_n\Delta w+(1-\mu)B_2 w]=0 \text{ on } \Gamma_1\end{aligned} \tag{4.19}$$

and the zero clamped boundary conditions on Γ_0. Multiplying the first equation by u, multiplying the second by $(1/2)w_t$, integrating over Ω, and adding the results yields

$$N(u,w)\equiv 0, \quad a(w,w)\equiv 0;$$

hence $w\equiv 0$ and $\epsilon(u)\equiv 0$. This combined with $\epsilon(u)\nu=0$ on Γ_1 and $u=0$ on Γ_0 gives $u\equiv 0, w\equiv 0$ as desired. Thus we have proved the assertion (4.13) when b_1 is injective.

We shall examine next the case when b_2 is injective. In this case we obtain that $w_t=\tilde{w}\equiv 0$ in Q. Therefore, $L_1(w,\tilde{w})\equiv 0$ and the variable $\tilde{u}$ satisfies

$$\tilde{u}_{tt}+b_1\tilde{u}_t-\operatorname{div}[\mathcal{C}\boldsymbol{\epsilon}(\tilde{u})]=0 \text{ in } Q, \tag{4.20}$$

$$\tilde{u}=0 \text{ on } \Sigma, \quad \boldsymbol{\epsilon}(\tilde{u})\nu=0 \text{ on } \Sigma_1. \tag{4.21}$$

The Holmgren type of uniqueness result for the elastic system (see, for instance, more precise results given in Theorem 1.2 and [8]) with the overdetermined boundary values gives that $\tilde{u}=0$ in Q_T. This, in turn, implies that both u_t and w_t are identically zero. As a result, we obtain the same static problem as in (4.19), which then yields the conclusion $u\equiv 0$ and $w\equiv 0$ as desired. Thus, the assertion (4.13) has been proved under the assumption of Part I of Lemma 4.1.

For Part II, the situation is more complicated and the assumption in Part II of Lemma 4.1 is also required. This is due to the fact that, in this case, our proof relies on an application of a new uniqueness result due to Isakov [9] in the context of a fully

dynamic von Karman system. This, in turn, requires the additional smoothness of the solutions and the coefficients corresponding to the system (4.10). The required regularity follows from Theorem 6.5 in section 6, which is proved under the additional assumption: $u_0 \in H^{1+\alpha}(\Omega), w_0 \in H^{2+\alpha}(\Omega)$ for $\alpha > 0$. Therefore, Theorem 1.3 in [9] applies to the linearized equation (4.10) and gives

$$\tilde{u} = \tilde{w} = 0.$$

Hence, u, w satisfy the static equation in (4.19) and we have $u = w = 0$. Thus we have established that under the assumptions of Parts I or II of Lemma 4.1 we always have $u = w = 0$. But this is a contradiction with the assumption made for Case 1. We shall next proceed to Case 2.

Case 2. $lot(u, w) = 0$. In this case, we do not need to distinguish between the Part I and Part II.

We define new variables

$$\hat{u}_n \equiv \frac{u_n}{c_n}, \quad \hat{w}_n \equiv \frac{w_n}{c_n},$$

where

$$c_n{}^2 \equiv lot(u_n, w_n) \to 0.$$

Thus we have

$$lot(\hat{u}_n, \hat{w}_n) = 1, \quad (1/c_n^2)P(u_n, w_n) \to 0, \tag{4.22}$$

which in turn implies

$$\begin{aligned} \hat{u}_{n,t} &\to 0 \text{ in } L_2(\Sigma), \\ \nabla \hat{w}_{n,t} &\to 0 \text{ in } L_2(\Sigma), \\ (1/c_n)g(u_{n,t}) &\to 0 \text{ in } L_2(\Sigma), \\ (1/c_n)h_1(D_n \hat{w}_{n,t}) &\to 0 \text{ in } L_2(\Sigma), \\ (1/c_n)h_2(D_\tau \hat{w}_{n,t}) &\to 0 \text{ in } L_2(\Sigma). \end{aligned} \tag{4.23}$$

Also, it is straightforward to verify that the new variables $\hat{u_n}, \hat{w_n}$ satisfy the system

$$\begin{aligned} &\hat{u}_{n,tt} + b_1 \hat{u}_t - \operatorname{div}[\mathcal{C}[\boldsymbol{\epsilon}(\hat{u}_n) + (1/c_n)f(\nabla w_n)]] = 0 \text{ in } \ Q, \\ &[I - \gamma\Delta]\hat{w}_{n,tt} + b_2 \hat{w}_t + D\Delta^2 \hat{w}_n - \operatorname{div}[\mathcal{C}[\boldsymbol{\epsilon}(\hat{u}_n) + (1/c_n)f(\nabla w_n)]\nabla w_n] = 0 \text{ in } \ Q \end{aligned} \tag{4.24}$$

with the Dirichlet boundary conditions on the uncontrolled part of the boundary Γ_0

$$\hat{u}_n = \hat{w}_n = \nabla \hat{w}_n = 0 \ \text{ on } \Sigma_0$$

and the dissipative boundary conditions on the controlled part of the boundary Γ_1

$$\begin{aligned} \mathcal{C}[\boldsymbol{\epsilon}(\hat{u}_n) + (1/c_n)f(\nabla w_n)]\nu &= -(1/c_n)g(u_{n,t}) \text{ in } \ \Sigma_1, \\ D[\Delta \hat{w}_n + (1-\mu)B_1 \hat{w}_n] &= -(1/c_n)h_1(D_n w_{n,t}) \text{ in } \ \Sigma_1, \\ D[D_n \Delta \hat{w}_n + (1-\mu)B_2 \hat{w}_n] - \gamma D_n \hat{w}_{n,tt} - [\mathcal{C}[\boldsymbol{\epsilon}(\hat{u}_n) + (1/c_n)f(\nabla w_n)]\nu \cdot \nabla w_n] & \\ = -(1/c_n)D_\tau h_2(D_\tau w_{n,t}). & \end{aligned} \tag{4.25}$$

Denoting

$$E_n \equiv E(u_n, w_n)$$

and recalling the stabilizability estimate in Lemma 3.1 together with (4.22) gives

(4.26)
$$(1/c_n^2)\left[E_n(0)+E_n(T)+\int_0^T E_n(t)dt\right]\le C(E_n(0))[(1/c_n^2)P(u_n,w_n)+lot(\hat{u}_n,\hat{w}_n)]\le M.$$

Elementary calculations show that (4.26) implies

$$\begin{aligned}|\hat{u}_n(t)|_{1,\Omega}&\le C,\\ |\hat{u}_{n,t}(t)|_{0,\Omega}&\le C,\\ |\hat{w}_n(t)|_{2,\Omega}&\le C,\\ |\hat{w}_{n,t}(t)|_{1,\Omega}&\le C.\end{aligned}\tag{4.27}$$

Hence, on a subsequence,

$$\hat{u}_n\to\hat{u}\text{ in }L_\infty(0,T;H^1(\Omega))\text{ weakly}\star,\tag{4.28}$$

$$\hat{u}_{n,t}\to\hat{u}_t\text{ in }L_\infty(0,T;L_2(\Omega))\text{ weakly}\star,\tag{4.29}$$

$$\hat{w}_n\to\hat{w}\text{ in }L_\infty(0,T;H^2(\Omega))\text{ weakly}\star,\tag{4.30}$$

$$\hat{w}_{n,t}\to\hat{w}_t\text{ in }L_\infty(0,T;H^1(\Omega))\text{ weakly}\star.\tag{4.31}$$

By the compactness of the lower-order term *lot* and (4.22) we infer that

$$lot(\hat{u}_n,\hat{w}_n)\to lot(\hat{u},\hat{w})=1.\tag{4.32}$$

We also notice that due to the uniform boundedness in (4.27)

$$w_n=\hat{w}_nc_n\to 0\text{ in }L_\infty(0,T;H^2(\Omega)).\tag{4.33}$$

Therefore, passing on the limit in the equation for $\hat{u}_n,\hat{w}_n$ (via routine arguments), using (4.31), (4.33), (4.23), and

$$|(1/c_n)f(\nabla w_n)|_{L_2(\Omega)}\le C|\nabla\hat{w}_n|_{L_4(\Omega)}|\nabla w_n|_{L_4(\Omega)}\to 0\tag{4.34}$$

leads to a decoupled system

$$\begin{aligned}\hat{u}_{tt}+b_1\hat{u}_t-\operatorname{div}[\mathcal{C}[\boldsymbol{\epsilon}(\hat{u})]&=0\text{ in }\ Q,\\ [I-\gamma\Delta]\hat{w}_{tt}+b_2\hat{w}_t+D\Delta^2\hat{w}&=0\ \text{ in }\ Q\end{aligned}\tag{4.35}$$

with Dirichlet/clamped boundary conditions on the uncontrolled part of the boundary Γ_0

$$\hat{u}=\hat{w}=\nabla\hat{w}=0\ \text{ on }\Sigma_0,$$

homogeneous boundary conditions on the controlled part of the boundary Γ_1

$$\begin{aligned}\mathcal{C}[\boldsymbol{\epsilon}(\hat{u})]\nu&=0\text{ in }\Sigma_1,\\ D[\Delta\hat{w}+(1-\mu)B_1\hat{w}]&=0\text{ in }\ \Sigma_1,\\ D[D_n\Delta\hat{w}+(1-\mu)B_2\hat{w}]&=0\ \text{ in }\ \Sigma_1,\end{aligned}\tag{4.36}$$

and the overdetermined boundary conditions

$$\hat{u}_t=0,\nabla\hat{w}_t=0\text{ on }\Sigma.$$

Denoting

$$\tilde{u} \equiv \hat{u}_t, \quad \tilde{w} \equiv \hat{w}_t,$$

we obtain that $\tilde{u}, \tilde{w}$ satisfy the same equation (4.35) with the boundary conditions

$$\tilde{u} = \tilde{w} = \nabla\tilde{w} = 0 \text{ on } \Sigma, \tag{4.37}$$

$$D_n\tilde{u} = \Delta\tilde{w} = D_n\Delta\tilde{w} = 0 \text{ on } \Sigma_1. \tag{4.38}$$

A standard Holmgren-type uniqueness result valid for these linear equations (see also [9]) yields

$$\tilde{u} \equiv 0, \quad \tilde{w} \equiv 0,$$

and going back to the static equation (4.19) we obtain

$$\hat{u} = 0, \quad \hat{w} = 0.$$

This is a contradiction of (4.32). The proof is thus completed. □

5. Completion of the proof of Theorem 1.2. By combining the results of Lemmas 3.1 and 4.1 we obtain Lemma 5.1.

LEMMA 5.1. *Let u, w be a regular solution to the original system. Then there exists a constant $T_0 > 0$ such that for any $T > T_0$,*

$$E(0) + E(T) + \int_0^T E(t)dt \le C_T(E(0)) \int_{\Sigma_1} [|u_t|^2 + |\nabla w_t|^2 + |g(u_t)|^2$$

$$+|h_1(D_n w_t)|^2 + |h_2(D_\tau w_t)|^2]d\Sigma_1 + \int_0^T [(b_1 u_t, u_t)_\Omega + (b_2 w_t, w_t)_\Omega]dt. \tag{5.1}$$

In what follows we shall denote

$$\Sigma_A \equiv \{(t,x) \in \Sigma_1 : |u_t| \le 1\}$$

and

$$\Sigma_B \equiv \Sigma_1 - \Sigma_A.$$

Recalling the definition of the function $\mathcal{G}$ we obtain

$$\int_{\Sigma_1} [|u_t|^2 + |g(u_t)|^2]d\Sigma_1 \le \int_{\Sigma_A} \mathcal{G}(u_t, g(u_t))d\Sigma_A + C\int_{\Sigma_B} g(u_t)u_t d\Sigma_B$$

$$\le \int_{\Sigma_1} [\mathcal{G}(u_t, g(u_t)) + Cu_t g(u_t)]d\Sigma_1. \tag{5.2}$$

A similar argument applies to the remaining feedback terms:

$$\int_{\Sigma_1} [|\nabla w_t|^2 + |h_1(D_n w_t)|^2 + |h_2(D_\tau w_t)|^2]d\Sigma_1 \le \int_{\Sigma_1} [C[D_n w_t h_1(D_n w_t)$$

$$+D_\tau w_t, h_2(D_\tau w_t)] + \mathcal{H}1(D_n w_t, h_1(D_n w_t)) + \mathcal{H}2(D_\tau w_t, h_2(D_\tau w_t))]d\Sigma_1. \tag{5.3}$$

Using Jensen's inequality we infer

$$\int_{\Sigma_1} [|u_t|^2 + |\nabla w_t|^2 + |g(u_t)|^2 + |h_1(D_n w_t)|^2 + |h_2(D_\tau w_t)|^2]d\Sigma_1$$

$$\le [CI + \mathcal{H}] \int_{\Sigma_1} [u_t g(u_t) + h_1(D_n w_t)D_n w_t + h_2(D_\tau w_t)D_\tau w_t]d\Sigma_1. \tag{5.4}$$

Denoting by

$$\mathcal{F} \equiv \int_{\Sigma_1} [u_t g(u_t) + h_1(D_n w_t) D_n w_t + h_2(D_\tau w_t) D_\tau w_t] d\Sigma_1 + \int_0^T [(b_1 u_t, u_t)_\Omega + (b_2 w_t, w_t)_\Omega] dt \tag{5.5}$$

we have that the inequality in (5.1) reads

$$E(0) + E(T) + \int_0^T E(t)dt \le C_T(E(0))[\mathcal{F} + \mathcal{H}(\mathcal{F})]. \tag{5.6}$$

Since the function $\mathcal{H} + \mathcal{I}$ is monotone, we can write

$$[I + \mathcal{H}]^{-1} E(T)/C_T(E(0)) \le \mathcal{F} = E(0) - E(T), \tag{5.7}$$

which in turn gives

$$p(E(T)) + E(T) \le E(0), \tag{5.8}$$

where the monotone function p is defined in section 1. Thus we have proved Lemma 5.2.

LEMMA 5.2. *Let u, w be a solution to the original equation. Then there exists a constant $T > 0$ such that*

$$p(E(T)) + E(T) \le E(0), \tag{5.9}$$

where the monotone function p is defined in section 1.

The final conclusion of Theorem 1.2 follows now from (5.9) and Lemma 3 in [19]. □

6. Regularity of the problem which is overdetermined on the boundary. We recall that for the proof of Lemma 4.1 in section 4 we have used the unique continuation results due to Isakov, which, however, require an additional regularity of the solutions to (4.14), (4.18), and (4.10). The purpose of this section is to establish the needed regularity. In fact, we shall show that solutions to these overdetermined on the boundary problems display an arbitrary level of smoothness. We shall begin our analysis with a simpler case of a semidynamic problem consisting of the system of equations given by (4.14), (4.18). (See the proof of Part I in Lemma 4.1.) Thus we are led to consider the following problem:

$$[I - \gamma\Delta]\tilde{w}_{tt} + b_2\tilde{w}_t + D\Delta^2\tilde{w} = L(u, \tilde{w}, w) \text{ in } Q \tag{6.1}$$

with the homogeneous boundary conditions on the boundary Γ

$$\nabla\tilde{w} = 0 \text{ on } \Gamma, \quad \Delta\tilde{w} = 0, \ D_n\Delta\tilde{w} = 0 \text{ on } \Gamma_1, \tag{6.2}$$

where we have used the notation

$$L(u, \tilde{w}, w) \equiv \mathcal{C}N(u, w)H(\tilde{w}) + 1/2\mathcal{C}(\nabla\tilde{w} \times \nabla w)H(w) + 1/2\mathcal{C}(\nabla w \times \nabla\tilde{w})H(w). \tag{6.3}$$

We know a priori (see (4.16)) that for any $\epsilon > 0$,

$$\begin{aligned} &u \in C(0, T; H^{2-\epsilon}(\Omega)), \quad w \in C(0, T; H^2(\Omega)), \quad \tilde{w} \in C(0, T; H^1(\Omega)), \\ &\tilde{w}_t \in C(0, T; L_2(\Omega)). \end{aligned} \tag{6.4}$$

Our first result follows in Lemma 6.1.

LEMMA 6.1. *Let $\tilde{w}$ be the solution to* (6.1) *with regularity specified above and the overdetermined boundary conditions given in* (6.2). *Then*

$$\tilde{w} \in C(0,T;H^2(\Omega)) \cap C^1(0,T;H^1(\Omega)). \tag{6.5}$$

Proof. The idea is to use Carleman's estimates, which are derived from [27] (see also [26]) and applied to equation (6.1). However, in order to do this we need to consider smoother solution than the ones given a priori. Therefore, we take a sequence of the initial data

$$\tilde{w}_{n,0} \in H_0^2(\Omega), \quad \tilde{w}_{n,1} \in H_0^1(\Omega)$$

such that

$$w_{n,0} \to \tilde{w}(0) \ \text{ in } \ H^1(\Omega), w_{n,1} \to \tilde{w}_t(0) \ \text{ in } \ L_2(\Omega). \tag{6.6}$$

Accounting for the regularity of the right-hand side in equation (6.1), i.e., term L which satisfies the estimate

$$|L(u,\tilde{w},w)|_{[H^1(\Omega)]'} \le C|\tilde{w}|_{2,\Omega}[|u|_{2-\epsilon,\Omega} + |w|^2_{2,\Omega}],$$

we infer by standard perturbation argument in linear semigroup theory that the solutions corresponding to the regularized initial data and denoted by $\tilde{w}_n(t)$ satisfy

$$\tilde{w}_n \in H_0^2(\Omega), \quad \tilde{w}_{n,t} \in H_0^1(\Omega), \tag{6.7}$$

$$\tilde{w}_n \to \tilde{w}(t) \ \text{ in } \ H^1(\Omega), \ \ \tilde{w}_{n,t} \to \tilde{w}_t(t) \ \text{ in } \ L_2(\Omega). \tag{6.8}$$

Since the term $L(u,w,\tilde{w}_n) \in C(0,T;H^{-1}(\Omega))$ we are in a position to apply Carleman's estimates in [27] to equation (6.1) satisfied by $\tilde{w}_n$. To do this, we need to introduce some notation. Let ϕ denotes a pseudoconvex function with respect to the real characteristics of the differential operator associated with the elastic and Kirchhoff system. In particular one can take

$$\phi(x) = |x - x_0|^2 - c(t - T/2)^2,$$

where the constant c is suitably small (see [27]). Let ψ be a nonnegative nondecreasing smooth real function with the following properties:

$$\psi(0) = 0, \quad \psi(x) > 0 \text{ for } x > 0, \tag{6.9}$$

$$\psi^{'}\psi^k/\psi \text{ is bounded on } R^+ \text{ for } k = 1,2,3. \tag{6.10}$$

We define

$$z \equiv \psi(\phi).$$

Applying the estimate of Theorem 2 in [27] (with $p = -1$) and noting that the anisotropic norms in the normal direction are irrelevant due to the zero overdetermined boundary conditions gives

$$\int_0^T |ze^{\tau\phi}\tilde{w}_{n,t}|^2_{1,\Omega} + |ze^{\tau\phi}\tilde{w}_n|^2_{2,\Omega}dt$$

$$\le C/\tau \int_0^T |ze^{\tau\phi}L(\tilde{w}_n,u,w)|^2_{-1,\Omega}dt + M_\tau(u,w,\tilde{w}_n), \tag{6.11}$$

where the constant τ can be arbitrarily large. The term $M_\tau(u, w, \tilde{w})$ is bounded by the a priori regularity of u, w, and, moreover, depends linearly on the lower-order norms of $\tilde{w}_n$, i.e.,

$$M_\tau(u, w, \tilde{w}_n) \leq M_{\tau,u,w} \int_0^T |\tilde{w}_n|^2_{1,\Omega} + |\tilde{w}_{n,t}|^2_{0,\Omega} dt. \tag{6.12}$$

Estimating the right-hand side of (6.11) gives

$$\begin{aligned} &\int_0^T |ze^{\tau\phi}\tilde{w}_{n,t}|^2_{1,\Omega} + |ze^{\tau\phi}\tilde{w}_n|^2_{2,\Omega} dt \\ &\leq C/\tau \int_0^T |ze^{\tau\phi}\tilde{w}_n|^2_{2,\Omega} dt [|w|^2_{L_\infty(0,T;H^2(\Omega))} + |u|_{L_\infty(0,T;H^{2-\epsilon}(\Omega))}] \\ &\qquad + M_\tau(u, w, \tilde{w}_n). \end{aligned} \tag{6.13}$$

Taking the constant τ to be large enough and recalling the a priori regularity in (6.4) gives

$$\int_0^T [|ze^{\tau\phi}\tilde{w}_{n,t}|^2_{1,\Omega} + |ze^{\tau\phi}\tilde{w}_n|^2_{2,\Omega}] dt \leq M_\tau(u, w, \tilde{w}_n). \tag{6.14}$$

Since ψ can be selected in such a way that

$$ze^{\tau\phi} \geq 1 \text{ on } [t_0, t_1] \in [0, T],$$

the inequality in (6.14) gives

$$\int_{t_0}^{t_1} |\tilde{w}_{n,t}|^2_{1,\Omega} + |\tilde{w}_n|^2_{2,\Omega} dt \leq M_\tau(u, w, \tilde{w}_n). \tag{6.15}$$

Define

$$\tilde{E}_n(t) \equiv |\tilde{w}_{n,t}(t)|^2_{1,\Omega} + |\tilde{w}_n(t)|^2_{2,\Omega}.$$

Noting that

$$\begin{aligned} &\int_s^{t_1} |L(w, u, \tilde{w}_n)|^2_{-1,\Omega} dt \\ &\leq C \int_s^{t_1} [|\tilde{w}_n|^2_{2,\Omega}] dt [|w|^2_{L_\infty(0,T);H^2(\Omega)} + |u|^2_{L_\infty(0,T);H^{2-\epsilon}(\Omega)}] + M_\tau(u, w, \tilde{w}_n), \end{aligned} \tag{6.16}$$

applying standard energy inequality to equation (6.1), and accounting for (6.15), we obtain for $t_0 \leq s \leq t_1$,

$$\begin{aligned} \tilde{E}_n(t_1) &\leq C\tilde{E}_n(s) + C \int_s^{t_1} |L(w, u\tilde{w}_n)|^2_{-1,\Omega} \\ &\leq C\tilde{E}_n(s) + M_\tau(u, w, \tilde{w}_n) + C \int_s^{t_1} \tilde{E}_n(t) dt \leq C\tilde{E}_n(s) + CM_\tau(u, w, \tilde{w}_n). \end{aligned} \tag{6.17}$$

Integrating the above inequality with respect to s yields

$$(t_1 - t_0)\tilde{E}_n(t_1) \leq C \int_{t_0}^{t_1} \tilde{E}_n(s) ds + CTM_\tau(u, w, \tilde{w}). \tag{6.18}$$

From (6.15) and (6.18) we infer

$$\tilde{E}_n(t_1) \leq C_T M_\tau(u, w, \tilde{w}_n). \tag{6.19}$$

Since the problem is linear in the variable $\tilde{w}$ and the term $M_\tau(u, w, \tilde{w}_n)$ depends on the lower-norm estimates for $\tilde{w}_n$, see (6.12), we apply the same argument to the sequence $\tilde{w}_n - \tilde{w}_m$. From (6.8) and (6.12) it follows that

$$M_\tau(u, w, \tilde{w}_n - \tilde{w}_m) \to 0. \tag{6.20}$$

Passing through the limit on the right-hand side of (6.19) (this time written for the difference between the two solutions) and accounting for (6.20) we conclude that $\tilde{w}_n(t_1), \tilde{w}_{n,t}(t_1)$ is a Cauchy sequence in $H^2(\Omega) \times H^1(\Omega)$ and it converges to $\tilde{w}(t_1), \tilde{w}_t(t_1)$. Standard semigroup argument yields the improved regularity for $\tilde{w}$:

$$\tilde{w} \in C(0, T; H^2(\Omega)), \tag{6.21}$$

$$\tilde{w}_t \in C(0, T; H^1(\Omega)). \qquad \square \tag{6.22}$$

In order to apply Isakov's uniqueness result in the context of equation (4.18), one needs to assert the regularity of solutions as well as the regularity of the coefficients of equation (4.18), which, in our case, amounts to the regularity of u, w. Since the variables $\tilde{u}, \tilde{w}$ are nothing but time derivatives of u, w, we have so far obtained

$$w_t \in C(0, T; H^2(\Omega)), \quad w_{tt} \in C(0, T; H^1(\Omega)). \tag{6.23}$$

In order to obtain a higher spatial regularity, we return to the original equation satisfied by u, w. Recalling the fact that the traces of velocities of u's and ∇w's are zero on Γ_1, we obtain

$$\begin{aligned}
\mathrm{div}\mathcal{C}[\boldsymbol{\epsilon}(u) + f(\nabla w)] &= 0 \text{ in } Q,\\
D\Delta^2 w = -[I - \gamma\Delta]w_{tt} - b_2 w_t + \mathrm{div}[\mathcal{C}[\boldsymbol{\epsilon}(u) + f(\nabla w)]\nabla w] &\text{ in } Q,\\
u = w = \nabla w &= 0 \text{ on } \Sigma_0,\\
\mathcal{C}[\boldsymbol{\epsilon}(u) + f(\nabla w)]\nu &= 0 \text{ in } \Sigma_1,\\
D[\Delta w + (1-\mu)B_1 w] &= 0 \text{ in } \Sigma_1,\\
D[D_n \Delta w + (1-\mu)B_2 w] &= 0 \text{ in } \Sigma_1.
\end{aligned} \tag{6.24}$$

By straightforward calculations and accounting for the result of Lemma 6.1 we infer that

$$\mathrm{div}[\mathcal{C} f(\nabla w)] \in C(0, T; H^{-\epsilon}(\Omega)), \tag{6.25}$$

$$[I - \gamma\Delta]w_{tt} + b_2 w_t + \mathrm{div}[\mathcal{C}[\boldsymbol{\epsilon}(u) + f(\nabla w)]\nabla w] \in C(0, T; H^{-1-\epsilon}(\Omega)). \tag{6.26}$$

From the regularity of the biharmonic problem, we further infer that

$$w \in C(0, T; H^{3-\epsilon}(\Omega)).$$

Hence,

$$f(\nabla w) \in C(0, T; H^1(\Omega)),$$

and going back to (6.24) we have that

$$\mathrm{div}[\mathcal{C} f(\nabla w)] \in C(0, T; L_2(\Omega)), \tag{6.27}$$

$$[I - \gamma\Delta]w_{tt} + \mathrm{div}[\mathcal{C}[\boldsymbol{\epsilon}(u) + f(\nabla w)]\nabla w] \in C(0, T; H^{-1}(\Omega)). \tag{6.28}$$

Since

$$f(\nabla w)|_{\Gamma_1} \in C(0,T;H^{1/2}(\Gamma_1)),$$

by reading off the boundary conditions in (6.24) we also obtain that

$$\mathcal{C}[\boldsymbol{\epsilon}(u)]|_{\Gamma_1} \in C(0,T;H^{1/2}(\Gamma_1)).$$

We are now in a position to apply elliptic estimates to the problem (6.24) and conclude that

$$u \in C(0,T;H^2(\Omega)), \quad w \in C(0,T;H^3(\Omega)). \tag{6.29}$$

Differentiating in time equation (6.1) and reiterating the same argument over and over we arrive at the conclusion that

$$u, w, \tilde{u}, \tilde{w} \in C^\infty(Q).$$

Thus we have obtained Theorem 6.2.

THEOREM 6.2. *Let u, w be a finite energy solution to the system* (4.7), (4.8) *with the overdetermined boundary data* (4.9) *such that $u_t \equiv 0$. Then*

$$u, w \in C^\infty(Q).$$

This gives us the improved regularity of the coefficients and of the solutions to equation (4.18), which, in turn, justifies the application, in section 4, of Isakov's Theorem 3 in [9]. From Theorem 3 in [9] we infer that the solution to (4.18) is identically zero, proving Corollary 6.3.

COROLLARY 6.3. *Let u, w be a finite energy solution to the system* (4.7), (4.8) *with the overdetermined boundary data* (4.9) *such that $u_t \equiv 0$. Then*

$$u, w \equiv 0 \text{ in } Q.$$

Now we consider the more difficult case, when Isakov's uniqueness property is used in the context of proving Lemma 4.1 under the assumptions of Part II. Application of Isakov's result requires that we show that the solutions and the coefficients of a fully dynamic system (4.10) are sufficiently smooth. Under the assumption of Part II, we have $E_\alpha \leq M$, which leads us to the limit solutions with the incrementally improved regularity (see [18]). This is to say that in equation (4.10) the coefficients u, w are in $C(0,T;H^{1+\alpha}(\Omega) \times H^{2+\alpha}(\Omega))$, where the parameter α may be arbitrarily small. Our goal is to show that this incremental improvement of regularity of the initial data implies that the solution to the problem (4.10) with the zero overdetermined boundary data is $C^\infty(Q)$. Thus we are led to consider the following problem. Let u, w be a finite energy solution corresponding to the original system with the additional regularity specified below, i.e., we assume the following.

ASSUMPTION 3. *We have that*

$$u \in C(0,T;H^{1+\alpha}(\Omega)), \quad w \in C(0,T;H^{2+\alpha}(\Omega)). \tag{6.30}$$

We consider the coupled system (see (4.10))

$$\begin{aligned} \tilde{u}_{tt} + b_1\tilde{u}_t - \operatorname{div}[\mathcal{C}[\boldsymbol{\epsilon}(\tilde{u})] &= L_1(\tilde{w}, w) \text{ in } Q, \\ [I - \gamma\Delta]\tilde{w}_{tt} + b_2\tilde{w}_t + D\Delta^2\tilde{w} &= L_2(\tilde{u}, u, \tilde{w}, w) \text{ in } Q, \end{aligned} \tag{6.31}$$

with the homogeneous boundary conditions on the boundary Γ

$$\tilde{u} = \nabla\tilde{w} = 0 \ \text{ on } \Gamma, \quad \Delta\tilde{w} = 0, \ D_n\Delta\tilde{w} = 0 \text{ on } \Gamma_1,$$

where we have used the notation

$$L_1(w,\tilde{w}) \equiv 1/2\mathrm{div}[\nabla\tilde{w}\times\nabla w] + 1/2\mathrm{div}[\nabla w\times\nabla\tilde{w}], \tag{6.32}$$

$$L_2(\tilde{u},u,\tilde{w},w) \equiv \mathrm{div}[\mathcal{C}(\boldsymbol{\epsilon}(\tilde{u}))\nabla w + \mathcal{C}(\epsilon(u))\nabla\tilde{w}$$

$$+1/2\mathcal{C}(\nabla\tilde{w}\times w)\nabla w + 1/2\mathcal{C}(\nabla w\times\nabla\tilde{w})\nabla w + \mathcal{C}f(\nabla w)\nabla\tilde{w}]. \tag{6.33}$$

Our result follows in Lemma 6.4.

LEMMA 6.4. *Under Assumption* 3, *stated above, any solution* $\tilde{u},\tilde{w}$ *to the problem* (6.31) *with the zero overdetermined boundary conditions*

$$\nabla\tilde{u} = 0, \ \nabla\tilde{w} = 0, \quad \Delta\tilde{w} = D_n\Delta\tilde{w} = 0 \text{ on } \Gamma_1$$

and such that a priori $\tilde{u},\tilde{w} \in C(0,T;L_2(\Omega)\times H^1(\Omega))$ *satisfies*

$$\tilde{u},\tilde{w} \in C(0,T;H^1(\Omega)\times H^2(\Omega)) \cap C^1(0,T;L_2(\Omega)\times H^1(\Omega)). \tag{6.34}$$

Proof. As before, the key idea is to use Carleman's estimates applied to the "regularized" problem, i.e., with "smooth" initial conditions such that (6.6) holds and, moreover,

$$\tilde{u}_{n,0} \in H_0^1(\Omega), \quad \tilde{u}_{n,1} \in L_2(\Omega), \tag{6.35}$$

$$\tilde{u}_{n,0} \to \tilde{u}(0) \text{ in } L_2(\Omega), \ \tilde{u}_{n,1} \to \tilde{u}_t(0) \text{ in } H^{-1}(\Omega). \tag{6.36}$$

In this case, the corresponding solutions $\tilde{u}_n,\tilde{w}_n$ satisfy (6.7), (6.8) and, moreover,

$$\tilde{u}_n(t) \in H_0^1(\Omega), \quad \tilde{u}_{n,t}(t) \in L_2(\Omega), \tag{6.37}$$

$$\tilde{u}_n(t) \to \tilde{u}(t) \text{ in } L_2(\Omega), \quad \tilde{u}_{n,t}(t) \to \tilde{u}_t(t) \text{ in } H^{-1}(\Omega). \tag{6.38}$$

With the improved regularity of regularized solutions and by virtue of Assumption 3 we easily verify that

$$L_1(w,\tilde{w}_n) \in C(0,T;L_2(\Omega)), \quad L_2(u,\tilde{u}_n,w,\tilde{w}_n) \in C(0,T;H^{-1}(\Omega)).$$

The above regularity allows us to apply Carleman's estimates given in [27]. Indeed, Theorem 2 in [27] applied separately to the Kirchhoff plate (6.31) and also to the system of elasticity (after decoupling into two wave equations and applying the estimates for the wave equation with $p=-1$ [25]) yields the following inequalities that are valid with any large constant τ:

$$\int_0^T |ze^{\tau\phi}\tilde{u}_{n,t}|^2_{0,\Omega} + |ze^{\tau\phi}\tilde{u}_n|^2_{1,\Omega}dt$$

$$\le C/\tau\int_0^T |ze^{\tau\phi}L_1(\tilde{w}_n,w)|^2_{0,\Omega}dt + M_\tau(u,w,\tilde{u}_n,\tilde{w}_n), \tag{6.39}$$

$$\int_0^T |ze^{\tau\phi}\tilde{w}_{n,t}|^2_{1,\Omega} + |ze^{\tau\phi}\tilde{w}_n|^2_{2,\Omega}dt$$

$$\le C/\tau\int_0^T |ze^{\tau\phi}L_2(\tilde{u}_n,\tilde{w}_n,u,w)|^2_{-1,\Omega}dt + M_\tau(u,w,\tilde{u}_n,\tilde{w}_n), \tag{6.40}$$

where the constant $M_\tau(u,w,\tilde{u},\tilde{w})$ is bounded by the a priori regularity of the solution $\tilde{u},\tilde{w}$ and u,w. More precisely,

$$M_\tau(u,w,\tilde{u}_n,\tilde{w}_n) \le M_{\tau,u,w}\int_0^T [|\tilde{u}_n|^2_{0,\Omega}+|\tilde{w}_n|^2_{1,\Omega}+|\tilde{w}_{n,t}|^2_{0,\Omega}]dt. \tag{6.41}$$

Estimating the terms on the right-hand sides of these two inequalities leads to

$$\begin{aligned}
&\int_0^T |ze^{\tau\phi}\tilde{u}_{n,t}|^2_{0,\Omega}+|ze^{\tau\phi}\tilde{u}_n|^2_{1,\Omega}dt \\
&\le C/\tau\int_0^T |ze^{\tau\phi}\tilde{w}_n|^2_{2,\Omega}dt|w|^2_{L_\infty(0,T);H^{2+\epsilon}(\Omega)}+M_\tau(u,w,\tilde{u}_n,\tilde{w}_n), \qquad (6.42)\\
&\int_0^T |ze^{\tau\phi}\tilde{w}_{n,t}|^2_{1,\Omega}+|ze^{\tau\phi}\tilde{w}_n|^2_{2,\Omega}dt \le C/\tau\int_0^T [|ze^{\tau\phi}\boldsymbol{\epsilon}(\tilde{u}_n)\nabla w|^2_{0,\Omega}+|ze^{\tau\phi}\epsilon(u)\nabla\tilde{w}_n|^2_{0,\Omega}\\
&\quad +|ze^{\tau\phi}(\nabla\tilde{w}_n\times\nabla w)\nabla w|^2_{0,\Omega}+|ze^{\tau\phi}(\nabla w\times\nabla w)\nabla\tilde{w}_n|^2_{0,\Omega}]dt+M_\tau(u,w,\tilde{u}_n,\tilde{w}_n)\\
&\le C/\tau\int_0^T [|ze^{\tau\phi}\tilde{u}_n|^2_{1,\Omega}+|ze^{\tau\phi}\tilde{w}_n|^2_{2,\Omega}]dt|w|^2_{L_\infty(0,T);H^{2+\epsilon}(\Omega)}\\
&\quad +\int_0^T |ze^{\tau\phi}\tilde{w}_n|^2_{2,\Omega}dt|u|^2_{L_\infty(0,T);H^{1+\epsilon}(\Omega)}+M_\tau(u,w,\tilde{u}_n,\tilde{w}_n). \qquad (6.43)
\end{aligned}$$

Combining (6.42) and (6.43) gives

$$\begin{aligned}
&\int_0^T |ze^{\tau\phi}\tilde{u}_{n,t}|^2_{0,\Omega}+|ze^{\tau\phi}\tilde{u}_n|^2_{1,\Omega}dt+\int_0^T |ze^{\tau\phi}\tilde{w}_{n,t}|^2_{1,\Omega}+|ze^{\tau\phi}\tilde{w}_n|^2_{2,\Omega}dt\\
&\le C(u,w)/\tau\left[\int_0^T |ze^{\tau\phi}\tilde{w}_n|^2_{2,\Omega}+|ze^{\tau\phi}\tilde{u}_n|^2_{1,\Omega}\right]dt+M_\tau(u,w,\tilde{u}_n,\tilde{w}_n). \qquad (6.44)
\end{aligned}$$

Taking the constant τ large enough gives

$$\int_0^T [|ze^{\tau\phi}\tilde{u}_{n,t}|^2_{0,\Omega}+|ze^{\tau\phi}\tilde{u}_n|^2_{1,\Omega}dt+|ze^{\tau\phi}\tilde{w}_{n,t}|^2_{1,\Omega}+|ze^{\tau\phi}\tilde{w}_n|^2_{2,\Omega}]dt \le M_\tau(u,w,\tilde{u}_n,\tilde{w}_n). \tag{6.45}$$

Since ψ can be selected in such a way that

$$ze^{\tau\phi}\ge 1 \text{ on } [t_0,t_1]\in[0,T],$$

the inequality in (6.45) gives

$$\int_{t_0}^{t_1} |\tilde{u}_{n,t}|^2_{0,\Omega}+|\tilde{u}_n|^2_{1,\Omega}dt+|\tilde{w}_{n,t}|^2_{1,\Omega}+|\tilde{w}_n|^2_{2,\Omega}dt \le CM_\tau(u,w,\tilde{u}_n,\tilde{w}_n). \tag{6.46}$$

Next define

$$\tilde{E}_n(t)\equiv |\tilde{u}_{n,t}(t)|^2_{0,\Omega}+|\tilde{u}_n(t)|^2_{1,\Omega}dt+|\tilde{w}_{n,t}(t)|^2_{1,\Omega}+|\tilde{w}_n(t)|^2_{2,\Omega}.$$

Noting that

$$\begin{aligned}
&\int_s^{t_1} |L_1(w,\tilde{w}_n)|^2_{0,\Omega}+|L_2(\tilde{u}_n,\tilde{w}_n,u,w)|^2_{-1,\Omega}]dt \\
&\le C\int_s^{t_1}[|\tilde{u}_n|^2_{1,\Omega}+|\tilde{w}_n|^2_{2,\Omega}]dt[|w|^2_{L_\infty(0,T);H^{2+\epsilon}(\Omega)}+|u|^2_{L_\infty(0,T);H^{1+\epsilon}(\Omega)}]+M_\tau(u,w,\tilde{u}_n,\tilde{w}_n), \qquad (6.47)
\end{aligned}$$

applying the standard energy inequality to equation (6.31), and accounting for (6.47) we obtain that

$$\tilde{E}_n(t_1) \leq C\tilde{E}_n(s) + C\int_s^{t_1} |L_1(w,\tilde{w}_n)|^2_{0,\Omega} + |L_2(\tilde{u}_n,\tilde{w}_n,u,w)|^2_{-1,\Omega}]dt$$

$$\leq C\tilde{E}_n(s) + CM_\tau(u,w,\tilde{u}_n,\tilde{w}_n) + C_{u,w}\int_s^{t_1} \tilde{E}_n(t)dt. \tag{6.48}$$

Integrating the above inequality with respect to s yields

$$(t_1-t_0)\tilde{E}_n(t_1) \leq C\int_{t_0}^{t_1} \tilde{E}_n(s)ds + CT\left[M_\tau(u,w,\tilde{u}_n,\tilde{w}_n) + \int_{t_0}^{t_1} \tilde{E}_n(t)dt\right]. \tag{6.49}$$

From (6.46) and (6.49) we infer that

$$\tilde{E}_n(t_1) \leq CM_\tau(u,w,\tilde{u}_n,\tilde{w}_n). \tag{6.50}$$

Applying the same argument to the difference of solutions $\tilde{u}_n - \tilde{u}_m,\ \tilde{w}_n - \tilde{w}_m$ and taking advantage of (6.38), (6.8), (6.41) we conclude that

$$\tilde{u}_n(t_1), \tilde{u}_{n,t}(t_1), \tilde{w}_n(t_1), \tilde{w}_{n,t}(t_1)$$

are Cauchy sequences in

$$H^1(\Omega) \times L_2(\Omega) \times H^2(\Omega) \times H^1(\Omega).$$

This leads, via standard semigroup argument, to the improved regularity for $\tilde{u}, \tilde{w}$:

$$\tilde{u} \in C(0,T;H^1(\Omega)), \quad \tilde{w} \in C(0,T;H^2(\Omega)), \tag{6.51}$$

$$\tilde{u}_t \in C(0,T;L_2(\Omega)), \quad \tilde{w}_t \in C(0,T;H^1(\Omega)). \qquad \square \tag{6.52}$$

In order to apply Isakov's uniqueness result, in the context of equation (4.10) we need to assert the regularity of solutions as well as the regularity of the coefficients of the equation, which, in our case, amounts to the regularity of u, w. Since the variables $\tilde{u}, \tilde{w}$ are the time derivatives of u, w, we have so far obtained

$$u_t \in C(0,T;H^1(\Omega)), \quad w_t \in C(0,T;H^2(\Omega)), \tag{6.53}$$

$$u_{tt} \in C(0,T;L_2(\Omega)), \quad w_{tt} \in C(0,T;H^1(\Omega)). \tag{6.54}$$

In order to obtain a higher spatial regularity, we return to the original equation satisfied by u, w. Recalling the fact that the traces of velocities of $u, \nabla w$ are zero on Γ_1, we obtain

$$\begin{aligned}
\operatorname{div}[\mathcal{C}[\boldsymbol{\epsilon}(u)]] &= u_{tt} + b_1u_t - \operatorname{div}[\mathcal{C}f(\nabla w)] \text{ in } Q,\\
D\Delta^2 w = -[I-\gamma\Delta]w_{tt} - b_2w_t + \operatorname{div}[\mathcal{C}[\boldsymbol{\epsilon}(u) + f(\nabla w)]\nabla w] &\text{ in } Q,\\
u = w = \nabla w &= 0 \text{ on } \Sigma_0,\\
\mathcal{C}[\boldsymbol{\epsilon}(u) + f(\nabla w)]\nu &= 0 \text{ in } \Sigma_1,\\
D[\Delta w + (1-\mu)B_1 w] &= 0 \text{ in } \Sigma_1,\\
D[D_n\Delta w + (1-\mu)B_2 w] &= 0 \text{ in } \Sigma_1.
\end{aligned} \tag{6.55}$$

By straightforward calculations and accounting for the result of Lemma 6.4 we infer that

$$u_{tt} + b_1u_t - \operatorname{div}[\mathcal{C}f(\nabla w)] \in C(0,T;H^{-\epsilon}(\Omega)), \tag{6.56}$$

$$[I-\gamma\Delta]w_{tt} + b_2w_t + \operatorname{div}[\mathcal{C}[\boldsymbol{\epsilon}(u) + f(\nabla w)]\nabla w] \in C(0,T;H^{-1-\epsilon}(\Omega)), \tag{6.57}$$

and from the regularity of the biharmonic problem we then infer that

$$w \in C(0,T;H^{3-\epsilon}(\Omega)).$$

Hence,

$$f(\nabla w) \in C(0,T;H^1(\Omega)),$$

and going back to (6.55) we have that

$$u_{tt} + b_1 u_t - \operatorname{div}[\mathcal{C} f(\nabla w)] \in C(0,T;L_2(\Omega)), \tag{6.58}$$

$$[I - \gamma\Delta]w_{tt} + b_2 w_t + \operatorname{div}[\mathcal{C}[\boldsymbol{\epsilon}(u) + f(\nabla w)]\nabla w] \in C(0,T;H^{-1}(\Omega)). \tag{6.59}$$

Since

$$f(\nabla w)|_{\Gamma_1} \in C(0,T;H^{1/2}(\Gamma_1)),$$

hence

$$\mathcal{C}[\boldsymbol{\epsilon}(u)]|_{\Gamma_1} \in C(0,T;H^{1/2}(\Gamma_1)).$$

We are now in a position to apply elliptic estimates to the problem (6.55), and we conclude that

$$u \in C(0,T;H^2(\Omega)), \quad w \in C(0,T;H^3(\Omega)). \tag{6.60}$$

Differentiating in time equation (6.31) and reiterating the same argument over and over we arrive at the conclusion that

$$u, w, \tilde{u}, \tilde{w} \in C^\infty(Q).$$

Thus we have proved Theorem 6.5.

THEOREM 6.5. *Let u, w be a finite energy solution of the system* (4.7), (4.8) *with the overdetermined boundary conditions* (4.9) *and such that*

$$|u(t)|_{1+\alpha,\Omega} + |w(t)|_{2+\alpha,\Omega} \leq \infty,$$

where $\alpha > 0$ can be arbitrary small. Then

$$u, w \in C^\infty(Q).$$

This gives us the improved regularity of the coefficients and solutions to equation (4.10) which is, in turn, required for the application of Theorem 3 in [9] in the context of the proof of Part II of Lemma 4.1. Theorem 3 in [9] then implies that the solution to (4.10) is identically zero, which proves Corollary 6.6.

COROLLARY 6.6. *Under the assumptions of Theorem* 6.5 *we have*

$$u, w \equiv 0 \text{ in } Q.$$

7. Appendix. Proof of Proposition 1.1.

7.1. Existence of regular solutions: Proof of Part 1 of Proposition 1.1. We shall use the nonlinear Galerkin method. Indeed, let h denote a parameter tending to zero and let $\mathcal{U}_h$ (resp., $\mathcal{W}_h$) be a finite-dimensional subspace of $H^2_{\Gamma_0}(\Omega)$ (resp., $H^3_{\Gamma_0}(\Omega)$), where the subindex Γ_0 indicates that the functions are subject to zero boundary conditions (see (1.1)) on Γ_0. We then denote

$$U_h \equiv \mathcal{U}_h \times \mathcal{U}_h, \quad V_h \equiv U_h \times W_h.$$

We consider the following semidiscrete approximation of the original problem (1.1), (1.2).

Given $(u_{h,0}, w_{h,0}, u_{h,1}, w_{h,1}) \in V_h \times V_h$, find $(u_h(t), w_h(t)) \in V_h$ such that $u_h(0) = u_{h,0}, w_h(0) = w_{h,0}, u_{h,t}(0) = u_{h,1}, w_{h,t}(0) = w_{h,1}$, and

$$(u_{h,tt}, \phi)_\Omega + (b_1 u_{h,t}, \phi)_\Omega + (\mathcal{C}N(u_h, w_h), \epsilon(\phi))_\Omega + \langle g(u_{h,t}), \phi\rangle_{\Gamma_1} = 0, \tag{7.1}$$
$$(w_{h,tt}, \psi)_\Omega + \gamma(\nabla w_{h,tt}, \nabla\psi)_\Omega + a(w_h, \psi) + (b_2 w_{h,t}, \psi)_\Omega$$

$$+(\mathcal{C}N(u_h, w_h), \nabla\psi \times \nabla w_h)_\Omega + \langle h_1(D_n w_{h,t}), D_n\psi\rangle_{\Gamma_1} + \langle h_2(D_\tau w_{h,t}), D_\tau\psi\rangle_{\Gamma_1} = 0 \tag{7.2}$$

for all $(\phi, \psi) \in V_h$.

Global existence and uniqueness of semidiscrete solutions follows from the fact that nonlinear terms are locally Lipschitz on V_h together with a standard a priori bound

$$E_h(t) \le E_h(0), \tag{7.3}$$

where

$$E_h(t) \equiv \int_\Omega |u_{h,t}|^2 + |w_{h,t}|^2 + \gamma|\nabla w_{h,t}|^2 d\Omega + a(w_h(t), w_h(t))$$
$$+ \int_\Omega [\mathcal{C}N(u_h(t), w_h(t)) \cdot N(u_h(t), w_h(t))]\, d\Omega. \tag{7.4}$$

Moreover, the solutions $(u_h(t), w_h(t))$ are $C^\infty([0, T_0), V_h)$ for some $T_0 > 0$. We shall show that Galerkin approximations defined by (7.2) are stable in higher norms. Indeed, differentiating (7.2) in time and denoting

$$\bar{u} \equiv u_{h,t}, \quad \bar{w} \equiv w_{h,t}, \quad N_t(u_h, w_h) \equiv \frac{d}{dt} N(u_h, w_h) = \epsilon(\bar{u}) + \frac{d}{dt} f(\nabla w_h)$$

we obtain the following equation satisfied for the new variables:

$$(\bar{u}_{tt}, \phi)_\Omega + (b_1\bar{u}_t, \phi)_\Omega + (\mathcal{C}N_t(u_h, w_h), \epsilon(\phi))_\Omega + \langle g'(u_{h,t})\bar{u}_t, \phi\rangle_{\Gamma_1} = 0, \tag{7.5}$$
$$(\bar{w}_{tt}, \psi)_\Omega + \gamma(\nabla\bar{w}_{tt}, \nabla\psi)_\Omega + a(\bar{w}_h, \psi) + (b_2\bar{w}_t, \psi)_\Omega + (\mathcal{C}N_t(u_h, w_h), \nabla\psi \times \nabla w_h)_\Omega$$
$$+(\mathcal{C}N(u_h, w_h), \nabla\psi \times \nabla\bar{w})_\Omega + \langle {h_1}'(D_n w_{h,t}) D_n\bar{w}_t, D_n\psi\rangle_{\Gamma_1}$$
$$+\langle {h_2}'(D_\tau w_{h,t}) D_\tau\bar{w}_t, D_\tau\psi\rangle_{\Gamma_1} = 0 \tag{7.6}$$

for all $(\phi, \psi) \in V_h$.

Taking $\phi = \bar{u}_t, \psi = \bar{w}_t$ yields

$$1/2\frac{d}{dt}[|\bar{u}_t|^2_\Omega + (\mathcal{C}N_t(u_h, w_h), N_t(u_h, w_h))_\Omega + |\bar{w}_t|^2_\Omega + \gamma|\nabla\bar{w}_t|^2_\Omega + a(\bar{w}, \bar{w})]$$
$$+(b_2\bar{w}_t, \bar{w}_t)_\Omega + (b_1\bar{u}_t, \bar{u}_t)_\Omega + \langle g'(u_{h,t})\bar{u}_t, \bar{u}_t\rangle_{\Gamma_1} + \langle {h_1}'(D_n w_{h,t}) D_n\bar{w}_t, D_n\bar{w}_t\rangle_{\Gamma_1}$$
$$+\langle {h_2}'(D_\tau w_{h,t}) D_\tau\bar{w}_t, D_\tau\bar{w}_t\rangle_{\Gamma_1}$$
$$= (\mathcal{C}N_t(u_h, w_h), \nabla\bar{w} \times \nabla\bar{w})_\Omega - (\mathcal{C}N(u_h, w_h), \nabla\bar{w}_t \times \nabla\bar{w})_\Omega = 0. \tag{7.7}$$

We denote

$$\bar{E}(t) \equiv |\bar{u}_t(t)|^2_\Omega + (\mathcal{C}N_t(u_h(t), w_h(t)), N_t(u_h(t), w_h(t)))_\Omega + |\bar{w}_t(t)|^2_\Omega + \gamma|\nabla\bar{w}_t(t)|^2_\Omega$$
$$+a(\bar{w}(t), \bar{w}(t)). \tag{7.8}$$

It is straightforward to verify that $\bar{E}(t)$ is bounded from above and below by the expression $C(|\bar{u}|_{1,\Omega}, |\bar{u}_t|_{0,\Omega}, |\bar{w}|_{2,\Omega}, |\bar{w}_t|_{1,\Omega})$. This fact will be used frequently without further mention.

Integrating (7.7) from 0 to t, integrating by parts the last term in (7.7), and recalling the monotonicity of boundary feedback yields the inequality

$$\bar{E}(t) \leq \bar{E}(0) + 6\int_0^t (\mathcal{C}N_t(u_h, w_h), f(\nabla \bar{w}))_\Omega dt - 2(\mathcal{C}N(u_h, w_h), f(\nabla \bar{w}))_\Omega|_0^t. \tag{7.9}$$

By Sobolev's embedding $L_4(\Omega) \subset H^{1/2}(\Omega)$ and the classical interpolation inequality $|w|^2_{1/2,\Omega} \leq C|w|_{1,\Omega}|w|_{0,\Omega}$, we obtain

$$|f(\nabla \bar{w})|_{0,\Omega} \leq C|\bar{w}|_{2,\Omega}|w_{h,t}|_{1,\Omega} \leq C[E_h(0)]^{1/2}|\bar{w}|_{2,\Omega}, \tag{7.10}$$

where in the last step we have used the a priori bound in (7.3). Combining (7.9) with (7.10) and using, again, the a priori bound in (7.3) gives

$$\begin{aligned} \bar{E}(t) &\leq \bar{E}(0) + CE_h(0)\int_0^t |N_t(u_h, w_h)|_{0,\Omega}|\bar{w}|_{2,\Omega}dt + \epsilon|\bar{w}(0)|^2_{2,\Omega} \\ &+\epsilon|\bar{w}(t)|^2_{2,\Omega} + C_\epsilon E_h(0)[|N(u_h(0), w_h(0))|^2_{0,\Omega} + |N(u_h(t), w_h(t))|^2_{0,\Omega}] \\ &\leq C\bar{E}(0) + CE_h(0)\int_0^t |N_t(u_h, w_h)|_{0,\Omega}|\bar{w}|_{2,\Omega}dt + \epsilon|\bar{w(t)}|^2_{2,\Omega} + C_\epsilon E_h(0)^3. \end{aligned} \tag{7.11}$$

Taking ϵ small enough, recalling the definition of $\bar{E}$, and applying Gronwall's inequality yields

$$\bar{E}(t) \leq C_T(E_h(0))[\bar{E}(0) + [E_h(0)]^3], \quad t \leq T, \tag{7.12}$$

where C_T does not depend on h.

In order to provide an effective estimate (independent on h) of the right side of inequality (7.12) we need to estimate $\bar{E}(0), E_h(0)$. To this end we make the following natural approximation/stability properties imposed on our approximating spaces. We assume that the following conditions hold.

(i) For any $u \in H^s_{\Gamma_0}(\Omega)$, there exists $\phi \in \mathcal{U}_h$ such that

$$|u - \phi|_{s,\Omega} \to 0 \text{ when } h \to 0, s \leq 2.$$

(ii) For any $w \in H^s_{\Gamma_0}(\Omega)$, there exist $\psi \in \mathcal{W}_h$ such that

$$|w - \psi|_{s,\Omega} \to 0 \text{ when } h \to 0, s \leq 3.$$

Moreover, there exist $p > 0$ such that the following "inverse approximation properties" hold

(iii)

$$|\phi|_{0,\Gamma} \leq Ch^{-p}|\phi|_{0,\Omega}, \quad |\nabla\psi|_{0,\Gamma} \leq Ch^{-p}|\psi|_{1,\Omega} \quad \text{for } \phi, \psi \in V_h,$$

(iv)

$$|D_1[u - \phi]|_{L_2(\Gamma)} \leq Ch^p|u|_{2,\Omega}, \quad |D_l[w - \psi]|_{L_2(\Gamma)} \leq Ch^p|w|_{l+1,\Omega}, \quad l = 1, 2,$$

where by D_i we have denoted the differential operator of order i.

We note that the approximation properties listed above are the standard ones for a variety of approximating subspaces including cubic splines defined on quasi-uniform

mesh. (This last property is guaranteed by the inverse approximation properties. In this case we have $p = 1/2$.)

Let us denote by C_0 a constant which depends on $|w_0|_{2,\Omega}, |w_1|_{1,\Omega}, |u_0|_{1,\Omega}, |u_1|_{0,\Omega}$ and by C_1 a constant depending on the higher-order norms of the initial data, i.e., $|w_0|_{3,\Omega}, |w_1|_{2,\Omega}, |u_0|_{2,\Omega}, |u_1|_{1,\Omega}$.

Now, we assume that $u_{h,0}, u_{h,1}, w_{h,0}, w_{h,1}$ are "good" approximations of the initial data; i.e., the approximation properties listed above are satisfied and, in particular,

$$\begin{aligned} &|u_0 - u_{h,0}|_{2,\Omega} \to 0 \ \text{ when } h \to 0, \ |u_1 - u_{h,1}|_{1,\Omega} \to 0 \ \text{ when } h \to 0, \\ &|w_0 - w_{h,0}|_{3,\Omega} \to 0 \ \text{ when } h \to 0, \ |w_1 - w_{h,1}|_{2,\Omega} \to 0 \ \text{ when } h \to 0. \end{aligned} \tag{7.13}$$

By stability of the estimates resulting from (7.13) and the regularity of the initial conditions we obtain

$$E_h(0) \le C_0.$$

Moreover, since $\bar{u}(0) = u_{h,1}, \bar{w}(0) = w_{h,1}$, we also have

$$|\bar{u}(0)|_{1,\Omega} \le |u_1|_{1,\Omega} \le C_1, |\bar{w}(0)|_{2,\Omega} \le |w_1|_{2,\Omega} \le C_1. \tag{7.14}$$

In order to estimate $|\bar{u}_t(0)|_{0,\Omega}$ we shall use the compatibility conditions. In fact from (7.2) we have

$$\begin{aligned} (\bar{u}_t(0), \phi)_{0,\Omega} \le |&(b_1 u_{h,1}, \phi)_\Omega + (\mathrm{div}\mathcal{C}N(u_{h,0}, w_{h,0}), \phi)_{0,\Omega} \\ &-\langle \mathcal{C}N(u_{h,0}, w_{h,0})\nu + g(u_{h,1}), \phi\rangle_{\Gamma_1}| \text{ for all } \phi \in U_h. \end{aligned} \tag{7.15}$$

Hence

$$\begin{aligned} (\bar{u}_t(0), \phi)_{0,\Omega} &\le C[|u_{h,0}|_{2,\Omega} + |w_{h,0}|_{3,\Omega} + |u_{h,1}|_{0,\Omega}]|\phi|_{0,\Omega} \\ &+C|N(u_{h,0}, w_{h,0})\nu + g(u_{h,1}) - N(u_0, w_0)\nu - g(u_1)|_{0,\Gamma_1}|\phi|_{0,\Gamma_1}, \end{aligned} \tag{7.16}$$

where we have used the compatibility relations (1.4) for the first equation. By Lipschitz continuity of the feedback g and the inverse approximation properties listed in (iii), (iv) we obtain

$$\begin{aligned} &|N(u_{h,0}, w_{h,0})\nu + g(u_{h,1}) - N(u_0, w_0)\nu - g(u_1)|_{0,\Gamma_1}|\phi|_{0,\Gamma_1} \\ &\le C[|u_{h,0} - u_0|_{1,\Gamma_1} + C_1|\nabla[w_{h,0} - w_0]|_{0,\Gamma_1} + |u_{h,1} - u_1|_{0,\Gamma}]h^{-p}||\phi|_{0,\Omega} \\ &\le C[|u_0|_{2,\Omega} + C_1|w_0|_{2,\Omega} + |u_1|_{1,\Omega}]h^p h^{-p}||\phi|_{0,\Omega} \le C[C_1 + C_1C_0]|\phi|_{0,\Omega}. \end{aligned} \tag{7.17}$$

Since ϕ is arbitrary in U_h, we obtain

$$|\bar{u}_t(0)|_{0,\Omega} \le C_1 + C_1C_0 \tag{7.18}$$

providing a desired an a priori bound for the initial datum $\bar{u}_t(0)$. In the same manner we obtain an a priori bound for the second variable, i.e.,

$$|\bar{w}_t(0)|_{1,\Omega} \le C_1 + C_0 + C_0C_1 = C(C_0, C_1). \tag{7.19}$$

Indeed,

$$\begin{aligned} (\bar{w}_t(0), \psi)_{1,\Omega} &= (\bar{w}_t(0), \psi)_\Omega + \gamma(\nabla\bar{w}_t(0), \nabla\psi)_\Omega \\ &= -a(w_h(0), \psi) - (b_2 w_{h,t}(0), \psi)_\Omega - (\mathcal{C}[N(u_h(0), w_h(0))\nabla w_h(0)], \nabla\psi)_\Omega \\ &\quad -\langle h_1(D_n\bar{w}(0)), D_n\psi\rangle_{\Gamma_1} - \langle h_2(D_\tau\bar{w}(0)), D_\tau\psi\rangle_{\Gamma_1}. \end{aligned} \tag{7.20}$$

Application of the first Green formula together with compatibility conditions (1.4) yield

$$(\bar{w}_t(0),\psi)_{1,\Omega} = a_1(w_h(0),\psi) - (b_2 w_{h,t}(0),\psi)_\Omega - (\mathcal{C}[N(u_h(0),w_h(0))\nabla w_h(0)],\nabla\psi)_\Omega \\ -\langle h_2(D_\tau\bar{w}(0)), D_\tau\psi\rangle_{\Gamma_1} + X = 0, \tag{7.21}$$

where we have denoted

$$X \equiv \langle D[\Delta w_h(0)+(1-\mu)B_1 w_h(0)-\Delta w_0-(1-\mu)B_1 w_0]-h_1(D_n w_{h,1})+h_1(D_n w_1), D_n\psi\rangle_{\Gamma_1}$$

and $a_1(w,\psi)\le C|w|_{3,\Omega}|\psi|_{1,\Omega}$. By using approximation properties listed in (iii), (iv), in the same manner as before, together with Lipschitz continuity of h_1 we obtain

$$|X| \le Ch^p|w_0|_{3,\Omega}h^{-p}|\psi|_{1,\Omega} + Ch^p|w_1|_{2,\Omega}h^{-p}|\psi|_{1,\Omega} \\ \le C_1|\psi|_{1,\Omega}. \tag{7.22}$$

A priori bounds in (7.14) together with the estimate (7.22) and the fact that h_2 is bounded from $H^{1/2}(\Gamma_1)$ into itself yield

$$(\bar{w}_t(0),\psi)_{1,\Omega} \le C_1|\psi|_{1,\Omega} + C[|w_h(0)|_{3,\Omega}|\psi|_{1,\Omega} + |\bar{w}(0)|_{0,\Omega}|\psi|_{0,\Omega} \\ +(|u_h(0)|_{2,\Omega}+|w_h(0)|_{2,\Omega})|w_h(0)|_{2,\Omega}|\psi|_{1,\Omega} + |h_2(D_\tau\bar{w}(0))|_{1/2,\Gamma_1}|D_\tau\psi|_{-1/2,\Gamma_1}] \\ \le C_1|\psi|_{1,\Omega} + C[|w_h(0)|_{3,\Omega}|\psi|_{1,\Omega} + |\bar{w}(0)|_{0,\Omega}|\psi|_{0,\Omega} \\ +(|u_h(0)|_{2,\Omega}+|w_h(0)|_{2,\Omega})|w_h(0)|_{2,\Omega}|\psi|_{1,\Omega} + |(D_\tau\bar{w}(0))|_{1/2,\Gamma_1}|D_\tau\psi|_{-1/2,\Gamma_1}] \\ \le C[|w_h(0)|_{3,\Omega}|\psi|_{1,\Omega} + |\bar{w}(0)|_{0,\Omega}|\psi|_{0,\Omega} + (|u_h(0)|_{2,\Omega}+|w_h(0)|_{2,\Omega})|w_h(0)|_{2,\Omega}|\psi|_{1,\Omega} \\ +|\bar{w}(0)|_{2,\Omega}|\psi|_{1/2,\Gamma_1} + C_1|\psi|_{1,\Omega} \le C[C_1+C_0+C_1C_0]|\psi|_{1,\Omega}. \tag{7.23}$$

This gives the desired result in (7.19).

Collecting (7.14), (7.18), (7.19) we conclude that $\bar{E}(0)\le C(C_0,C_1)$, and by (7.12)

$$\bar{E}(t)\le C(C_0,C_1).$$

Hence the a priori bounds

$$|u_{h,t}(t)|_{1,\Omega} + |u_h(t)|_{1,\Omega} + |u_{h,tt}(t)|_{0,\Omega} + |w_{h,t}(t)|_{2,\Omega} + |w_h(t)|_{2,\Omega} + |w_{h,tt}(t)|_{1,\Omega} \le C \tag{7.24}$$

hold uniformly in h for all $t<T$ where T is arbitrary.

We can now extract convergent subsequences (denoted by the same symbol) such that

$$u_h \to u \text{ weakly}^\star \text{ in } L_\infty(0,T;H^1(\Omega)), u_{h,t}\to u_t \text{ weakly}^\star \text{ in } L_\infty(0,T;H^1(\Omega)), \\ u_{h,tt}\to u_{tt} \text{ weakly}^\star \text{ in } L_\infty(0,T;L_2(\Omega)), \\ w_h \to w \text{ weakly}^\star \text{ in } L_\infty(0,T;H^2(\Omega)), w_{h,t}\to w_t \text{ weakly}^\star \text{ in } L_\infty(0,T;H^2(\Omega)), \\ w_{h,tt}\to w_{tt} \text{ weakly}^\star \text{ in } L_\infty(0,T;H^1(\Omega)). \tag{7.25}$$

Now, we can pass with the limit on the original semidiscrete form of equation (7.2). Indeed, this is possible due to weak continuity of the von Karman nonlinearity and due to the fact that from (7.25) we have the *strong* convergence of the boundary traces

$$u_{h,t}|_\Gamma \to u_t|_\Gamma, \quad \nabla w_{h,t}|_\Gamma \to \nabla w_t|_\Gamma \quad \text{strongly in } L_\infty(0,T;H^{1/2-\epsilon}(\Gamma)). \tag{7.26}$$

Thus, the assumption imposed on nonlinear feedback, which implies the continuity of g, h_i from $L_\infty(0,T;H^\alpha(\Gamma_1)), \alpha \le 1/2$ into itself, allows us to conclude that

$$g(u_{h,t}) \to g(u_t),$$
$$h_1(D_n w_{h,t}) \to h_1(D_n w_t);\ h_2(D_\tau w_{h,t}) \to h_2(D_\tau w_t) \text{ strongly in } L_2(\Sigma). \tag{7.27}$$

Passing through the limit we conclude that u, w satisfy the weak form of the original equation (1.1) and, moreover, they display the regularity

$$u \in C(0,T;H^1(\Omega)), u_t \in C(0,T;H^1(\Omega)), u_{tt} \in L_\infty(0,T;L_2(\Omega)),$$
$$w \in C(0,T;H^2(\Omega)), w_t \in C(0,T;H^2(\Omega)), w_{tt} \in L_\infty(0,T;H^1(\Omega)). \tag{7.28}$$

In order to obtain higher regularity in the space variable, we proceed as usual by reading off the elliptic regularity. Indeed, from (7.28) for each $0 \le t \le T$ we now have

(7.29)
$$\operatorname{div}[\mathcal{C}N(u,w)] \in L_2(\Omega), u_t|\Gamma_1 \in H^{1/2}(\Gamma_1), f(\nabla w) \in H^{1-\epsilon}(\Omega), f(\nabla w)|_{\Gamma_1} \in H^{1/2-\epsilon}(\Gamma_1).$$

Hence

(7.30)
$$\operatorname{div}[\mathcal{C}\boldsymbol{\epsilon}(u)] \in H^{-\epsilon}(\Omega), \mathcal{C}[\boldsymbol{\epsilon}(u)]\nu = -f(\nabla w)\nu - g(u_t) \in H^{1/2-\epsilon}(\Gamma_1), u = 0 \text{ on } \Gamma_0.$$

By standard elliptic theory we conclude that

$$u(t) \in C(0,T;H^{2-\epsilon}(\Omega)). \tag{7.31}$$

As for the variable w we shall use standard, by now, semigroup formulation of the underlying PDE. Indeed, we rewrite (see [4], [1]) the original equation for w as

(7.32)
$$\mathcal{A}w = -[I + \gamma A_N]w_{tt} - b_2 w_t - \mathcal{A}G_1 h_1(D_n w_t) - \mathcal{A}G_2 D_\tau h_2(D_\tau w_t) + F(u,w),$$

where we have used the following operators:

$$\mathcal{A}w = D\Delta^2 w, w \in D(\mathcal{A}),$$
(7.33)
$$D(\mathcal{A}) = \{w \in H^4_{\Gamma_0}; [\Delta w + (1-\mu)B_1 w] = 0, [D_n \Delta w + (1-\mu)B_2 w] = 0 \text{ on } \Gamma_1\},$$
$$G_i : L_2(\Gamma_1) \to L_2(\Omega) \text{ are given by}$$
$$\Delta^2 G_i g = 0 \text{ in } \Omega,\ G_i g = 0, D_n G_i g = 0 \text{ on } \Gamma_0, \tag{7.34}$$
$$[\Delta + (1-\mu)B_1]G_1 g = g \text{ on } \Gamma_1,\ [D_n\Delta + (1-\mu)B_2]G_1 g = 0 \text{ on } \Gamma_1,$$
$$[\Delta + (1-\mu)B_1]G_2 g = 0 \text{ on } \Gamma_1,\ [D_n\Delta + (1-\mu)B_2]G_2 g = g \text{ on } \Gamma_1, \tag{7.35}$$
$$A_N w = -\Delta w,\ w \in D(A_N) = \left\{ w \in H^2_{\Gamma_0}(\Omega); w = 0, \frac{\partial w}{\partial \nu} = 0;\ \text{on } \Gamma_1 \right\}, \tag{7.36}$$
$$F(u,w) = \operatorname{div}\mathcal{C}N(u,w)\nabla w + \mathcal{A}G_2 \mathcal{C}N(u,w)\nu\nabla w. \tag{7.37}$$

Taking advantage of the improved regularity for u (see (7.31)) we obtain

$$N(u,w)\nabla w \in H^{1-\epsilon}(\Omega); N(u,w)\nabla w|_{\Gamma_1} \in H^{1/2-\epsilon}(\Gamma_1). \tag{7.38}$$

By elliptic regularity we have

$$G_2 : H^{1/2-4\epsilon}(\Gamma_1) \to H^{4-4\epsilon}(\Omega) \subset D(\mathcal{A}^{7/8-\epsilon}).$$

Thus,

$$F(u,w) \in H^{-\epsilon}(\Omega) + D(\mathcal{A}^{1/8+\epsilon})' \subset D(\mathcal{A}^{1/4})' \sim D({A_N}^{1/2})'. \tag{7.39}$$

We write $w = w_1 + w_2$, where

$$\mathcal{A}w_1 = -[I + \gamma A_N]w_{tt} - b_2 w_t + F(u,w), \tag{7.40}$$

$$\mathcal{A}w_2 = -\mathcal{A}G_1 h_1(D_n w_t) - \mathcal{A}G_2 D_\tau h_2(D_\tau w_t). \tag{7.41}$$

By (7.39) and the fact that $w_{tt} \in H^1(\Omega) = D({A_N}^{1/2})$, $A_N w_{tt} \in D({A_N}^{1/2})' = D(\mathcal{A}^{1/4})'$, we conclude that $\mathcal{A}w_1 \in D(\mathcal{A}^{1/4})'$. Hence,

$$w_1 \in D(\mathcal{A}^{3/4}) \in H^3(\Omega). \tag{7.42}$$

On the other hand, $w_2 = -G_1 h_1(D_n w_t) - G_2 D_\tau h_2(D_\tau w_t)$. By the regularity of Green's maps G_i together with the fact that $h_1(D_n w_t) \in H^{1/2}(\Gamma_1), D_\tau h_2(D_\tau w_t) \in H^{-1/2}(\Gamma_1)$ we also have

$$G_1 h_1(D_n w_t) + G_2 D_\tau h_2(D_\tau w_t) \in H^3(\Omega). \tag{7.43}$$

Hence

$$w_2(t) \in H^4(\Omega) \oplus H^3(\Omega) \in H^3(\Omega). \tag{7.44}$$

Combining (7.42) and (7.44) gives the desired assertion

$$w \in C(0,T,H^3(\Omega)),$$

and going back to (7.28) we further improve the regularity by ϵ, obtaining $u \in C(0,T;H^2(\Omega))$, as desired for the proof of the existence of regular solutions. ∎

7.2. Uniqueness of regular solutions. To complete the proof of the first part of Proposition 1.1, we need to establish the uniqueness of the solutions. The proof of the uniqueness is standard, and it relies on the fact that nonlinear terms are locally Lipschitz with respect to the topology considered for regular solutions. In fact, one can prove an even stronger uniqueness result which is stated below.

LEMMA 7.1. *The solutions are unique in the space*

$$X_\epsilon \equiv C(0,T;H^{1+\epsilon}(\Omega)) \cap C^1(0,T;L_2(\Omega)) \times C(0,T;H^{2+\epsilon}(\Omega)) \cap C^1(0,T;H^1(\Omega)),$$

where ϵ is an arbitrary positive number.

Proof. Let $\tilde{u} \equiv u_1 - u_2, \tilde{w} \equiv w_1 - w_2$ with $(u_1,w_1),(u_2,w_2)$ be two potential solutions living in the space X_ϵ. Then $\tilde{u},\tilde{w}$ satisfy the system of equations

$$\begin{aligned}(\tilde{u}_{tt},\phi)_\Omega + (b_1 u_t,\phi)_\Omega + (\mathcal{C}[\boldsymbol{\epsilon}(\tilde{u})],\epsilon(\phi))_\Omega + \langle g(u_{1,t}) - g(u_{2,t}),\phi\rangle_{\Gamma_1} \\ = -(\mathcal{C}[f(\nabla w_1) - f(\nabla w_2)],\boldsymbol{\epsilon}(\phi))_\Omega\end{aligned} \tag{7.45}$$

$$\begin{aligned}(\tilde{w}_{tt},\psi)_\Omega + \gamma(\nabla\tilde{w}_{tt},\nabla\psi)_\Omega + a(\tilde{w},\psi) + (b_2\tilde{w}_t,\psi)_\Omega \\ +\langle(h_1(D_n w_{1,t}) - (h_1(D_n w_{2,t})), D_n\psi\rangle_{\Gamma_1} + \langle(h_2(D_\tau w_{1,t}) - (h_2(D_\tau w_{2,t})), D_\tau\psi\rangle_{\Gamma_1} \\ = -(\mathcal{C}[\boldsymbol{\epsilon}(u_1) + f(\nabla w_1)]\nabla w_1 - \mathcal{C}[\boldsymbol{\epsilon}(u_2) + f(\nabla w_2)]\nabla w_2,\nabla\psi)_\Omega\end{aligned} \tag{7.46}$$

for all $\phi \in [H^1(\Omega)]^2; \psi \in H^1(\Omega)$.

The following regularity can be easily shown:

$$|f(\nabla w_1) - f(\nabla w_2)|_{0,\Omega} \le C|\tilde{w}|_{1,\Omega}[|w_1|_{2+\epsilon_1,\Omega} + |w_2|_{2+\epsilon_1,\Omega}] \tag{7.47}$$

and

$$\left|\frac{d}{dt}[f(\nabla w_1)-f(\nabla w_2)]\right|_{0,\Omega}\le C|\tilde{w}_t|_{1,\Omega}[|w_1|_{2+\epsilon_1,\Omega}+|w_2|_{2+\epsilon_1,\Omega}]. \tag{7.48}$$

Integration by parts gives

$$\begin{aligned}\int_0^t(\mathcal{C}[f(\nabla w_1)-f(\nabla w_2)],\boldsymbol{\epsilon}(\tilde{u}_s))_\Omega ds&=(\mathcal{C}[f(\nabla w_1)-f(\nabla w_2)],\boldsymbol{\epsilon}(\tilde{u}_t))_\Omega|_0^t\\&\quad-\int_0^t\left(\frac{d}{ds}\mathcal{C}[f(\nabla w_1)-f(\nabla w_2)],\boldsymbol{\epsilon}(\tilde{u})\right)_\Omega ds.\end{aligned} \tag{7.49}$$

From (7.48)

$$\begin{aligned}\int_0^t\left(\frac{d}{ds}\mathcal{C}[f(\nabla w_1)-f(\nabla w_2)],\boldsymbol{\epsilon}(\tilde{u})\right)_\Omega ds&\le C\int_0^t|\tilde{u}|^2_{1,\Omega}ds+\int_0^t|\tilde{w}_t|^2_{1,\Omega}[|w_1|_{2+\epsilon_1,\Omega}\\&\quad+|w_2|_{2+\epsilon_1,\Omega)}]^2ds.\end{aligned} \tag{7.50}$$

On the other hand

$$(\mathcal{C}[f(\nabla w_1)-f(\nabla w_2)],\boldsymbol{\epsilon}(\tilde{u}_t))_\Omega|_0^t\le\epsilon_0|\tilde{u}(t)|^2_{1,\Omega}+C_{\epsilon_0}|f(\nabla w_1(t))-f(\nabla w_2(t)|^2_{0,\Omega}. \tag{7.51}$$

But

$$\begin{aligned}|f(\nabla w_1(t,x))-f(\nabla w_2(t,x))|&\le C|\nabla\tilde{w}(t,x)||\nabla w_i(t,x)|\\&=C\left|\int_0^t\nabla\tilde{w}_t(t,x)ds\right|[|\nabla w_1(t,x)|+|\nabla w_2(t,x)|]\\&\le C_T\left[\int_0^t|\nabla\tilde{w}_t(s,x)|^2ds\right]^{1/2}[|w_1(t)|_{2+\epsilon_1,\Omega}+|w_2(t)|_{2+\epsilon_1,\Omega}].\end{aligned} \tag{7.52}$$

Hence

$$|f(\nabla w_1(t))-f(\nabla w_2(t))|^2_{0,\Omega}\le C_T\int_0^t|\tilde{w}_s|^2_{1,\Omega}ds[|w_1(t)|^2_{2+\epsilon_1,\Omega}+|w_2(t)|^2_{2+\epsilon_1,\Omega}]. \tag{7.53}$$

Combining (7.51), (7.53) gives

$$\begin{aligned}(\mathcal{C}[f(\nabla w_1)-f(\nabla w_2)],\boldsymbol{\epsilon}(\tilde{u}_t))_\Omega|_0^t&\le\epsilon_0|\tilde{u(t)}|^2_{1,\Omega}+C_{\epsilon_0,T}\int_0^t|\tilde{w}_s|^2_{1,\Omega}ds[|w_i(t)|^2_{2+\epsilon_1,\Omega}]\\&\quad+[|w_i(t)|^2_{2+\epsilon_1,\Omega}].\end{aligned} \tag{7.54}$$

Setting in (7.45) $\phi\equiv\tilde{u_t}$, taking advantage of inequalities in (7.50), (7.54), and recalling the monotonicity of g yields

$$|\tilde{u}_t|^2_{0,\Omega}+|\tilde{u}|^2_{1,\Omega}\le C_{\epsilon_0,T}\int_0^t|\tilde{w}_t|^2_{1,\Omega}[|w_1|^2_{2+\epsilon,\Omega}+|w_2|^2_{2+\epsilon,\Omega}]+|\tilde{u}|^2_{1,\Omega}ds+\epsilon_0|\tilde{u}|^2_{1,\Omega}. \tag{7.55}$$

Similarly,

$$\begin{aligned}|[\boldsymbol{\epsilon}(u_1)-\boldsymbol{\epsilon}(u_2)]\nabla w_1|_{0,\Omega}&\le C|\tilde{u}|_{1,\Omega}|w_1|_{2+\epsilon},\\|\boldsymbol{\epsilon}(u_2)[\nabla w_1-\nabla w_2]|_{0,\Omega}&\le C|\tilde{w}|_{2,\Omega}|u|_{W^{1,\infty}(\Omega)}\le C|\tilde{w}|_{2,\Omega}|u_2|_{1+\epsilon,\Omega},\end{aligned} \tag{7.56}$$

$$|[f(\nabla w_1)\nabla w_1 - f(\nabla w_2)\nabla w_2]|_{0,\Omega} \le C|\tilde{w}|_{2,\Omega}[|w_1|^2_{2+\epsilon,\Omega} + |w_2|^2_{2+\epsilon,\Omega}]. \tag{7.57}$$

From (7.46) (where we set $\psi = \tilde{w}_t$) and (7.57) and accounting for the monotonicity of h_i we obtain the estimate

$$|\tilde{w}_t|^2_{1,\Omega} + |\tilde{w}|^2_{2,\Omega} \le C\int_0^t |\tilde{w}|^2_{2,\Omega}[|u_2|_{1+\epsilon,\Omega} + |\tilde{u}|_{1,\Omega}|w_1|_{2+\epsilon,\Omega}]dt. \tag{7.58}$$

Combining Gronwall's lemma gives $\tilde{u} \equiv 0, \tilde{w} \equiv 0$, as desired. □

7.3. Uniqueness of weak solutions: Proof of Part II of Proposition 1.1. Here we adapt the method proposed in [24] where Marguerre–Vlasov equations were considered with the zero Dirichlet boundary data. We note that the presence of the damping on the boundary as well as the fact that the boundary conditions are dynamic and of the higher order (as treated in this paper) add substantial technical difficulties with respect to the homogeneous Dirichlet case.

To begin with, let us introduce the operators providing for the semigroup representation of the part of the system corresponding to the longitudinal displacements u. Indeed, let A_0 be the generator corresponding to the system of elasticity, and let G_0 be the corresponding Green map defined as below:

$$A_0 u \equiv -\mathrm{div}\mathcal{C}\boldsymbol{\epsilon}(u); D(A_0) = \{u \in H^2_{\Gamma_0}(\Omega); \mathcal{C}\boldsymbol{\epsilon}(u)\nu = 0 \text{ on } \Gamma_1\},$$

$$G_0 g \equiv g \text{ iff } \mathrm{div}\mathcal{C}\boldsymbol{\epsilon}(v) = 0 \text{ in } \Omega,\ v = 0 \text{ on } \Gamma_0,\ \mathcal{C}\epsilon(v)\nu = g \text{ on } \Gamma_1.$$

Let $\tilde{u} \equiv u_1 - u_2, \tilde{w} \equiv w_1 - w_2$, where u_1, w_1 and u_2, w_2 are two potential solutions of finite energy (i.e., weak solutions). Using the definitions of operators A_0, G_0 given above and $M, \mathcal{A}, G_i$ introduced in the previous section we rewrite the original PDE as an abstract second-order system defined on $[D(A_0)]' \times [D(\mathcal{A})]'$ (see [1], [4], [16]):

$$\begin{aligned} &\tilde{u}_{tt} + b_1\tilde{u}_t + A_0\tilde{u} + A_0G_0(G_0^*A_0\tilde{u}_t) = f_1(\tilde{w}, w_i),\\ &M\tilde{w}_{tt} + b_2\tilde{w}_t + \mathcal{A}\tilde{w} + \mathcal{A}G_1(G_1^*\mathcal{A}\tilde{w}_t) + \mathcal{A}G_2(D_\tau^2 G_2^*\mathcal{A}\tilde{w}_t) = f_2(\tilde{u}, \tilde{w}, u_i, w_i), \end{aligned} \tag{7.59}$$

where the forcing terms f_i are defined as follows:

$$\begin{aligned} f_1(\tilde{w}, w_i) &\equiv \mathrm{div}\mathcal{C}[f(\nabla w_1) - f(\nabla w_2)] - A_0G_0G_0^*A_0\mathcal{C}[f(\nabla w_1) - f(\nabla w_2)]\nu,\\ f_2(\tilde{u}, \tilde{w}, u_i, w_i) &\equiv \mathrm{div}[\mathcal{C}N(u_1, w_1)\nabla w_1 - \mathcal{C}N(u_2, w_2)\nabla w_2]\\ &\quad + \mathcal{A}G_2G_2^*\mathcal{A}[\mathcal{C}N(u_1, w_1)\nabla w_1 - \mathcal{C}N(u_2, w_2)\nabla w_2]\nu. \end{aligned} \tag{7.60}$$

Since the damping is linear, we have assumed, without the loss of generality, that $g_0, h_{i0} = 1; i = 1, 2$.

Denoting

$$A_1 \equiv \begin{bmatrix} 0 & I \\ -A_0 & -A_0G_0G_0^*A_0 - b_1 \end{bmatrix}, \tag{7.61}$$

$A_1 : H_1 \to H_1$ with $H_1 \equiv [D({A_0}^{1/2})]^2 \times [L_2(\Omega)]^2$,

$$D(A_1) = \{(u_1, u_2) \in [D({A_0}^{1/2})]^2 \times [D({A_0}^{1/2})]^2; u_1 + G_0G_0^*A_0u_2 \in D(A_0)\},$$

$$A_2 \equiv \begin{bmatrix} 0 & I \\ -M^{-1}\mathcal{A} & -M^{-1}(\mathcal{A}G_1G_1^*\mathcal{A} + \mathcal{A}G_2D_\tau^2G_2^*\mathcal{A}_0 + b_2) \end{bmatrix}, \tag{7.62}$$

$A_2 : H_2 \to H_2$ with $H_2 \equiv D(\mathcal{A}^{1/2}) \times D(A_N)^{1/2}$,

$$D(A_2) = \{(w_1, w_2) \in D(\mathcal{A}^{1/2}) \times D(\mathcal{A}^{1/2}); \mathcal{A}(w_1 + G_1 G_1^* \mathcal{A} w_2 + G_2 D_\tau^2 G_2^* \mathcal{A} \tilde{w}_2) \in [D(A_N{}^{1/2})]'\}, \tag{7.63}$$

we rewrite (7.59) as

$$\begin{pmatrix} \tilde{u}(t) \\ \tilde{u}_t(t) \end{pmatrix}_t = A_1 \begin{pmatrix} \tilde{u}(t) \\ \tilde{u}_t(t) \end{pmatrix} + \begin{pmatrix} 0 \\ f_1(\tilde{w}, w_i) \end{pmatrix}, \tag{7.64}$$

$$\begin{pmatrix} \tilde{w}(t) \\ \tilde{w}_t(t) \end{pmatrix}_t = A_2 \begin{pmatrix} \tilde{w}(t) \\ \tilde{w}_t(t) \end{pmatrix} + \begin{pmatrix} 0 \\ M^{-1} f_2(\tilde{u}, \tilde{w}, u_i, w_i) \end{pmatrix}. \tag{7.65}$$

It is well known that the operators A_1, (resp., A_2) are generators of a contraction semigroup on the spaces $[D(A_0{}^{1/2})]^2 \times [L_2(\Omega)]^2$ and $D(\mathcal{A}^{1/2}) \times D(A_N{}^{1/2})$, respectively. By the standard semigroup argument we also know that these operators generate strongly continuous semigroups on the dual spaces $[D(A_1^*)]'$ and $[D(A_2^*)]'$, respectively, where the duality is, as usual, with respect to the H_1, H_2 topology. This implies the following negative norm estimates:

$$\begin{aligned} |(\tilde{u}, \tilde{u}_t)|_{[D(A_1^*)]'} &\le C \int_0^t |(0, f_1)|_{[D(A_1^*)]'} dt, \\ |(\tilde{w}, \tilde{w}_t)|_{[D(A_2^*)]'} &\le C \int_0^t |(0, M^{-1} f_2)|_{[D(A_1^*)]'} dt. \end{aligned} \tag{7.66}$$

On the other hand, by direct computations of $[A_i^*]^{-1}$ we obtain

$$\begin{aligned} |(u_1, u_2)|^2_{[D(A_1^*)]'} &= |u_1|^2_{0,\Omega} + |A_0{}^{-1/2} u_2 + A_0{}^{1/2} G_0 G_0^* A_0^* u_1|^2_{0,\Omega}, \\ |(w_1, w_2)|^2_{[D(A_2^*)]'} &= |M^{1/2} w_1|^2_{0,\Omega} + |\mathcal{A}^{-1/2} M w_2 + \mathcal{A}^{1/2} [G_1 G_1^* \mathcal{A} w_1 + G_2 D_\tau^2 G_2^* \mathcal{A} w_1]|^2_{0,\Omega}. \end{aligned} \tag{7.67}$$

Thus, in particular,

$$\begin{aligned} |\tilde{u}(t)|^2_{0,\Omega} &\le C \int_0^t |A_0^{-1/2} f_1|^2_{0,\Omega} dt, \\ |\tilde{w}(t)|^2_{1,\Omega} &\le C \int_0^t |\mathcal{A}^{-1/2} f_2|^2_{0,\Omega} dt. \end{aligned} \tag{7.68}$$

Inequalities in (7.68) form the basis for subsequent analysis.

Let $\phi \in [L_2(\Omega)]^2$. We compute

$$(A_0^{-1/2} f_1, \phi)_{0,\Omega} = -(\mathcal{C}[f(\nabla w_1) - f(\nabla w_2)], \boldsymbol{\epsilon}(A_0^{-1/2} \phi))_{0,\Omega}. \tag{7.69}$$

Hence,

$$|A_0^{-1/2} f_1|_{0,\Omega} \le C |f(\nabla w_1) - f(\nabla w_2)|_{0,\Omega}. \tag{7.70}$$

Let P_n be the orthogonal projection on the subspace spanned by n eigenvectors of $\mathcal{A}$ and let $Q_n = I - P_n$. (One could also take a projection on the subspace spanned by the eigenvectors of the biharmonic operator with clamped boundary conditions; this particular choice is not critical for the arguments.) The following "logarithmic"

estimate resulting from Sobolev's embedding and the Holder inequality is known [24], [2]:

$$|(P_n f)g|_{0,\Omega} \le C\mathrm{lg}(1+\lambda_n)^{1/2}|f|_{0,\Omega}|g|_{1,\Omega}, \tag{7.71}$$

where λ_n is an eigenvalue of $\mathcal{A}$ and the constant C is independent of n. Applying (7.71) and abusing notation slightly by using the projection operator applied to a vector function (meaning the projection is applied to each component) we obtain

$$\begin{aligned}|(f(\nabla w_1)-f(\nabla w_2))|_{0,\Omega} &\le C|\nabla\tilde{w}\times\nabla(w_1+w_2)|_{0,\Omega} \le C|(P_n\nabla\tilde{w})\times\nabla(w_1+w_2)|_{0,\Omega}\\ &+|(Q_n\nabla\tilde{w})\times\nabla(w_1+w_2)|_{0,\Omega} \le C\mathrm{lg}(1+\lambda_n)^{1/2}|\tilde{w}|_{1,\Omega}[|w_1|_{2,\Omega}+|w_2|_{2,\Omega}]\\ &+C|Q_n\nabla\tilde{w}|_{\epsilon,\Omega}[|w_1|_{2,\Omega}+|w_2|_{2,\Omega}].\end{aligned} \tag{7.72}$$

On the other hand we have

$$\begin{aligned}|Q_n\nabla\tilde{w}|_{\epsilon,\Omega} &\le C|\mathcal{A}^{1/4\epsilon}Q_n\nabla\tilde{w}|_{0,\Omega} \le C|\mathcal{A}^{1/4(\epsilon-\beta_0)}Q_n|_{\mathcal{L}(\mathcal{L}_2(\Omega))}|\nabla\tilde{w}|_{\beta_0,\Omega}\\ &\le C{\lambda_n}^{1/4(\epsilon-\beta_0)}|\nabla\tilde{w}|_{\beta_0,\Omega}| \le C{\lambda_n}^{1/4(\epsilon-\beta_0)}|\tilde{w}|_{3/2,\Omega},\end{aligned} \tag{7.73}$$

where $\beta_0 < 1/2$.

Combining (7.72), (7.73), (7.70) gives

$$\begin{aligned}|(A_0^{-1/2}f_1|_{0,\Omega} &\le C|(f(\nabla w_1)-f(\nabla w_2)|_{0,\Omega}\\ &\le C[\mathrm{lg}(1+\lambda_n)^{1/2}|\tilde{w}|_{1,\Omega}+{\lambda_n}^{-\beta}][|w_1|_{2,\Omega}+|w_2|_{2,\Omega}]\\ &\le C(E(0))\mathrm{lg}(1+\lambda_n)^{1/2}|\tilde{w}|_{1,\Omega}+C(E(0)){\lambda_n}^{-\beta},\end{aligned} \tag{7.74}$$

where $\beta < 1/8$ and $E(0)$ denotes the initial energy of weak solutions. The estimate for f_2 is carried out next. With $\psi \in L_2(\Omega)$ we have

$$(\mathcal{A}^{-1/2}f_2,\psi)_{0,\Omega} = (\mathcal{C}N(u_1,w_1)\nabla w_1 - \mathcal{C}N(u_2,w_2)\nabla w_2, \nabla\mathcal{A}^{-1/2}\psi)_{0,\Omega}. \tag{7.75}$$

We shall compute the right-hand side of (7.75). By using (7.71) we obtain

$$\begin{aligned}(\boldsymbol{\epsilon}(u_2)\nabla\tilde{w},\nabla\mathcal{A}^{-1/2}\psi)_{0,\Omega} &\le C|u_2|_{1,\Omega}|\nabla\tilde{w}\times\nabla\mathcal{A}^{-1/2}\psi|_{0,\Omega}\\ \le C|u_2|_{1,\Omega}[\log(1+\lambda_n)^{1/2}|\tilde{w}|_{1,\Omega}|\nabla\mathcal{A}^{-1/2}\psi|_{1,\Omega} &+ {\lambda_n}^{-\beta}[|w_1|_{2,\Omega}+|w_2|_{2,\Omega}]|\nabla\mathcal{A}^{-1/2}\psi|_{1,\Omega}]\\ &\le C[\mathrm{lg}(1+\lambda_n)^{1/2}|\tilde{w}|_{1,\Omega}+{\lambda_n}^{-\beta}[|w_1|_{2,\Omega}+|w_2|_{2,\Omega}]]|u_2|_{1,\Omega}|\psi|_{0,\Omega}].\end{aligned} \tag{7.76}$$

By the divergence theorem we have

$$\begin{aligned}(\boldsymbol{\epsilon}(\tilde{u})\nabla w_1,\nabla\mathcal{A}^{-1/2}\psi)_{0,\Omega} &= \langle\tilde{u},(1/2\nabla w_1\times\nabla\mathcal{A}^{-1/2}\psi)+(1/2\nabla w_1\times\nabla\mathcal{A}^{-1/2}\psi)^T\nu\rangle_{0,\Gamma_1}\\ &-(\tilde{u},\mathrm{div}((1/2\nabla w_1\times\nabla\mathcal{A}^{-1/2}\psi)+(1/2\nabla w_1\times\nabla\mathcal{A}^{-1/2}\psi)^T))_\Omega.\end{aligned} \tag{7.77}$$

Define

$$K \equiv 1/2\nabla w_1\times\nabla\mathcal{A}^{-1/2}\psi + (1/2\nabla w_1\times\nabla\mathcal{A}^{-1/2}\psi)^T.$$

Simple calculations and (7.71) imply

$$\begin{aligned}|(\tilde{u},\mathrm{div}K)_{0,\Omega}| &\le |(P_n\tilde{u},\mathrm{div}K)_{0,\Omega}+(Q_n\tilde{u},\mathrm{div}K)_{0,\Omega}]\\ \le C[|\tilde{u}|_{0,\Omega}\mathrm{lg}(1+\lambda_n)^{1/2}|w_1|_{2,\Omega}|\nabla\mathcal{A}^{-1/2}\psi|_{1,\Omega} &+ |\tilde{u}|_{1,\Omega}{\lambda_n}^{-\beta}|w_1|_{2,\Omega}|\nabla\mathcal{A}^{-1/2}\psi|_{1,\Omega}]\\ &\le C[|\tilde{u}|_{0,\Omega}\mathrm{lg}(1+\lambda_n)^{1/2}+|\tilde{u}|_{1,\Omega}{\lambda_n}^{-\beta}]|w_1|_{2,\Omega}|\psi|_{0,\Omega}.\end{aligned} \tag{7.78}$$

Also, by the trace theorem we have

$$\langle \tilde{u}, K\nu\rangle_\Gamma \le C|\tilde{u}|_{0,\Gamma_1}|K|_{1/2+\epsilon,\Omega} \le C|\tilde{u}|_{0,\Gamma_1}|w_1|_{2,\Omega}|\psi|_{0,\Omega}. \tag{7.79}$$

The crucial boundary term $|\tilde{u}|_{0,\Gamma_1}$ will be estimated later.

Finally the term $(f(\nabla w_1)\nabla w_1 - f(\nabla w_2)\nabla w_2, \nabla\mathcal{A}^{-1/2}\psi)_{0,\Omega}$ is estimated directly as

$$(f(\nabla w_1)\nabla w_1 - f(\nabla w_2)\nabla w_2, \nabla\mathcal{A}^{-1/2}\psi)_{0,\Omega} \le C|\tilde{w}|_{1,\Omega}[|w_1|^2_{2,\Omega} + |w_1|^2_{2,\Omega}]|\psi|_{0,\Omega}. \tag{7.80}$$

Collecting (7.75)–(7.80) yields

$$\begin{aligned}(\mathcal{A}^{-1/2}f_2,\psi)_{0,\Omega} &\le C\lg(1+\lambda_n)^{1/2}[|\tilde{u}|_{0,\Omega} + |\tilde{w}|_{1,\Omega}]\\ &\cdot[|w_1|_{2,\Omega} + |w_2|_{2,\Omega} + |u_2|_{1,\Omega} + |u_1|_{1,\Omega} + |w_1|^2_{2,\Omega} + |w_2|^2_{2,\Omega}]|\psi|_{0,\Omega}\\ &+C|\tilde{u}|_{0,\Gamma_1}|w_1|_{2,\Omega}|\psi|_{0,\Omega} + C{\lambda_n}^{-\beta}[|w_1|^2_{2,\Omega} + |w_2|^2_{2,\Omega} + |u_2|^2_{1,\Omega} + |u_1|^2_{1,\Omega}]|\psi|_{0,\Omega}.\end{aligned} \tag{7.81}$$

Hence

$$\begin{aligned}\int_0^t |\mathcal{A}^{-1/2}f_2|^2_{0,\Omega}ds &\le C(E(0))\lg(1+\lambda_n)\int_0^t [|\tilde{u}|^2_{0,\Omega} + |\tilde{u}|^2_{0,\Gamma_1} + |\tilde{w}|^2_{1,\Omega}]ds\\ &+C(E(0))\int_0^t |\tilde{u}|^2_{0,\Gamma_1}ds + {\lambda_n}^{-2\beta}C_T(E(0)).\end{aligned} \tag{7.82}$$

The crucial next step is to prove the following trace estimate

$$\int_0^t |\tilde{u}|^2_{0,\Gamma_1}ds \le C(E(0))\lg(1+\lambda_n)\int_0^t |\tilde{w}|^2_{1,\Omega}ds + C_T(E(0)){\lambda_n}^{-2\beta}. \tag{7.83}$$

To prove (7.83), the idea is to use the energy estimates for a new variable u^* defined as

$$u^*(t) \equiv \int_0^t \tilde{u}ds.$$

Denoting

$$\tilde{f} \equiv \mathcal{C}[f(\nabla w_1) - f(\nabla w_2)], \quad f^*(t) \equiv \int_0^t \tilde{f}ds,$$

we obtain the following equation satisfied by the new variable u^*:

$${u^*}_{tt} + A_0u^* + b_1u^*_t + A_0G_0(G_0^*A_0u^*_t) = \mathrm{div}f^* - A_0G_0G_0^*A_0f^*\nu. \tag{7.84}$$

Multiplying this equation by u^*_t and integrating by parts yield

$$|u^*_t(t)|^2_{0,\Omega} + |u^*(t)|^2_{1,\Omega} + \int_0^t |u^*_t(s)|^2_{0,\Gamma_1}ds \le C_T\int_0^t (f^*, \boldsymbol{\epsilon}(u^*_t))_{0,\Omega}ds. \tag{7.85}$$

When we integrate by parts in time the last term in the above inequality yields

$$|u^*_t(t)|^2_{0,\Omega} + |u^*(t)|^2_{1,\Omega} + \int_0^t |u^*_t(s)|^2_{0,\Gamma_1}ds \le C\left[\int_0^t |f^*_t|_{0,\Omega}|u^*|_{1,\Omega}ds + |f^*(t)|_{0,\Omega}|u^*(t)|_{1,\Omega}\right]. \tag{7.86}$$

Hence, in particular,

$$\int_0^t |\tilde{u}(s)|^2_{0,\Gamma_1} = \int_0^t |u_t^*(s)|^2_{0,\Gamma_1} ds \le C_T \left[\int_0^t |f_t^*|^2_{0,\Omega} ds + |f^*(t)|^2_{0,\Omega} \right]. \tag{7.87}$$

On the other hand,

$$\begin{aligned} \int_0^t |f_t^*|^2_{0,\Omega} ds &= \int_0^t |\tilde{f}|^2_{0,\Omega} d \le C \int_0^t |\nabla \tilde{w} \times \nabla w_i|^2_{0,\Omega} ds \\ &\le C \int_0^t |P_n \nabla \tilde{w} \times \nabla w_i|^2_{0,\Omega} + |Q_n \nabla \tilde{w} \times \nabla w_i|^2_{0,\Omega} ds \\ &\le C \lg(1+\lambda_n) \int_0^t |\tilde{w}|^2_{1,\Omega} |w_i|^2_{2,\Omega} ds + C {\lambda_n}^{-2\beta} \int_0^t |\tilde{w}|^2_{2,\Omega} |w_i|^2_{2,\Omega} ds. \end{aligned} \tag{7.88}$$

Similarly,

$$\begin{aligned} |f^*(t)|^2_{0,\Omega} &\le C_T \int_0^t |\tilde{f}|^2_{0,\Omega} ds \\ &\le C \lg(1+\lambda_n) \int_0^t |\tilde{w}|^2_{1,\Omega} |w_i|^2_{2,\Omega} ds + C {\lambda_n}^{-2\beta} \int_0^t |\tilde{w}|^2_{2,\Omega} |w_i|^2_{2,\Omega} ds. \end{aligned} \tag{7.89}$$

Collecting (7.87)–(7.89) leads to the desired conclusion in (7.83). From (7.82) and (7.83) we conclude

$$\int_0^t |\mathcal{A}^{-1/2} f_2|^2_{0,\Omega} ds \le C(E(0)) \lg(1+\lambda_n) \int_0^t [|\tilde{u}|^2_{0,\Omega} + + |\tilde{w}|^2_{1,\Omega}] ds + {\lambda_n}^{-2\beta} C(E(0)). \tag{7.90}$$

From (7.74) and (7.90) we obtain

$$\begin{aligned} \int_0^t [|\mathcal{A}^{-1/2} f_2|^2_{0,\Omega} + |{A_0}^{-1/2} f_1|^2_{0,\Omega}] ds \le C(E(0)) [\lg(1+\lambda_n)] \int_0^t [|\tilde{u}|^2_{0,\Omega} + |\tilde{w}|^2_{1,\Omega}] \\ + {\lambda_n}^{-\beta} C_T(E(0)), \end{aligned} \tag{7.91}$$

and combining this with (7.68) and Gronwall's inequality, we have

$$|\tilde{u}(t)|^2_{0,\Omega} + |\tilde{w}(t)|^2_{1,\Omega} \le C_T(E(0)) {\lambda_n}^{-2\beta} (1+\lambda_n)^{C(E(0))t}. \tag{7.92}$$

Taking $t < T_0$, where T_0 is sufficiently small yields the conclusion of the lemma for $t < T_0$. Applying the "boost trap" argument completes the proof of the second part of Proposition 1.1. □

REFERENCES

[1] M. Bradley and I. Lasiecka, *Local exponential stabilization for a nonlinearly perturbed von Karman plate*, Nonlinear Anal., 18 (1992), pp. 333–343.

[2] A. B. de Monvel and I. Chueshov, *Uniqueness theorem for weak solutions of von Karman evolution equations*, J. Math. Anal. Appl., to appear.

[3] A. Favini, M. A. Horn, I. Lasiecka, and D. Tataru, *Global existence, uniqueness and regularity of solutions to a von Karman system with nonlinear boundary dissipation*, Differential Integral Equations, 9 (1996), pp. 267–294.

[4] A. Favini and I. Lasiecka, *Wellposedness and regularity of second order abstract equations arising in hyperbolic-like problems with nonlinear boundary conditions*, Osaka J. Math., 32 (1995), pp. 721–752.

[5] M. A. Horn, *Sharp trace regularity of the traces to solutions of dynamic elasticity*, J. Math. Systems, Estim. Control, 8 (1998), pp. 217–219.
[6] M. A. Horn and I. Lasiecka, *Uniform decay of weak solutions to a von Karman plate with nonlinear dissipation*, Differential Integral Equations, 7 (1994), pp. 885–908.
[7] M. A. Horn and I. Lasiecka, *Global stabilization of a dynamic von Karman plate with nonlinear boundary feedback*, Appl. Math. Optim., 31 (1995), pp. 57–84.
[8] V. Isakov, *A nonhyperbolic Cauchy problem for $\Box_b \Box_c$ and its applications to elasticity theory*, Comm. Partial Differential Equations, XXXIX (1986), pp. 747–767.
[9] V. Isakov, *On uniqueness in a lateral Cauchy Problem with multiple characteristics*, J. Differential Equations, 97 (1997), pp. 134–147.
[10] V. Komornik, *Exact Controllability and Stabilization. The Multipliers Method*, Masson, Paris, 1994.
[11] J. Lagnese, *Boundary Stabilization of Thin Plates*, SIAM, Philadelphia, PA, 1989.
[12] J. Lagnese, *Modeling and stabilization of nonlinear plates*, Internat. Ser. Numer. Math., 100 (1991), pp. 247–264.
[13] J. Lagnese, *Uniform asymptotic energy estimates for solutions of the equations of dynamic plane elasticity with nonlinear dissipation at the boundary*, Nonlinear Anal., 16 (1991), pp. 35–54.
[14] J. Lagnese and G. Leugering, *Uniform stabilization of a nonlinear beam by nonlinear boundary feedback*, J. Differential Equations, 91 (1991), pp. 355–388.
[15] J. Lagnese and J. Lions, *Modelling Analysis and Control of Thin Plates*, Masson, Paris, 1988.
[16] I. Lasiecka, *Existence and uniqueness of the solutions to second order abstract equations with nonlinear and nonmonotone boundary conditions*, Nonlinear Anal., 23 (1994), pp. 797–823.
[17] I. Lasiecka, *Finite dimensionality of attractors associated with von Karman plate equations and boundary damping*, J. Differential Equations, 117 (1995), pp. 357–389.
[18] I. Lasiecka, *Intermediate solutions to full von Karman system of dynamic nonlinear elasticity*, Appl. Anal., (1998), to appear.
[19] I. Lasiecka and D. Tataru, *Uniform boundary stabilization of semilinear wave equations with nonlinear boundary damping*, J. Differential Integral Equations, 6 (1993), pp. 507–533.
[20] I. Lasiecka and R. Triggiani, *Uniform stabilization of the wave equation with Dirichlet and Neumann feedback control without geometrical conditions*, Appl. Math. Optim., 25 (1992), pp. 189–224.
[21] I. Lasiecka and R. Triggiani, *Sharp trace estimates of solutions to Kirchhoff and Euler Bernoulli equations*, Appl. Math. Optim., 28 (1993), pp. 277–306.
[22] J. Puel and M. Tucsnak, *Boundary stabilization for the von Karman equations*, SIAM J. Control, 33 (1996), pp. 255–273.
[23] J. Puel and M. Tucsnak, *Global existence for the full von Karman system*, Appl. Math. Optim., 34 (1996), pp. 139–161.
[24] V. I. Sedenko, *On uniqueness of the generalized solutions of initial boundary value problem for Marguerre-Vlasov nonlinear oscillations of the shallow shells*, Russian Izv., North-Caucasus Region, Ser. Natural Sciences, 1-2 (1994).
[25] D. Tataru, *Private communication*, 1997.
[26] D. Tataru, *A-priori Pseudoconvexity Energy Estimates in Domains with Boundary and Exact Controllability for Conservative P.D.E's*, Ph.D. thesis, University of Virginia, Charlottesville, VA, 1992.
[27] D. Tataru, *A-priori estimates of Carleman's type in domains with boundary*, J. Math. Pure Appl., 73 (1994), pp. 355–387.
[28] M. A. Horn, *Implications of sharp trace regularity results on boundary stabilization of the system of linear elasticity*, J. Math. Anal. Appl., (1998), to appear.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 4, pp. 1423–1447, July 1998

015

APPROXIMATE FILTER FOR THE CONDITIONAL LAW OF A PARTIALLY OBSERVED PROCESS IN NONLINEAR FILTERING*

A. GEGOUT-PETIT†

Abstract. We study a problem of singular perturbations for a special class of nonlinear filtering problems in which the dimension of the signal process is 2 and only one of the two components of this process is observed. We propose an approximate filter of finite dimension for the observed part. Using this filter, we construct an approximate filter of infinite dimension for the nonobserved part. This filter solves a Zakai-type equation whose spatial variable dimension is 1 even though the spatial variable dimension of the Zakai equation solved by the exact filter is 2. The method used gives the order of the approximation error.

Key words. nonlinear filtering, singular perturbations, partially observed process, Zakai's equation

AMS subject classifications. 93E11, 60G35, 60F99

PII. S0363012995287040

Introduction. This paper considers an asymptotic problem in nonlinear filtering when the signal observes only one component of X. In recent years, nonlinear filtering with high signal-to-noise ratio has been studied in numerous publications, some of whose fundamental results we will now recall before presenting our work.

Consider the following filtering problem:

$$\begin{cases} X_t = X_0 + \displaystyle\int_0^t f(X_s)ds + \int_0^t g(X_s)dV_s, \\ Y_t = \displaystyle\int_0^t h(X_s)ds + \varepsilon W_t, \end{cases} \tag{1}$$

where X_t is a nonobserved vectorial process and Y_t is the observation.

The filtering problem consists of computing the best approximation of the law of X_t using the observation of Y up to instant t: it is the conditional law of X_t given $\mathcal{Y}_t = \sigma\{Y_s, 0 \le s \le t\}$. In the nonlinear case, the solution of this problem has infinite dimension. Indeed, the unnormalized conditional law of X_t given $\mathcal{Y}_t$ satisfies a parabolic partial differential equation (PDE) called Zakai's equation (see Zakai [16] or Pardoux [5], [6]). As is often the case in practice, we will suppose throughout our work that ε is small. We are therefore dealing with the case of high signal-to-noise ratio.

We will often use the notation $\hat{X}_t \stackrel{\Delta}{=} E[X_t/\mathcal{Y}_t]$.

Filtering with high signal-to-noise ratio.

Case where h is one to one. When $\varepsilon = 0$, $dY_t = h(X_t)dt$, and if h is one to one, X is perfectly observed. We expect that X_t is "almost $\mathcal{Y}_t$-measurable" when ε is small in the sense that $(X_t - E[X_t/\mathcal{Y}_t])$ is small in the L^p-norm. In fact, suppose

*Received by the editors June 2, 1995; accepted for publication (in revised form) June 3, 1997; published electronically May 28, 1998.
http://www.siam.org/journals/sicon/36-4/28704.html
†L.A.T.P. URA CNRS 225, C.M.I. Université de Provence, 39, rue Joliot Curie, 13453 MARSEILLE Cedex 13, France (petit@dlp.ems-cachan.fr).

$\hat{X}_t - X_t = O(\varepsilon^k)$ (we will define the meaning of $O(\varepsilon^k)$ rigorously in Definition 1.2). For any function φ smooth enough, we have

$$\varphi(X_t) = \varphi(\hat{X}_t) + \varphi'(\hat{X}_t)(X_t - \hat{X}_t) + \frac{1}{2}\varphi''(\xi_t)(X_t - \hat{X}_t)^2$$

with $\xi_t \in [X_t, \hat{X}_t]$. Taking the conditional expectation of this expression, we get

$$E[\varphi(X_t)/\mathcal{Y}_t] = \varphi(\hat{X}_t) + O(\varepsilon^{2k}). \tag{2}$$

In this case we see that if we have a good approximation of $\hat{X}_t$, we also get a good estimation of the whole conditional law since it is "concentrated" around $\hat{X}_t$ and the filtering problem is considerably simplified as there is no need to compute the whole conditional law of X_t given $\mathcal{Y}_t$ (i.e., $E[\varphi(X_t)/\mathcal{Y}_t]$ with φ varying in a large class of functions) but merely an approximation of $\hat{X}_t$. We will therefore try to find an approximate filter for $\hat{X}_t$. There is an extensive literature on the existence of approximate filters of finite dimension in the case of a high signal-to-noise ratio, and the list of articles referenced in this paper is far from exhaustive. We have mostly used the following papers by Picard: [7], [8], [9], [10].

An approximate filter is a $\mathcal{Y}_t$-measurable process here denoted by M_t, defined by a finite number of equations. We will use only first-order filters here, but it is possible to define second-order filters like the extended Kalman filter, which uses an approximation of the conditional variance $E[(X_t - \hat{X}_t)^2/\mathcal{Y}_t]$, or third-order filters [12]. For example, under sufficient assumptions, the process M_t defined below is an approximate filter of the process X_t defined in (1):

$$M_t = m_0 + \int_0^t f(M_s)ds + \int_0^t \frac{1}{\varepsilon}(h')^{-1}(h'gg^*h'^*)^{\frac{1}{2}}(M_s)(dY_s - h(M_s)ds), \tag{3}$$

and we get the following result:

$$X_t - M_t = O(\sqrt{\varepsilon}). \tag{4}$$

This estimation is easily established (we will give the proof for a particular case in the proof of Lemma 1.3), and we deduce immediately that

$$\hat{X}_t - M_t = O(\sqrt{\varepsilon}), \tag{5}$$

which can be improved to get

$$\hat{X}_t - M_t = O(\varepsilon). \tag{6}$$

This new estimation is much more difficult to establish than the previous one: it will require techniques of time reversal of diffusion processes (Picard [7]), techniques of derivation with respect to the initial condition or derivation in the Wiener space (Picard [9]), or fine techniques of PDEs (Bensoussan [1]). (In some particular cases, such as the semilinear case, or for higher-order filters, we get estimations of order $\varepsilon^{\frac{3}{2}}$ or even of order ε^2 (Picard [7], [10])). For regular φ functions we can then, using (2), estimate $E[\varphi(X_t)/\mathcal{Y}_t]$ by $\varphi(M_t)$ and the error is of order ε.

Case where h is not one to one. The function h can be locally, but not globally, one to one. In this work we will consider the case where h is not one to one because it observes only some of the components of X. As expected, the components of X which are nonobserved are not $\mathcal{Y}_t$-measurable when $\varepsilon = 0$. We distinguish two different cases.

1) The case where the conditional variance of any component of X converges to 0 with ε. Therefore, there exist approximate filters for X. The process X is said to be observable.

2) The case where X is not perfectly observed when $\varepsilon = 0$. In this case, we can construct approximate filters for the observed components of X, but no approximate filter can exist for the other components.

Zeitouni and Dembo [17] give cases of observability of the system. Picard [11] shows that the detectability of the system is a sufficient criterion for the conditional variance to converge to 0 with ε. Except for the linear detectable case where the terms of the conditional variance matrix can be estimated (the variance satisfies a Riccati equation; see [6]), it seems that the only method to prove the convergence of the variance to 0 with ε is to construct an approximate filter for X. In [15] using a formal asymptotic development for the conditional law, Yaesh, Bobrovsky, and Schuss give detectability criteria when the dimension of X is 2 and when Y, whose dimension is 1, observes only a component of X. Approximate filters are given for the observed part of X.

In the case 2) (we can suppose without restriction that X^1 is the observed part of X, and X^2 the nonobserved one), Takeuchi and Akashi [14], using a theorem of martingale convergence, prove that $E[\varphi(X_t)/\mathcal{Y}_t]$ converges in probability to $E[\varphi(X_t)/\mathcal{X}_t^1]$ ($\mathcal{X}_t^1 = \sigma(X_s^1, 0 \le s \le t)$) when ε converges to 0. Sachs [13] obtains the same result in the linear case. In particular, $E[\varphi(X_t^2)/\mathcal{Y}_t]$ converges to $E[\varphi(X_t^2)/\mathcal{X}_t^1]$. When $\varepsilon = 0$, X^1 is perfectly observed and the conditional law of X^2 given $\mathcal{X}_t^1$ satisfies the associated Zakai equation. This equation is a parabolic PDE whose partial variable dimension is the dimension of X^2. As seen above, approximate filters of finite dimension do not exist for X^2, but we can expect that there exists an approximate filter which satisfies a Zakai-type equation whose spatial variable has dimension smaller than that of the exact Zakai equation satisfied by $E[\varphi(X_t)/\mathcal{Y}_t]$. The advantage of this filter is that the numerical computations of the Zakai equation are easier.

In this work, we give such an approximate filter for X^2 in the case $\dim(X^1) = \dim(X^2) = \dim(Y) = 1$. To define it, we use an approximate filter of finite dimension for X^1. The idea is to replace X^1 by M in the Zakai equation satisfied by $E[\varphi(X_t^2)/\mathcal{X}_t^1]$ (which is equal to $E[\varphi(X_t^2)/\mathcal{Y}_t]$ when $\varepsilon = 0$). The method shows that the rate of convergence is of order ε.

This paper is organized as follows: in section 1, we give the assumptions and we define the approximate filter for X^1 as well as the approximate filter for X^2 and the Zakai-type equation it satisfies. In section 2, we give an expression for the difference between the exact and the approximate filters as a function of $(X^1 - M)$. In section 3, we recall results about the rate of convergence of M. We deduce the order of the rate of convergence of our problem.

1. Approximate filters definitions.

1.1. Assumptions. Let $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\ge 0}, P)$ be a filtered probability space. All the filtrations we are using here are completed and right-continuous. If $(X_t)_{t\ge 0}$ is a

process, we call $\mathcal{X}_t$ its natural completed filtration $\mathcal{X}_t = \sigma(X_s, 0 \le s \le t)$.

We consider the filtering problem:

$$\begin{cases} X_t^1 = X_0^1 + \displaystyle\int_0^t f_1(X_s^1, X_s^2)ds + V_t^1, \\ X_t^2 = X_0^2 + \displaystyle\int_0^t f_2(X_s^1, X_s^2)ds + V_t^2, \\ Y_t = \displaystyle\int_0^t h(X_s^1)ds + \varepsilon W_t, \end{cases} \tag{7}$$

where X^1, X^2, and Y are scalar. We note that the observation-function h only depends on X^1.

Assumptions.

H.1. V_t^1, V_t^2, and W_t are scalar-independent Wiener processes.

H.2. X_0^1 is deterministic.

H.3. X_0^2 is independent of V_t^1, V_t^2, and W_t.

H.4. f_1 and f_2 are bounded and have derivatives of order 3 which are bounded.

H.5. Let $g(x^1, x^2) = \int_0^{x^1} f_1(u, x^2)du$. We suppose that g and its derivatives are bounded. (This is true if, for example, the support of f_1 relative to x^1 is compact.)

H.6. h is C_b^2 and there exists $0 < \alpha < \delta$ such that $\forall x \in \mathbb{R},\ 0 < \alpha < h'(x) < \delta$.

1.2. An approximate filter for X^1. We define a new process $(M_t, 0 \le t < \infty)$ which is an approximate filter for X_t^1 (cf. Lemma 1.3):

$$\begin{cases} M_0 = X_0^1, \\ dM_t = \dfrac{1}{\varepsilon}(dY_t - h(M_t)dt). \end{cases} \tag{8}$$

REMARK 1.1. *The filtrations $\mathcal{Y}_t$ and $\mathcal{M}_t$ are equal $\forall t \ge 0$. We will use the two notations interchangeably.*

We will see that $(X^1 - M)$ is of order $\sqrt{\varepsilon}$, but only on each time interval $[t_0, \infty]$. This is why we give the following definition.

DEFINITION 1.2. *Let Z_t be an adapted process defined on Ω. Let us suppose that Z depends on ε. We say that Z is of order ε^k (and note that $Z_t = O(\varepsilon^k)$) if there exist $\varepsilon_0 > 0$ and $r \in \mathbb{R}$ (r could be negative) such that $\forall t_0 > 0$ and $1 \le q < \infty$, $\exists C > 0$ such that for all $\varepsilon < \varepsilon_0$,*

$$\sup_{t_0 \le t < \infty} \|Z_t\|_q \le C\varepsilon^k \quad \text{and} \quad \sup_{0 \le t < \infty} \|Z_t\|_q \le C\varepsilon^r. \tag{9}$$

Let us denote by $\| \ \|_p$ the norm of $L^p(\Omega, \mathcal{F}, P)$. Likewise we will say that Z_t is of order ε^k under the probability $\tilde{P}$ (introduced at (10)) if (9) is true for $\| \ \|_p^{\sim}$, the norm of $L^p(\Omega, \mathcal{F}, \tilde{P})$.

We will see later that M is an approximate filter of order ε for X^1. For the moment, we only use the following result, whose proof is easier.

LEMMA 1.3. *Under the probability P we have the following two assertions:*

i) $(X_t^1 - M_t) = O(\sqrt{\varepsilon})$,

ii) $E[X_t^1/\mathcal{Y}_t] - M_t = O(\sqrt{\varepsilon})$.

Proof. The conditional expectation is an L^p-contraction, so i) implies ii). Using assumption H.6, i) is equivalent to $h(X_t^1) - h(M_t) = O(\sqrt{\varepsilon})$. Using Itô's formula, we obtain

$$d(h(X_t^1) - h(M_t)) = -\frac{h'(M_t)}{\varepsilon}(h(X_t^1) - h(M_t))dt + h'(X_t^1)dV_t^1 - h'(M_t)dW_t + \left(f_1(X_t^1, X_t^2)h'(X_t^1) + \frac{1}{2}(h''(X_t^1) - h''(M_t))\right)dt,$$

$$\begin{aligned} h(X_t^1) - h(M_t) = {} & \exp\left(-\frac{1}{\varepsilon}\int_0^t h'(M_s)ds\right)(h(X_0^1) - h(M_0)) \\ & + \int_0^t e^{-\frac{1}{\varepsilon}\int_s^t h'(M_r)dr}\left(f_1(X_s^1, X_s^2)h'(X_s^1) + \frac{1}{2}(h''(X_s^1) - h''(M_s))\right)ds \\ & + \int_0^t e^{-\frac{1}{\varepsilon}\int_s^t h'(M_r)dr}(h'(X_s^1)dV_s^1 - h'(M_s)dW_s). \end{aligned}$$

Again using H.6 ($h'(x) > \alpha$), the first term is of order $e^{-\frac{c}{\varepsilon}}$ for $t \geq t_0$, the second is of order ε, and the last is of order $\sqrt{\varepsilon}$. □

REMARK 1.4. *We will need later the fact that X_0^1 is deterministic. In the last equation, we see that we do not need the assumption $M_0 = X_0^1$ to establish Lemma* 1.3. *The error at time $t = 0$ disappears because of the exponential function. The filter therefore has a "short memory."*

1.3. Change of probability. Let $\Gamma_t = \exp(\int_0^t f_1(X_s^1, X_s^2)dX_s^1 - \frac{1}{2}\int_0^t (f_1(X_s^1, X_s^2))^2 ds)$. We can define a new probability $\tilde{P}$ on $\mathcal{F}_t$ by

$$\left.\frac{d\tilde{P}}{dP}\right|_{\mathcal{F}_t} = \Gamma_t^{-1}. \tag{10}$$

Under $\tilde{P}$, X_t^1 is a Wiener process which does not depend on $\mathcal{V}^2 \vee \mathcal{W}$.

If we define

$$g(x^1, x^2) = \int_0^{x_1} f_1(u, x^2)du \tag{11}$$

and apply Itô's formula to the process $g(X_t^1, X_t^2)$, Γ_t can be rewritten in the following way:

$$\begin{aligned} \Gamma_t = \exp\Bigg(& g(X_t^1, X_t^2) - g(X_0^1, X_0^2) - \int_0^t [\mathcal{L}_{(X_s^1)}g(X_s^1, .)](X_s^2)ds \\ & - \frac{1}{2}\int_0^t \frac{\partial f_1}{\partial x_1}(X_s^1, X_s^2)ds - \int_0^t \frac{\partial g}{\partial x_2}(X_s^1, X_s^2)dV_s^2 \\ & - \frac{1}{2}\int_0^t (f_1(X_s^1, X_s^2))^2 ds\Bigg), \end{aligned} \tag{12}$$

with

$$\mathcal{L}_{(y)}\varphi(x) = \frac{1}{2}\varphi''(x) + \varphi'(x)f_2(y, x). \tag{13}$$

Under this new expression of Γ we see that we can "freeze" X^1 in the Γ_t definition. For a function $x \in \mathcal{C}([0,t])$, let us define

$$X_t^2(x) \triangleq X_0^2 + \int_0^t f_2(x_s, X_s^2(x))ds + V_t^2,$$

$$\begin{aligned}\Gamma_t(x) = \exp\Bigg(& g(x_t, X_t^2(x)) - g(x_0, X_0^2) - \int_0^t \mathcal{L}_{(x_s)} g(x_s, .)(X_s^2(x))ds \\ & - \frac{1}{2}\int_0^t \frac{\partial f_1}{\partial x_1}(x_s, X_s^2(x))ds - \int_0^t \frac{\partial g}{\partial x_2}(x_s, X_s^2(x))dV_s^2 \\ & - \frac{1}{2}\int_0^t (f_1(x_s, X_s^2(x)))^2 ds\Bigg). \end{aligned} \tag{14}$$

The assumptions on f_1 will make Γ_t bounded in the L^p norm. We deduce that the order of a process (in the sense of Definition 1.2) is the same under the two probabilities P and $\tilde{P}$.

LEMMA 1.5. *Under the assumptions of the Introduction, for a fixed $T > 0$, $\forall p \in \mathbb{R}$, there exists $c(p)$ such that $\forall t \leq T$ and $\forall x \in \mathcal{C}([0,T])$,*

$$E[(\Gamma_t(x))^p] \leq c(p). \tag{15}$$

Proof. We use H.4 and H.5 to show that the first exponential below is bounded for all $x \in \mathcal{C}([0,T])$:

$$\begin{aligned}(\Gamma_t(x))^p = \exp p\Bigg(& g(x_t, X_t^2(x)) - g(x_0, X_0^2) - \int_0^t \mathcal{L}_{(x_s)} g(x_s, .)(X_s^2(x))ds \\ & - \frac{1}{2}\int_0^t \frac{\partial f_1}{\partial x_1}(x_s, X_s^2(x))ds - \frac{1}{2}\int_0^t (f_1(x_s, X_s^2(x)))^2 ds \\ & + \frac{p}{2}\int_0^t \left(\frac{\partial g}{\partial x_2}(x_s, X_s^2(x))\right)^2 ds\Bigg) \\ & \times \exp\left(-p\int_0^t \frac{\partial g}{\partial x_2}(x_s, X_s^2(x))dV_s^2 - \frac{p^2}{2}\int_0^t \left(\frac{\partial g}{\partial x_2}(x_s, X_s^2(x))\right)^2 ds\right) \\ \leq c(p,T)\exp & \left(-p\int_0^t \frac{\partial g}{\partial x_2}(x_s, X_s^2(x))dV_s^2 - \frac{p^2}{2}\int_0^t \left(\frac{\partial g}{\partial x_2}(x_s, X_s^2(x))\right)^2 ds\right). \end{aligned} \tag{16}$$

$c(p,T)$ does not depend on the function x. The last term is an exponential martingale whose expectation is equal to 1. ∎

LEMMA 1.6. *Let $(Z_t)_{t\leq 0}$ be a process defined on Ω, $T > 0$. The following two properties are equivalent:*

i) $\forall t \leq T, \quad Z_t = O(\varepsilon^k)$ *under* P,

ii) $\forall t \leq T, \quad Z_t = O(\varepsilon^k)$ *under* $\tilde{P}$.

Proof of Lemma 1.6. We use Lemma 1.5:

$$\begin{aligned} E[(Z_t)^p] &= \tilde{E}[(Z_t)^p \Gamma_t] \\ &\le \sqrt{\tilde{E}[(Z_t)^{2p}]}\sqrt{\tilde{E}[(\Gamma_t)^2]} \\ &\le \sqrt{\tilde{E}[(Z_t)^{2p}]}\sqrt{E[(\Gamma_t)]} \\ &\le c(T)\varepsilon^{kp} \end{aligned}$$

for t large enough. □

REMARK 1.7. *The results of Lemma* 1.3 *are true under the probability* $\tilde{P}$. *From now on, when we say that a process is of order* ε^k, *we mean under both* P *and* $\tilde{P}$.

1.4. An approximate filter for X^2.

1.4.1. Definition. Let us consider the filtering problem (7) at the limit case $\varepsilon = 0$. Because h is one to one relative to X^1, X^1 is perfectly observed and (7) is reduced to the filtering problem when X^1 is the observation and X^2 the nonobserved process. The last change of probability now becomes natural because we need to work under the probability which makes the observation (X^1 if $\varepsilon = 0$) a Brownian motion.

Under the classical filtering results (see [16] and [6]), for a function $\varphi \in C_b^2(\mathbb{R})$, the process $\tilde{E}[\varphi(X_t^2)\Gamma_t/\mathcal{X}_t^1]$ satisfies the well-known Zakai equation

(17)
$$\begin{aligned} &\tilde{E}[\varphi(X_t^2)\Gamma_t/\mathcal{X}_t^1] \\ &\quad = E[\varphi(X_0^2)] + \int_0^t \tilde{E}[\mathcal{L}_{(X_s^1)}\varphi(X_s^2)\Gamma_s/\mathcal{X}_s^1]ds + \int_0^t \tilde{E}[(\varphi(X_s^2)f_1(X_s^1, X_s^2))\Gamma_s/\mathcal{X}_s^1]dX_s^1. \end{aligned}$$

To give a heuristic explanation about the method we use to construct the approximate filter $\mu_t(\varphi)$ for $E[\varphi(X_t^2)/\mathcal{Y}_t]$, we can say that, when ε is small, by the results [13], [14] explained in the Introduction, we can expect $\tilde{E}[\varphi(X_t^2)\Gamma_t/\mathcal{X}_t^1]$ and an unnormalized version of $E[\varphi(X_t^2)/\mathcal{Y}_t]$ to be close. This leads us to replace X^1 by M in $\tilde{E}[\varphi(X_t^2)\Gamma_t/\mathcal{X}_t^1]$ in order to define the filter. Because of a sort of continuity of $\tilde{E}[\varphi(X_t^2)\Gamma_t/\mathcal{X}_t^1]$ relative to X^1 the process obtained will converge to the unnormalized version of $E[\varphi(X_t^2)/\mathcal{Y}_t]$.

Let us define

$$\breve{X}_t^2 \triangleq X^2(M), \tag{18}$$

$$\breve{\Gamma}_t \triangleq \exp\left(\int_0^t f_1(M_s, \breve{X}_s^2)dM_s - \frac{1}{2}\int_0^t (f_1(M_s, \breve{X}_s^2))^2 ds\right). \tag{19}$$

Applying Itô's formula to $g(M_t, \breve{X}_t^2)$ as in (12), we see from (14) that $\breve{\Gamma}_t = \Gamma_t(M_t)$.

Let φ be a function from $\mathbb{R}$ to $\mathbb{R}$ and

$$\mu_t(\varphi) \triangleq \frac{\tilde{E}[\varphi(\breve{X}_t^2)\breve{\Gamma}_t/\mathcal{M}_t]}{\tilde{E}[\breve{\Gamma}_t/\mathcal{M}_t]}. \tag{20}$$

We will show that $\mu_t(\varphi)$ is an approximate filter for $E[\varphi(X_t^2)/\mathcal{Y}_t]$.

1.4.2. Equation satisfied by the filter μ_t. When φ is twice-differentiable, the process $\mu_t(\varphi)$ satisfies the Kushner–Stratonovitch-type equation, as in the following proposition.

PROPOSITION 1.8. *For $\varphi \in \mathcal{C}_b^2(\mathbb{R})$,*

$$\mu_t(\varphi) = E[X_0^2] + \int_0^t \mu_s(\mathcal{L}_{(M_s)}\varphi)ds + \int_0^t [\mu_s(f_1(M_s,.)\varphi) - \mu_s(f_1(M_s,.))\mu_s(\varphi)][dM_s - \mu_s(f_1(M_s,.))ds], \tag{21}$$

where $\mathcal{L}$ is as defined in (13).

REMARK 1.9. *As mentioned in the Introduction, this equation no longer depends on X^1, although the Kushner–Stratonovitch equation for the exact filter $E[\varphi(X_t^2)/\mathcal{Y}_t]$ is coupled with $E[\varphi(X_t^1)/\mathcal{Y}_t]$.*

Proof. We first derive the associated Zakai-type equation using the two following lemmas.

LEMMA 1.10. *Let $\sigma_t(\varphi) = \tilde{E}[\varphi(\breve{X}_t^2)\breve{\Gamma}_t/\mathcal{M}_t]\ \forall \varphi \in \mathcal{C}_b^2(\mathbb{R})$. Then*

$$\sigma_t(\varphi) = \sigma_0(\varphi) + \int_0^t \sigma_s(\mathcal{L}_{(M_s)}\varphi)ds + \int_O^t \sigma_s(f_1(M_s,.)\varphi)dM_s. \tag{22}$$

REMARK 1.11. *Equation* (22) *is identical to* (17)*, with X^1 replaced by M.*

Proof of Lemma 1.10. As for the proof of the exact Zakai equation (see Pardoux [6, Theorem 2.3.3]), by Itô's formula we have

$$\breve{\Gamma}_t\varphi(\breve{X}_t^2) = \varphi(X_0^2) + \int_0^t \mathcal{L}_{(M_s)}\varphi(\breve{X}_s^2)\breve{\Gamma}_s ds + \int_0^t \breve{\Gamma}_s\varphi'(\breve{X}_s^2)dV_s^2 + \int_0^t \varphi(\breve{X}_s^2)\breve{\Gamma}_s f_1(M_s,\breve{X}_s^2)dM_s.$$

We take the conditional expectation $\tilde{E}[\dots/\mathcal{Y}_t]$ of this equation. We use the result (see, for example, [6, Lemma 2.2.4]) to pass the conditional expectation through the integral. We obtain

—$\tilde{E}[\int_0^t \varphi(\breve{X}_s^2)\breve{\Gamma}_s f_1(M_s,\breve{X}_s^2)dM_s/\mathcal{Y}_t] = \int_0^t \tilde{E}[\varphi(\breve{X}_s^2)\breve{\Gamma}_s f_1(M_S,\breve{X}_s^2)/\mathcal{Y}_t]dM_s$ because of Remark 1.1;

—$\tilde{E}[\int_0^t \varphi'(\breve{X}_s^2)\breve{\Gamma}_s dV_s^2/\mathcal{Y}_t] = 0$, using the independence of V^2 and Y_t under $\tilde{P}$;

—$\tilde{E}[\int_0^t \mathcal{L}_{(M_s)}\varphi\breve{\Gamma}_s ds/\mathcal{Y}_t] = \int_0^t \tilde{E}[\mathcal{L}_{(M_s)}\varphi\breve{\Gamma}_s/\mathcal{Y}_t]ds$.

The proof of Lemma 1.10 will be completed if we can replace $\mathcal{Y}_t$ and $\mathcal{Y}_s$ in the conditional expectations. Unlike the proof of the exact Zakai equation, we do not work here under the probability $\tilde{P}$ which turns Y_t into a Brownian motion. The filtration $\mathcal{Y}_t$ does not satisfy $\mathcal{Y}_t = \mathcal{Y}_s \vee \mathcal{Y}_s^t$, where $\mathcal{Y}_s$ and $\mathcal{Y}_s^t$ are independent ($\mathcal{Y}_s^t = \sigma\{Y_r - Y_s,\ s \le r \le t\}$). However, we have the following lemma.

LEMMA 1.12. *For any map $F : \mathbb{R}^2 \to \mathbb{R}$, we have*

$$\tilde{E}[F(M_s,\breve{X}_s^2)\breve{\Gamma}_s/\mathcal{Y}_t] = \tilde{E}[F(M_s,\breve{X}_s^2)\breve{\Gamma}_s/\mathcal{Y}_s].$$

Proof of Lemma 1.12. $F(M_t,\breve{X}_t^2)\breve{\Gamma}_t$ depends only on $\sigma(X_0^2, V_s^2\ s \le t)$ and $\mathcal{Y}_t$. Using the independence of these two filtrations under $\tilde{P}$, the properties of the conditional expectation allow us to write

$$\tilde{E}[F(M_s,\breve{X}_s^2)\breve{\Gamma}_s/\mathcal{Y}_t] = \tilde{E}[F(x_s, X_s^2(x))\Gamma_s(x)]|_{(x=M)} = \tilde{E}[F(M_s,\breve{X}_s^2)\breve{\Gamma}_s/\mathcal{Y}_s].$$

$\breve{\Gamma}_s = \Gamma_s(M)$ is justified by (14). □

To show Proposition 1.8, we only need to normalize the above process as is done for the exact Kushner–Stratonovitch equation (see [6, Theorem 2.3.7]), and we get (21). □

2. $\mu_t(\varphi) - E[\varphi(X_t^2)/\mathcal{Y}_t]$ as a function of $(X^1 - M)$. We need to show the convergence of $\mu_t(\varphi)$ to $E[\varphi(X_t^2)/\mathcal{Y}_t]$. We first prove that the processes $\mu_t(\varphi) - E[\varphi(X_t^2)/\mathcal{Y}_t]$ and $\tilde{E}[\varphi(X_t^2)\Gamma_t/\mathcal{Y}_t] - \tilde{E}[\varphi(\breve{X}_t^2)\breve{\Gamma}_t/\mathcal{Y}_t]$ have the same rate of convergence.

PROPOSITION 2.1. *Let $T < \infty$. Let us suppose that $\forall t \le T$ and for any function $\varphi \in \mathcal{C}_b^2(\mathbb{R})$, $\exists k$ such that the following assertion holds:*

$$\tilde{E}[\varphi(X_t^2)\Gamma_t/\mathcal{Y}_t] - \tilde{E}[\varphi(\breve{X}_t^2)\breve{\Gamma}_t/\mathcal{Y}_t] = O(\varepsilon^k).$$

Then, $\forall \varphi \in \mathcal{C}^2(\mathbb{R})$, we have

$$\mu_t(\varphi) - E[\varphi(X_t^2)/\mathcal{Y}_t] = O(\varepsilon^k).$$

Proof. We will use the Kallianpur–Striebel formula [6]:

$$\begin{aligned}
&E[|E[\varphi(X_t^2)/\mathcal{Y}_t] - \mu_t(\varphi)|^p] \\
&\quad = E\left[\left|\frac{\tilde{E}[\varphi(X_t^2)\Gamma_t/\mathcal{Y}_t]}{\tilde{E}[\Gamma_t/\mathcal{Y}_t]} - \frac{\tilde{E}[\varphi(\breve{X}_t^2)\breve{\Gamma}_t/\mathcal{Y}_t]}{\tilde{E}[\breve{\Gamma}_t/\mathcal{Y}_t]}\right|^p\right] \\
&\quad \le KE\left[\left|\frac{\tilde{E}[\varphi(X_t^2)\Gamma_t/\mathcal{Y}_t] - \tilde{E}[\varphi(\breve{X}_t^2)\breve{\Gamma}_t/\mathcal{Y}_t]}{\tilde{E}[\Gamma_t/\mathcal{Y}_t]}\right|^p\right] \\
&\qquad + KE\left[\left|\frac{\tilde{E}[\varphi(\breve{X}_t^2)\breve{\Gamma}_t/\mathcal{Y}_t]}{\tilde{E}[\breve{\Gamma}_t/\mathcal{Y}_t]}\frac{1}{\tilde{E}[\Gamma_t/\mathcal{Y}_t]}(\tilde{E}[\Gamma_t/\mathcal{Y}_t] - \tilde{E}[\breve{\Gamma}_t/\mathcal{Y}_t])\right|^p\right]. \qquad (23)
\end{aligned}$$

The first term of this expression is bounded by the Cauchy–Schwarz inequality:

$$\begin{aligned}
&E\left[\left|\frac{\tilde{E}[\varphi(X_t^2)\Gamma_t/\mathcal{Y}_t] - \tilde{E}[\varphi(\breve{X}_t^2)\breve{\Gamma}_t/\mathcal{Y}_t]}{\tilde{E}[\Gamma_t/\mathcal{Y}_t]}\right|^p\right] \\
&\quad \le \sqrt{E\left[\left(\frac{1}{\tilde{E}[\Gamma_t/\mathcal{Y}_t]}\right)^{2p}\right]}\sqrt{E[(\tilde{E}[\varphi(X_t^2)\Gamma_t/\mathcal{Y}_t] - \tilde{E}[\varphi(\breve{X}_t^2)\breve{\Gamma}_t/\mathcal{Y}_t])^{2p}]} \\
&\quad = O(\varepsilon^{kp}).
\end{aligned}$$

Indeed, the last equality is true because of the Jensen inequality applied to the convex function $x \to (\frac{1}{x})^{1-2p}$ on $\mathbb{R}^+$. We get

$$\begin{aligned}
E\left[\left(\frac{1}{\tilde{E}[\Gamma_t/\mathcal{Y}_t]}\right)^{2p}\right] &= \tilde{E}\left[\Gamma_t(\tilde{E}[\Gamma_t/\mathcal{Y}_t])^{-2p}\right] \\
&= \tilde{E}[(\tilde{E}[\Gamma_t/\mathcal{Y}_t])^{1-2p}] \\
&\le \tilde{E}[(\Gamma_t)^{1-2p}] \\
&= E[(\Gamma_t)^{-2p}],
\end{aligned}$$

and the last expression does not depend on ε.

Remarking that

$$\frac{\tilde{E}[\varphi(\breve{X}_t^2)\breve{\Gamma}_t/\mathcal{Y}_t]}{\tilde{E}[\breve{\Gamma}_t/\mathcal{Y}_t]} \leq \|\varphi\|_\infty,$$

we obtain in the same way the order of the second term of (23). The proof is now complete. □

In order to compute the rate of convergence of $\mu_t(\varphi)$ to $E[\varphi(X_t^2)/\mathcal{Y}_t]$, we will express $\tilde{E}[\varphi(X_t^2)\Gamma_t/\mathcal{Y}_t]-\tilde{E}[\varphi(\breve{X}_t^2)\breve{\Gamma}_t/\mathcal{Y}_t]$ as a function of (X^1-M). To give the general idea of the proof without using computations that are too big, let us first establish this expression for the process $\tilde{E}[X_t^2 - \breve{X}_t^2/\mathcal{Y}_t]$ (in this case the computations are easier). We will later give an expression for $\tilde{E}[\Gamma_t-\breve{\Gamma}_t/\mathcal{Y}_t]$, and finally for $\tilde{E}[\varphi(X_t^2)\Gamma_t-\varphi(\breve{X}_t^2)\breve{\Gamma}_t/\mathcal{Y}_t]$.

2.1. Order of $\tilde{E}[X_t^2 - \breve{X}_t^2/\mathcal{Y}_t]$.

LEMMA 2.2. *Let $T>0$. Then, $\forall 0 \leq t \leq T$, we have the estimation*

$$X_t^2 - \breve{X}_t^2 = O(\sqrt{\varepsilon}).$$

Proof. We use assumption H.4, Gronwall's lemma, and Lemma 1.3. □

This first estimation shows that $\breve{X}_t^2$ converges to X_t^2 when ε converges to 0. But we will see that if we condition by $\mathcal{Y}_t$, the rate of convergence will be better.

Let us define the following directional derivative by

$$Z_t^2(x,y) = \lim_{\lambda\to 0} \frac{X_t^2(x+\lambda y) - X_t^2(x)}{\lambda}. \tag{24}$$

We have an exact formula for $Z_t^2(x,y)$, given in Lemma 2.3.

LEMMA 2.3.

$$Z_t^2(x,y) = \int_0^t \exp\left(\int_s^t \frac{\partial f_2}{\partial x_2}(x_u, X_u^2(x))du\right) \frac{\partial f_2}{\partial x_1}(x_s, X_s^2(x))y_s ds. \tag{25}$$

Proof. As there is no λ in the stochastic integral, the computations are easy:

$$\begin{aligned} Z_t^2(x,y) &= \lim_{\lambda\to 0} \int_0^t \frac{f_2(x_s+\lambda y_s, X_s^2(x+\lambda y)) - f_2(x_s, X_s^2(x))}{\lambda} ds \\ &= \int_0^t \frac{\partial f_2}{\partial x_1}(x_s, X_s^2(x))y_s + \frac{\partial f_2}{\partial x_2}(x_s, X_s^2(x)) Z_s^2(x,y) ds. \end{aligned}$$

We have

$$\frac{d}{dt} Z_t^2(x,y) = \frac{\partial f_2}{\partial x_1}(x_t, X_t^2(x))y_t + \frac{\partial f_2}{\partial x_2}(x_t, X_t^2(x)) Z_t^2(x,y),$$

and the result follows. □

Using Lemma 1.3 and the exact formula for Z_t^2, we can deduce the following corollary.

COROLLARY 2.4. *Let $T>0$. Then, $\forall t \leq T$ and $\lambda \in [0,1]$, we have*

$$Z_t^2(M + \lambda(X^1 - M), X^1 - M) = O(\sqrt{\varepsilon}).$$

To make the computations simpler, let us note that

$$L_{(s,t)}(x) = \exp\left(\int_s^t \frac{\partial f_2}{\partial x_2}(x_u, X_u^2(x))du\right). \tag{26}$$

For $x,\ y \in \mathcal{C}([0,T],\ \mathbb{R})$ and $\lambda \in [0,1]$, let us note that $\rho_t(\lambda) = X_t^2(y + \lambda(x-y))$; the difference $X_t^2(x) - X_t^2(y)$ can be written as

$$\begin{aligned} X_t^2(x) - X_t^2(y) &= X_t^2(y + (x-y)) - X_t^2(y) \\ &= \rho_t(1) - \rho_t(0) \\ &= \int_0^1 \rho_t'(\lambda)d\lambda \\ &= \int_0^1 Z_t^2(y + \lambda(x-y), (x-y))d\lambda \end{aligned}$$

because $\rho_t'(\lambda) = Z_t^2(y + \lambda(x-y), x-y)$.

LEMMA 2.5. *Let $x,\ y \in \mathcal{C}([0,T],\ \mathbb{R})$. We have*

$$|Z_t^2(x + \lambda y, y) - Z_t^2(x,y)| \le c(T)\left(\int_0^t (y_s)^2 ds + \left(\int_0^t |y_s|ds\right)^2\right).$$

Proof. Let $x,\ y$ be two functions of $\mathcal{C}([0,T],\ \mathbb{R})$ and let us denote $\gamma_t(l) \triangleq Z_t^2(x + ly, y)$. Rolle's theorem (or the Taylor–Lagrange theorem) gives

$$Z_t^2(x + \lambda y, y) - Z_t^2(x,y) = \gamma_t(\lambda) - \gamma_t(0) = \lambda\, \gamma_t'(l)$$

with $l \in [0, \lambda]$. We have

$$\begin{aligned} |\gamma_t'(l)| = \Bigg| &\int_0^t \left(\int_s^t \left[\frac{\partial^2 f_2}{\partial x_1 \partial x_2}(x_u + ly_u, X_u^2(x+ly))y_u + \frac{\partial^2 f_2}{\partial^2 x_2}(\ldots)Z_u^2(x+ly, y)\right]du\right) \\ &\times L_{(s,t)}(x+ly)\frac{\partial f_2}{\partial x_1}(x_s + ly_s, X_s^2(x+ly))y_s ds \\ &+ \int_0^t L_{(s,t)}(x+ly)\left(\frac{\partial^2 f_2}{\partial x_1^2}(\ldots)y_s + \frac{\partial^2 f_2}{\partial x_1 \partial x_2}(\ldots)Z_s^2(x+ly, y)\right)y_s ds\Bigg| \\ \le c(T)&\left(\int_0^t (y_s)^2 ds + \left(\int_0^t |y_s|ds\right)^2\right). \qquad \square \end{aligned} \tag{27}$$

Lemmas 1.3 and 2.5 imply the following result.

LEMMA 2.6. *Let $T > 0$. Then, $\forall t \le T$, we have*

$$Z_t^2(M + \lambda(X^1 - M), X^1 - M) = Z_t^2(M, X^1 - M) + O(\varepsilon).$$

PROPOSITION 2.7. *Let $T > 0$. $\forall t \le T$,*

$$\begin{aligned} &\tilde{E}[X_t^2 - \breve{X}_t^2/\mathcal{Y}_t] \\ &\quad = \tilde{E}\left[\int_0^t L_{(s,t)}(M)\frac{\partial f_2}{\partial x_1}(M_s, X_s^2(M))(X_s^1 - M_s)ds/\mathcal{Y}_t\right] + O(\varepsilon). \end{aligned} \tag{28}$$

Proof.

$$\begin{aligned}
&\tilde{E}[X_t^2 - \breve{X}_t^2/\mathcal{Y}_t] \\
&\quad = \tilde{E}\left[\int_0^1 Z_t^2(M + \lambda(X^1 - M), X^1 - M)d\lambda/\mathcal{Y}_t\right] \\
&\quad = \tilde{E}[Z_t^2(M, X^1 - M)/\mathcal{Y}_t] + O(\varepsilon) \\
&\quad = \tilde{E}\left[\int_0^t L_{(s,t)}(M)\frac{\partial f_2}{\partial x_1}(M_s, X_s^2(M))(X_s^1 - M_s)ds/\mathcal{Y}_t\right] + O(\varepsilon).
\end{aligned}$$

Using Lemma 1.3, we see that $\tilde{E}[X_t^2 - \breve{X}_t^2/\mathcal{Y}_t]$ is of order $\sqrt{\varepsilon}$. We will later improve this estimation and see that the rate of convergence is of order ε. □

2.2. Order of $\tilde{E}[\Gamma_t - \breve{\Gamma}_t/\mathcal{Y}_t]$. As seen above, $\Gamma_t - \breve{\Gamma}_t = \Gamma_t(X^1) - \breve{\Gamma}_t(M)$. $\Gamma_t(x)$ depends on x directly and via X_t^2 (see (14)). One can, however, write

$$\begin{aligned}
\Gamma_t(x) - \Gamma_t(y) &= \Gamma_t(y + (x - y)) - \Gamma_t(y) \\
&= \sigma_t(1) - \sigma_t(0) \\
&= \int_0^1 \sigma_t'(\lambda)d\lambda \qquad (29)
\end{aligned}$$

with

$$\sigma_t(\lambda) = \Gamma_t(y + \lambda(x - y)). \qquad (30)$$

We need to differentiate $\Gamma_t(y + \lambda(x - y))$ with respect to λ. We have already noticed the differentiability of X_t^2. The problem will come from the stochastic integral, but using the Kunita method we can show the following result.

PROPOSITION 2.8. *For $\lambda \in]-1, 1[$, the derivative of the stochastic integral*

$$\int_0^t \frac{\partial g}{\partial x_2}(x_s^1 + \lambda y_s, X_s^2(x + \lambda y))dV_s^2$$

with respect to λ exists and is given by

$$\frac{\partial}{\partial\lambda}\left(\int_0^t \frac{\partial g}{\partial x_2}(x_s^1 + \lambda y_s, X_s^2(x + \lambda y))dV_s^2\right) = \int_0^t \frac{\partial}{\partial\lambda}\left(\frac{\partial g}{\partial x_2}(x_s^1 + \lambda y_s, X_s^2(x + \lambda y))\right)dV_s^2.$$

This result is proposed in an exercise in Kunita ([3, exercise 3.1.5]) and is proved in our specific case in Gégout-Petit [2, Part I, Appendix]. We don't give the proof here.

Then

$$\sigma_t'(\lambda) = \Gamma_t(y + \lambda(x - y))\frac{\partial}{\partial\lambda}\log(\Gamma_t(y + \lambda(x - y))). \qquad (31)$$

Write

$$\begin{aligned}
\psi^1(x, y) &= -\left(\frac{1}{2}\frac{\partial^2 f_1}{\partial x_2^2} + \frac{\partial f_1}{\partial x_2}f_2 + \frac{\partial g}{\partial x_2}\frac{\partial f_2}{\partial x_1} + \frac{1}{2}\frac{\partial^2 f_1}{\partial x_1^2} + \frac{\partial f_1}{\partial x_1}f_1\right)(x, y), \\
\psi^2(x, y) &= -\left(\frac{1}{2}\frac{\partial^3 g}{\partial x_2^3} + \frac{\partial^2 g}{\partial x_2^2}f_2 + \frac{\partial g}{\partial x_2}\frac{\partial f_2}{\partial x_2} + \frac{1}{2}\frac{\partial^2 f_1}{\partial x_1\partial x_2} + \frac{\partial f_1}{\partial x_2}f_1\right)(x, y).
\end{aligned}$$

We obtain

$$\begin{aligned}
&\frac{\partial}{\partial\lambda}\Big(\log(\Gamma_t(y+\lambda(x-y)))\Big)\\
&\quad=-f_1(y_0+\lambda(x_0-y_0),X_0^2)(x_0-y_0)\\
&\qquad+f_1(y_t+\lambda(x_t-y_t),X_t^2(y+\lambda(x-y)))(x_t-y_t)\\
&\qquad+\frac{\partial g}{\partial x_2}(y_t+\lambda(x_t-y_t),X_t^2(y+\lambda(x-y)))\\
&\qquad\times Z_t^2(y+\lambda(x-y),x-y)\\
&\qquad+\int_0^t\psi_1(y_s+\lambda(x_s-y_s),X_s^2(y+\lambda(x-y)))(x_s-y_s)ds\\
&\qquad+\int_0^t\psi_2(y_s+\lambda(y_s-y_s),X_s^2(y+\lambda(x-y)))\\
&\qquad\times Z_s^2(y+\lambda(x-y),x-y)ds\\
&\qquad+\int_0^t\frac{\partial f_1}{\partial x_2}(y_s+\lambda(x_s-y_s),X_s^2(y+\lambda(x-y)))(x_s-y_s)dV_s^2\\
&\qquad-\int_0^t\frac{\partial^2 g}{\partial x_2^2}(y_s+\lambda(x_s-y_s),X_s^2(y+\lambda(x-y)))\\
&\qquad\times Z_s^2(y+\lambda(x-y),x-y)dV_s^2.
\end{aligned}$$

LEMMA 2.9. *Let us note that*

(32)

$$\begin{aligned}
&\Psi^{1,t,\lambda}(y,x-y)\\
&\quad=(f_1(y_t+\lambda(x_t-y_t),X_t^2(y+\lambda(x-y)))-f_1(y_t,X_t^2(y)))(x_t-y_t)\\
&\qquad+\frac{\partial g}{\partial x_2}(y_t+\lambda(x_t-y_t),X_t^2(y+\lambda(x-y)))Z_t^2(y+\lambda(x-y),x-y)\\
&\qquad-\frac{\partial g}{\partial x_2}(y_t,X_t^2(y))Z_t^2(y,x-y)\\
&\qquad+\int_0^t(\psi_1(y_s+\lambda(x_s-y_s),X_s^2(y+\lambda(x-y)))-\psi_1(y_s,X_s^2(y)))(x_s-y_s)ds\\
&\qquad+\int_0^t\psi_2(y_s+\lambda(x_s-y_s),X_s^2(y+\lambda(x-y)))Z_s^2(y+\lambda(x-y),x-y)\\
&\qquad-\psi_2(y_s,X_s^2(y))Z_s^2(y,x-y)ds\\
&\qquad+\int_0^t\left(\frac{\partial f_1}{\partial x_2}(y_s+\lambda(x_s-y_s),X_s^2(y+\lambda(x-y)))-\frac{\partial f_1}{\partial x_2}(y_s,X_s^2(y))\right)\\
&\qquad\times(x_s-y_s)dV_s^2\\
&\qquad-\int_0^t\left(\frac{\partial^2 g}{\partial x_2^2}(y_s+\lambda(x_s-y_s),X_s^2(y+\lambda(x-y)))Z_s^2(y+\lambda(x-y),x-y)\right.\\
&\qquad\qquad\left.-\frac{\partial^2 g}{\partial x_2^2}(y_s,X_s^2(y))Z_s^2(y,x-y)\right)dV_s^2.
\end{aligned}$$

For $\lambda \in [0,1]$ the following equality holds:

$$\begin{aligned}
&\frac{\partial}{\partial\lambda}\Big(\log(\Gamma_t(y+\lambda(x-y)))\Big)\\
&\quad= f_1(y_0+\lambda(x_0-y_0),X_0^2)(x_0-y_0)\\
&\qquad+ f_1(y_t,X_t^2(y))(x_t-y_t)+\frac{\partial g}{\partial x_2}(y_t,X_t^2(y))Z_t^2(y,x-y)\\
&\qquad+\int_0^t \psi_1(y_s,X_s^2(y))(x_s-y_s)ds+\int_0^t \psi_2(y_s,X_s^2(y))Z_s^2(y,x-y)ds\\
&\qquad+\int_0^t \frac{\partial f_1}{\partial x_2}(y_s,X_s^2(y))(x_s-y_s)dV_s^2\\
&\qquad-\int_0^t \frac{\partial^2 g}{\partial x_2^2}(y_s,X_s^2(y))Z_s^2(y,x-y)dV_s^2\\
&\qquad+\Psi^{1,t,\lambda}(y,x-y).
\end{aligned}$$

Thanks to Rolle's theorem, there exists

$$\Theta_y^{\lambda(x-y)} \in [\log(\Gamma_t(y+\lambda(x-y))),\log(\Gamma_t(y))] \tag{33}$$

such that

$$\Gamma_t(y+\lambda(x-y)) = \Gamma_t(y)+\exp(\Theta_y^{\lambda(x-y)})[\log(\Gamma_t(y+\lambda(x-y)))-\log(\Gamma_t(y))]. \tag{34}$$

REMARK 2.10. *Thanks to* (33) *and Lemma* 1.5 *we can claim that for a fixed $T>0$, $\forall p\in\mathbb{R}$, there exists $c(p)$ such that $\forall t\le T$ and $\forall x\in\mathcal{C}([0,T])$,*

$$E[(\exp(\Theta_y^{\lambda(x-y)}))^p] \le c(p). \tag{35}$$

We now use (29), (31), Lemma 2.9, and (34) in the case $x=X^1$ and $y=M$ to write $\Gamma_t-\breve{\Gamma}_t$ in the following way:

(36)

$$\begin{aligned}
&\Gamma_t-\breve{\Gamma}_t\\
&\quad=\Gamma_t(M+(X^1-M))-\Gamma_t(M)\\
&\quad=\int_0^1 \Gamma_t(M+\lambda(X^1-M))\frac{\partial}{\partial\lambda}\log(\Gamma_t(M+\lambda(X^1-M)))d\lambda\\
&\quad=-\Gamma_t(M+\lambda(X^1-M))f_1(M_0\lambda+(X_0^1-M_0),X_0^2)(X_0^1-M_0)\\
&\qquad+\Gamma_t(M)f_1(M_t,X_t^2(M))(X_t^1-M_t)\\
&\qquad+\Gamma_t(M)\frac{\partial g}{\partial x_2}(M_t,X_t^2(M))Z_t^2(M,(X^1-M))\\
&\qquad+\Gamma_t(M)\int_0^t \psi_1(M_s,X_s^2(M))(X_s^1-M_s)+\psi_2(M_s,X_s^2(M))Z_s^2(M,X^1-M)ds\\
&\qquad+\int_0^1 [\exp(\Theta_M^{\lambda(X^1-M)})[\log(\Gamma_t(M+\lambda(X^1-M)))-\log(\Gamma_t(M))]]\\
&\qquad\times\Big[f_1(M_t,X_t^2(M))(X_t^1-M_t)+\frac{\partial g}{\partial x_2}(M_t,X_t^2(M))Z_t^2(M,X^1-M)
\end{aligned}$$

$$
\begin{aligned}
&+ \int_0^t \psi_1(M_s, X_s^2(M))(X_s^1 - M_s) + \psi_2(M_s, X_s^2(M)) Z_s^2(M, X^1 - M) ds \\
&+ \int_0^t \left[\frac{\partial f_1}{\partial x_2}(M_s, X_s^2(M))(X_s^1 - M_s) \right. \\
&\qquad \left. - \frac{\partial^2 g}{\partial x_2^2}(M_s, X_s^2(M)) Z_s^2(M, X^1 - M) \right] dV_s^2 \Bigg] d\lambda \\
&+ \Gamma_t(M) \times \int_0^t \left[\frac{\partial f_1}{\partial x_2}(M_s, X_s^2(M))(X_s^1 - M_s) \right. \\
&\qquad \left. - \frac{\partial^2 g}{\partial x_2^2}(M_s, X_s^2(M)) Z_s^2(M, X^1 - M) \right] dV_s^2 \\
&+ \int_0^1 \Gamma_t(M + \lambda(X^1 - M)) \Psi_t^{1,t,\lambda}(X^1, X^1 - M) d\lambda.
\end{aligned}
$$

We now take the $\mathcal{Y}_t$-conditional expectation of this formula and then propose to show that all the terms of $\tilde{E}[\Gamma_t - \breve{\Gamma}_t / \mathcal{Y}_t]$ are small because $(X_t^1 - M_t)$ is small with ε (Lemma 1.3). Without assumption H.2 and equation (8), the first term need not be small. We group the terms of order ε, using the following estimations.

LEMMA 2.11. *Fix $T < \infty$. $\forall t < T$, $\forall \lambda \in [0,1]$, we have*

$$
\kappa_t^\lambda \triangleq \log(\Gamma_t(M + \lambda(X^1 - M))) - \log(\Gamma_t(M)) = O(\sqrt{\varepsilon}) \tag{37}
$$

and

$$
\Psi_t^{1,t,\lambda}(X^1, X^1 - M) = O(\varepsilon). \tag{38}
$$

Sketch of the proof. We use the following results:

—$X^1 - M = O(\sqrt{\varepsilon})$ (Lemma 1.3);

—$Z_t^2(M + \lambda(X^1 - M), X^1 - M) = O(\sqrt{\varepsilon})$ (Corollary 2.4);

—For all F from $\mathbb{R}^2$ into $\mathbb{R}$, bounded with bounded derivatives, $F(X_t^1, X_t^2) - F(M_t, \breve{X}_t^2) = O(\sqrt{\varepsilon})$.

All the terms of $\Psi_t^{1,t,\lambda}(X^1, X^1 - M)$ (see (32)) are products of two terms of order $\sqrt{\varepsilon}$, and by Lemma 2.6, $Z_t^2(M + \lambda(X^1 - M), X^1 - M) - Z_t^2(M, X^1 - M) = O(\varepsilon)$. □

Some terms of (36) are clearly of order ε. Indeed, if we use Jensen's and Cauchy–Schwarz's inequalities, we get

$$
\begin{aligned}
&\tilde{E}[|\tilde{E}[\exp(\Theta_M^{\lambda(X^1 - M)})(\log(\Gamma_t(M + \lambda(X^1 - M))) - \log(\Gamma_t(M))) \\
&\qquad \times f_1(M_t, X_t^2(M))(X_t^1 - M_t)/\mathcal{Y}_t]|^p] \\
&\quad = \tilde{E}[|\tilde{E}[\exp(\Theta_M^{\lambda(X^1 - M)}) \kappa_t^\lambda f_1(M_t, X_t^2(M))(X_t^1 - M_t)/\mathcal{Y}_t]|^p] \\
&\quad \le (\tilde{E}[(\exp(\Theta_M^{\lambda(X^1 - M)}) \kappa_t^\lambda f_1(M_t, X_t^2(M)))^{2p}])^{\frac{1}{2}} (\tilde{E}[(X_t^1 - M_t)^{2p}])^{\frac{1}{2}}.
\end{aligned}
$$

Also, using Lemma 2.11 and Remark 2.10, we get

$$
\int_0^1 \tilde{E}[\exp(\Theta_M^{\lambda(X^1 - M)}) \kappa_t^\lambda f_1(M_t, X_t^2(M))(X_t^1 - M_t)/\mathcal{Y}_t] d\lambda = O(\varepsilon).
$$

By the same method we get

$$\begin{aligned}\int_0^1 \tilde{E}\Big[&\exp(\Theta_M^{\lambda(X^1-M)})\kappa_t^\lambda \times \Big(\frac{\partial g}{\partial x_2}(M_t, X_t^2(M))Z_t^2(M, X^1-M)\\ &+ \int_0^t \psi_1(M_s, X_s^2(M))(X_s^1 - M_s) + \psi_2(M_s, X_s^2(M))Z_s^2(M, X^1-M)ds\\ &+ \int_0^t \frac{\partial f_1}{\partial x_2}(M_s, X_s^2(M))(X_s^1 - M_s)\\ &- \frac{\partial^2 g}{\partial x_2^2}(M_s, X_s^2(M))Z_s^2(M, X^1-M)dV_s^2\Big)/\mathcal{Y}_t\Big]d\lambda\\ &= O(\varepsilon)\end{aligned}$$

and

$$\tilde{E}\left[\int_0^1 \Gamma_t(M + \lambda(X^1 - M))\Psi_t^{1,t,\lambda}(X^1, X^1 - M)d\lambda/\mathcal{Y}_t\right] = O(\varepsilon).$$

PROPOSITION 2.12. *For a fixed $T > 0$ and $\forall t \le T$ we have*

$$\begin{aligned}&\tilde{E}[\Gamma_t - \breve{\Gamma}_t/\mathcal{Y}_t]\\ &\quad= \tilde{E}[\Gamma_t(M)f_1(M_t, X_t^2(M))(X_t^1 - M_t)|\mathcal{Y}_t]\\ &\quad+ \tilde{E}\left[\Gamma_t(M)\int_0^t \psi_1(M_s, X_s^2(M))(X_s^1 - M_s)ds/\mathcal{Y}_t\right]\\ &\quad+ \tilde{E}\left[\Gamma_t(M)\int_0^t \psi_2(M_s, X_s^2(M))\int_0^s L_{(u,t)}(M)\frac{\partial f_2}{\partial x_1}(M_u, X_u^2(M))(X_u^1 - M_u)duds/\mathcal{Y}_t\right]\\ &\quad+ \tilde{E}\left[\Gamma_t(M)\frac{\partial g}{\partial x_2}(M_t, X_t^2(M))\int_0^t L_{(s,t)}(M)\frac{\partial f_2}{\partial x_1}(M_s, X_s^2(M))(X_s^1 - M_s)ds/\mathcal{Y}_t\right]\\ &\quad+ \tilde{E}\Big[\Gamma_t(M)\int_0^t \frac{\partial f_1}{\partial x_2}(M_s, X_s^2(M))(X_s^1 - M_s)\\ &\qquad- \frac{\partial^2 g}{\partial x_2^2}(M_s, X_s^2(M))Z_s^2(M, X^1 - M)dV_s^2/\mathcal{Y}_t\Big]\\ &\quad+ O(\varepsilon).\end{aligned}$$

COROLLARY 2.13.

$$\tilde{E}[\Gamma_t - \breve{\Gamma}_t/\mathcal{Y}_t] = O(\sqrt{\varepsilon}).$$

2.3. Order of $\tilde{E}[\varphi(X_t^2)\Gamma_t - \varphi(\breve{X}_t^2)\breve{\Gamma}_t/\mathcal{Y}_t]$.

PROPOSITION 2.14. *Let $\varphi \in \mathcal{C}_b^2(\mathbb{R})$. For a fixed $T > 0$ and $\forall t \le T$, we have the following expression for the difference between the two filters:*

$$\begin{aligned}&\tilde{E}[\varphi(X_t^2)\Gamma_t - \varphi(\breve{X}_t^2)\breve{\Gamma}_t/\mathcal{Y}_t]\\ &\quad= \tilde{E}[\varphi(X_t^2(M))\Gamma_t(M)f_1(M_t, X_t^2(M))(X_t^1 - M_t)/\mathcal{Y}_t] \qquad (39)\\ &\quad+ \tilde{E}\left[\varphi(X_t^2(M))\Gamma_t(M)\int_0^t \psi_1(M_s, X_s^2(M))(X_s^1 - M_s)ds/\mathcal{Y}_t\right] \qquad (40)\end{aligned}$$

$$+\tilde{E}\left[\varphi(X_t^2(M))\Gamma_t(M)\int_0^t \psi_2(M_s,X_s^2(M))\right.$$

$$\left.\times\int_0^s L_{(u,s)}(M)\frac{\partial f_2}{\partial x_1}(M_u,X_u^2(M))(X_u^1-M_u)duds/\mathcal{Y}_t\right] \tag{41}$$

$$+\tilde{E}\left[\Gamma_t(M)\left(\varphi(X_t^2(M))\frac{\partial g}{\partial x_2}(M_t,X_t^2(M))+\varphi'(X_t^2(M))\right)\right.$$

$$\left.\times\int_0^t L_{(s,t)}(M)\frac{\partial f_2}{\partial x_1}(M_s,X_s^2(M))(X_s^1-M_s)ds/\mathcal{Y}_t\right] \tag{42}$$

$$+\tilde{E}\left[\Gamma_t(M)\varphi(X_t^2(M))\int_0^t\left[\frac{\partial f_1}{\partial x_2}(M_s,X_s^2(M))(X_s^1-M_s)\right.\right.$$

$$\left.\left.-\frac{\partial^2 g}{\partial x_2^2}(M_s,X_s^2(M))Z_s^2(M,X^1-M)\right]dV_s^2\right] \tag{43}$$

$$+O(\varepsilon).$$

Proof. By the method used in the previous section, we write

$$\Gamma_t(x)\varphi(X_t^2(x))-\Gamma_t(y)\varphi(X_t^2(y))=\nu_t(1)-\nu_t(0)$$
$$=\int_0^1\nu_t'(\lambda)d\lambda$$

with

$$\nu_t(\lambda)=\Gamma_t(y+\lambda(x-y))\varphi(X_t^2(y+\lambda(x-y))), \tag{44}$$

with σ defined by (30):

$$\nu_t'(\lambda)=\Gamma_t(y+\lambda(x-y))\varphi'(X_t^2(y+\lambda(x-y)))Z_t^2(y+\lambda(x-y),x-y)$$
$$+\sigma_t'(\ldots)\varphi(X_t^2(\ldots)).$$

If we proceed as in the previous sections, we get the required result. □

All the terms of the difference in Proposition 2.14 are clearly of order $\sqrt{\varepsilon}$. In order to give a finer estimation of the rate of convergence we now recall other results about filtering.

3. Rate of convergence of $\mu_t(\varphi)$ to $E[\varphi(X_t^2)/\mathcal{Y}_t]$.

3.1. M_t, approximate filter of order ε. If we study our system under the probability $\tilde{P}$, we have

$$\begin{cases}X_t^1,\\ Y_t=\displaystyle\int_0^t h(X_s^1)ds+\varepsilon W_t,\end{cases} \tag{45}$$

where X_t^1 is a Wiener process independent of W. The drift is therefore "erased" and X^1 no longer depends on X^2. So, h being one to one in X^1, we can apply Picard's results [9, Main Theorem of section 2] and we have the following proposition.

PROPOSITION 3.1. *In the sense of Definition* 1.2 *we have*

$$\tilde{E}[X_t^1/\mathcal{Y}_t]-M_t^1=O(\varepsilon). \tag{46}$$

REMARK 3.2. *The proof of this result is more difficult than the case* $\sqrt{\varepsilon}$. *As seen in the proof of Lemma* 1.3, *to show* $\sqrt{\varepsilon}$ *order we first show that* $(X^1 - M) = O(\sqrt{\varepsilon})$ *and, by contraction, the same estimation for* $\tilde{E}[X_t^1/\mathcal{Y}_t] - M_t^1$ *follows. By computing the conditional variance of* X^1 *(in the linear detectable case, the computations are explicit; for the nonlinear case, see Picard* [7]*), we notice that the order of convergence for* $X^1 - M$ *is exactly* $\sqrt{\varepsilon}$. *The conditional expectation improves the convergence, but makes the majoration more difficult. The method used in* [9] *uses techniques of differentiation on a Wiener space.*

As seen in Definition 1.2, the convergence is not good when t is near 0. Even in the case when X_0^1 is deterministic and $X_0^1 = M_0$, this problem remains. But as Picard remarked it [9] (see the discussion after the main theorem of section 2), the proof of his main theorem applied to the case $X_0^1 = M_0$ gives estimates of $\tilde{E}[X_t^1/\mathcal{Y}_t] - M_t$ for small t. Indeed, we have the following result.

PROPOSITION 3.3. *In the case when* $X_0^1 = M_0$ $\forall q \geq 1$, $\exists C > 0$ *such that*

$$\sup_{0\leq t\leq\sqrt{\varepsilon}} \|\tilde{E}[X_t^1/\mathcal{Y}_t] - M_t\|_q \leq C\sqrt{\varepsilon}, \qquad \sup_{\sqrt{\varepsilon}\leq t<\infty} \|\tilde{E}[X_t^1/\mathcal{Y}_t] - M_t\|_q \leq C\varepsilon. \tag{47}$$

3.2. Order of convergence of filter. We can now estimate all the terms obtained for $\tilde{E}[\varphi(X_t^2)\Gamma_t - \varphi(\check{X}_t^2)\check{\Gamma}_t/\mathcal{Y}_t]$ in Proposition 2.14. We will deal separately with (i) term (39), which is a special case, (ii) the Lebesgue integrals (40), (41), and (42), and (iii) the stochastic integrals (43).

We will use the results of the previous section. An important property used in the proofs below is the fact that $\mathcal{V}^2$ and $\mathcal{X}^1 \vee \mathcal{Y}$ are independent under $\tilde{P}$.

Estimation of term (39).

PROPOSITION 3.4. *In the sense of Definition* 1.2 *we have*

$$\tilde{E}[\varphi(X_t^2(M))\Gamma_t(M)f_1(M_t, X_t^2(M))(X_t^1 - M_t)/\mathcal{Y}_t] = O(\varepsilon).$$

Proof.

$$\begin{aligned}
&\tilde{E}[\varphi(X_t^2(M))\Gamma_t(M)f_1(M_t, X_t^2(M))(X_t^1 - M_t)/\mathcal{Y}_t] \\
&\quad = \tilde{E}[\tilde{E}[\varphi(X_t^2(M))\Gamma_t(M)f_1(M_t, X_t^2(M))/\mathcal{X}_t^1 \vee \mathcal{Y}_t](X_t^1 - M_t)/\mathcal{Y}_t] \\
&\quad = \Phi_t^1(M)\tilde{E}[(X_t^1 - M_t)|\mathcal{Y}_t].
\end{aligned}$$

The fact that $\mathcal{V}_t^2$ and $\mathcal{X}_t^1 \vee \mathcal{Y}_t$ are independent allows us to write

$$\tilde{E}[\varphi(X_t^2(M))\Gamma_t(M)f_1(M_t, X_t^2(M))/\mathcal{X}_t^1 \vee \mathcal{Y}_t] = \Phi_t^1(M), \tag{48}$$

$$\Phi_t^1(M) = \tilde{E}[\Gamma_t(x)f_1(x, X_t^2(x))]|_{x=M}. \tag{49}$$

Using Lemma 1.5 and the Cauchy–Schwarz inequality we get, $\forall x \in \mathcal{C}([0,t],\ \mathbb{R})$, $\forall p \geq 1$,

$$\begin{aligned}
\tilde{E}[|\Gamma_t(M)f_1(M, X_t^2(M))|^p] &\leq \sqrt{\tilde{E}[f_1^{2p}(M, X_t^2(M))]}\sqrt{\tilde{E}[(\Gamma_t(M))^{2p}]} \\
&\leq c'(p).
\end{aligned} \tag{50}$$

(50) and (47) allow us to conclude the proof. □

REMARK 3.5. *The law of* V^2 *being the same under* P *and under* $\tilde{P}$, *we can also take the expectation under* P *in* (49).

3.2.1. Estimation of the Lebesgue integrals.

Estimation of terms (40), (42). These two terms are a priori not the same because of the non-$\mathcal{F}_s$-measurability of $L_{(s,t)}(x)$ in (42). Using (26) we can note, however, that

$$L_{(s,t)}(x) = \exp\left(\int_s^t \frac{\partial f_2}{\partial x_2}(x_u, X_u^2(x))du\right) = L_{(0,t)}(x)(L_{(0,s)}(x))^{-1}.$$

The two terms can then be represented by

$$\tilde{E}[\Phi_t(M_\cdot, X_\cdot^2(M)) \int_0^t \Psi(M_s, X_s^2(M))(X_s^1 - M_s)ds/\mathcal{Y}_t],$$

where Φ_t looks like $\Gamma_t(M)$. In other words, Φ_t is a function of the paths of M and $X^2(M)$ from 0 to t, bounded either in L^∞ or in all L^p with $1 \le p < \infty$. The function Ψ is $\mathcal{F}_s$-measurable and bounded.

Let us first study the convergence of $\tilde{E}[X_s^1/\mathcal{Y}_t]$ to M_s. We have $(X_s^1 - M_s^1) = O(\sqrt{\varepsilon})$, which implies $\tilde{E}[X_s^1/\mathcal{Y}_t] - M_s^1 = O(\sqrt{\varepsilon})$. It seems that the order $\sqrt{\varepsilon}$ cannot be improved.[1]

Nevertheless, when we integrate $X^1 - M$ from 0 to t, we get a term of order ε. We obtain

$$\tilde{E}\left[\int_0^t (X_s^1 - M_s^1)ds/\mathcal{Y}_t\right] = O(\varepsilon).$$

We will not show this result but a stronger one because of Φ and Ψ, which appear in term 2.

PROPOSITION 3.6. *For all $0 \le t \le T$, we have*

$$\tilde{E}\left[\Phi_t(M_\cdot, X_\cdot^2(M)) \int_0^t \Psi(M_s, X_s^2(M))(X_s^1 - M_s)ds/\mathcal{Y}_t\right] = O(\varepsilon). \tag{51}$$

Proof.

$$\tilde{E}\left[\left|\tilde{E}\left[\Phi_t(M_\cdot, X_\cdot^2(M)) \int_0^t \Psi(M_s, X_s^2(M))(X_s^1 - M_s)ds/\mathcal{Y}_t\right]\right|^p\right]$$

$$\le \left(\tilde{E}\left[\left|\Phi_t(M_\cdot, X_\cdot^2(M))\right|^{2p}\right]\right)^{\frac{1}{2}} \left(\tilde{E}\left[\left|\int_0^t \Psi(M_s, X_s^2(M))(X_s^1 - M_s)ds\right|^{2p}\right]\right)^{\frac{1}{2}}. \tag{52}$$

The first square root is bounded.

Let us assume for a moment the following proposition.

PROPOSITION 3.7. *When $(F_s)_{0\le s\le t}$ is adapted to the filtration $\mathcal{Y}_s \vee \mathcal{V}_s^2$ and in L^p for $1 \le p < \infty$, then in the sense of Definition* 1.2 *we have*

$$\int_0^s F_s(X_s^1 - M_s)ds = O(\varepsilon).$$

[1] This conjecture is due to J. Picard (private communication) and is based on the following considerations (the first one is proved and the last two are heuristic).

1) The order of $X_s^1 - E[X_s^1/\mathcal{Y}_s]$ is exactly $O(\sqrt{\varepsilon})$ (Picard [7, Corollary 6.2]).

2) The information given by the future after s is not better than that given by the past, so as in 1) we have $E[X_s^1/\mathcal{Y}_s] - E[E[X_s^1/\mathcal{Y}_s]/\mathcal{Y}_s^t] = O(\sqrt{\varepsilon})$.

3) Since $\mathcal{Y}_s^t$ and $\mathcal{Y}_s$ are independent given Y_s and $\mathcal{Y}_t = \mathcal{Y}_s \vee \mathcal{Y}_s^t$, we can conjecture that $\tilde{E}[X_s^1/\mathcal{Y}_s]$ and $E[E[X_s^1/\mathcal{Y}_s]/\mathcal{Y}_s^t]$ are very closed and that $\tilde{E}[X_s^1/\mathcal{Y}_t] - \tilde{E}[X_s^1/\mathcal{Y}_s] = O(\sqrt{\varepsilon})\ \forall t > s$.

It follows that the second square root of (52) is of order ε. This shows Proposition 3.6. □

Proof of Proposition 3.7. We write the Taylor formula at point M_s for $h(X_s^1)$ and we get

$$h(X_s^1) - h(M_s) = (X_s^1 - M_s)h'(M_s) + \frac{1}{2}h''(\theta_s)(X_s^1 - M_s)^2 \quad \text{with } \theta_s \in [X_s^1, M_s],$$

$$(X_s^1 - M_s) = \frac{1}{h'(M_s)}(h(X_s^1) - h(M_s)) + \frac{h''(\theta_s)}{2h'(M_s)}(X_s^1 - M_s)^2. \tag{53}$$

By assumption H.6 and Lemma 1.3, we have

$$\frac{h''(\theta_s)}{2h'(M_s)}(X_s^1 - M_s)^2 = O(\varepsilon).$$

Then

$$(X_s^1 - M_s) - \frac{1}{h'(M_s)}(h(X_s^1) - h(M_s)) = O(\varepsilon). \tag{54}$$

The proof of Lemma 1.3 shows that the estimation is good for all $s \geq 0$. Taking the conditional expectation given $\mathcal{Y}_s$ under $\tilde{P}$ in equation (53), we also have

$$\tilde{E}[h(X_s^1)/\mathcal{Y}_s] - h(M_s) = h'(M_s)(\tilde{E}[X_s^1/\mathcal{Y}_s] - M_s) + O(\varepsilon).$$

By Proposition 3.3 and assumption H.6, we have $\tilde{E}[X_s^1/\mathcal{Y}_s] - M_s = O(\varepsilon)$, which implies

$$h(M_s) = \tilde{E}[h(X_s^1)/\mathcal{Y}_s] + O(\varepsilon). \tag{55}$$

We have seen (Proposition 3.3) that the expectation of $|h(M_s) - \tilde{E}[h(X_s^1)/\mathcal{Y}_s]|^p$ is not bounded by $C\varepsilon^p$ for $s \geq 0$ but only for $s \geq \sqrt{\varepsilon}$. For $s \leq \sqrt{\varepsilon}$, it is bounded by $C(\sqrt{\varepsilon})^p$. We integrate an error $\sqrt{\varepsilon}$ during a time $\sqrt{\varepsilon}$. We thus get an error of order ε. By (54) and (55) we have

$$\begin{aligned}
&\int_0^t F_s(X_s^1 - M_s)ds \\
&\quad = \int_0^t \frac{F_s}{h'(M_s)}[(h(X_s^1) - \tilde{E}[h(X_s^1)/\mathcal{Y}_s])ds + \varepsilon dW_s] - \underbrace{\int_0^t \frac{F_s}{h'(M_s)}\varepsilon dW_s}_{=O(\varepsilon)} + O(\varepsilon) \\
&\quad = \int_0^t \frac{F_s}{h'(M_s)}(dY_s - \tilde{E}[h(X_s^1)/\mathcal{Y}_s]ds) + O(\varepsilon). \qquad (56)
\end{aligned}$$

We have used the assumptions on F_s and the fact that

$$\frac{F_s}{h'(M_s)} \in L^p \quad \text{for } 1 \leq p < \infty.$$

In filtering theory, the process $I_t \overset{\Delta}{=} Y_t - \int_0^t \tilde{E}[h(X_s^1)/\mathcal{Y}_s]ds$ is called the innovation process, and under the probability P, $(\frac{I_t}{\varepsilon})_{0 \leq t \leq T}$ is a $\mathcal{Y}_t$-Wiener process (Pardoux [6]). Again using the fact that $\mathcal{Y} \vee \mathcal{X}^1$ and $\mathcal{V}^2$ are independent under $\tilde{P}$, we get without difficulty the following proposition.

PROPOSITION 3.8. *Under* $\tilde{P}$, $(\frac{I_t}{\varepsilon})_{0\le t\le T}$ *is a* $\mathcal{Y}_t \vee \mathcal{V}_t^2$ *Brownian motion.*
The process

$$\frac{F_s}{h'(M_s)}$$

being $\mathcal{Y}_s \vee \mathcal{V}_s^2$ measurable, then

$$\int_0^t \frac{F_s}{h'(M_s)} dI_s$$

is a stochastic integral. We can therefore use the Burkholder–Davis–Gundy inequalities:

$$\begin{aligned}\tilde{E}\left[\left|\int_0^t \frac{F_s}{h'(M_s)} dI_s\right|^p\right] &\le c(p)\tilde{E}\left[\left(\int_0^t \left(\frac{F_s}{h'(M_s)}\right)^2 \varepsilon^2 ds\right)^{\frac{p}{2}}\right] \\ &\le c'(p)\varepsilon^p.\end{aligned}$$

The result follows. □

REMARK 3.9. *In the study of* (39) *the conditioning by* $\mathcal{Y}_t$ *is indispensable in obtaining the order* ε. *For* (40) *and* (42), *it is not necessary.*

Estimation of term (41). Again using $L_{(u,s)} = L_{(0,s)}(L_{(u,s)})^{-1}$, we can write (41) as

$$\tilde{E}\left[\Phi_t(M_\cdot, X_\cdot^2(M)) \int_0^t \Psi(M_s, X_s^2(M)) \int_0^s \psi(M_u, X_u^2(M))(X_u^1 - M_u) du ds / \mathcal{Y}_t\right].$$

Φ is bounded in L^p, and Ψ and ψ are bounded. This term is also of order ε.

We use the same techniques as for the computation of terms (40) and (42). By Hölder's and Jensen's inequalities,

$$\begin{aligned}&\tilde{E}\left[\left|\Phi_t(M_\cdot, X_\cdot^2(M)) \int_0^t \Psi(M_s, X_s^2(M)) \int_0^s \psi(M_u, X_u^2(M))(X_u^1 - M_u) du ds\right|^p\right] \\ &\le c(p,t)\tilde{E}\left[\int_0^t |\Phi_t(M_\cdot, X_\cdot^2(M))\Psi(M_s, X_s^2(M)) \int_0^s \psi(M_u, X_u^2(M))(X_u^1 - M_u) du|^p ds\right] \\ &\le \left(\tilde{E}\left[\int_0^t |\Phi_t(M_\cdot, X_\cdot^2(M))\Psi(M_s, X_s^2(M))|^{2p} ds\right]\right)^{\frac{1}{2}} \\ &\qquad \times \left(\int_0^t \tilde{E}\left[\left|\int_0^s \psi(M_u, X_u^2(M))(X_u^1 - M_u) du\right|^{2p} ds\right]\right)^{\frac{1}{2}}.\end{aligned}$$

The first square root is bounded and, using Proposition 3.7, the second one is bounded by $C\varepsilon^p$.

3.2.2. Estimation of the stochastic integrals.
PROPOSITION 3.10.

$$\begin{aligned}&\tilde{E}\left[\Gamma_t(M) \int_0^t \frac{\partial f_1}{\partial x_2}(M_s, X_s^2(M))(X_s^1 - M_s) dV_s^2 / \mathcal{Y}_t\right] \\ &\quad - \tilde{E}\left[\Gamma_t(M) \int_0^t \frac{\partial^2 g}{\partial x_s^2}(M_s, X_s^2(M)) Z_s^2(M, X^1 - M) dV_s^2 / \mathcal{Y}_t\right] = O(\varepsilon).\end{aligned}$$

Proof. We use here the techniques of differentiation on the Wiener space (Malliavin calculus). Let us denote by D_s the operator of differentiation with respect to the perturbation of the Brownian motion, V_s^2. We can apply an integration-by-parts formula thanks again to the fact that $\mathcal{Y}_t$ and $\mathcal{V}_t^2$ are independent. (See Ocone [4] or Pardoux [6, chapter 5.])

$$\begin{aligned}&\tilde{E}\left[\Gamma_t(M)\int_0^t \frac{\partial f_1}{\partial x_2}(M_s, X_s^2(M))(X_s^1 - M_s)dV_s^2/\mathcal{Y}_t\right]\\ &\quad = \tilde{E}\left[\int_0^t D_s\Gamma_t(M)\frac{\partial f_1}{\partial x_2}(M_s, X_s^2(M))(X_s^1 - M_s)ds/\mathcal{Y}_t\right]. \qquad (57)\end{aligned}$$

Let us write the expressions for $D_sX_t^2$ and $D_s\Gamma_t(M)$. Using (18), we get

$$D_s\breve{X}_t^2 = \exp\left(\int_s^t \frac{\partial f_2}{\partial x_s}(M_u, \breve{X}_u^2)du\right). \qquad (58)$$

Using (19), we also have

$$\begin{aligned}D_s\Gamma_t(M) = \Gamma_t(M)(D_0\breve{X}_s^2)^{-1}&\left[\int_0^t \left(\frac{\partial f_1}{\partial x_2}(M_u, \breve{X}_u^2) - f_1\frac{\partial f_1}{\partial x_2}(M_u, \breve{X}_u^2)\right)D_0\breve{X}_u^2 dM_u\right.\\ &\left. - \int_0^s \left(\frac{\partial f_1}{\partial x_2}(M_u, \breve{X}_u^2) - f_1\frac{\partial f_1}{\partial x_2}(M_u, \breve{X}_u^2)\right)D_0\breve{X}_u^2 dM_u\right]. \qquad (59)\end{aligned}$$

(57) now becomes

(60)

$$\begin{aligned}&\tilde{E}\left[\Gamma_t(M)\int_0^t \frac{\partial f_1}{\partial x_2}(M_s, X_s^2(M))(X_s^1 - M_s)dV_s^2/\mathcal{Y}_t\right]\\ &= \tilde{E}\left[\Gamma_t(M)\left(\int_0^t \left(\frac{\partial f_1}{\partial x_2}(M_u, \breve{X}_u^2) - f_1\frac{\partial f_1}{\partial x_2}(M_u, \breve{X}_u^2)\right)D_0\breve{X}_u^2 dM_u\right)\right.\\ &\qquad \left.\times \int_0^t (D_0\breve{X}_s^2)^{-1}\frac{\partial f_1}{\partial x_2}(M_s, X_s^2(M))(X_s^1 - M_s)ds/\mathcal{Y}_t\right]\\ &\quad - \tilde{E}\left[\Gamma_t(M)\int_0^t (D_0\breve{X}_s^2)^{-1}\left(\int_0^s \left(\frac{\partial f_1}{\partial x_2}(M_u, \breve{X}_u^2) - f_1\frac{\partial f_1}{\partial x_2}(M_u, \breve{X}_u^2)\right)D_0\breve{X}_u^2 dM_u\right)\right.\\ &\qquad \left.\times \frac{\partial f_1}{\partial x_2}(M_s, X_s^2(M))(X_s^1 - M_s)ds/\mathcal{Y}_t\right].\end{aligned}$$

Except for $(X_s^1 - M_s)$, this expression is only a function on M and V^2. Provided the dM_u-integrals in (60) are bounded in L^p, the terms of this equation look like the Lebesgue integrals estimated in the previous section. $\tilde{E}[\Gamma_t(M)\int_0^t \frac{\partial f_1}{\partial x_2}(M_s, X_s^2(M))(X_s^1 - M_s)dV_s^2/\mathcal{Y}_t]$ will be of order ε if we show the following lemma holds.

LEMMA 3.11. *Let v be a bounded function from $\mathbb{R}^2$ to $\mathbb{R}$. Let us suppose that its derivative with respect to the second variable is bounded. There exists $\varepsilon_0 > 0$ such that*

$$\forall 0 \le t \le T,\ \forall p \ge 1,\ \exists C_p \text{ such that } \forall \varepsilon < \varepsilon_0, \quad \tilde{E}\left[\left|\int_0^t v(M_s, \breve{X}_s^2)dM_s\right|^p\right] \le C_p.$$

Proof of Lemma 3.11. Let us define $\Upsilon(x_1,x_2)=\int_0^{x_1} v(u,x_2)du$. By Itô's formula,

$$\int_0^t v(M_s,\breve{X}_s^2)dM_s=\Upsilon(M_t,\breve{X}_t^2)-\Upsilon(X_0^1,X_0^2)-\int_0^t\frac{\partial\Upsilon}{\partial x_2}(M_s,\breve{X}_s^2)dX_s^2,$$

we obtain

$$\begin{aligned}
&\tilde{E}\left[\left|\int_0^t v(M_s,\breve{X}_s^2)dM_s\right|^p\right]\\
&\le c(p)\left(\tilde{E}[|\Upsilon(M_t,\breve{X}_t^2)|^p]+\tilde{E}[|\Upsilon(X_0^1,X_0^2)|^p+\int_0^t\tilde{E}\left[\left|\frac{\partial\Upsilon}{\partial x_2}(M_s,\breve{X}_s^2)f_2(M_s,\breve{X}_s^2)\right|^p\right]ds\right.\\
&\qquad\left.+\tilde{E}\left[\left|\int_0^t\frac{\partial\Upsilon}{\partial x_2}(M_s,\breve{X}_s^2)dV_s^2\right|^p\right]\right)\\
&\le c(p)\left(\tilde{E}[|\Upsilon(M_t,\breve{X}_t^2)|^p]+\tilde{E}[|\Upsilon(X_0^1,X_0^2)|^p+\int_0^t\tilde{E}\left[\left|\frac{\partial\Upsilon}{\partial x_2}(M_s,\breve{X}_s^2)f_2(M_s,\breve{X}_s^2)\right|^p\right]ds\right.\\
&\qquad\left.+c'(p)\tilde{E}\left[\left|\int_0^t\left(\frac{\partial\Upsilon}{\partial x_2}(M_s,\breve{X}_s^2)\right)^2ds\right|^{\frac{p}{2}}\right]\right).\qquad(61)
\end{aligned}$$

We have used the Burkholder–Davis–Gundy inequalities.

Using Lemma 1.3, we also have

$$\begin{aligned}
\tilde{E}[|\Upsilon(M_t,\breve{X}_t^2)|^p]&\le\|v\|_\infty^p\tilde{E}[|M_t|^p]\\
&\le c(p)\|v\|_\infty^p(\tilde{E}[|M_t-X_t^1|^p]+\tilde{E}[|X_t^1|^p])\\
&\le C(p)(C(\varepsilon_0)+\tilde{E}[\sup_{0\le t\le T}|X_t^1|^p]).\qquad(62)
\end{aligned}$$

The parameter ε_0 is the parameter which appears in Definition 1.2. X_t^1 being a Brownian motion under probability $\tilde{P}$, $\tilde{E}[\sup_{0\le t\le T}|X_t^1|^p]$ no longer depends on ε. By the same method, we show the existence of a constant $C(\varepsilon_0,p,T)$ such that

$$\forall\varepsilon<\varepsilon_0,\ \forall 0\le t\le T,\quad\tilde{E}\left[\left|\frac{\partial\Upsilon}{\partial x_2}(M_t,\breve{X}_t^2)\right|^p\right]\le C(\varepsilon_0,p,T).\qquad(63)$$

f_2 being bounded (see H.5), we find that

$$\tilde{E}\left[\left|\int_0^t v(M_s,\breve{X}_s^2)dM_s\right|^p\right]\le c(\varepsilon_0,p,T).\qquad\square$$

By the same method we find

$$\tilde{E}\left[\Gamma_t(M)\int_0^t\frac{\partial^2 g}{\partial x_2^2}(M_s,X_s^2(M))Z_s^2(M,X^1-M)dV_s^2/\mathcal{Y}_t\right]=O(\varepsilon).$$

Proposition 3.10 is thus proved. $\square$

We can now conclude with the following theorem.

THEOREM 3.12. *Under the assumptions of section* 1.1, $\forall\varphi\in\mathcal{C}_b^2(\mathbb{R})$, *we have*

$$\tilde{E}[\varphi(X_t^2)\Gamma_t-\varphi(\breve{X}_t^2)\breve{\Gamma}_t|\mathcal{Y}_t]=O(\varepsilon).\qquad(64)$$

COROLLARY 3.13. *Under the assumptions of section* 1.1, $\forall\varphi\in\mathcal{C}_b^2(\mathbb{R})$, *we have*

$$\mu_t(\varphi)-E[\varphi(X_t^2)/\mathcal{Y}_t]=O(\varepsilon).\qquad(65)$$

The corollary is the direct consequence of Theorem 3.12 and Proposition 2.1.

4. Conclusion. In this work we have used a method which gives an approximation of the order of convergence between the approximate filter and the exact filter. The main element used to obtain this approximation is the integration by parts of section 1.3, which makes it possible to express $\mu_t(\varphi) - E[\varphi(X_t^2)/\mathcal{Y}_t]$ as a function of $(X^1 - M)$. It is easy to notice that the method extends without difficulty to the case where there is a coefficient of diffusion $\sigma(X^1)$ in front of dV^1 and a coefficient $\sigma(X^2)$ in front of dV^2. We can also generalize this result to the case where the dimension of X^2 is greater than 1, but where f_1 is a potential in X^2.

In the case when the integration by parts is not valid (in particular, the case when the dimension of X^1 is greater than 1, which would be useful for numerical applications), it seems possible to construct an approximate filter for the conditional law of X^2 given $\mathcal{Y}_t$ of the same type as μ_t, and it is very likely that this filter converges when ε tends to 0. However, the method used in this paper to compute the rate of convergence of such an approximate filter cannot be used. It remains an open problem to prove the convergence.

Acknowledgments. Proposition 3.7, which improves the order of convergence, is due to J. Picard. This work is part of a Ph.d. Thesis under the direction of E. Pardoux at the Université de Provence (France).

REFERENCES

[1] A. Bensoussan, *On some approximation techniques in nonlinear filtering*, in Stochastic Differential Systems, Stochastic Control and Applications (Minneapolis 1986), Springer-Verlag, New York, 1988.

[2] A. Gégout-Petit, *Filtrage d'un processus partiellement observé et Equations Differentielles Stochastiques Rétrogrades Réfléchies dans un convexe*, Thèse, Univ. de Provence, France, 1995.

[3] H. Kunita, *Stochastic Flows and Stochastic Differential Equations*, Cambridge University Press, Cambridge, UK, 1990.

[4] D. Ocone, *Malliavin calculus and stochastic integral representation of functionals of diffusion processes*, Stochastics, 12 (1984), pp. 161–185.

[5] E. Pardoux, *Equations du filtrage non linéaire de la prédiction et du lissage*, Stochastics, 6 (1982), pp. 193–231.

[6] E. Pardoux, *Filtrage non linéaire et équations aux dérivées partielles stochastiques associées*, in Ecole de Probabilité de St-Flour XIX-1989, Lectures Notes in Math. 1464, Springer-Verlag, Berlin, 1991, pp. 67–163.

[7] J. Picard, *Nonlinear filtering of one-dimensional diffusions in the case of high signal-to-noise ratio*, SIAM J. Appl. Math., 46 (1986), pp. 1098–1125.

[8] J. Picard, *Filtrage des diffusions vectorielles faiblement bruitées*, in Analysis and Optimization of Systems, Lecture Notes in Control and Inform. Sci. 83, Springer-Verlag, New York, 1986.

[9] J. Picard, *Non linear filtering and smoothing with high signal-to-noise ratio*, in Stochastic Processes in Physics and Engineering, D. Reidel, Dordrecht, the Netherlands, 1986.

[10] J. Picard, *Asymptotic study of estimation problems with small observation noise*, in Stochastic Modelling and Filtering, Lecture Notes in Control and Inform. Sci. 91, Springer-Verlag, New York, 1987.

[11] J. Picard, *Efficiency of the extended Kalman filter for nonlinear systems with small noise*, SIAM J. Appl. Math, 51 (1991), pp. 843–885.

[12] J. Picard, *Estimation of the quadratic variation of nearly observed semimartingales with application to filtering*, SIAM J. Control Optim. 31 (1993), pp. 494–517.

[13] S. S. Sachs, *Asymptotic Analysis of Linear Filtering Problems*, Ph.D. Thesis, Case Western Reserve University, Cleveland, OH, 1980.

[14] Y. Takeuchi and H. Akashi, *On the gap between deterministic and stochastic ordinary differential equations*, Ann. Probab., 6 (1978), pp. 19–41.

[15] I. Yaesh, B. Z. Bobrovsky, and Z. Schuss, *Asymptotic analysis of the optimal filtering problem for two-dimensional diffusions measured in a low noise channel*, SIAM J. Appl. Math., 50 (1990), pp. 1134–1155.

[16] M. Zakai, *On the optimal filtering of diffusions processes*, Z. Wahr. Verw. Geb., 11 (1969), pp. 230–243.
[17] O. Zeitouni and A. Dembo, *On the maximal achievable accuracy in nonlinear filtering problems*, IEEE Trans. Automat. Control, 33 (1988), pp. 965–967.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 4, pp. 1448–1472, July 1998

016

APPROXIMATION OF INFINITE-DIMENSIONAL LINEAR PROGRAMMING PROBLEMS WHICH ARISE IN STOCHASTIC CONTROL*

MARTA SUSANA MENDIONDO† AND RICHARD H. STOCKBRIDGE†

Abstract. We study a general approximation scheme for infinite-dimensional linear programming (LP) problems which arise naturally in stochastic control. We prove that the optimal value of the approximating problems converges to the value of the original LP problem. For the controls, we show that if the approximating optimal controls converge, the limiting control is an optimal control for the original LP problem.

As an application of this theory, we present numerical approximations to the LP formulation of stochastic control problems in continuous time. We study long-term average and discounted control problems. For the example for which the theoretical solution is known, our approximation results are very accurate.

Key words. linear programming, stochastic control, numerical approximation, long-term average criterion, discounted criterion

AMS subject classifications. 49M35, 93E20, 93E25

PII. S0363012996313367

1. Introduction.

This paper addresses the task of solving linear programming problems of the following form:

$$P_0: \quad \begin{cases} \text{minimize} \displaystyle\int c(x,u)\,\mu(dx\times du) \\ \text{subject to} \displaystyle\int Af(x,u)\mu(dx\times du)=0 \quad \forall f\in\mathcal{D}(A), \\ \qquad\qquad \mu \text{ is a probability measure,} \end{cases}$$

where c denotes the cost function, A is an operator on functions f, and $\mathcal{D}(A)$ denotes the domain of the operator. (Refer to section 2 for a formal definition of P_0.) Observe that optimization occurs over the space of measures satisfying the given constraints, which is typically infinite-dimensional. Often explicit solutions are difficult to determine due to this infinite-dimensionality. For such problems, it is necessary to make some reduction to finite dimensions.

This paper considers a general approach to approximating P_0 by a sequence of linear programs:

$$P_n: \quad \begin{cases} \text{minimize} \displaystyle\int c_n(x,u)\,\mu_n(dx\times du) \\ \text{subject to} \displaystyle\int A_nf_n(x,u)\mu_n(dx\times du)=0 \quad \forall f_n\in\mathcal{D}(A_n), \\ \qquad\qquad \mu_n \text{ is a probability measure} \end{cases}$$

*Received by the editors December 9, 1996; accepted for publication (in revised form) October 7, 1997; published electronically May 28, 1998. This research was partially supported by NSF grant DMS-9404990.

http://www.siam.org/journals/sicon/36-4/31336.html

†Department of Statistics, University of Kentucky, Lexington, KY 40506–0027 (marta@ms.uky.edu, stockb@ms.uky.edu).

where the function f_n approximates f, the operator A_n approximates the operator A, the cost function c_n is related to the original cost function c, and $\mathcal{D}(A_n)$ denotes the domain of the operator A_n. The optimal values of the approximating problems P_n will be shown to be approximately optimal for the original problem P_0. The application of the general approximation scheme will be to obtain linear programming problems P_n which are finite-dimensional and hence accessible to computational solutions.

Linear programming problems of the form P_0 arise as an equivalent formulation of stochastic control problems. The idea of reformulating problems in a space of measures was introduced by Young [23] for calculus of variations problems and has been applied to nonstochastic control problems by Rubio [17, 18]. The use of linear programming to solve stochastic control problems has been studied for more than three decades. Early work concentrated on control problems in a discrete setting: discrete state space, discrete control space, and/or discrete time (see, for example, [3, 4, 15, 21, 22]). Recently, Hernández-Lerma, Hennet, and Lasserre [11] considered discrete time Markov decision processes with Borel state and control spaces under a long-run expected average cost criterion. Working in continuous time, Kurtz [12] and Stockbridge [19] studied control problems when the state space E and control space U are allowed to be locally compact, complete, and separable metric spaces. Stockbridge showed that stochastic control problems involving the long-term average cost can be reformulated as LP problems. The equivalence depended upon the existence of stationary solutions but was limited in that the optimal control was not characterized. This equivalence has been improved in several ways independently by Bhatt and Borkar [2] and Kurtz and Stockbridge [13]. Both papers characterize the optimal control and also extend the LP formulation to infinite horizon discounted and finite horizon control problems. Kurtz and Stockbridge [13] also extend the results to first passage control problems, and Bhatt and Borkar [2] relax the local compactness of the state space.

An LP formulation has been used by Heinricher and Stockbridge [9] to solve a stochastic control problem for processes modeling the wear of a system. Also, Ghosh, Arapostathis, and Marcus [6] used an LP formulation to show existence of an optimal solution for the ergodic control problem but then used dynamic programming to express the solution in terms of the Hamilton–Jacobi–Bellman equation.

This paper studies a general approximation of these types of LP problems with the goal of providing numerical methods of solution. In our example, the reduction to finite dimensions uses the Markov chain approximations studied extensively by Kushner and Dupuis [14] and others. They use dynamic programming on these approximations to solve control problems, whereas our approach relies on linear programming methods. They justify the results using weak convergence techniques and work with measures on the space of stochastic processes, whereas our justification uses weak convergence directly on measures on the state and control spaces. Some other papers which study finite-dimensional approximations of linear programming problems are [20, 8].

The paper is organized as follows. In section 2 we show that, given a linear programming problem P_0, we can define approximating problems P_n, whose optimal solutions converge to the optimal solution for P_0. Convergence of the controls is also considered. Section 3 examines finite-dimensional LP approximations to the original infinite-dimensional LP problem for two examples. Both examples are modifications of the bounded follower problem studied by Beneš, Shepp, and Witsenhausen [1]. The first one uses a long-term average criterion for which the solution is known; the second

one uses an infinite horizon discounted cost criterion.

2. Mathematical formulations and theoretical results. This section is divided into three sections. Section 2.1 contains the formal definition of the linear programming problem P_0. Section 2.2 defines the approximating LP problems. Section 2.3 provides results about the existence of the solutions to the approximating linear programming problems P_n and their convergence to the solution of P_0. Section 2.3.1 deals with convergence of the values, while section 2.3.2 considers convergence of the controls.

2.1. The original LP problem. In this section we formally define the LP problem.

Denote the state space by E and the control space by U. We assume E and U are compact, metric spaces. We denote the distance between points in E and U by $|\cdot - \cdot|$ and on the product space by $d(\cdot,\cdot)$. Denote by $\mathcal{P}(E)$ and $\mathcal{P}(U)$ the space of probability measures on E and U, respectively. Also, let $C(E)$ and $C(E\times U)$ denote the spaces of continuous functions on E and $E\times U$, respectively, and let $M(E\times U)$ denote the space of measurable functions on $E\times U$.

Let $A : \mathcal{D}(A) \to C(E\times U)$, where $\mathcal{D}(A) \subset C(E)$ denotes the domain of the operator A.

2.1.1. Conditions on the generator A and cost function c. We assume that A satisfies the following conditions:

(C1) $\mathcal{D}(A)$ is an algebra which is dense in $C(E)$, and

(C2) there exists some reference measure π on $(E,\mathcal{B}(E))$ such that if $\mu\in\mathcal{P}(E\times U)$ satisfies the stationarity condition $\int_{E\times U} Af(x,u)\mu(dx\times du) = 0$ for each $f\in\mathcal{D}(A)$, $\mu_0(\cdot)=\mu(\cdot\times U)$ is absolutely continuous with respect to π.

Condition (C1) is used in [13] to prove that many stochastic control problems can be equivalently written as LP problems of the form P_0. Condition (C2) implies that any set that has π-measure zero also has μ_0-measure zero for each stationary distribution μ.

Let $c(x,u) : E\times U\to\mathbb{R}$ be a continuous function.

2.1.2. Statement of the original LP problem. We need to place one additional restriction on the optimization problem concerning the types of controls over which optimization occurs. In order to state the condition, we need the following observation and terminology.

Observe that, for any $\mu\in\mathcal{P}(E\times U)$, we can decompose μ as $\mu(dx\times du) = \eta(x,du)\mu_0(dx)$ in which μ_0 is the marginal of μ on E and η is the regular conditional distribution on U given x under μ_0. We refer to η as a relaxed control.

We make the following restriction. *The only relaxed controls η we consider for the optimization problem satisfy*

(R1) η, as a measure-valued function of the state, is continuous almost everywhere (in the Prohorov metric (see [5, section 3.1])) with respect to the reference measure π of (C2).

This is a continuity restriction about the relaxed controls and is necessary for the weak convergence arguments we use in this paper. From a practical point of view, (R1) is not a serious limitation. The measure π is typically absolutely continuous with respect to Lebesgue measure so sets of positive π measure will be uncountable. Implementation of controls having infinitely many discontinuities is impossible. In addition, it is often possible to show a priori that the optimal control for the unrestricted problem satisfies (R1). When this is so, there is no loss of generality.

Let $\mathcal{D}_\eta$ denote the set of $x \in E$ where $\eta(x, \cdot)$ is discontinuous as a function of x. We define the linear programming problem P_0 as

$$P_0: \begin{cases} \text{minimize } \displaystyle\int_{E\times U} c(x,u)\,\mu(dx\times du) \\ \\ \text{subject to } \displaystyle\int_{E\times U} Af(x,u)\mu(dx\times du) = 0 \quad \forall f \in \mathcal{D}(A), \\ \\ \qquad\qquad \mu \in \mathcal{P}(E\times U) \text{ satisfies (R1) for some cond. dist. } \eta. \end{cases}$$

To simplify notation, $\mu = \eta\mu_0$ is used to indicate $\mu(dx \times du) = \eta(x, du)\mu_0(dx)$, and for a function $g \in M(E \times U)$ and a probability measure $\mu \in \mathcal{P}(E \times U)$, let

$$\langle g, \mu\rangle = \int_{E\times U} g(x,u)\mu(dx\times du)\,.$$

We define feasibility and optimality as follows:
Feasibility: A probability measure $\mu \in \mathcal{P}(E\times U)$ is a feasible point for P_0 if $\langle Af, \mu\rangle = 0$ for each $f \in \mathcal{D}(A)$ and μ has decomposition $\eta\mu_0$ for some η satisfying (R1). We denote the set of P_0-feasible points by $\mathcal{A}$.
Optimality: μ^* is an optimal solution for P_0 if $\mu^* \in \mathcal{A}$ and for each $\mu \in \mathcal{A}$

$$\langle c, \mu^*\rangle \le \langle c, \mu\rangle.$$

Kurtz and Stockbridge [13] show that the conditional distribution $\eta^*(x, du)$ of an optimizing μ^* identifies an optimal control for the stochastic control problem.

2.2. The approximating LP problems. Throughout the remainder of this paper we adopt the following notation. For $n \ge 0$, let E_n and U_n be compact metric spaces. We assume that for each n there exist measurable functions

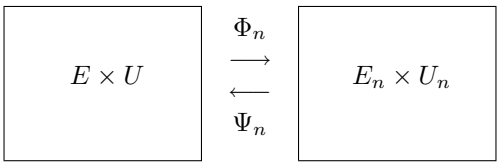
$$\psi_n^1 : E_n \to E, \qquad \psi_n^2 : U_n \to U,$$

$$\phi_n^1 : E \to E_n, \qquad \phi_n^2 : U \to U_n$$

such that

(C3) $\sup_{u\in U} \left|u - \psi_n^2\left(\phi_n^2(u)\right)\right| \to 0$ as $n \to \infty$ and
(C4) $\left|x - \psi_n^1\left(\phi_n^1(x)\right)\right| \to 0$ as $n \to \infty$ for each $x \in E$.
We denote

(2.1) $\Psi_n = (\psi_n^1, \psi_n^2) : E_n \times U_n \to E \times U$ and $\Phi_n = (\phi_n^1, \phi_n^2) : E \times U \to E_n \times U_n$:



For each n, we transfer the cost function c to the approximating space $E_n \times U_n$ by defining $c_n = c \circ \Psi_n$. Note that $c_n : E_n \times U_n \to \mathbb{R}$.

Given $\mu \in \mathcal{P}(E \times U)$, we define $\widehat{\mu}^n := \mu \circ \Phi_n^{-1}$, the distribution on $E_n \times U_n$ induced by Φ_n. Equivalently, $\widehat{\mu}^n$ can be defined as the distribution on $E_n \times U_n$ such that for each continuous h

$$\int_{E_n \times U_n} h(y,v)\widehat{\mu}^n(dy \times dv) = \int_{E \times U} h\left(\phi_n^1(x), \phi_n^2(u)\right) \mu(dx \times du)\,. \tag{2.2}$$

Similarly, given $\mu^n \in \mathcal{P}(E_n \times U_n)$, we define $\overline{\mu^n} := \mu^n \circ \Psi_n^{-1}$ as the distribution on $E \times U$ which satisfies

$$\int_{E \times U} h(x,u)\overline{\mu^n}(dx \times du) = \int_{E_n \times U_n} h\left(\psi_n^1(y), \psi_n^2(v)\right) \mu^n(dy \times dv) \tag{2.3}$$

for each $h \in C(E \times U)$.

The notation for the measures in this paper can at times appear quite cumbersome since we use the marginals of measures, measures on $E_n \times U_n$ induced by measures on $E \times U$, and measures on $E \times U$ induced by measures on $E_n \times U_n$ in various combinations. We have adopted the following conventions, hoping to aid the reader.

- For a measure μ on a product space, the marginal on the first component (E or E_n) is denoted μ_0.
- For a measure $\mu \in \mathcal{P}(E \times U)$, the induced measure on $E_n \times U_n$ is denoted by placing "$\widehat{\ }^n$" over the μ. The "hat" is to indicate that the measure "comes from" a measure on $E \times U$.
- For a measure $\mu \in \mathcal{P}(E_n \times U_n)$, the induced measure on $E \times U$ is denoted by placing "$\overline{\ }$" over the μ. The "bar" is to indicate that the measure "comes from" a measure on $E_n \times U_n$.
- The use of $*$ with a measure indicates that it is an optimal measure for the linear programming problem.

An example of this notation is $\widehat{\overline{\mu_0^n}}$, which represents the marginal of the measure $\widehat{\overline{\mu^n}}$ on $E \times U$, which is induced by the measure $\widehat{\mu}^n$ on $E_n \times U_n$, that is in turn induced by the measure μ on $E \times U$.

2.2.1. Definition of the approximating LP problem. Given E_n, U_n, and Ψ_n as defined above, for each $n \geq 1$, let

$$A_n : \mathcal{D}(A_n) \subset C(E_n) \to C(E_n \times U_n)$$

be such that, for each $f \in \mathcal{D}(A)$, there exists $f_n \in \mathcal{D}(A_n)$ satisfying

(C5) $\sup_{y \in E_n} \left|f_n(y) - f\left(\psi_n^1(y)\right)\right| \to 0$ as $n \to \infty$ and

(C6) $\sup_{(y,v) \in E_n \times U_n} |A_n f_n(y,v) - Af\left(\Psi_n(y,v)\right)| \to 0$ as $n \to \infty$.

We assume

(C7) for each n, for each relaxed control $\widehat{\eta}^n : E_n \times \mathcal{B}(U_n) \to [0,1]$, there exists a probability measure $\widehat{\mu}_0^n \in \mathcal{P}(E_n)$ such that

$$\int_{E_n} \int_{U_n} A_n f_n(y,v)\widehat{\eta}^n(y,dv)\widehat{\mu}_0^n(dy) = 0 \quad \forall f_n \in \mathcal{D}(A_n)\,;$$

i.e., each relaxed control induces some stationary distribution $\widehat{\mu}_0^n$ on the states.

In our application, the spaces E_n and U_n are a discretization of E and U, respectively, and the function Ψ_n is the embedding of $E_n \times U_n$ into $E \times U$. Φ_n maps intervals into a single point in the interval. Thus, conditions (C3) and (C4) are easily verified. As for conditions (C5), (C6), and (C7), we approximate a diffusion process by a finite-state Markov chain so these conditions can also be readily verified.

Another way of satisfying these conditions is to take $E_n = E$, $U_n = U$, Φ_n and Ψ_n to be the identity and $f_n = f$ and approximate the generator A uniformly by a sequence A_n. For example, if A is a diffusion operator, A_n can be an approximating diffusion in which the drift and diffusion coefficients uniformly approximate the corresponding coefficients of A.

The approximating LP problem is given by

$$P_n: \quad \begin{cases} \text{minimize } \langle c_n, \mu^n \rangle \\ \\ \text{subject to } \quad \langle A_n f_n, \mu^n \rangle = 0 \quad \forall f_n \in \mathcal{D}(A_n), \\ \\ \qquad\qquad \mu^n \in \mathcal{P}(E_n \times U_n). \end{cases}$$

As before, we say that a probability measure μ^n is P_n-feasible if $\langle A_n f_n, \mu^n \rangle = 0$ for each $f_n \in \mathcal{D}(A_n)$, and we denote the collection of P_n-feasible points by $\mathcal{A}_n$. We say $\overset{*n}{\mu}$ is P_n-optimal if $\overset{*n}{\mu} \in \mathcal{A}_n$ and $\langle c_n, \overset{*n}{\mu} \rangle \leq \langle c_n, \mu^n \rangle$ for each $\mu^n \in \mathcal{A}_n$.

2.3. Convergence results. In this section we state and prove the results that justify the use of the solution of P_n to approximate the solution of P_0. We first consider the convergence of the optimal values and then the convergence of the optimal controls.

2.3.1. Convergence of the values. In this section we show the following results. First, a convergent sequence of P_n-feasible points converges to a P_0-feasible point. Then, given any P_0-feasible point, we can construct a sequence of P_n-feasible points that converges to it, and finally, if those points are P_n-optimal solutions, any limit is P_0-optimal.

THEOREM 1. *Let $\{\mu^n\}$ be a sequence of P_n-feasible points. Define $\overline{\mu^n}(dx \times du) \in \mathcal{P}(E \times U)$ by* (2.3)*; i.e., for each continuous h*

$$\int_{E\times U} h(x,y)\overline{\mu^n}(dx \times dy) = \int_{E_n \times U_n} h\left(\psi_n^1(y), \psi_n^2(v)\right) \mu^n(dy \times dv).$$

If there exists a $\mu \in \mathcal{P}(E \times U)$ and some subsequence $\{n_k\}$ such that $\overline{\mu^{n_k}} \Rightarrow \mu$, then

$$\int_{E\times U} Af d\mu = 0 \quad \forall f \in \mathcal{D}(A)\,;$$

i.e., μ is P_0-feasible.

Proof. Let $f \in \mathcal{D}(A)$, and let μ be as in the statement of the theorem. Without loss of generality, assume the entire sequence converges in distribution to μ; i.e., $\overline{\mu^n} \Rightarrow \mu$. Observe that $\int A_n f_n d\mu^n = 0$ since μ^n is P_n-feasible, so

$$\int_{E\times U} Af(x,y)\mu(dx \times du) = \int_{E\times U} Af(x,u)\mu(dx \times du)$$
$$- \int_{E_n \times U_n} Af\left(\psi_n^1(y), \psi_n^2(v)\right) \mu^n(dy \times dv)$$

$$
\begin{aligned}
&+\int_{E_n\times U_n} Af\left(\psi_n^1(y),\psi_n^2(v)\right)\mu^n(dy\times dv)\\
&-\int_{E_n\times U_n} A_nf_n(y,v)\mu^n(dy\times dv)\\
&\quad=\left\{\int_{E\times U} Af(x,u)\mu(dx\times du)-\int_{E\times U} Af(x,u)\overline{\mu^n}(dx\times du)\right\}\\
&+\left\{\int_{E_n\times U_n}\left[Af\left(\psi_n^1(y),\psi_n^2(v)\right)-A_nf_n(y,v)\right]\mu^n(dy\times dv)\right\}.
\end{aligned}
$$

Since Af is (bounded and) continuous and $\overline{\mu^n}\Rightarrow\mu$, then

$$\int_{E\times U} Af(x,u)\mu(dx\times du)-\int_{E\times U} Af(x,u)\overline{\mu^n}(dx\times du)\to 0\,.$$

Also, by condition (C6),

$$\sup_{(y,v)\in E_n\times U_n}\left|Af\left(\psi_n^1(y),\psi_n^2(v)\right)-A_nf_n(y,v)\right|\to 0\,,$$

so

$$\int_{E_n\times U_n}\left[Af\left(\psi_n^1(y),\psi_n^2(v)\right)-A_nf_n(y,v)\right]\mu^n(dy\times dv)\to 0\,.$$

Therefore,

$$\int_{E\times U} Af(x,u)\mu(dx\times du)=0\,;$$

i.e., μ is P_0-feasible. □

The second theorem in this section shows that given any P_0-feasible point $\mu=\eta\mu_0$, there exists a distribution $\nu_0\in\mathcal{P}(E)$ and P_n-feasible points $\widehat{\mu}^n$ such that their induced measure $\widehat{\overline{\mu^n}}$ on $E\times U$ converges to $\eta\nu_0$. This result does not assume uniqueness of the stationary distribution on E for the control η. When the control η has a unique stationary distribution, then $\mu_0=\nu_0$ (see Corollary 3).

THEOREM 2. *Let μ be a P_0-feasible point having some decomposition $\mu(dx\times du)=\eta(x,du)\mu_0(dx)$ with η satisfying the restriction* (R1). *For each n, define the relaxed control $\widehat{\eta}^n:E_n\times\mathcal{B}(U_n)\to[0,1]$ satisfying*

$$\int_{U_n} h(v)\widehat{\eta}^n(y,dv)=\int_U h(\phi_n^2(u))\eta(\psi_n^1(y),du)\qquad\forall h\in C(U),$$

and let $\widehat{\mu}_0^n\in\mathcal{P}(E_n)$ be the stationary distribution satisfying condition (C7) *with control $\widehat{\eta}^n$. Then there exists a distribution $\nu_0\in\mathcal{P}(E)$ such that*

(i) *$\eta(x,du)\nu_0(dx)$ is P_0-feasible and*

(ii) *for each continuous function h on $E\times U$,*

$$\int_{E_n}\int_{U_n} h\left(\psi_n^1(y),\psi_n^2(v)\right)\widehat{\eta}^n(y,dv)\widehat{\mu}_0^n(dy)\to\int_E\int_U h(x,u)\eta(x,du)\nu_0(dx)\,.$$

Proof. Let μ, η, $\widehat{\eta}^n$, and $\widehat{\mu}_0^n$ be as in the statement of the theorem. Let

$$\widehat{\mu}^n(dy\times dv)=\widehat{\eta}^n(y,dv)\widehat{\mu}_0^n(dy),$$

and define $\overline{\widehat{\mu}^n} \in \mathcal{P}(E \times U)$ as in (2.3) so that for all $h \in C(E \times U)$,

$$\int_{E\times U} h(x,u)\overline{\widehat{\mu}^n}(dx \times du) = \int_{E_n \times U_n} h\left(\psi_n^1(y), \psi_n^2(v)\right) \widehat{\mu}^n(dy \times dv).$$

Let $\overline{\widehat{\mu}_0^n}(dx) = \overline{\widehat{\mu}^n}(dx \times U)$ be the state marginal of $\overline{\widehat{\mu}^n}$. Observe that

$$\begin{aligned}
\int_{E\times U} h(x,u)\overline{\widehat{\mu}^n}(dx \times du) &= \int_{E_n}\int_{U_n} h\left(\psi_n^1(y), \psi_n^2(v)\right) \widehat{\mu}^n(dy \times dv) \\
&= \int_{E_n}\int_{U_n} h\left(\psi_n^1(y), \psi_n^2(v)\right) \widehat{\eta}^n(y, dv)\widehat{\mu}_0^n(dy) \\
&= \int_{E_n}\int_{U} h\left(\psi_n^1(y), \psi_n^2(\phi_n^2(u))\right) \eta\left(\psi_n^1(y), du\right) \widehat{\mu}_0^n(dy) \\
&= \int_{E}\int_{U} h\left(x, \psi_n^2(\phi_n^2(u))\right) \eta(x, du)\overline{\widehat{\mu}_0^n}(dx)\,. \qquad (2.4)
\end{aligned}$$

Since $E \times U$ is compact, $\{\overline{\widehat{\mu}^n}\}$ is tight and hence relatively compact. Thus, there exists a subsequence $\{n_k\}$ and a $\nu \in \mathcal{P}(E \times U)$ such that $\overline{\widehat{\mu}^{n_k}} \Rightarrow \nu$; without loss of generality, we can assume that $\overline{\widehat{\mu}^n} \Rightarrow \nu$; i.e.,

$$\int_{E\times U} h(x,u)\overline{\widehat{\mu}^n}(dx \times du) \to \int_{E\times U} h(x,u)\nu(dx \times du)\,. \qquad (2.5)$$

Letting ν_0 denote the state marginal of ν, it immediately follows that $\overline{\widehat{\mu}_0^n} \Rightarrow \nu_0$. Since h is continuous on the compact set $E \times U$, given $\epsilon > 0$ $\exists \delta > 0$ such that $d\left((x_1, y_1), (x_2, y_2)\right) < \delta$ implies

$$|h(x_1, u_1) - h(x_2, u_2)| < \epsilon\,.$$

By condition (C3), $\left|u - \psi_n^2\left(\phi_n^2(u)\right)\right| \to 0$ uniformly as $n \to \infty$, so there exists $N > 0$ such that $\forall n \geq N$,

$$\left|u - \psi_n^2\left(\phi_n^2(u)\right)\right| < \delta\,.$$

Then $\forall n \geq N$,

$$\begin{aligned}
&\left|\int_E\int_U h\left(x, \psi_n^2\left(\phi_n^2(u)\right)\right) \eta(x, du)\overline{\widehat{\mu}_0^n}(dx) - \int_E\int_U h(x,u)\eta(x, du)\overline{\widehat{\mu}_0^n}(dx)\right| \\
&\leq \int_E\int_U \left|h\left(x, \psi_n^2\left(\phi_n^2(u)\right)\right) - h(x,u)\right| \eta(x, du)\overline{\widehat{\mu}_0^n}(dx) < \epsilon\,.
\end{aligned}$$

Since ϵ is arbitrary,

$$\lim_{n\to\infty}\left|\int_E\int_U h\left(x, \psi_n^2\left(\phi_n^2(u)\right)\right) \eta(x, du)\overline{\widehat{\mu}_0^n}(dx) - \int_E\int_U h(x,u)\eta(x, du)\overline{\widehat{\mu}_0^n}(dx)\right| = 0,$$

and so, by (2.4),

$$\lim_{n\to\infty}\left|\int_{E\times U} h(x,u)\overline{\widehat{\mu}^n}(dx \times du) - \int_E\int_U h(x,u)\eta(x, du)\overline{\widehat{\mu}_0^n}(dx)\right| = 0.$$

Then, by (2.5), we conclude that

$$\lim_{n\to\infty}\int_E\int_U h(x,u)\eta(x, dx)\overline{\widehat{\mu}_0^n}(dx) = \int_{E\times U} h(x,u)\nu(dx \times du)\,.$$

Recall, $\mathcal{D}_\eta$ is the set of discontinuity points of η. We claim that $\nu_0(\mathcal{D}_\eta) = 0$. To see this, let $f \in \mathcal{D}(A)$ be arbitrary but fixed, and let $h(x,u) = Af(x,u)$. Then

$$\int_{E_n}\int_{U_n} Af\left(\psi_n^1(y),\psi_n^2(v)\right)\widehat{\eta}^n(y,dv)\widehat{\mu}_0^n(dy) \to \int_E\int_U Af(x,u)\nu(dx\times du)\,. \tag{2.6}$$

Also, condition (C6) on the generators A_n implies

$$\int_{E_n}\int_{U_n} \left|Af\left(\psi_n^1(y),\psi_n^2(v)\right) - A_nf_n(y,v)\right|\widehat{\eta}^n(y,dv)\widehat{\mu}_0^n(dy) \to 0,$$

and so

$$\left|\int_{E_n}\int_{U_n} Af\left(\psi_n^1(y),\psi_n^2(v)\right)\widehat{\eta}^n(y,dv)\widehat{\mu}_0^n(dy) - \int_{E_n}\int_{U_n} A_nf_n(y,v)\widehat{\eta}^n(y,dv)\widehat{\mu}_0^n(dy)\right| \to 0.$$

Since $\widehat{\mu}_0^n$ is chosen so that the second term is zero, (2.6) implies

$$\int_E\int_U Af(x,u)\nu(dx\times du) = 0.$$

Thus, ν is P_0-feasible, and condition (C2) implies $\nu_0 \ll \pi$, and therefore $\nu_0(\mathcal{D}_\eta) = 0$, proving the claim.

The continuous mapping theorem [5, Corollary 3.1.9] implies

$$\begin{aligned}\int_E\int_U h(x,u)\eta(x,du)\nu_0(dx) &= \lim_{n\to\infty}\int_E\int_U h(x,u)\eta(x,du)\overline{\widehat{\mu}_0^n}(dx)\\ &= \int_{E\times U} h(x,u)\nu(dx\times du)\\ &= \int_E\int_U h(x,u)\overline{\eta}(x,du)\nu_0(dx)\,.\end{aligned}$$

Taking $h(x,u) = h_1(x)h_2(u)$, for $h_1 \in C(E)$ and $h_2 \in C(U)$,

$$\int_E h_1(x)\left(\int_U h_2(u)\eta(x,du)\right)\nu_0(dx) = \int_E h_1(x)\left(\int_U h_2(u)\overline{\eta}(x,du)\right)\nu_0(dx).$$

Since this is true for each $h_1 \in C(E)$, it follows that for each $h_2 \in C(U)$,

$$\int_U h_2(u)\eta(x,du) = \int_U h_2(u)\overline{\eta}(x,du) \text{ a.e. (almost everywhere) } \nu_0,$$

and thus

$$\eta(x,\cdot) = \overline{\eta}(x,\cdot) \text{ a.e. } \nu_0\,.$$

This shows that ν has decomposition

$$\nu(dx\times du) = \eta(x,du)\nu_0(dx). \tag{2.7}$$

From the definition of $\overline{\widehat{\mu}^n}$, (2.5), and (2.7), the result is established. □

Corollary 3. *Suppose η is a relaxed control satisfying restriction* (R1). *Suppose also there is a unique $\mu_0 \in \mathcal{P}(E)$ such that for each $f \in \mathcal{D}(A)$,*

$$\int_E\int_U Af(x,u)\,\eta(x,du)\,\mu_0(dx) = 0.$$

Define $\widehat{\eta}^n$ *and* $\widehat{\mu}_0^n$ *as in Theorem* 2. *Then*

$$\int_{E_n}\int_{U_n} h(\psi_n^1(y),\psi_n^2(v))\widehat{\eta}^n(y,dv)\widehat{\mu}_0^n(dy) \to \int_E\int_U h(x,u)\eta(x,du)\mu_0(dx).$$

Proof. The fact that μ_0 is unique implies that μ_0 is the ν_0 of Theorem 2. The result now follows from Theorem 2. □

Finally, we only need to show that the limit of optimal solutions in P_n is an optimal solution in P_0.

THEOREM 4. *Suppose* $\overset{*n}{\mu}$ *is an optimal solution for* P_n, *and as in* (2.3), *let* $\overline{\overset{*n}{\mu}} = \overset{*n}{\mu} \circ \Psi_n^{-1}$. *Suppose that, for each* η *satisfying* (R1), *there is a unique stationary distribution* μ_0 *such that* $\langle Af, \eta\mu_0\rangle = 0$ *for all* $f \in \mathcal{D}(A)$. *Then any* μ^* *which is a weak limit of* $\{\overline{\overset{*n_k}{\mu}}\}$ *for some subsequence* $\{n_k\}$ *is* P_0*-optimal.*

Proof. By Theorem 1, μ^* is a P_0-feasible point. Let μ be any other P_0-feasible point, and let $\{\widehat{\mu}^n = \widehat{\eta}^n\widehat{\mu}_0^n\}$ be the P_n-feasible points given in the statement of Theorem 2. By Corollary 3, the induced measures $\overline{\widehat{\mu}^n} \Rightarrow \mu$. By the Skorohod representation theorem, there exists a sequence of $E\times U$-valued random variables $\{(X_n,U_n)\}$ and an $E\times U$-valued random variable (X,U) such that $X_n \to X$ almost surely (a.s.) and $U_n \to U$ a.s. This implies that $\int c\, d\overline{\widehat{\mu}^n} \to \int c\, d\mu$. In a similar manner, it follows that $\int c\, d\,\overline{\overset{*n}{\mu}} \to \int c\, d\mu^*$. Thus,

$$\begin{aligned}
\int_{E\times U} c(x,u)\mu(dx\times du) &= \lim_{n\to\infty}\int_{E\times U} c(x,u)\overline{\widehat{\mu}^n}(dx\times du)\\
&= \lim_{n\to\infty}\int_{E_n\times U_n} c\left(\psi_n^1(y),\psi_n^2(v)\right)\widehat{\mu}^n(dy\times dv)\\
&\geq \lim_{n\to\infty}\int_{E_n\times U_n} c\left(\psi_n^1(y),\psi_n^2(v)\right)\ \overset{*n}{\mu}(dy\times dv)\\
&= \lim_{n\to\infty}\int_{E\times U} c(x,u)\ \overline{\overset{*n}{\mu}}(dx\times du)\\
&= \int_{E\times U} c(x,u)\mu^*(dx\times du)\,. \qquad \square
\end{aligned} \tag{2.8}$$

2.3.2. Convergence of controls. This section presents results concerning the controls. We show that, given any control $\eta \in \mathcal{P}(U)$ satisfying restriction (R1), the induced controls $\overline{\widehat{\eta}^n}$ converge in distribution to η for almost all x. We also show that a limit of P_n-optimal controls is a P_0-optimal control.

THEOREM 5. *Given any* $\eta : E\times\mathcal{B}(U) \to [0,1]$ *satisfying restriction* (R1), *define* $\widehat{\eta}^n : E_n\times\mathcal{B}(U_n) \to [0,1]$ *such that* $\forall g \in C(U)$,

$$\int_{U_n} g(v)\widehat{\eta}^n(y,dv) = \int_U g\left(\phi_n^2(u)\right)\eta\left(\psi_n^1(x),du\right),$$

and define $\overline{\widehat{\eta}^n} : E\times\mathcal{B}(U) \to [0,1]$ *such that* $\forall h \in C(U)$,

$$\int_U h(u)\overline{\widehat{\eta}^n}(x,du) = \int_{U_n} h\left(\psi_n^2(v)\right)\widehat{\eta}^n\left(\phi_n^1(x),dv\right).$$

Then

$$\overline{\widehat{\eta^n}}(x,\cdot) \Rightarrow \eta(x,\cdot) \qquad a.e.\ \pi.$$

Proof. Let $h \in C(U)$. Select $\epsilon > 0$ arbitrarily. Since h is continuous on a compact space, find $\delta > 0$ such that $|h(u_1) - h(u_2)| \le \epsilon/2$ whenever $|u_1 - u_2| \le \delta$. By condition (C3), find N_1 such that for all $n \ge N_1$, $|u - \psi_n^2(\phi_n^2(u))| \le \delta$. Then for all $n \ge N_1$,

$$\begin{aligned}
&\left| \int_U h(u) \overline{\widehat{\eta^n}}(x,du) - \int_U h(u)\eta(x,du) \right| \\
&= \left| \int_U h\left(\psi_n^2\left(\phi_n^2(u)\right)\right) \eta\left(\psi_n^1\left(\phi_n^1(x)\right), du\right) - \int_U h(u)\eta(x,du) \right| \\
&= \left| \int_U \left[h\left(\psi_n^2\left(\phi_n^2(u)\right)\right) - h(u)\right] \eta\left(\psi_n^1\left(\phi_n^1(x)\right), du\right) \right. \\
&\quad \left. + \int_U h(u)\eta\left(\psi_n^1\left(\phi_n^1(x)\right), du\right) - \int_U h(u)\eta(x,du) \right| \\
&\le \frac{\epsilon}{2} + \left| \int_U h(u)\eta\left(\psi_n^1\left(\phi_n^1(x)\right), du\right) - \int_U h(u)\eta(x,du) \right|.
\end{aligned}$$

By condition (C4), $\left|x - \psi_n^1\left(\phi_n^1(x)\right)\right| \to 0$ as $n \to \infty$. Letting $x_n = \psi_n^1\left(\phi_n^1(x)\right)$, we have $x_n \to x$. Since η is continuous a.e. π, $\eta(x_n,\cdot) \to \eta(x,\cdot)$ in the Prohorov metric for a.e. $x\,(d\pi)$. Therefore,

$$\eta(x_n,\cdot) \Rightarrow \eta(x,\cdot) \qquad \text{a.e. } \pi,$$

or equivalently,

$$\lim_{n\to\infty} \int_U h(u)\eta(x_n,du) = \int_U h(u)\eta(x,du) \qquad \text{a.e. } \pi\,.$$

So we can select N_2 such that $\forall n \ge N_2$,

$$\left| \int_U h(u)\eta(x_n,du) - \int_U h(u)\eta(x,du) \right| < \frac{\epsilon}{2}\,,$$

and taking $N = \max(N_1, N_2)$,

$$\left| \int_U h(u) \overline{\widehat{\eta^n}}(x,du) - \int_U h(u)\eta(x,du) \right| < \epsilon \quad \forall n \ge N.$$

Therefore,

$$\overline{\widehat{\eta^n}}(x,\cdot) \Rightarrow \eta(x,\cdot) \qquad \text{a.e. } \pi\,. \qquad \square \tag{2.9}$$

Next we show that a limit of the P_n-optimal controls is an optimal control for the P_0-problem.

THEOREM 6. *Let* $\overset{*n}{\eta} : E_n \times \mathcal{B}(U_n) \to [0,1]$ *be an optimal control for* P_n*, i.e., there exists some* $\overset{*n}{\mu_0} \in \mathcal{P}(E_n)$ *such that for each* $\mu^n \in \mathcal{A}_n$*,*

$$\int_{E_n} \int_{U_n} c_n(y,v)\ \overset{*n}{\eta}(y,dv)\, \overset{*n}{\mu_0}(dy) \le \int_{E_n \times U_n} c_n(y,v)\mu^n(dy \times dv).$$

Let $\overline{\overset{*n}{\eta}} : E \times \mathcal{B}(U) \to [0,1]$ *be the induced control satisfying*

$$\int_U h(u)\ \overline{\overset{*n}{\eta}}(x, du) = \int_{U_n} h\left(\psi_n^2(v)\right)\ \overset{*n}{\eta}\left(\phi_n^1(x), dv\right)$$

for each $h \in \overline{C}(U)$*, and let* $\overline{\overset{*n}{\mu_0}} \in \mathcal{P}(E)$ *be the induced measure satisfying*

$$\int_E h(x)\overline{\overset{*n}{\mu_0}}(dx) = \int_{E_n} h\left(\psi_n^1(y)\right)\ \overset{*n}{\mu_0}(dy)$$

for each $h \in \overline{C}(E)$*. Suppose there exists* $\eta : E \times \mathcal{B}(U) \to [0,1]$ *which, as a measure-valued function of* x*, is continuous a.e.* π*,* η *induces a unique stationary distribution* μ_0 *on the state space* E*, and* $\overline{\overset{*n}{\eta}}(x, \cdot) \Rightarrow \eta(x, \cdot)$ *for almost every* x *with respect to* π*. Then* $\overline{\overset{*n}{\mu_0}} \Rightarrow \mu_0$ *and* $\eta(x, du)\mu_0(dx)$ *is* P_0*-optimal.*

Proof. Define $\overline{\overset{*n}{\mu}}(dx \times du) = \overline{\overset{*n}{\eta}}(x, du)\overline{\overset{*n}{\mu_0}}(dx)$. Since $E \times U$ is compact, $\{\ \overline{\overset{*n}{\mu}}\ \}$ is tight, and hence $\overline{\overset{*n}{\mu}} \Rightarrow \tilde{\mu}$ along a subsequence for some $\tilde{\mu} \in \mathcal{P}(E \times U)$. Let $\tilde{\mu}_0$ be the marginal of $\tilde{\mu}$, and decompose $\tilde{\mu}$ as

$$\tilde{\mu}(dx \times du) = \tilde{\eta}(x, du)\tilde{\mu}_0(dx)$$

for some regular conditional distribution $\tilde{\eta}$ on U given x. Since $\tilde{\mu}$ is the limit of P_n-feasible solutions, it is P_0-feasible by Theorem 1.

Since by assumption $\tilde{\mu}_0 \ll \pi$, then $\tilde{\mu}_0(\mathcal{D}_\eta) = 0$. Using the argument which produces (2.7), $\eta(x, \cdot)$ is a version of $\tilde{\eta}(x, \cdot)$. Since $\tilde{\mu}$ is P_0-feasible and has decomposition $\tilde{\mu}(dx \times du) = \eta(x, du)\tilde{\mu}_0(dx)$, the uniqueness of the stationary distribution induced by η implies $\tilde{\mu}_0 = \mu_0$. Therefore, for each $h \in C(E \times U)$,

$$\int_E \int_U h(x, u)\ \overline{\overset{*n}{\eta}}(x, du)\overline{\overset{*n}{\mu_0}}(dx) \to \int_E \int_U h(x, u)\eta(x, du)\mu_0(dx)$$

along the subsequence. By Theorem 4, $\mu = \eta\mu_0$ is P_0-optimal. ∎

3. Numerical solutions using finite-dimensional LP approximations. In this section we approximate infinite-dimensional LP problems with finite-dimensional problems by approximating a jump-diffusion process with a sequence of finite state Markov chains. The convergence results of the previous section indicate that the optimal values for the approximations will be close to the optimal value of the original problem. In addition, the numerical results are consistent with convergence of the optimal controls.

3.1. The bounded follower problem. One of the stochastic control processes studied by Beneš, Shepp, and Witsenhausen [1] is the bounded follower problem. The state of this process is $x + w_t - \xi_t$, in which w is a Brownian motion process, $x + w_t$ gives the location in $\mathbb{R}^1$ at time t of an object, and ξ_t gives the location at time t of something which attempts to follow the object. In this, ξ is an absolutely continuous process whose rate of change is bounded between θ_0 and θ_1. The objective is to minimize the expected discounted square of the difference between the locations of the object and the follower:

$$E\left[\int_0^\infty e^{-\alpha t}(x + w_t - \xi_t)^2\, dt\right].$$

This problem has as its optimal control

$$\dot{\xi}_t = \begin{cases} \theta_0, & x + w_t - \xi_t < \delta, \\ \theta_1, & x + w_t - \xi_t \geq \delta, \end{cases}$$

where δ is determined from the parameters α, θ_0, and θ_1. We refer the reader to [1] for the specific expression for δ, but when $\theta_0 = -\theta_1$, the switch point δ equals 0.

For our examples we modify the stochastic process of the difference in locations $x - \xi + w$ by truncating the state space to an interval $[-b, b]$ and having the process stick at the boundary for an exponential (λ) amount of time, after which it jumps to zero. For simplicity, we take $b = 1$ and $\theta_0 = -1$ and $\theta_1 = 1$. Specifically, when the process is in the interval $(-1, 1)$, it follows the stochastic differential equation

$$dx_t = u_t dt + \sigma dw_t,$$

where w is a standard Brownian motion process and u is a nonanticipating process with $-1 \leq u_t \leq 1$.

3.2. Long-term average criterion. The first example we investigate uses a long-term average criterion for which the exact optimal control and optimal cost are known.

3.2.1. The original LP problem. The state space E, control space U, cost function c, and generator A are

$$E = [-1, 1], \qquad U = [-1, 1], \qquad c(x, u) = x^2, \text{ and}$$

$$Af(x, u) = \left[uf'(x) + \frac{\sigma^2}{2} f''(x) \right] I_{(-1,1)}(x) + \lambda \left[f(0) - f(x) \right] I_{\{\pm 1\}}(x), \tag{3.1}$$

where $\mathcal{D}(A) = C^2(E)$.

The original LP problem P_0 is

$$P_0 : \begin{cases} \text{minimize} & \displaystyle\int_{[-1,1]\times[-1,1]} x^2 \mu(dx \times du) \\ \text{subject to} & \displaystyle\int_{[-1,1]\times[-1,1]} \left\{ \left[uf'(x) + \frac{\sigma^2}{2} f''(x) \right] I_{(-1,1)}(x) \right. \\ & \left. \quad + \lambda \left[f(0) - f(x) \right] I_{\{\pm 1\}}(x) \right\} \mu(dx \times du) = 0 \quad \forall f \in C^2(E), \\ & \mu \in \mathcal{P}(E \times U). \end{cases}$$

Observe that each regular conditional distribution η on the control space specifies a nondegenerate diffusion. It is well known (see [7, section 18, Theorem 1]) that the stationary distribution of a nondegenerate diffusion (with η specified) is absolutely continuous with respect to Lebesgue measure. Thus, condition (C2) is satisfied in this example with the measure π consisting of Lebesgue measure on the interval $(-1, 1)$ and placing unit point masses at the endpoints $\{-1, 1\}$.

To verify the uniqueness of the stationary distribution μ_0 on the state for a given η, we display μ_0. The reader is referred to the appendix of [16] for the verification.

The measure μ_0 has the density in the interval $(-1,1)$

$$p(x) = \begin{cases} p_+(x), & x \geq 0, \\ p_-(x), & x < 0, \end{cases}$$

in which, letting $\overline{u}(x) = \int_U u\,\eta(x,du)$,

$$p_+(x) = \int_x^1 2\lambda K_1 e^{\int_y^x 2\overline{u}(r)dr} dy,$$

$$p_-(x) = \int_{-1}^x 2\lambda K_2 e^{\int_y^x 2\overline{u}(r)dr} dy.$$

In this the constant $K_1 = \frac{c_3}{c_1c_4+c_2c_3}$ and $K_2 = \frac{c_4}{c_1c_4+c_2c_3}$, where

$$c_1 = 1 + 2\lambda \int_{-1}^0 \int_{-1}^x e^{\int_y^x 2\overline{u}(r)dr} dydx,$$

$$c_2 = 1 + 2\lambda \int_0^1 \int_0^x e^{\int_y^x 2\overline{u}(r)dr} dydx,$$

$$c_3 = 2\lambda \int_{-1}^0 e^{\int_y^0 2\overline{u}(r)dr} dydx,$$

$$c_4 = 2\lambda \int_0^1 e^{\int_y^0 2\overline{u}(r)dr} dydx.$$

The masses of μ_0 at the endpoints $\{\pm 1\}$ are $\mu_0(\{-1\}) = K_2$ and $\mu_0(\{1\}) = K_1$.

3.2.2. The exact solution. Helmes and Stockbridge [10] have shown that this formulation is equivalent to the P_0 problem stated in section 3.1. They have also found that the exact optimal solution to this problem takes the form

$$u^*(x) = -\text{sign}(x) \quad \text{when} \quad \lambda < 5.55471$$

and for $\lambda \geq 5.55471$

$$u^*(x) = \begin{cases} -1, & -1 \leq x < -a, \\ 1, & -a \leq x < 0, \\ 0, & x = 0, \\ -1, & 0 \leq x < a, \\ 1, & a \leq x < 1, \end{cases} \tag{3.2}$$

where the switch point a is a function of λ and is the solution to the equation

$$a^2 + a + \left(r - \frac{1}{2}\right)(e^{2a} - 1) = 0,$$

where

$$r = \frac{\frac{1}{12}[-8a^3 - 6a^2 - 6a + 6a^2e^{(2a-2)} - 6ae^{(2a-2)} + 3e^{(2a-2)} + 3e^{2a} - 2] + \frac{1}{\lambda}}{\frac{1}{2}[-4a + e^{2a} + e^{(2a-2)}] + \frac{1}{\lambda}}.$$

Values of a for some selected values of λ are given in Table 1.

TABLE 1
Values of the switch point a.

λ	a
10	.88706
20	.80607
100	.72990
1000	.71066

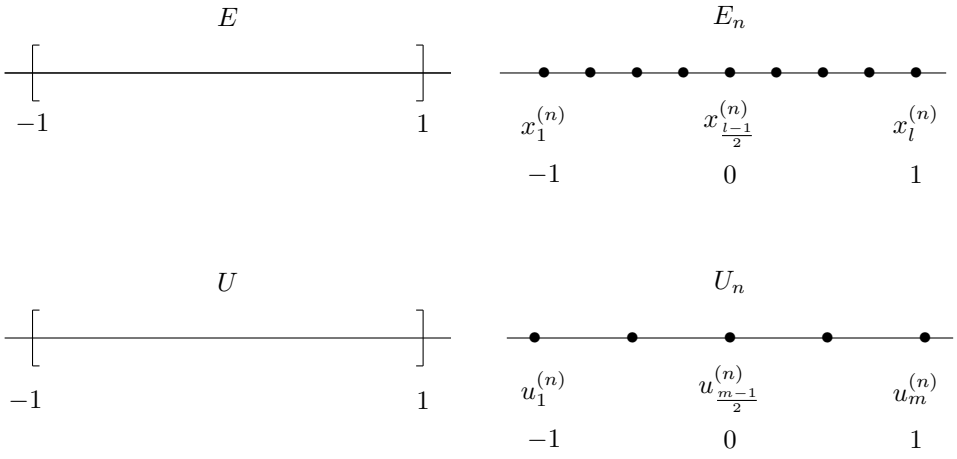
3.2.3. The approximating LP problems P_n. Let $l = l(n)$ and $m = m(n)$ be odd integers such that $l, m \to \infty$ as $n \to \infty$. (This allows different discretizations for the E and U spaces.) Let $h_l = 2/(l-1)$ and $h_m = 2/(m-1)$, and note that h_l and $h_m \to 0$ as $n \to \infty$. Define

$$x_1^{(n)} = -1, \quad x_i^{(n)} = -1 + (i-1)h_l \quad (i = 2, \ldots, l-1), \quad x_l^{(n)} = 1;$$

$$u_1^{(n)} = -1, \quad u_i^{(n)} = -1 + (i-1)h_m \quad (i = 2, \ldots, m-1), \quad u_m^{(n)} = 1,$$

and set

$$E_n = \left\{x_1^{(n)}, x_2^{(n)}, \ldots, x_l^{(n)}\right\} \qquad \text{and} \qquad U_n = \left\{u_1^{(n)}, u_2^{(n)}, \ldots, u_m^{(n)}\right\}.$$



Define

$$\phi_n^1(x) = x_i^{(n)} \quad \text{for } \left(x_i^{(n)} - \frac{1}{2}h_l\right) \le x < \left(x_i^{(n)} + \frac{1}{2}h_l\right), \ \ i = 1, 2, \ldots, l, \quad x \in E,$$

$$\phi_n^2(u) = u_i^{(n)} \quad \text{for } \left(u_i^{(n)} - \frac{1}{2}h_m\right) \le u < \left(u_i^{(n)} + \frac{1}{2}h_m\right), \ \ i = 1, 2, \ldots, m, \quad u \in U,$$

$$\psi_n^1\left(x_i^{(n)}\right) = x_i^{(n)} \quad \text{for } x_i^{(n)} \in E_n,$$

$$\psi_n^2\left(u_i^{(n)}\right) = u_i^{(n)} \quad \text{for } u_i^{(n)} \in U_n,$$

and $\Phi_n : E \times U \to E_n \times U_n$ and $\Psi_n : E_n \times U_n \to E \times U$ by (2.1). Note that Ψ_n is the identity and embeds $E_n \times U_n$ in $E \times U$. Since $c_n = c \circ \Psi_n$, then $c_n(x_i^{(n)}, u_j^{(n)}) = (x_i^{(n)})^2$. Also, for all n, define $f_n \equiv f|_{E_n}$ and $A_n f_n$ using the finite difference approximations for the derivatives of Af as in section 5.3 of Kushner and Dupuis [14],

$$\begin{aligned} A_n f_n(x_i^{(n)}, u_j^{(n)}) = \Bigg[& \frac{\sigma^2}{2} \frac{f_n\left(x_i^{(n)} + h_l\right) + f_n\left(x_i^{(n)} - h_l\right) - 2f_n\left(x_i^{(n)}\right)}{h_l^2} \\ & + u_j^{(n)^+} \left(f_n(x_i^{(n)} + h_l) - f_n(x_i^{(n)})\right)/h_l \\ & + u_j^{(n)^-} \left(f_n(x_i^{(n)}) - f_n(x_i^{(n)} - h_l)\right)/h_l \Bigg] I_{(-1,1)}(x_i^{(n)}) \\ & + \lambda \left[f_n(0) - f_n(x_i^{(n)})\right] I_{\{\pm 1\}}(x_i^{(n)}), \end{aligned} \tag{3.3}$$

where $u_j^{(n)^+} = u_j^{(n)} \vee 0$ and $u_j^{(n)^-} = -u_j^{(n)} \vee 0$.

The approximating LP problem P_n is therefore

$$P_n : \begin{cases} \text{minimize} & \displaystyle\sum_{i=1}^{l}\sum_{j=1}^{m} \left(x_i^{(n)}\right)^2 \mu^n(x_i^{(n)}, u_j^{(n)}) \\ \text{subject to} & \displaystyle 0 = \sum_{i=1}^{l}\sum_{j=1}^{m} \Bigg\{ \left(\frac{\sigma^2}{2h_l^2} + \frac{u_j^{(n)^+}}{h_l}\right) I_{(-1,1)}(x_i^{(n)}) f_n(x_i^{(n)} + h_l) \\ & \qquad + \left(\frac{\sigma^2}{2h_l^2} - \frac{u_j^{(n)^-}}{h_l}\right) I_{(-1,1)}(x_i^{(n)}) f_n(x_i^{(n)} - h_l) \\ & \qquad + \lambda I_{\{\pm 1\}}(x_i^{(n)}) f_n(0) \\ & \qquad - \Bigg[\left(\frac{\sigma^2}{h_l^2} + \frac{u_j^{(n)^+}}{h_l} - \frac{u_j^{(n)^-}}{h_l}\right) I_{(-1,1))}(x_i^{(n)}) \\ & \qquad\quad + \lambda I_{\{\pm 1\}}(x_i^{(n)}) \Bigg] f_n(x_i^{(n)}) \Bigg\} \mu^n(x_i^{(n)}, u_j^{(n)} \\ & \qquad\qquad \forall f_n \in \mathcal{D}(A_n), \\ & \mu^n \in \mathcal{P}(E_n \times U_n). \end{cases}$$

Conditions (C3)–(C6) are easily verified. In order to verify condition (C7), we rewrite the generator $A_n f_n$ in (12) as

$$\begin{aligned}A_n f_n(x_i^{(n)}, u_j^{(n)}) &= \left(\frac{\sigma^2}{2h_l^2} + \frac{u_j^{(n)^+}}{h_l}\right) I_{(-1,1)}(x_i^{(n)}) f_n(x_i^{(n)} + h_l) \\ &\quad + \left(\frac{\sigma^2}{2h_l^2} - \frac{u_j^{(n)^-}}{h_l}\right) I_{(-1,1)}(x_i^{(n)}) f_n(x_i^{(n)} - h_l) \\ &\quad + \lambda I_{\{\pm 1\}}(x_i^{(n)}) f_n(0) \\ &\quad - \left[\left(\frac{\sigma^2}{h_l^2} + \frac{u_j^{(n)^+}}{h_l} - \frac{u_j^{(n)^-}}{h_l} I_{(-1,1)}(x_i^{(n)})\right) + \lambda I_{\{\pm 1\}}(x_i^{(n)})\right] f_n(x_i^{(n)}).\end{aligned}$$

Let

$$\gamma(x_i^{(n)}) = \left(\frac{\sigma^2}{h_l^2} + \frac{u_j^{(n)^+}}{h_l} + \frac{u_j^{(n)^-}}{h_l}\right) I_{(-1,1)}(x_i^{(n)}) + \lambda I_{\{\pm 1\}}(x_i^{(n)}),$$

and define

$$\begin{aligned}P\left(x_{i+1}^{(n)} | x_i^{(n)}, u_j^{(n)}\right) &= \frac{\frac{\sigma^2}{2h_l^2} + \frac{u_j^{(n)^+}}{h}}{\gamma(x_i^{(n)})} I_{(-1,1)}(x_i^{(n)}), \\ P\left(x_{i-1}^{(n)} | x_i^{(n)}, u_j^{(n)}\right) &= \frac{\frac{\sigma^2}{2h_l^2} + \frac{u_j^{(n)^-}}{h}}{\gamma(x_i^{(n)})} I_{(-1,1)}(x_i^{(n)}), \\ P\left(0 | x_i^{(n)}, u_j^{(n)}\right) &= \frac{\lambda}{\gamma(x_i^{(n)})} I_{\{\pm 1\}}(x_i^{(n)}), \\ P\left(x_k^{(n)} | x_i^{(n)}, u_j^{(n)}\right) &= 0, \qquad k \neq i-1, i+1, \frac{l-1}{2}.\end{aligned}$$

Note that for each $x_i^{(n)}$, $u_j^{(n)}$,

$$P\left(x_k^{(n)} | x_i^{(n)}, u_j^{(n)}\right) \geq 0 \quad \forall k = 1, \ldots, l$$

and

$$\sum_{k=1}^{l} P\left(x_k^{(n)} | x_i^{(n)}, u_j^{(n)}\right) = 1,$$

and therefore we can write

$$A_n f_n(x_i^{(n)}, u_j^{(n)}) = \gamma(x_i^{(n)}) \sum_{k=1}^{l} \left[f_n(x_k^{(n)}) - f_n(x_i^{(n)})\right] P\left(x_k^{(n)} | x_i^{(n)}, u_j^{(n)}\right) \quad \forall x_i^{(n)}, u_j^{(n)}.$$

This has the form of a generator for a finite state, continuous time Markov chain. Since $\sigma^2 > 0$, the chain is irreducible for each fixed control η so there exists a unique stationary distribution on E, and condition (C7) is verified.

3.2.4. Reduction of the P_n-problem constraints. We now show that it is sufficient to consider a finite collection of indicator functions f_i, $i = 1, \ldots, l$. Let

$$g(x) = \begin{cases} (1-x^2)^3, & -1 \le x \le 1, \\ 0 & \text{otherwise}, \end{cases}$$

and define the collection of functions f_i in $\mathcal{D}(A)$ as

$$f_i(x) = g\left(\frac{x - x_i^{(n)}}{h_l/2}\right) \qquad \forall x \in [-1,1]\,.$$

Then $f_i(x) \in C^2(E)$ and $f_i(x)|_{E_n} = I_{\{x_i^{(n)}\}}(x)$ for $x \in E_n$. Denote the restriction of f_i to E_n by $f_i^{(n)}$. Since any function on E_n can be expressed as a linear combination of these indicator functions, it is sufficient to consider the stationarity constraints using only $f_i^{(n)}$, $i = 1, \ldots, l$. Thus, the constraints given in P_n reduce to

∘ For $f_i^{(n)}(x) = I_{\{x_i^{(n)}\}}(x)$ when $x_i^{(n)} \neq 1, -1, 0$,

$$0 = \sum_{j=1}^{m} \left[\left(\frac{\sigma^2}{2h_l^2} + \frac{u_j^{(n)^+}}{h_l} \right) \mu^n(x_{i-1}^{(n)}, u_j^{(n)}) + \left(\frac{\sigma^2}{2h_l^2} - \frac{u_j^{(n)^-}}{h_l} \right) \mu^n(x_{i+1}^{(n)}, u_j^{(n)}) \right.$$
$$\left. - \left(\frac{\sigma^2}{2h_l^2} + \frac{u_j^{(n)^+}}{h_l} - \frac{u_j^{(n)^-}}{h_l} \right) \mu^n(x_i^{(n)}, u_j^{(n)}) \right].$$

∘ For $f_i^{(n)}(x) = I_{\{x_i^{(n)}\}}(x)$, when $x_i^{(n)} = 0$,

$$0 = \sum_{j=1}^{m} \left[\left(\frac{\sigma^2}{2h_l^2} + \frac{u_j^{(n)^+}}{h_l} \right) \mu^n(x_{i-1}^{(n)}, u_j^{(n)}) + \left(\frac{\sigma^2}{2h_l^2} - \frac{u_j^{(n)^-}}{h_l} \right) \mu^n(x_{i+1}^{(n)}, u_j^{(n)}) \right.$$
$$- \left(\frac{\sigma^2}{2h_l^2} + \frac{u_j^{(n)^+}}{h_l} - \frac{u_j^{(n)^-}}{h_l} \right) \mu^n(x_i^{(n)}, u_j^{(n)}) + \lambda \mu^n(-1, u_j^{(n)})$$
$$\left. + \lambda \mu^n(1, u_j^{(n)}) \right].$$

∘ For $f_i^{(n)}(x) = I_{\{x_i^{(n)}\}}(x)$, when $x_i^{(n)} = -1$ (i.e., when $i = 1$),

$$0 = \sum_{j=1}^{m} \left[\left(\frac{\sigma^2}{2h_l^2} - \frac{u_j^{(n)^-}}{h_l} \right) \mu^n(x_{i+1}^{(n)}, u_j^{(n)}) - \lambda \mu^n(x_i^{(n)}, u_j^{(n)}) \right].$$

∘ For $f_i^{(n)}(x) = I_{\{x_i^{(n)}\}}(x)$, $x_i^{(n)} = 1$ (i.e., when $i = l$),

$$0 = \sum_{j=1}^{m} \left[\left(\frac{\sigma^2}{2h_l^2} + \frac{u_j^{(n)+}}{h_l} \right) \mu^n(x_{i-1}^{(n)}, u_j^{(n)}) - \lambda \mu^n(x_i^{(n)}, u_j^{(n)}) \right].$$

In addition to the constraints for stationarity, there is the constraint that μ^n be a probability measure:

$$\sum_{j=1}^{m}\sum_{i=1}^{l}\mu^n(x_i^{(n)},u_j^{(n)})=1 \qquad \text{and} \qquad \mu^n(x_i^{(n)},u_j^{(n)})\geq 0 \quad \forall i,j\,.$$

3.2.5. P_n-optimal solution. The P_n problem is solved using a program written in SAS which utilizes the LP procedure. Since in this example the optimal control only takes the values -1, 0 and 1 in U even when $h_m = 0.1$, all the numerical approximations presented here therefore use $h_m = 0.5$.

Figure 1 shows the optimal control obtained for the discretized problem P_n ($h_l = .10$, $\lambda = 100$) represented by the dots, and the induced control in the interval $[-1, 1]$ represented by the solid line.

Figures 2–4 illustrate how the optimal control changes as a function of λ. When $\lambda = 5$, the optimal control is $-\text{sign}(x)$, while for $\lambda = 20$ and $\lambda = 100$ the optimal control is of the form (3.2), where the switch location a depends on λ and a decreases as λ increases.

Table 2 gives the interval in which the positive "switch point" occurs for P_n. The results for the negative switch point follow by symmetry. Note that the P_0 switch point a falls within the switch interval for P_n.

TABLE 2
"Switch point" as a function of λ, $h = .01$.

		$\lambda = 10$	$\lambda = 20$	$\lambda = 100$	$\lambda = 1000$
P_n	$h_l = .05$	(.85, .90)	(.80, .85)	(.70, .75)	(.70, .75)
switch	$h_l = .02$	(.88, .90)	(.80, .82)	(.72, .74)	(.70, .72)
interval	$h_l = .01$	(.88, .89)	(.80, .81)	(.72, .73)	(.71, .72)
P_0	switch	.88706	.80607	.72990	.71066

TABLE 3
Optimal values for the long-term average cost.

		$\lambda = 5$	$\lambda = 10$	$\lambda = 100$
	$h_l = .20$	.218321	.173058	.113664
	$h_l = .10$	.208359	.167671	.118411
P_n	$h_l = .05$	.201667	.162920	.119131
	$h_l = .02$	.197067	.160184	.118735
	$h_l = .01$	.195432	.159153	.118452
P_0		.193746	.158037	.117969

The values for the cost function for three values of λ and five values of h_l are given in Table 3. Note that as the mesh size decreases, the P_n-optimal values well-approximate the P_0-optimal value.

3.3. Discounted cost criterion. For this example, we examine a discounted cost of the processes considered in the previous section. Note that this is the same cost criterion used by Beneš, Shepp, and Witsenhausen [1] in the bounded follower problem, though the dynamics have been modified. In this section, α denotes the discount rate.

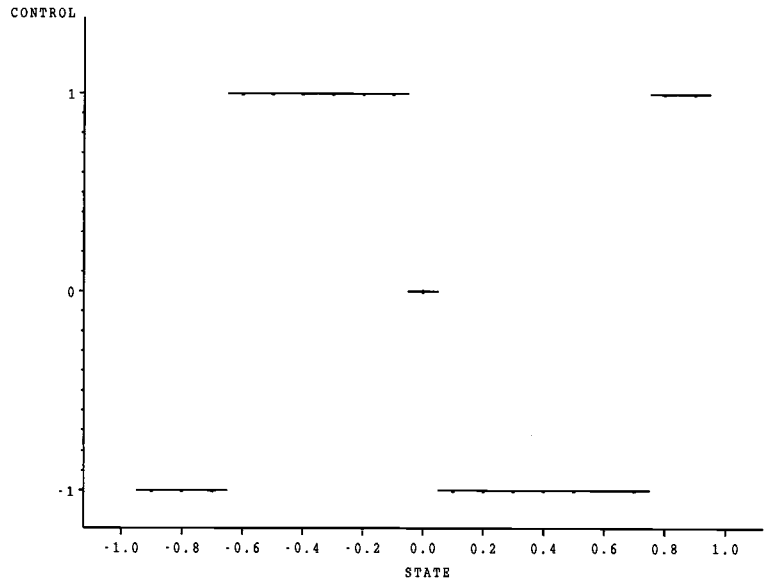


FIG. 1. *Long-term average (LTA) optimal control induced by* P_n, $h_l = .10$, $\lambda = 100$.

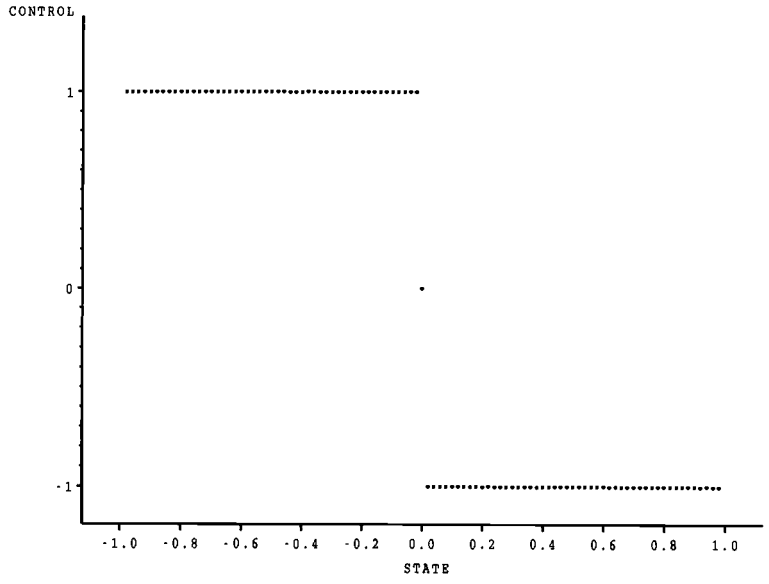


FIG. 2. P_n *optimal control for LTA,* $\lambda = 5$, $h = .02$.

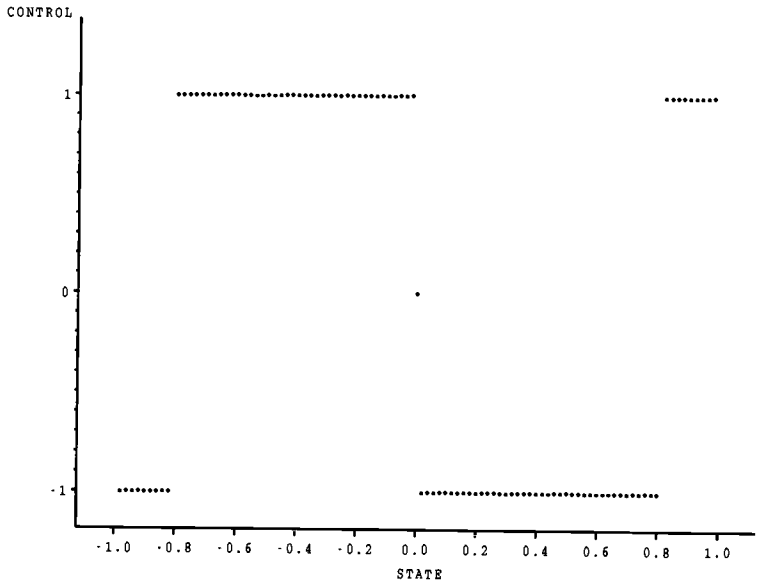


FIG. 3. *P_n optimal control for LTA, $\lambda = 20$, $h = .02$.*

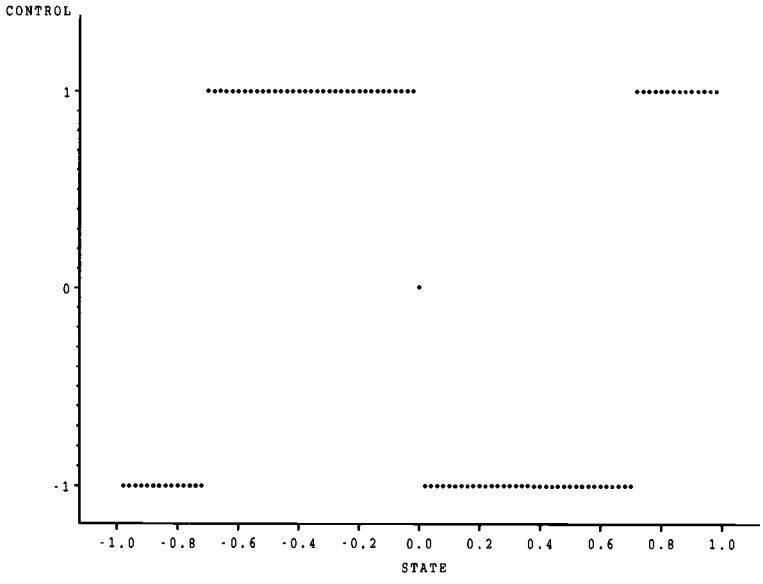


FIG. 4. *P_n optimal control for LTA, $\lambda = 1000$, $h = .02$.*

3.3.1. The original LP problem. Kurtz and Stockbridge [13] show that the LP formulation of the discounted control problem adjusts the cost by a factor of α and introduces a jump term into the generator of the process. The original LP problem P_0 is

$$P_0 : \begin{cases} \text{minimize} & \dfrac{1}{\alpha}\displaystyle\int_{[-1,1]\times[-1,1]} x^2 \mu(dx \times du) \\ \text{subject to} & \displaystyle\int_{[-1,1]\times[-1,1]} \left\{ \left[uf'(x) + \frac{\sigma^2}{2} f''(x) \right] I_{(-1,1)}(x) \right. \\ & \qquad + \lambda \left[f(0) - f(x) \right] I_{\{\pm 1\}}(x) \\ & \qquad \left. + \alpha \left[f(0) - f(x) \right] \right\} \mu(dx \times du) = 0 \\ & \qquad\qquad \forall f \in C^2(E), \\ & \mu \in \mathcal{P}(E \times U). \end{cases}$$

3.3.2. The approximating LP problem P_n. Using the same Markov chain approximations, we have

$$P_n : \begin{cases} \text{minimize} & \dfrac{1}{\alpha} \displaystyle\sum_{i=1}^{l} \sum_{j=1}^{m} \left(x_i^{(n)} \right)^2 \mu^n(x_i^{(n)}, u_j^{(n)}) \\ \text{subject to} & 0 = \displaystyle\sum_{i=1}^{l} \sum_{j=1}^{m} \left\{ \left(\frac{\sigma^2}{2h_l^2} + \frac{u_j^{(n)^+}}{h_l} \right) I_{(-1,1)}(x_i^{(n)}) f_n(x_i^{(n)} + h_l) \right. \\ & \qquad + \left(\dfrac{\sigma^2}{2h_l^2} - \dfrac{u_j^{(n)^-}}{h_l} \right) I_{(-1,1)}(x_i^{(n)}) f_n(x_i^{(n)} - h_l) \\ & \qquad + \lambda I_{\{\pm 1\}}(x_i^{(n)}) f_n(0) \\ & \qquad - \left[\left(\dfrac{\sigma^2}{h_l^2} + \dfrac{u_j^{(n)^+}}{h_l} - \dfrac{u_j^{(n)^-}}{h_l} \right) I_{(-1,1)}(x_i^{(n)}) \right. \\ & \qquad \left. + \lambda I_{\{\pm 1\}}(x_i^{(n)}) \right] f_n(x_i^{(n)}) \\ & \qquad \left. + \alpha \left[f_n(0) - f_n(x_i^{(n)}) \right] \right\} \mu^n(x_i^{(n)}, u_j^{(n)}) \quad \forall f_n \in \mathcal{D}(A_n), \\ & \mu^n \in \mathcal{P}(E_n \times U_n). \end{cases}$$

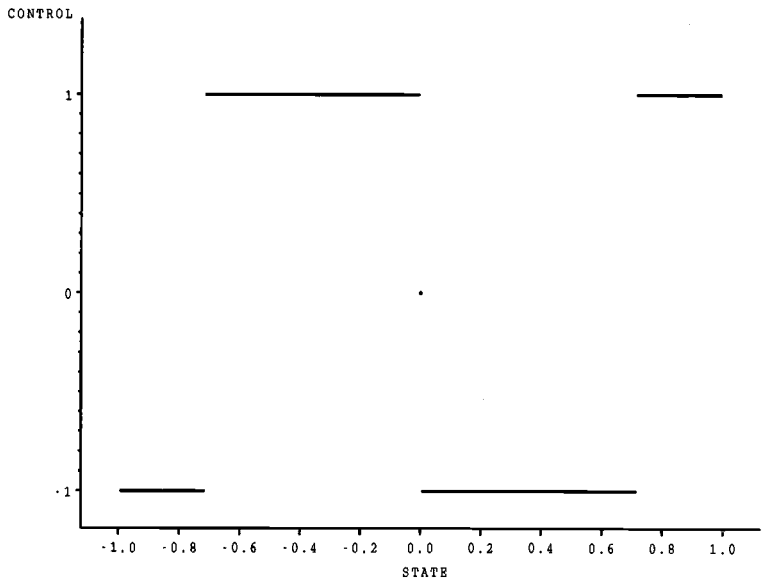


FIG. 5. *P_n optimal control for the discounted problem, $\lambda = 1000$, $\alpha = 0.3$, $h_l = .01$.*

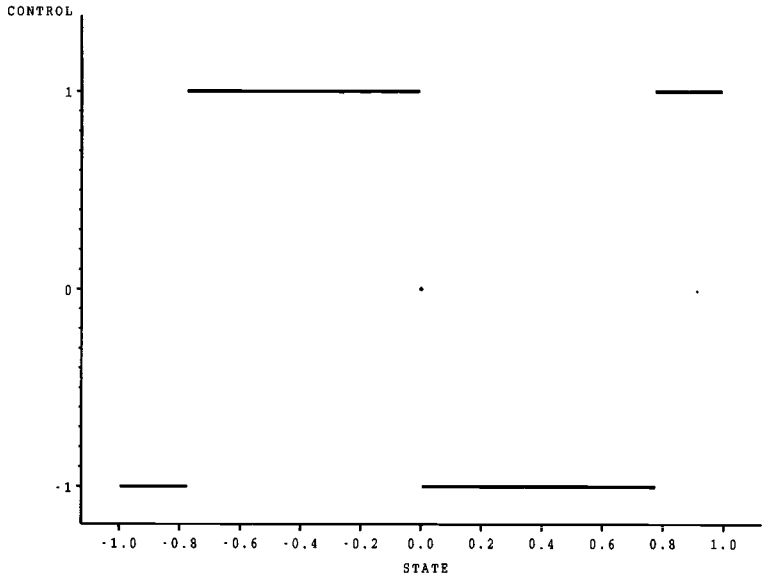


FIG. 6. *P_n optimal control for the discounted problem, $\lambda = 1000$, $\alpha = 20.0$, $h_l = .01$.*

The same reduction to a finite number of constraints can be adopted using the functions $\{f_i^{(n)}\}$ for the long-term average LP problem P_n.

3.3.3. P_n-optimal solutions. The numerical solution to the discounted problem takes the same form as the one for the long-term average problem. In the case of $\lambda \leq 5.55471$, the optimal control is the $u(x) = -\text{sign}(x)$ for every value of α.

Figures 5 and 6 illustrate the effect of α on the optimal control when $\lambda = 1000$. In these examples, $h_l = .01$ and α takes the values 0.3 and 20. Note that the "switch point" for the control that occurs in the (.71, .72) interval in Table 2 ($\alpha = 0$) does not change when α takes the value .3, but it changes to (.77, .78) when α takes the value 20. As the discount rate increases, the switch point moves closer to the boundary. It should be noted that a value of $\alpha = 0.3$ corresponds to an annual inflation rate of 35% and $\alpha = 20$ corresponds to a *daily* inflation rate of 5.6%.

One approach to solving long-term average control problems is to begin with the discounted problem, normalize the cost by multiplying by α, and let α go to zero. Table 4 gives the values of the α-normalized cost function for three values of λ. Note that as $\alpha \to 0$, the α-normalized cost function for the discounted problem converges to the cost function for the long-term average problem. (These values are calculated using $h = .01$.)

TABLE 4
α-normalized P_n-values.

	$\lambda = 5$	$\lambda = 10$	$\lambda = 1000$
$\alpha = 20$	.801494	.670752	.053619
$\alpha = 10$	.681642	.523903	.064713
$\alpha = 5$	.543192	.389026	.080350
$\alpha = 1$	.301076	.216418	.104629
$\alpha = .3$	.230261	.177057	.110719
$\alpha = .1$	.207339	.165193	.112596
$\alpha = .05$	.201449	.162582	.113025
$\alpha = 0$	.201667	.159153	.113559

Acknowledgments. The authors would like to thank Professor Tom Kurtz for helpful discussions on the results in Theorem 2 and the referees for their helpful suggestions and additional references.

REFERENCES

[1] V.E. Beneš, L.A. Shepp, and H.S. Witsenhausen, *Some solvable stochastic control problems*, Stochastics, 4 (1980), pp. 39–83.
[2] A.G. Bhatt and V.S. Borkar, *Occupation measures for controlled Markov processes: Characterization and optimality*, Ann. Probab., 24 (1996), pp. 1531–1562.
[3] E.V. Denardo, *On linear programming in a Markov decision problem*, Management Sci., 16 (1970), pp. 281–288.
[4] C. Derman, *On sequential decisions and Markov chains*, Management Sci., 9 (1962), pp. 16–24.
[5] S.N. Ethier and T.G. Kurtz, *Markov Processes: Characterization and Convergence*, John Wiley, New York, 1986.
[6] M.K. Ghosh, A. Arapostathis, and S.I. Marcus, *Optimal control of switching diffusions with application to flexible manufacturing systems*, SIAM J. Control Optim., 31 (1993), pp. 1183–1204.
[7] I.I. Gihman and A.V. Skorohod, *Stochastic Differential Equations*, Springer-Verlag, New York, 1972.

[8] R. González and E. Rofman, *An algorithm to obtain the maximum solution of the Hamilton-Jacobi equation*, in Optimization Techniques (Proc. 8th IFIP Conf., Wurzburg, 1977), Part I, Lecture Notes in Control and Inform. Sci. 6, Springer-Verlag, New York, 1978, pp. 109–116.

[9] A.C. Heinricher and R.H. Stockbridge, *Optimal control and replacement with state-dependent failure rate: An invariant measure approach*, Ann. Appl. Probab., 3 (1993), pp. 380–402.

[10] K. Helmes and R.H. Stockbridge, *Numerical comparison of controls and verification of optimality for stochastic control problems*, submitted.

[11] O. Hernández-Lerma, J.C. Hennet, and J.B. Lasserre, *Average cost Markov decision processes: Optimality conditions*, J. Math. Anal. Appl., 158 (1991), pp. 396–406.

[12] T.G. Kurtz, *Martingale problems for controlled processes*, in Stochastic Modelling and Filtering, Lecture Notes in Control and Inform. Sci. 91, Springer-Verlag, Berlin, 1987, pp. 75–90.

[13] T.G. Kurtz and R.H. Stockbridge, *Existence of Markov controls and characterization of optimal Markov controls*, SIAM J. Control Optim., 36 (1998), pp. 609–653.

[14] H.J. Kushner and P.G. Dupuis, *Numerical Methods for Stochastic Control Problems in Continuous Time*, Springer-Verlag, New York, 1992.

[15] A.S. Manne, *Linear programming and sequential decisions*, Management Sci., 6 (1960), pp. 259–267.

[16] M.S. Mendiondo, *Approximation of Infinite-Dimensional Linear Programming Problems*, Ph.D. thesis, University of Kentucky, Lexington, KY, 1997.

[17] J.E. Rubio, *Solution of nonlinear control problems in Hilbert spaces by means of linear programming techniques*, J. Optim. Theory Appl., 30 (1980), pp. 643–661.

[18] J.E. Rubio, *The global control of nonlinear diffusion equations*, SIAM J. Control Optim., 33 (1995), pp. 308–322.

[19] R.H. Stockbridge, *Time-average control of martingale problems: A linear programming formulation*, Ann. Probab., 18 (1990), pp. 206–217.

[20] A.M. Vershik, *Some remarks on the infinite-dimensional problems of linear programming*, Russian Math. Surveys, 29 (1970), pp. 117–124.

[21] H.M. Wagner, *On the optimality of pure strategies*, Management Sci., 6 (1960), pp. 268–269.

[22] P. Wolfe and G.B. Dantzig, *Linear programming in a Markov chain*, Oper. Res., 10 (1962), pp. 702–710.

[23] L.C. Young, *On approximation by polygons in the calculus of variations*, Proc. Roy. Soc. London, Ser. A, 141 (1933), pp. 325–341.

SIAM J. SCI. COMPUT.
Vol. 19, No. 1, pp. 227–245, January 1998

016

DYNAMIC THICK RESTARTING OF THE DAVIDSON, AND THE IMPLICITLY RESTARTED ARNOLDI METHODS*

ANDREAS STATHOPOULOS†, YOUSEF SAAD†, AND KESHENG WU†

Abstract. The Davidson method is a popular preconditioned variant of the Arnoldi method for solving large eigenvalue problems. For theoretical as well as practical reasons the two methods are often used with restarting. Frequently, information is saved through approximated eigenvectors to compensate for the convergence impairment caused by restarting. We call this scheme of retaining more eigenvectors than needed "thick restarting" and prove that thick restarted, nonpreconditioned Davidson is equivalent to the implicitly restarted Arnoldi. We also establish a relation between thick restarted Davidson and a Davidson method applied on a deflated system. The theory is used to address the question of which and how many eigenvectors to retain and motivates the development of a dynamic thick restarting scheme for the symmetric case, which can be used in both Davidson and implicit restarted Arnoldi. Several experiments demonstrate the efficiency and robustness of the scheme.

Key words. Davidson method, Arnoldi method, Lanczos method, implicit restarting, deflation, eigenvalue, preconditioning

AMS subject classification. 65F15

PII. S1064827596304162

1. Introduction.

The computation of a few eigenpairs of large, sparse, eigenvalue problems $Ax = \lambda x$ is central to many scientific applications [19]. The Arnoldi method and its equivalent in the symmetric case, the Lanczos method, have been the traditional approach to solving these problems. Preconditioning, through some shift-and-invert technique [22], is frequently employed to improve robustness. A different approach is followed by the generalized Davidson (GD) method [8, 16, 6] which is a popular preconditioned variant of the Lanczos iteration. Instead of using a three-term recurrence to build an orthonormal basis for the Krylov subspace, the GD algorithm obtains the next basis vector by explicitly orthogonalizing the preconditioned residual $(M - \lambda I)^{-1}(A - \lambda I)x$ against the existing basis. A straightforward extension to the nonsymmetric case has also been studied in [21]. When $M = A$, the preconditioned residual yields back x, thus providing no improvement. The Jacobi–Davidson (JD) modification, proposed in [23], suggests that the proper way to precondition the residual is through an operator with range orthogonal to x. The GD and its JD modification can be regarded as two ways of improving convergence and robustness at the expense of a more complicated step.

Often, eigenvalue problems are very large and ill conditioned. As a result, eigenvalue methods require a large number of steps and need to save all the vector iterates for extracting the eigenvectors. Such cases exhibit overwhelming storage requirements. In addition, the Lanczos and Arnoldi processes, which traditionally had been considered without restarting, are plagued by orthogonality problems and spurious solutions. For the above reasons many restarting variants are used in practice [7, 18, 24, 2]. The GD method improves convergence and solves some of the aforementioned problems

*Received by the editors May 14, 1996; accepted for publication (in revised form) February 3, 1997. This research was supported in part by NSF grants DMR–9217287 and ASC 95-04038, and the Minnesota Supercomputer Institute.

http://www.siam.org/journals/sisc/19-1/30416.html

†Department of Computer Science, University of Minnesota, 4-192 EE/CSci Bldg., Minneapolis, MN 55455-0154 (andreas@cs.umn.edu, saad@cs.umn.edu, kewu@mail.cs.umn.edu).

through orthogonalization and preconditioning. However, the number of iterations can still grow very large and cause similar storage problems. The problem is actually aggravated in the symmetric case, where the better theoretical framework and software has led researchers to consider matrices of huge size that allow only a few vectors to be stored. The GD method also can be restarted every time the basis contains m vectors (GD(m)). If the l lowest eigenvalues are needed, the l lowest Ritz values are computed at the mth step, and their corresponding Ritz vectors are used as initial guesses for the restarted GD iteration.

Truncating the Krylov sequence is expected to impair the convergence rate of the method. There are two main reasons: the new vectors entering the basis repeat some of the information that was discarded when restarting, and the Rayleigh–Ritz procedure does not minimize over the whole Krylov subspace. There has been much discussion about the problems caused by restarted methods for both linear systems and eigenvalue problems [27, 20, 24]. Some methods tend to save additional information at each restart [14, 3, 11]. For the Davidson method, Murray, Racine, and Davidson [17] and Van Lenthe and Pulay [29] have proposed restarting with two vectors per required Ritz vector with some success. In an effort to minimize execution time, Crouzeix, Philippe, and Sadkane [6] have proposed a dynamically chosen size m.

Recently, "implicit restarting" has gained popularity as a means of improving convergence of the restarted Arnoldi procedure [24]. By using $p = m - k$ steps of the implicit QR algorithm on the Hessenberg matrix, the basis is truncated down to k vectors. It turns out that the k new basis vectors can be considered the Arnoldi vectors obtained from a polynomially transformed starting vector. This is the basis of the popular eigenvalue package ARPACK [13]. Preconditioners for eigenvalue problems usually vary between steps, in which case the implicitly restarted Arnoldi (IRA(k, m)) is not straightforward to apply. Further, in case of the GD(m) where the residual is preconditioned, the Ritz vectors cannot be described with a polynomial of A. Clearly, a new restarting scheme is needed.

In this paper, we study an extension to the IRA(k, m) technique for the GD(m), which we call "thick restarting" and denote by GD(k, m), and which depends on an integer parameter k. GD(k, m) restarts with k Ritz vectors instead of the l wanted ones, where $l \leq k < m$. The principle idea is mentioned by Kosugi in [11], Sleijpen and van der Vorst in [23], and Morgan in [15]. In the literature, the benefits of IRA(k, m) are studied in relation to the polynomially transformed initial vector. This paper addresses the question of which and how many Ritz vectors should be kept. The theory presented motivates a dynamic strategy of thick restarting that can be used in both IRA(k, m) and GD(k, m). Although the results are proved for the nonpreconditioned case, the idea of thick restarting is readily applicable to the preconditioned GD(k, m) and similar behavior is expected. Compared with IRA(k, m), GD(k, m) can also assume any number of initial guesses and/or enhancements of the basis through arbitrary vectors during the procedure.

After briefly presenting the IRA(k, m) and GD(k, m) algorithms in section 2, in section 3 we prove as an extension to [15] that in the absence of preconditioning, and for arbitrary targeting scheme of GD(k, m), the IRA(k, m) using the Ritz values as shifts and GD(k, m) are equivalent, in the sense that their basis vectors span exactly the same space. In section 4 a theorem is proved that relates the IRA(k, m), and thus GD(k, m), with an Arnoldi process applied on an approximately deflated initial vector. This extends the ideas that appeared recently in [28]. In section 5, a dynamic choice of k is derived for the symmetric case, where the rate of convergence is described

by well-known bounds. In section 6, numerical experiments on matrices from the Harwell–Boeing collection demonstrate the effectiveness of GD(k, m).

2. The restarted Arnoldi and Davidson methods. Throughout this paper we assume that the matrix A is diagonalizable, of order N, with eigenpairs (λ_i, x_i). We look for l outermost eigenpairs (e.g., lowest or highest in the symmetric case). The Arnoldi and Davidson methods use a basis size of $m > l$. The following descriptions of the algorithms serve for establishing the notation. For theoretical and implementation details refer to [24, 13, 8, 16, 6, 23]. For all quantities the superscripts in parentheses denote the corresponding restarting step. These superscripts are dropped whenever there is no ambiguity.

Restarted Arnoldi's method in its simplest form can be expressed as follows.

ALGORITHM 2.1 Restarted Arnoldi.

0. *Start: Choose initial unit vector* $v^{(0)}$
1. *For* $s = 0, 1, \ldots$ *Do*
2. $\quad v_1 = v^{(s)}$, $V_1^{(s)} = \{v_1\}$
3. $\quad$ *For* $j = 1, \ldots, m$ *Do*
4. $\quad\quad h_{ij} = (Av_j, v_i),\ i = 1, \ldots, j,$
5. $\quad\quad w_j = Av_j - \sum_{i=1}^{j} h_{ij} v_i$
6. $\quad\quad h_{j+1,j} = \|w_j\|_2$, *if* $h_{j+1,j} = 0$ *stop.*
7. $\quad\quad v_{j+1} = w_j / h_{j+1,j}$
8. $\quad$ *Enddo*
9. $\quad$ *Compute the wanted eigenpairs* $(\mu_i^{(s)}, y_i^{(s)})$ *of* $H_m^{(s)} = (h_{i,j})$
   $\quad$ *and the Ritz vectors* $x_i^{(s)} = V_m^{(s)} y_i^{(s)}$, *where* $V_m^{(s)} = \{v_1, \ldots, v_m\}$
10. $\quad v^{(s+1)} = \sum c_i x_i^{(s)}$, *for some* c_i, *and the wanted* $x_i^{(s)}$
11. *Enddo*

The algorithm builds a Hessenberg matrix, from which the approximate eigenpairs are extracted through the Rayleigh–Ritz procedure. For the symmetric case, $H_m^{(s)}$ is a tridiagonal matrix, and a three-term recurrence replaces the above orthogonalization step. A linear combination of the wanted Ritz vectors are used to restart the algorithm. Such a restarting strategy, however, may discard a lot of information and result in degradation of the convergence rate.

Implicitly restarted Arnoldi applies the implicit QR algorithm with the $m - l$ unwanted eigenvalues as shifts to the Hessenberg matrix and uses the generated orthogonal transformations to truncate the basis down to l vectors. Therefore, it avoids the need to restart with a single vector which captures the information for all l eigenvectors. The number of vectors in the new basis after restart may also be larger than l, say k. For the rest of the paper we assume that $l \le k < m$, $p = m - k$, and IRA(k, m) denotes the associated method. An outline of the IRA(k, m) algorithm follows.

ALGORITHM 2.2 Implicitly restarted Arnoldi.

0. *Start: Choose initial vector* $v_1^{(0)}$
1. *Build an initial Arnoldi iteration of* k *steps:* $(V_k^{(0)}, H_k^{(0)})$
2. *For* $s = 0, 1, \ldots$ *Do*
3. $\quad$ *Test for convergence*
4. $\quad$ *Extend* $V_k^{(s)}$ *to* $k + p$ *vectors, taking* p *more Arnoldi steps:* $(V_{k+p}^{(s)}, H_{k+p}^{(s)})$
5. $\quad$ *Choose shifts* $\mu_i,\ i = 1, \ldots, p$
6. $\quad H_{k+p} = Q^T H_{k+p}^{(s)} Q$, *with* Q *the orthogonal matrix obtained through*
   $\quad$ *the implicit QR algorithm with* $\mu_i, i = 1, \ldots, p$ *shifts*

7. *Define* $V_k^{(s+1)} = (V_{k+p}^{(s)} Q)\begin{pmatrix} I_k \\ 0 \end{pmatrix}$, *and*
8. $H_k^{(s+1)} = \begin{pmatrix} I_k & 0 \end{pmatrix} H_{k+p} \begin{pmatrix} I_k \\ 0 \end{pmatrix}$
9. *Enddo*

The power of the IRA(k, m) lies in the following two properties. First,

$$v_1^{(s+1)} = \psi(A) v_1^{(s)} = \prod_{i=1}^{p} (A - \mu_i I) v_1^{(s)}, \tag{2.1}$$

for any choice of shifts μ_i, not limited to the exact shifts (Ritz values), and thus the new Arnoldi iteration starts with a polynomially transformed initial vector. Second, the vectors $v_2^{(s+1)}, \ldots, v_k^{(s+1)}$ can be considered the Arnoldi vectors of the Arnoldi process started with $v_1^{(s+1)}$. Thus, no matrix–vector multiplications are needed for the first k Arnoldi vectors. Among various interpretations, IRA(k, m) can be considered a truncation of the QR algorithm for dense matrices as well as an efficient and robust implementation of the subspace iteration with polynomial transformations.

The Davidson method first appeared as a diagonally preconditioned version of the Lanczos method for the symmetric eigenproblem. Extensions, to both general preconditioners and to the nonsymmetric case, have been given since. The following describes the algorithm for the symmetric case. For the nonsymmetric case, line 5 should also include the computation of the last row of the projection matrix $T_j^{(s)}$. MGS denotes the modified Gram–Schmidt procedure.

ALGORITHM 2.3 Generalized Davidson.

0. *Choose initial unit vectors* $U_l^{(0)} = \{u_1^{(0)}, \ldots, u_l^{(0)}\}$
1. *For* $s = 0, 1, \ldots$ *Do*
2. $\quad w_i^{(s)} = A u_i^{(s)},\ i = 1, \ldots, l-1$
3. $\quad$ *For* $j = l, \ldots, m$ *Do*
4. $\quad\quad w_j^{(s)} = A u_j^{(s)}$.
5. $\quad\quad t_{i,j} = (w_j^{(s)}, u_i^{(s)}),\ i = 1, \ldots, j$, *the last column of* $T_j^{(s)}$
6. $\quad\quad$ *Compute some wanted eigenpair, say* (μ_1, z_1) *of* $T_j^{(s)}$.
7. $\quad\quad x_1 = U_j^{(s)} z_1$ *and* $r = Ax_1 - \mu_1 x_1$, *the Ritz vector and its residual*
8. $\quad\quad$ *Test* $\|r\|$ *for convergence. If satisfied target a new vector.*
9. $\quad\quad$ *Solve* $M_{(s,j)} t = r$, *for* t.
10. $\quad\quad b_{j+1}^{(s)} = MGS(U_j^{(s)}, t)$
11. $\quad$ *Enddo*
12. $\quad$ *Set* $U_k^{(s+1)} = \{x_1, \ldots, x_k\},\ k < m$, *and restart*
13. *Enddo*

The preconditioning is performed by solving the equation at step 9, with $M_{(s,j)}$ approximating $(A - \mu_1 I)$ in some sense. In [23] Sleijpen and van der Vorst show that for stability, robustness, as well as efficiency, the operator $M_{(s,j)}$ should have a range orthogonal to x. This is the Jacobi–Davidson (JD) method, and it solves approximately the projected correction equation ($\|x_1\|_2 = 1$)

$$(I - x_1 x_1^T)(A - \mu_1 I)(I - x_1 x_1^T)\ t = (I - x_1 x_1^T)(\mu_1 I - A) x_1 = -r.$$

The projections can be easily applied if an iterative linear solver is used. For preconditioners which approximate A directly, such as incomplete factorizations and approximate inverses, the above orthogonality condition is enforced through an equivalent formulation known as Olsen method. Since the purpose of this paper is the study of

restarting strategies, we use the general description of GD, and the results are valid whether step 9 is performed through JD or otherwise.

A Davidson step is more expensive than that of the Lanczos and Arnoldi algorithms, to allow for preconditioning. In addition, the Davidson algorithm can start with any number of initial vectors and include in the basis any extra information that can be available during the execution. The targeted eigenpair (i.e., the one chosen for preconditioning) may vary in different steps, allowing for a variable targeting scheme. Finally, it can restart with the approximate eigenvectors, so it does not share the problems of the original restarted Arnoldi. As in IRA(k,m), the Davidson method can also restart with more Ritz vectors than needed. This version is "thick restarting" and we denote it by GD(k,m), where l,k, and m are defined as in IRA(k,m). In the following section, we show that IRA(k,m) and GD(k,m) are equivalent in the nonpreconditioned case, but GD(k,m) offers all the aforementioned advantages and extensions.

3. Thick and implicit restarting. It is known that the Lanczos and the Davidson methods are equivalent when no preconditioning is used. However, this has been pointed out only for the nonrestarted case, where one eigenvalue is sought [16]. Recently, the equivalence of the IRA(k,m) with an Arnoldi method restarting with a Ritz vector and augmented by $k-1$ Ritz vectors has been shown [15]. In this section we prove that, in the nonpreconditioned case, if GD(k,m) and IRA(k,m) start with the same initial vector, they are equivalent for any targeting scheme of GD(k,m).

The first lemma is an extension of Lemma 3.10 in [24], and it is the basis for the equivalence proof. Note that the implicit QR algorithm is applied to any diagonalizable matrix H.

LEMMA 3.1. *Let $\lambda(H)=\{\lambda_1,\ldots,\lambda_k\}\cup\{\mu_1,\ldots,\mu_p\}$ be a disjoint partition of the eigenvalue set of a diagonalizable matrix H. Let $Q=Q_1Q_2\cdots Q_p$, where Q_i is the orthogonal matrix implicitly defined by the shift μ_i in the implicit QR algorithm on H. Then, the first k columns of Q span the same space as the k eigenvectors y_i of H associated with the eigenvalues $\lambda_i,\ i=1,\ldots,k$.*

Proof. After p steps of the implicit QR algorithm, it holds that

$$QR=Q_1Q_2\cdots Q_pR_p\cdots R_2R_1=\prod_{i=1}^{p}(H-\mu_i),$$

where Q_iR_i is the QR decomposition of $H_i-\mu_i$ at the ith step, and $R=(r_{ij})=R_p\cdots R_2R_1$ denotes an upper triangular matrix. Since the shifts $\mu_i,\ i=1,\ldots,p$, are eigenvalues of H, QR is a rank k matrix, and if the decompositions are performed with traditional column pivoting, $r_{ii}\neq 0,\ i=1,\ldots,k$, and $r_{ii}=0,\ i=k+1,\ldots,k+p$. For Hessenberg matrices it is shown in [24] that $q_1=Qe_1$ is in the span of $\{y_1,\ldots,y_k\}$. Using a similar argument, if $e_1=\sum_{j=1}^{k+p}\xi_jy_j$,

$$QRe_1=q_1r_{11}=\sum_{j=1}^{k}\xi_j\prod_{i=1}^{p}(\lambda_j-\mu_i)y_j,$$

and $q_1\in\operatorname{span}\{y_1,\ldots,y_k\}$. Inductively, let $q_1,\ldots,q_s\in\operatorname{span}\{y_1,\ldots,y_k\}$. If $e_{s+1}=\sum_{j=1}^{k+p}\xi_{s,j}y_j$,

$$QRe_{s+1}=\sum_{j=1}^{s}r_{j,s+1}q_j+r_{s+1,s+1}q_{s+1}=\sum_{j=1}^{k}\xi_{s,j}\prod_{i=1}^{p}(\lambda_j-\mu_i)y_j,$$

and since $r_{s+1,s+1} \neq 0$, $q_{s+1} \in \text{span}\{y_1, \ldots, y_k\}$. Since Q is an orthogonal matrix, its first k columns are independent and therefore $\text{span}\{q_1, \ldots, q_k\} = \text{span}\{y_1, \ldots, y_k\}$. ∎

In the special case where the matrix H is the Hessenberg matrix obtained from the Arnoldi procedure, an immediate consequence is the following.

LEMMA 3.2. *If at step s the basis vectors $U_m^{(s)}$ and $V_m^{(s)}$ of GD(k,m) and IRA(k,m), respectively, span the same space, then, after restarting both methods,*

$$\text{span}(V_k^{(s+1)}) = \text{span}(U_k^{(s+1)}).$$

Proof. From the assumption, the Ritz vectors are the same for both methods at the end of the sth step, and after restarting, $U_k^{(s+1)}$ contains the k chosen ones, say X_k. If $Q(1:k)$ are the first k columns of the orthogonal matrix of Lemma 3.1, we have $V_k^{(s+1)} = V_m^{(s)} Q(1:k) = V_m^{(s)} Y(1:k) C = X_k C$, where $Y(1:k)$ are the chosen k eigenvectors of the Hessenberg matrix H_m, and C is some $k \times k$ coefficient matrix. ∎

The above shows the equivalence of the two methods at restart. To conclude the proof we need the following proposition which describes the residuals of the Ritz vectors of the Arnoldi procedure [19].

PROPOSITION 3.3. *At the jth step of inner Arnoldi loop, let y_i be the ith eigenvector of H_j associated with the eigenvalue λ_i, and x_i be the Ritz approximate eigenvector $x_i = V_j y_i$. Then,*

$$(A - \lambda_i I) x_i = h_{j+1,j} e_j^H y_i \; v_{j+1}.$$

THEOREM 3.4. *If GD(k,m) without preconditioning and IRA(k,m) are executed with the same initial vector $v^{(0)}$, and at each restarting the p shifts used in IRA(k,m) are the Ritz values of the Ritz vectors discarded by GD(k,m), then the basis vectors produced by the two methods span the same space, for any targeting scheme of GD(k,m), and thus the methods are equivalent.*

Proof. If the two methods start with the same initial vector and no restarting is used, the vectors built are identical. This is an immediate consequence of Proposition 3.3, for any selection of targets in GD(k,m). This is well established in the literature (see [16, 21]).

For the general case, a simple induction on the number s of restarts is used. From the above, it follows that for $s = 0$, the bases built by IRA(k,m) and GD(k,m) satisfy $V_m^{(0)} = U_m^{(0)}$.

Let, for $s > 0$, $\text{span}(V_m^{(s)}) = \text{span}(U_m^{(s)})$. After restarting both methods, and from Lemma 3.2, $\text{span}(V_k^{(s+1)}) = \text{span}(U_k^{(s+1)})$. As a result, at this k step, the Ritz vectors for both methods are the same, and because of Proposition 3.3, the next expansion vectors for both methods are parallel. Thus it holds, $\text{span}(V_{k+1}^{(s+1)}) = \text{span}(U_{k+1}^{(s+1)})$, and inductively

$$\text{span}(V_m^{(s+1)}) = \text{span}(U_m^{(s+1)}). \qquad ∎$$

A few comments are in order. Lemma 3.1 can be applied to the Hessenberg matrices built by Krylov subspace methods, if these are diagonalizable. This assumption is always satisfied by the tridiagonal matrices built in the symmetric case. This justifies the use of this result in Lemma 3.2 for the nonpreconditioned case.

Further, Lemma 3.1 applies to any non-Hessenberg diagonalizable matrix, and although Lemma 3.2 discusses the IRA(k,m) method, it is true for all methods that

use an implicit restarting scheme. Consequently, implicit restarting can be applied to the projection, full matrix T obtained from the preconditioned basis vectors of GD(k,m). If exact shifts are used, it produces a sequence of vectors that span the same space with the required Ritz vectors. Several numerical examples, however, have shown that this can be an unstable process. The reason is traced back to the forward numerical instability of the QR process. Treatments of the problem have been developed [12], but we find it inexpensive and stable to thick restart with the orthogonal (or orthogonalized in the nonsymmetric case) Ritz vectors.

In the preconditioned case, the application of implicit restarting does not result in a polynomial transformation as in (2.1). Specifically, let $U = \{u_1^{(0)}, \ldots, u_m^{(0)}\}$ be the GD(k,m) basis before restarting, with decomposition $AU = UT + E$ and $U^H E = 0$. Following steps similar to Lemma 3.1 and to those in [24], the first basis vector $u_1^{(1)}$ after the implicit restarting can be expressed as $u_1^{(1)} = \psi(UU^H AUU^H)u_1^{(0)} = U\psi(T)U^H u_1^{(0)}$, where ψ is the polynomial having the implicit shifts as roots. The polynomial transformation involves the projected matrix on the space spanned by U and not the full rank matrix A. An arbitrary choice of shifts may lead to a different polynomial, but there are no clear advantages for doing so.

Finally, in the unlikely case where preconditioning produces a defective projection matrix, both implicit and thick restarting may fail as described earlier. Working with the Schur vectors rather than the Ritz vectors provides a stable solution to the problem. The algorithm has been proposed recently in [10] and consists of a slight modification to the GD(k,m): instead of finding the eigendecomposition of T, a Schur decomposition is computed and the diagonal elements of the upper triangular matrix are used as shifts in the implicit restarting procedure. In this way the algorithm computes a partial Schur decomposition of A. Note that thick restarting can still be applied keeping more Schur vectors than needed.

4. The deflation connection. Krylov methods for linear systems, such as conjugate gradient (CG) and GMRES, demonstrate a superlinear convergence at later iterations. One explanation of this phenomenon is the convergence of the outermost eigenpairs of the matrix, so that each method behaves as if deflation has occurred, resulting in faster convergence. Such observations have appeared as early as in [5], but actual quantification of the behavior appears in [20] and [27, 28]. In the latter papers, the optimality of the CG and GMRES polynomials is employed to relate each method after some iterations with a similar process of the same method on a deflated residual.

Results similar to [28] cannot be applied directly to the residual and eigenvalues in the nonsymmetric Arnoldi, since there is no optimality principle. In the following, we extend the results found in [28] to the Arnoldi method by considering the distance of some eigenvector from the Arnoldi–Krylov subspace. Again, preconditioning is not considered since the space that it creates is not a Krylov subspace. This general result is used later in the context of thick/implicit restarting to justify the expected benefits and to help provide a good choice of k.

For simplicity, let A be a diagonalizable matrix, $X^{-1}AX = \Lambda = \text{diag}(\lambda_i)$, of order N. The results in this section can be extended naturally to the Jordan form of A, following the methodology in [28]. However, the presentation is more involved. Let $v = X\xi$ be the expansion of the starting Arnoldi vector to the eigenvector basis. Also let $\mathcal{K}_k(v)$ be the Krylov subspace of dimension k generated by v. Define three numbers satisfying $l < k' \leq k$, where $k-1$ is the number of steps that a nonrestarted Arnoldi method takes starting from v. We can assume an eigenvalue ordering so that

the first l ones are wanted, and the eigenvalues $l+1,\ldots,k'$ are well approximated by the $k-1$ steps of Arnoldi. Let μ_i be the k Ritz values from this $\mathcal{K}_k(v)$ space. At this point we let the Arnoldi process take p more steps and build the space $\mathcal{K}_{k+p}(v)$. The following shows the ordering of these numbers:

Define $\mathcal{D}^{(k)}$ a diagonal matrix with elements

$$\mathcal{D}^{(k)}_{jj} = \begin{cases} 0, & \text{for } j = l+1,\ldots,k', \\ \prod_{i=l+1}^{k'} \frac{\lambda_j-\lambda_i}{\lambda_j-\mu_i}, & \text{for } j \le l \text{ or } j = k'+1,\ldots,N. \end{cases} \tag{4.1}$$

Assuming the above definitions we have the following theorem.

THEOREM 4.1. *Let x_j be an eigenvector to be approximated from the Krylov subspace $\mathcal{K}_{k+p}(v)$, and $\tilde{x}_j$ be the corresponding Ritz vector from $\mathcal{K}_k(v)$, whose components of $x_{l+1},\ldots,x_{k'}$ have been removed. If these Krylov subspaces can be built, then for any $j = 1,\ldots,l$*

$$\operatorname{dist}(x_j, \mathcal{K}_{k+p}(v)) \le |1 - \mathcal{D}^{(k)}_{jj}| + \|X\mathcal{D}^{(k)}X^{-1}\| \operatorname{dist}(x_j, \mathcal{K}_p(\tilde{x}_j)).$$

Proof. At step $k-1$ of the Arnoldi procedure, the Ritz vector x'_j from $\mathcal{K}_k(v)$ has the following expression:

$$x'_j = q_j(A)v/\|q_j(A)v\|, \quad \text{with}$$
$$q_j(t) = \prod_{i=1, i\ne j}^{k} (t-\mu_i).$$

We define $h(t)$ a polynomial of degree $k-1$ as

$$h(t) = \prod_{i=l+1}^{k'} \frac{(t-\lambda_i)}{(t-\mu_i)} q_j(t).$$

Note that the eigencomponents $l+1,\ldots,k'$ of the vector $h(A)v$ are annihilated. If $\tilde{\xi}_i = 0$ for $i = l+1,\ldots,k'$ and $\tilde{\xi}_i = \xi_i$ otherwise, then $\tilde{x}_j = \frac{1}{\phi} X q_j(\Lambda)\tilde{\xi}$, where ϕ is a normalization factor. Since any vector in $\mathcal{K}_{k+p}(v)$ can be expressed as a polynomial of A applied on v, if π^* is some polynomial of degree p, and e_j is the jth orthocanonical vector, we have

$$\begin{aligned} \operatorname{dist}(x_j, \mathcal{K}_{k+p}(v)) &= \min_{q, deg(q)=k+p-1} \|x_j - q(A)v\| \\ &\le \|x_j - \pi^*(A)h(A)X\xi\| \\ &= \|x_j - X\pi^*(\Lambda)\mathcal{D}^{(k)}q_j(\Lambda)\xi\| \\ &= \|x_j - X\mathcal{D}^{(k)}X^{-1}X\pi^*(\Lambda)q_j(\Lambda)\tilde{\xi}\| \\ &= \|x_j - X\mathcal{D}^{(k)}X^{-1}\pi^*(A)\phi\tilde{x}_j\| \\ &= \|x_j - X\mathcal{D}^{(k)}X^{-1}(Xe_j - Xe_j + \pi^*(A)\phi\tilde{x}_j)\| \\ &\le \|x_j - \mathcal{D}^{(k)}_{jj}x_j\| + \|X\mathcal{D}^{(k)}X^{-1}\|\|x_j - \pi^*(A)\phi\tilde{x}_j\|. \end{aligned} \tag{4.2}$$

The result follows by choosing $\pi^* = \frac{1}{\phi}\pi_d$, where π_d is the polynomial that minimizes the distance of x_j from $\mathcal{K}_p(\tilde{x}_j)$, and assuming $\|x_j\| = 1$. ∎

The term $\|X\mathcal{D}^{(k)}X^{-1}\|$ is bounded as follows [28]:

$$\|X\mathcal{D}^{(k)}X^{-1}\| < k_2(X) \max_{j \neq l+1,\ldots,k'} \prod_{i=l+1}^{k'} \frac{\lambda_j - \lambda_i}{\lambda_j - \mu_i} = k_2(X)F_k,$$

where $k_2(X) = \|X\|\|X^{-1}\|$ is the condition number of the matrix X. If k is large enough, then the approximations μ_i converge to λ_i for $i = 1, \ldots, k'$. Thus, $F_k \to 1$, and $|1 - \mathcal{D}_{jj}^{(k)}| \to 0$. Even when these are not accurately converged, provided that $\mathcal{O}(\text{dist}) < \mathcal{O}(|1-\mathcal{D}_{jj}^{(k)}|)$, the distance behaves similarly to the distance from a deflated Krylov subspace. It should be noted that the above bound is rather pessimistic, since $\mathcal{D}^{(k)}$ converges to a part of the identity matrix and thus $X\mathcal{D}^{(k)}X^{-1}$ converges to a spectral projector.

4.1. Deflation in IRA (k, m). Theorem 4.1 can be applied to the $k+p$ vectors at the end of an IRA(k, m) step. As previously, l eigenpairs are needed, k pairs are retained after each restart, and $p = m - k$ additional vectors are built. Theorem 4.1 applies with the same l, k, p, and $k' = k$:

$$\text{dist}(x_j, \mathcal{K}_{k+p}(v^{(s)})) \leq |1 - \mathcal{D}_{jj}^{(k,s)}| + \|X\mathcal{D}^{(k,s)}X^{-1}\| \text{ dist}(x_j, \mathcal{K}_p(\tilde{x}_j)).$$

Note that the space $\mathcal{K}_k(v^{(s)})$ contains exactly the wanted k Ritz vectors at the end of the previous $s-1$ step. From the comments in section 2, the Krylov space $\mathcal{K}_{k+p}(v^{(s)})$ is built implicitly by only p steps. Therefore, Theorem 4.1 relates the p steps of the deflated method, to p, rather than $k + p$ steps of the original method.

The diagonal elements of $\mathcal{D}^{(k,s)}$ depend on two parameters: k is the number of the initial Krylov steps, and s is the restarting step on which the theorem is applied. Since k in IRA(k, m) is bounded, the reason for convergence of $\mathcal{D}_{jj}^{(k,s)}$ is assumed by s, the step number. It has been proved for the symmetric case, and under certain assumptions for the nonsymmetric case [24], that the retained eigenpairs in IRA(k, m) converge. Thus, $F_k^{(s)} \to 1$ and $|1 - \mathcal{D}_{jj}^{(k,s)}| \to 0$, as $s \to \infty$. After several restarts, the IRA(k, m) method builds a space close to the one built by an IRA(k, m) applied on a system deflated from the eigencomponents $l + 1, \ldots, k$. Because of Theorem 3.4, the GD(k, m) performs in a similar way.

The above results suggest that there are advantages in keeping more vectors at each restart, i.e., using a thicker restart. If only the wanted eigenpairs $(1, \ldots, l)$ are retained at restart, the method does not demonstrate the deflation behavior for any other eigenpairs. At every restarting the current approximations of eigenpairs $(l + 1, \ldots, k + p)$ are annihilated, and thus they do not converge. Frequently, some eigenvalues close to the wanted ones or close to the other end of the spectrum are relatively well approximated before restarting, and if retained, they would have converged soon. Even more undesirable is the fact that these approximations will slowly reappear in the Krylov subspace, since their approximations are not accurate enough to completely annihilate the corresponding eigenvectors. Therefore, thick restarting should almost always be beneficial.

5. Dynamic thick restarting in the symmetric case. In this section we restrict the discussion to the symmetric case where explicit bounds for convergence rates are known. Two difficulties are associated with thick restarting: the choice of

which eigenpairs to retain and how many of them. It is well known that the Arnoldi method constructs vectors with strong components in the direction of the extreme eigenvectors (associated with extreme eigenvalues) and, therefore, close to the few wanted ones. Sleijpen and van der Vorst in [23] argue that the restarted Arnoldi method repeats the information for these extreme eigenpairs that are dispensed in previous iterations, and they propose keeping $l+1,\ldots,k$ eigenvalues closest to the wanted ones. A similar strategy is followed in the implicit restarting of the ARPACK code. We denote this special case of GD(k,m) as TR(k), implying the basis size m.

The preceding discussion suggests that thick restarting should aim at improving the convergence of the method through deflation. TR(k) attempts to increase the gap of the wanted eigenvalues from the rest of spectrum by keeping nearby eigenpairs. The same objective is followed by subspace iteration where the number of vectors determines the rate of convergence. Since IRA(k,m) can be interpreted as an efficient way to perform subspace iteration [12], similar restarting considerations hold. However, convergence depends on the gap ratios of the eigenvalues and, therefore, the other end of the spectrum is also of importance. A more general form of thick restarting would be TR(L,R), where L lowest (leftmost) and R highest (rightmost) eigenvectors are kept.

We need to address the issue of choosing optimal restarting parameters. In ARPACK, k is chosen dynamically, starting from a relatively small number and increasing it every time an eigenvalue converges. This attempts to maintain a "constant" gap, and it is slightly different from the strategy reported in [24], where values of k close to $m/2$ usually gave the best results.

Because of the deflation relation, the thicker the restarting, the larger the part of the spectrum that is deflated. However, the basis size m is limited, and if too many vectors are retained when restarting, the Lanczos process cannot effectively build additional basis vectors. A dynamic choice of the parameters L and R should be able to capture this trade-off. For the Lanczos procedure, convergence is governed by a term involving a Chebyshev polynomial. If p Lanczos steps are taken, the error of the ith eigenvalue involves the following term:

$$\frac{1}{T_p^2(1+2\gamma_i)}, \quad \text{with } \gamma_i = \frac{\lambda_i - \lambda_{i+1}}{\lambda_{i+1} - \lambda_N}.$$

γ_i is the gap ratio of the ith eigenvalue, and for small gap ratios (i.e., difficult problems) the above term behaves as

$$\frac{1}{T_p^2(1+2\gamma_i)} \approx 2e^{-2p\sqrt{\gamma_i}}. \tag{5.1}$$

The L and R thick restarting parameters should maximize the deflated gap ratio $\gamma_i = (\lambda_i - \lambda_{L+1})/(\lambda_{L+1} - \lambda_{N-R})$ and also maximize the number of new Lanczos steps $p = m - L - R$. The trade-off is captured by minimizing the error approximation equation (5.1). Since the actual eigenvalues are not known, the m approximate Ritz values (μ_i) before restarting should be used to estimate the spectrum. Thus, assuming the l lowest eigenpairs are sought, L and R are obtained dynamically by maximizing the following expression:

$$\max_{L=l,\ldots,m,\ R=0,\ldots,m-l,\ L+R<m} (m-L-R)\sqrt{\frac{\lambda_i - \lambda_{L+1}}{\lambda_{L+1} - \lambda_{m-R}}}.$$

We implement a combination of the dynamic restarting and the TR(L) schemes. Similarly to subspace iteration and ARPACK, we keep at least $L' > l$ vectors from the side of the required eigenpairs to guarantee an increased separation gap. In the experiments in the next section the value $L' = 10$ is chosen. The dynamic scheme is adopted for the rest of the vectors, maximizing the above expression for $L = L', \ldots, m$. In this way, we capture the benefits from both strategies. It has been observed that if some unwanted eigenvector has converged, it is usually beneficial to include it in restarting, since this information may be slowly repeated. We do not consider this option and let the dynamic choice of L and R take care of such cases.

For the nonsymmetric GD(k, m) a similar expression may be maximized, where the Ritz values are ordered according to the required objective, i.e., largest modulus, largest real part, etc. Often, this ordering corresponds to the outermost eigenvalues of the spectrum that the Arnoldi method approximates first, and thus similar deflation arguments can be made. However, this may not always be true, and the choice is more ad hoc because of lack of general expressions for convergence rates. The dynamic strategy can also be used in case of preconditioning, although its effects are expected to be less pronounced for two reasons. First, the spectrum of the varying operator is transformed by the preconditioners and, second, the preconditioning equation usually targets one specific eigenvector for correction, offering little improvement to the rest of the eigenvectors. Often, however, the use of less efficient preconditioners does not affect the eigenvalue order significantly, and thick restarting can perform as well in this case. Finally, dynamic thick restarting can be used in both GD(k, m) and in the IRA(k, m) of the ARPACK package.

6. Numerical experiments. In the first part of this section we give a small artificial example which demonstrates the increasing effect of deflation in thick restart TR(k). In the second part, we present results from a large number of tests on the symmetric matrices of the Harwell–Boeing collection [9]. The GD(k, m) code is based on a program published in [25] and the extensions proposed in [26]. It implements a variable block generalized Davidson method, using the reverse communication protocol for matrix–vector multiplication and preconditioning operations. Robust shifting and the Olsen strategy, which is equivalent to the Jacobi–Davidson approach in exact arithmetic [23], are adopted in preconditioning. In the third and fourth parts, the dynamic strategy is used to provide the shifts to the IRA(k, m) of the ARPACK implementation. Results from standard nonsymmetric cases are reported in the third part. In the last part, comparisons with the original ARPACK code, and with the ARPACK code using Leja shifts [2] in the symmetric case, facilitate a discussion on the effects of the basis size.

6.1. Deflation works. The GD(k, m) is applied on an artificially generated diagonal matrix of order 100 and elements:

$$A_{jj} = \begin{cases} j/55, & \text{for } j = 1, \ldots, 8, \\ 19/55 + j/55, & \text{for } j = 9, \ldots, 16, \\ j - 16, & \text{for } j = 17, \ldots, 100. \end{cases} \tag{6.1}$$

The lowest eigenvalues of this matrix are grouped in two clusters of eight equidistant eigenvalues each. The separation between the two groups is equal to the separation of the second group from eigenvalue 17. Figure 6.1 depicts the lowest part of this spectrum. We look for the lowest eigenvalue and allow for 20 basis vectors in all versions of GD(k, m). The history of the logarithm of the eigenvalue error is plotted in Figure 6.2 for various restarting thicknesses of TR(k).

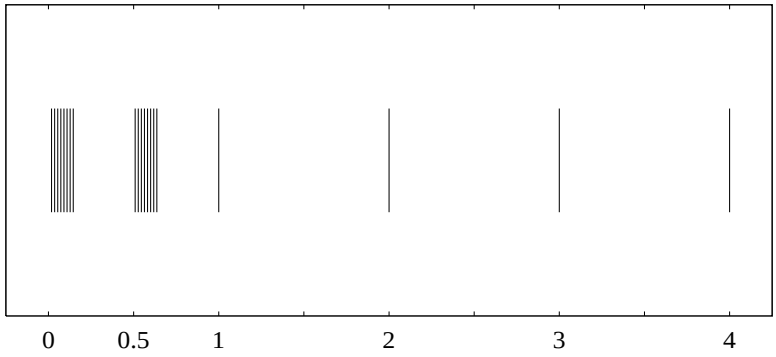


FIG. 6.1. *The lowest* 20 *eigenvalues of the* 100×100 *matrix. The first two clusters contain eight equidistant eigenvalues each. The rest of the* 80 *eigenvalues are the integers from* 5 *up to* 84.

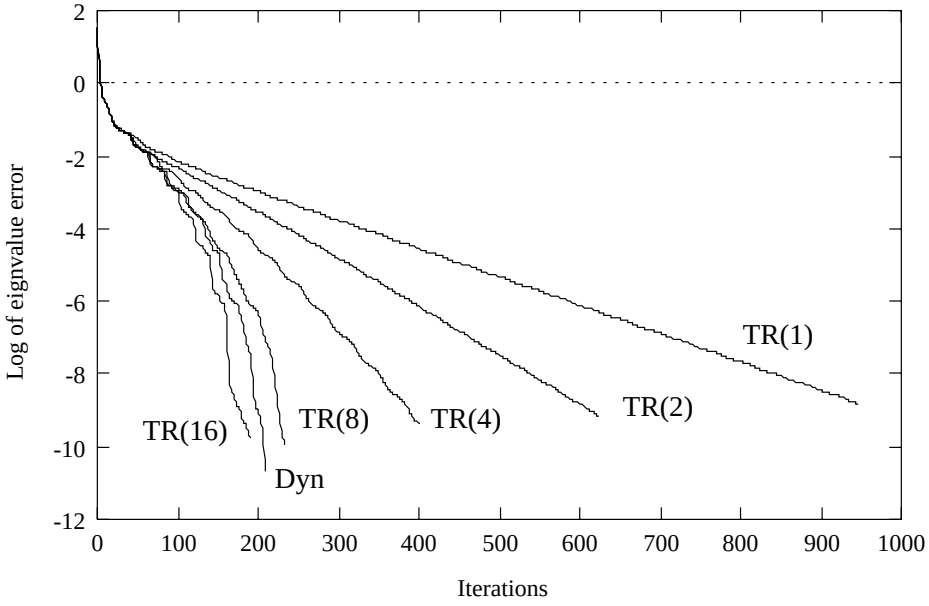


FIG. 6.2. *Effects of thick restarting to the convergence of the generalized Davidson. No preconditioning is used, and the lowest eigenvalue is sought. TR(k) denotes GD(k,*20*).*

As expected, the poor separation of the lowest eigenvalue results in a very slow original GD(20) (or TR(1)) method. A very good approximation of the second eigenvalue is available quite early, and thus when retained (TR(2)), the convergence rate improves by 30%, and similarly with TR(4) and TR(8). The superlinear convergence is more evident in TR(8). In early iterations, higher eigenvalues are not well approximated and TR(8) behaves similarly to TR(1) and TR(2). Later, as better approximations for eigenvalues 2–4 appear, TR(8) is similar to TR(4), and as higher eigenvalues settle down, TR(8) exhibits a concave convergence curve.

Methods TR(k), with $8 < k < 16$, are similar to TR(8) since there is no significant improvement to the deflated gap ratio. In theory, TR(16) should be different because of the large separation between eigenvalues 16 and 17. In practice, however, TR(16) does not perform significantly better than TR(8). The reason is that the Krylov

subspace is of dimension 20, and it is difficult for the 16th Ritz eigenvalue to converge. The dynamic thick restarting, shown as Dyn in the figure, takes advantage of both ends of the spectrum and performs better than TR(8) and close to TR(16), requiring no prior knowledge about the spectrum.

6.2. Harwell–Boeing tests. To confirm the theoretical benefits of thick and dynamic thick restarting, a wide variety of tests have been performed on the symmetric matrices from the Harwell–Boeing collection. This includes a set of 67 matrices with orders ranging from 48 to 15,439. Some of matrices have been derived from eigenvalue problems, but for almost all of them, the lowest end of the spectrum is very poorly conditioned, making them particularly hard test problems. The higher end of the spectrum usually consists of well separated, very large eigenvalues, providing a good test for easy or intermediate problems.

We have compared three different versions of GD(k, m) for both the lower and the higher part of the spectrum. Five eigenvalues are sought and the basis size m for all GD methods is 20. An eigenpair is considered converged when the norm of its residual is less than $10^{-12}\|A\|_F$, where $\|A\|_F$ is the Frobenius norm of its matrix. For the highest eigenvalues only the nonpreconditioned versions of GD(k, m) are considered, while for the lowest ones we consider diagonal and approximate inverse preconditioning. The former is computed at every step as $(\mathrm{diag}(A) - \mu)^{-1}$, and the latter is only computed once as the approximate inverse of A [4]. Since most of the matrices are positive definite, this is a relatively powerful preconditioner.

In Table 6.1, the results from the lower part of the spectrum are reported. A maximum number of 5000 matrix–vector multiplications is allowed. The table does not include any of the diagonal matrices. As it is easily seen, TR(11) outperforms the original Davidson method (TR(5)), except for BCSSTK22. It is usually several times faster, and offers better robustness, converging for six additional matrices. Further, dynamic thick restarting improves both the robustness and the speed in almost all cases. Sometimes the reduction in the matrix–vector multiplication number can be as high as 50 to 70% over TR(11). With diagonal preconditioning TR(11) still outperforms TR(5) in both convergence and robustness. Dynamic thick restarting improves convergence even further, although the improvements are not as impressive as in the nonpreconditioned case. On average, the approximate inverse preconditioner is better than the diagonal one but with several exceptions since it depends on the characteristics of the matrix. Dynamic thick restarting still performs much better than the original approach, and it is relatively faster and more robust than TR(12). However, as mentioned in the previous section, in those cases where approximate inverse works well, the differences between thick and dynamic thick restarting diminish because of the higher quality preconditioner.

Similar behavior of the methods is shown in Table 6.2, where the five largest eigenpairs are required. Dynamic thick restarting improves on the performance of TR(10) which in turn improves on the performance of TR(5). However, the few steps required for the problems in this table do not yield the same impressive improvements as in Table 6.1.

6.3. The effect of the basis size. The dynamic thick restarting strategy, developed for the GD(k, m), can also be used to provide the shifts to the ARPACK code through the supplied reverse communication protocol. Results from this implementation when seeking one lowest eigenpair of the Harwell–Boeing collection appear in Table 6.3. Two tests are performed, one with basis size of 25 and one with basis size of 10. The dynamic restarting significantly improves the speed and robustness of

TABLE 6.1

Comparison of thick (TR(L)) and dynamic thick restarting (Dyn) with original Davidson (TR(5)) on symmetric Harwell–Boeing matrices, with diagonal and approximate inverse preconditioners. The number of matrix–vector multiplications is reported, with a maximum of 5000. *Five smallest eigenvalues are sought. The GD codes use basis size of* 20.

	No preconditioning			Diagonal preconditioning			Approximate inverse		
Matrix	TR(5)	TR(11)	Dyn	TR(5)	TR(10)	Dyn	TR(5)	TR(12)	Dyn
BCSSTK01	-	1675	360	288	132	124	264	96	108
BCSSTK02	-	209	204	-	194	190	188	89	92
NOS4	321	178	171	405	261	244	127	90	91
BCSSTK03	-	-	-	-	3697	1225	-	4699	1685
BCSSTK04	-	-	1905	-	189	188	-	208	221
BCSSTK22	4054	-	1626	-	931	721	-	320	300
LUND A	-	2017	727	858	271	250	3623	394	349
LUND B	-	-	1347	774	396	349	909	381	338
BCSSTK05	1174	975	612	1322	465	409	358	247	251
BCSSTK07	-	-	-	-	-	1401	-	-	3158
BCSSTM07	-	-	3171	1018	406	363	-	2390	1195
NOS5	-	2016	921	2659	1401	819	837	387	354
662 BUS	-	-	-	3220	1482	902	699	307	291
NOS6	-	-	-	-	-	1434	-	-	-
685 BUS	-	-	1793	2473	987	763	486	272	267
NOS7	-	-	-	200	216	194	128	109	94
GR 30 30	259	228	229	248	224	221	204	146	143
NOS3	2179	620	458	2096	878	664	524	253	258
BCSSTK09	-	1206	721	2283+	1508	964	3291	363	352
BCSSTK10	-	-	-	-	-	2808	-	2093	1076
BCSSTM10	498	226	207	448	258	250	3266	3189	2636
BCSSTK27	-	-	-	-	-	3307	-	-	3017
BCSSTM27	-	4455	1689	-	4304	1768	-	636	509
BCSSTK14	-	-	-	-	-	2136	-	-	3723
BCSSTM13	-	-	-	381	285	269	291	183	177
BCSSTK21	-	-	-	-	2568+	1141	1776	877	601
BCSSTK16	3962	1333	676	2410	905	663	752	331	317
BCSSTK18	-	-	-	-	-	3098	-	-	-
BCSSTM25	-	-	-	62	64	55	40	38	37

+ denotes that one eigenpair has been skipped

the native restarting scheme of ARPACK, which for one eigenvalue is the equivalent with thick restart of half the basis size. What is more interesting is that dynamic restarting seems much less sensitive to reduction of the basis size. Similar insensitivity to the basis size has recently been demonstrated through the use of Leja points as shifts in IRA(k, m) [2]. We have implemented the Leja shifts restarting strategy as outlined in [2], and the results appear in Table 6.3. For the small basis size, dynamic thick restarting and Leja shifts are comparable. However, as the basis size increases, the dynamic strategy is more efficient and even more robust. Although Leja shifts may be better for extremely small spaces (less than five vectors), they are harder to implement and they are more expensive to compute.

Experience with the dynamic thick restarting has shown that most of the vectors are retained at every restart, and only three or four are annihilated. The range of the annihilated ones varies from step to step. Figure 6.3 shows the range of eigenvalues which the filtering polynomial covers, as well as the shifts of this polynomial, at every restart for a typical case. We have observed that it is important to have both a small degree polynomial at every restart (i.e., only few eigenvalues annihilated) and to also vary the range from where these shifts are chosen. Therefore, TR(16) does

TABLE 6.2

Comparison of thick (TR(10)) and dynamic thick restarting (Dyn) with original Davidson (TR(5)) on Harwell–Boeing matrices. The number of matrix–vector multiplications is reported. Five largest eigenvalues are sought. The GD codes use basis size of 20.

	No preconditioning		
Matrix	TR(5)	TR(10)	Dyn
BCSSTK01	57	42	38
BCSSTK02	62	49	52
NOS4	176	107	114
BCSSTK03	51	44	43
BCSSTK04	103	84	78
BCSSTK22	106	71	65
LUND A	195	124	120
LUND B	92	66	68
BCSSTK05	81	67	66
NOS1	257	147	133
PLAT362	165	111	114
BCSSTK06	332	114	109
BCSSTK07	332	114	109
BCSSTM07	240	172	155
NOS5	210	117	111
662 BUS	65	54	55
NOS6	123	91	87
685 BUS	31	30	30
NOS7	88	65	68
BCSSTK19	113	100	92
GR 30 30	502	451	396

	No preconditioning		
Matrix	TR(5)	TR(10)	Dyn
NOS2	2236	906	520
NOS3	194	156	150
BCSSTK08	36	35	33
BCSSTK09	316	236	206
BCSSTK10	146	94	90
BCSSTM10	443	151	137
1138 BUS	84	73	75
BCSSTK27	129	89	81
BCSSTM27	130	96	87
BCSSTK11	441	220	200
BCSSTM12	164	115	129
BCSSTK14	195	73	75
PLAT1919	102	94	100
ZENIOS	53	48	48
BCSSTK24	112	118	121
BCSSTK21	1144	418	335
BCSSTK15	-	1374	328
BCSSTK16	99	83	83
BCSSTK17	82	67	62
BCSSTK18	166	86	86
BCSSTK25	45	59	44

not perform as well as dynamic restarting, even though, on average, it retains the same number of vectors. Also, if we force the dynamic restarting to annihilate more than five or six shifts at every restart, the scheme does not perform as well either. The efficiency of the dynamic thick restarting may be attributed to the fact that the filtering polynomial is of low degree and seems to select the best region to dampen, without growing fast outside these regions. The efficient use of the Leja shifts in the ARPACK also exhibits analogous requirements.

Finally, we should point out that the above results compare the number of matrix–vector multiplications of the methods. This is an acceptable performance metric if the matrix–vector operation is expensive. Since, on average, thick restarting uses more vectors in the basis than the original Davidson, its Davidson step is also more expensive. Although improvements like the ones in Table 6.1 justify any increase in the expense of the Davidson step, for easier cases a less aggressive choice of restarting might be more effective.

6.4. Thick restarting in the nonsymmetric case. As in the symmetric case, we can likewise use the dynamic thick restarting scheme to provide the shifts to the nonsymmetric ARPACK code. Results from this implementation applied on the nonsymmetric matrices of the test matrix collection of eigenvalue problems of Bai et al. [1] appear in Table 6.4. All the matrices stem from standard eigenvalue problems, except ODEP400A which is included because it is close to symmetric. Since for almost all examples the rightmost eigenpairs are of interest, we look for five eigenpairs with largest real parts. The convergence threshold for ARPACK is set to 10^{-12}, and a maximum of 5000 matrix–vector multiplications is allowed.

The shifts for thick restarting are chosen similarly to the symmetric case. First, we order the Ritz values according to their real parts. The dynamic scheme works on

TABLE 6.3

Implementation of the Leja shifts and dynamic thick restarting for the ARPACK code. Native is the restarting scheme used internally by ARPACK, Leja(k) refers to implicit restarting with k Leja shifts, and Dyn is the dynamic thick restarting. The number of matrix–vector multiplications is reported for two tests with basis sizes 10 *and* 25 *on Harwell–Boeing matrices. One lowest eigenvalue is sought.*

	ARPACK					
	Basis size of 10			Basis size of 25		
Matrix	Native	Leja(3)	Dyn	Native	Leja(5)	Dyn
BCSSTK01	-	3805	3922	1637	1309	341
BCSSTK02	530	235	198	129	134	124
NOS4	220	136	166	116	114	120
BCSSTM03	1165	1696	298	265	1014	90
BCSSTK04	-	-	-	-	-	2013
BCSSTK22	-	1132	1222	1520	1124	999
BCSSTM22	240	166	149	103	104	89
LUND A	-	2461	1644	2079	1774	759
LUND B	-	1990	2777	3002	1404	1150
BCSSTK05	2810	727	874	766	609	588
BCSSTM06	4675	1462	494	792	529	243
NOS5	-	1123	1546	1494	864	880
BCSSTM20	-	-	-	-	-	896
494 BUS	-	-	-	-	-	3634
662 BUS	-	1642	1547	2443	1429	1108
685 BUS	-	2482	2515	1962	1819	700
NOS3	1210	388	492	402	334	348
BCSSTK09	1140	367	419	337	309	304
BCSSTM10	420	214	274	181	164	174
BCSSTM27	-	2698	1931	2781	1509	1461
BCSSTM11	65	196	41	25	25	25
BCSSTM13	-	4285	3471	-	4459	2995
ZENIOS	60	58	56	90	84	80
BCSSTK16	-	946	1063	1000	774	712
BCSSTK25	-	-	-	-	-	1399

these real parts, yielding the numbers L and R on the real axis. We then supply the corresponding Ritz values as shifts to ARPACK, requiring that conjugate Ritz values are either annihilated together or kept together.

The results show that the thick restarted versions improve efficiency and robustness of the native scheme of ARPACK, and that thicker restarting schemes achieve better efficiencies. This is expected by analogy with the subspace iteration method. The dynamic thick restarting is not uniformly better than the rest as in the symmetric case. In fact, it seems comparable to TR(20) which, on average, keeps the same number of vectors as the dynamic one. As mentioned in section 5, the extreme eigenpairs chosen by the dynamic scheme are based on the ordering of the real parts of the Ritz values and may not always represent the extreme eigenpairs approximated well by the Arnoldi method. In spite of this, dynamic thick restarting is still the most robust of the methods used and shows that the efficiency of the one-sided thick restarting can be improved.

7. Conclusions. Restarting is a necessary technique for solving large eigenvalue problems, which may cause significant convergence deterioration. In this paper we consider a class of restarting techniques which, at every restart, retain more Ritz vectors than needed, and we denote it as "thick restarting." The GD(k, m) and IRA(k, m) are proved to be equivalent in the absence of preconditioning and a relation is given between thick restarted Davidson and a Davidson method applied on a

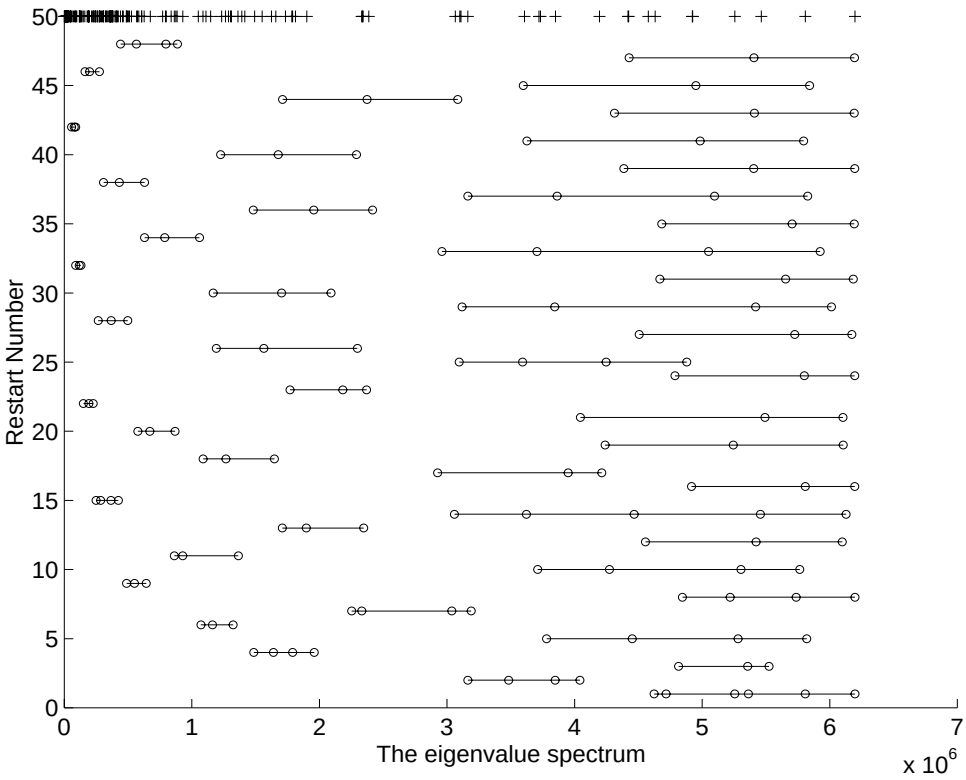


FIG. 6.3. *The range annihilated by the filtering polynomial of dynamic thick restarting at every restart. Each interval includes the m-L-R Ritz values, depicted as circles, which are picked for annihilation by the dynamic scheme. The example matrix is BCSSTK*05 *from Harwell–Boeing and basis size of* 20 *is used. The crosses on the top of the graph represent the location of the eigenvalues on the real axis.*

TABLE 6.4

Implementation of thick and dynamic thick restarting for the nonsymmetric ARPACK code. Native (Nat) is the restarting scheme used internally by ARPACK, (12) *and* (20) *are one-sided thick restarting with* 12 *and* 20 *vectors, respectively, and Dyn is the dynamic thick restarting. The number of matrix–vector multiplications is reported for the test-matrix collection for eigenvalue problems. Five eigenvalue with largest real parts are sought.*

	ARPACK			
Matrix	Nat	(12)	(20)	Dyn
BWM200	558	207	185	180
BWM2000	-	-	-	3999
CDDE5	357	272	252	237
DW2048	681	675	815	495
DW8192	-	-	-	4942
DWA512	118	116	116	113
DWB512	350	298	315	267
GRCAR200	2606	698	524	572
LOP163	383	279	214	242
ODEP400A	1683	837	1005	704
OLM100	548	357	255	316
OLM1000	-	-	-	3602
OLM500	4303	2514	1867	1622
PDE225	343	281	234	254
PDE2961	192	140	124	130

	ARPACK			
Matrix	Nat	(12)	(20)	Dyn
QH768	-	2935	751	881
RDB1250	610	454	145	139
RDB1250L	524	513	436	449
RDB2048	887	181	185	170
RDB2048L	755	615	588	598
RDB3200L	842	738	736	729
RDB450	376	259	90	85
RDB450L	343	295	280	309
RDB800L	429	421	354	391
RW136	170	128	109	108
RW496	247	179	164	168
RW5151	743	514	406	473
TOLS90	-	-	330	1295
TUB100	318	181	154	165
TUB1000	-	4042	3730	1696

deflated system. These theoretical results imply that retaining more outermost Ritz pairs can enhance convergence.

For the symmetric case, the results can be interpreted as an effort to increase the gap ratio for the required eigenvalues. Since the number of basis vectors is limited, the actual objective is to maximize the error reduction between restarts. This gives rise to a dynamic thick restarting technique which applies to IRA(k, m) and to the preconditioned GD(k, m). The extensive numerical experiments demonstrate the efficiency and robustness of the dynamic thick restarting and show that the robustness carries over to the nonsymmetric case. In addition, this scheme seems to be much less sensitive to smaller Krylov subspace dimensions and can be extremely beneficial in very large eigenvalue problems.

Acknowledgments. We are grateful to the referees and to H. A. van der Vorst whose insightful comments improved the presentation significantly. We also acknowledge R. B. Lehoucq for long, revealing discussions on implicit restarting, as well as E. Chow for kindly providing us with his approximate inverse preconditioning code.

REFERENCES

[1] Z. Bai, D. Day, J. Demmel, and J. Dongarra, *A Test Matrix Collection for Non-Hermitian Eigenvalue Problems*, Tech. report, Department of Mathematics, University of Kentucky, Lexington, KY, 1996.

[2] D. Calvetti, L. Reichel, and D. Sorensen, *An implicitly restarted Lanczos method for large symmetric eigenvalue problems*, Electron. Trans. Numer. Anal., 2 (1994), pp. 1–21.

[3] A. Chapman and Y. Saad, *Deflated and Augmented Krylov Subspace Techniques*, Tech. report 95-181, Supercomputing Institute, University of Minnesota, Minneapolis, MN, 1995.

[4] E. Chow and Y. Saad, *Approximate inverse preconditioners via sparse-sparse iteration*, SIAM J. Sci. Comput., to appear.

[5] P. Concus, G. H. Golub, and D. P. O'Leary, *A Generalized Conjugate Gradient Method for the Numerical Solution of Elliptic Partial Differential Equations*, Academic Press, New York, 1976.

[6] M. Crouzeix, B. Philippe, and M. Sadkane, *The Davidson method*, SIAM J. Sci. Comput., 15 (1994), pp. 62–76.

[7] J. Cullum and R. A. Willoughby, *Lanczos Algorithms for Large Symmetric Eigenvalue Computations*, Vol. 2, Programs of Progr. Sci. Comput. 4, Birkhäuser Boston, Boston, MA, 1985.

[8] E. R. Davidson, *The iterative calculation of a few of the lowest eigenvalues and corresponding eigenvectors of large real-symmetric matrices*, J. Comput. Phys., 17 (1975), pp. 87–94.

[9] I. S. Duff, R. G. Grimes, and J. G. Lewis, *Sparse matrix test problems*, ACM Trans. Math. Software, (1989), pp. 1–14.

[10] D. R. Fokkema, G. L. G. Sleijpen, and H. A. van der Vorst, *Jacobi-Davidson Style QR and QZ Algorithms for the Partial Reduction of Matrix Pencils*, Tech. report 941, Department of Mathematics, University of Utrecht, the Netherlands, 1996; SIAM J. Sci. Comput., to appear.

[11] N. Kosugi, *Modification of the Liu-Davidson method for obtaining one or simultaneously several eigensolutions of a large real-symmetric matrix*, J. Comput. Phys., 55 (1984), pp. 426–436.

[12] R. B. Lehoucq, *Analysis and Implementation of an Implicitly Restarted Arnoldi Iteration*, Ph.D. thesis, TR95-13, Department of Computational and Applied Mathematics, Rice University, Houston, TX, 1995.

[13] R. B. Lehoucq, D. C. Sorensen, and P. Vu, *An Implementation of the Implicitly Restarted Arnoldi Iteration that Computes Some of the Eigenvalues and Eigenvectors of a Large Sparse Matrix*, Tech. report, University of Tennessee, Knoxville, TN, Netlib, 1995.

[14] R. B. Morgan, *A restarted GMRES method augmented with eigenvectors*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 1154–1171.

[15] R. B. Morgan, *On restarting the Arnoldi method for large nonsymmetric eigenvalue problems*, Math. Comput., 65 (1996), pp. 1213–1230.

[16] R. B. Morgan and D. S. Scott, *Generalizations of Davidson's method for computing eigenvalues of sparse symmetric matrices*, SIAM J. Sci. Comput., 7 (1986), pp. 817–825.
[17] C. W. Murray, S. C. Racine, and E. R. Davidson, *Improved algorithms for the lowest eigenvalues and associated eigenvectors of large matrices*, J. Comput. Phys., 103 (1992), pp. 382–389.
[18] Y. Saad, *Chebyshev acceleration techniques for solving nonsymmetric eigenvalue problems*, Math. Comp., 42 (1984), pp. 567–588.
[19] Y. Saad, *Numerical Methods for Large Eigenvalue Problems*, Halstead Press, New York, 1992.
[20] Y. Saad, *Analysis of Augmented Krylov Subspace Methods*, Tech. report 95-176, Supercomputing Institute, University of Minnesota, Minneapolis, MN, 1995.
[21] M. Sadkane, *Block-Arnoldi and Davidson methods for unsymmetric large eigenvalue problems*, Numer. Math., 64 (1993), pp. 195–211.
[22] D. S. Scott, *The advantages of inverted operators in Rayleigh-Ritz approximations*, SIAM J. Sci. and Statist. Comput., 3 (1982), pp. 68–75.
[23] G. L. G. Sleijpen and H. A. van der Vorst, *A Jacobi-Davidson iteration method for linear eigenvalue problems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 401–425.
[24] D. C. Sorensen, *Implicit application of polynomial filters in a K-step Arnoldi method*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 357–385.
[25] A. Stathopoulos and C. Fischer, *A Davidson program for finding a few selected extreme eigenpairs of a large, sparse, real, symmetric matrix*, Comput. Phys. Comm., 79 (1994), pp. 268–290.
[26] A. Stathopoulos, Y. Saad, and C. Fischer, *Robust preconditioning of large, sparse, symmetric eigenvalue problems*, J. Comput. Appl. Math., 64 (1995), pp. 197–215.
[27] A. Van der Sluis and H. A. van der Vorst, *The rate of convergence of conjugate gradients*, Numer. Math., 48 (1986), pp. 543–560.
[28] H. A. van der Vorst and C. Vuik, *The superlinear convergence behavior of GMRES*, J. Comput. Appl. Math., 48 (1993), pp. 327–341.
[29] J. Van Lenthe and P. Pulay, *A space-saving modification of Davidson's eigenvector algorithm*, J. Comput. Chem., 11 (1990), pp. 1164–1168.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 4, pp. 1481–1484, July 1998

018

CORRECTIONS TO "A TARGET RECOGNITION PROBLEM: SEQUENTIAL ANALYSIS AND OPTIMAL CONTROL"*

MARK H. A. DAVIS† AND MOHAMMAD FARID‡

Abstract. In the above-mentioned paper [*SIAM J. Control Optim.*, 34 (1996), pp. 2116–2132], despite the correct results, the proof of Theorem 4.6 has a subtle mistake. In the following paragraphs the correct proof, including Lemmas 4.7, 4.8, and 4.9, are given.

Key words. optimal control, viscosity solutions, variational inequality, dynamic programming

AMS subject classifications. 49J40, 49L25, 62C10, 62K05, 62L10, 93C15

PII. S0363012997320304

THEOREM 4.6. *Consider the following variational inequality for all* $(x,\pi) \in \mathbb{R}^n \times [0,1]$:

$$\max\left\{\sup_{u\in U}[-\ell(x,u) - g(x,u)\cdot D_x V(x,\pi)],\ V(x,\pi) - \psi(x,\pi)\right\} = 0, \tag{1}$$

and make all the assumptions of Theorem 4.2*; then if* $\psi(x,\pi) \geq 0\ \forall\ (x,\pi) \in \mathbb{R}^n \times [0,1]$ *there is at most one viscosity solution of* (1).

Without loss of generality we drop π in all the following arguments. We will need the following lemmas.

LEMMA 4.7. *Define* $H(x,p) := \sup_{u\in U}[-\ell(x,u) - g(x,u)\cdot p]$*; then* $H(x,p)$ *is a convex function with respect to* p.

Proof. This is easily verified. □

LEMMA 4.8. *Define*

$$\tilde{H}(x,r,p) := \max\{H(x,p),\ r - \psi(x)\},$$

where

$$\ell(x,u) \geq \delta > 0 \quad \forall\ (x,u) \in \mathbb{R}^n \times U,$$
$$\psi(x) \geq 0 \quad \forall\ x \in \mathbb{R}^n.$$

Then $\tilde{H}(x,V,DV)$ *has a strict subsolution* [BP88]*; i.e.,*

$$\exists w \in C^1(\mathbb{R}^n) \cap BUC(\mathbb{R}^n), B > 0, \quad \text{s.t. } \tilde{H}(x,w,Dw) \leq -\beta < 0 \quad \text{in } \mathbb{R}^n.$$

Proof. It is easily seen that $w(x) \equiv -\delta$ is a strict subsolution; i.e., $\max\{H(x,0),\ -\delta - \psi(x)\} \leq -\delta < 0$. □

LEMMA 4.9. *Assume that* $v_1(x)$ *is a viscosity subsolution of* $\tilde{H}(x,V,DV) = 0$. *Then* $v_\eta = \eta v_1 + (1-\eta)w$ *is a viscosity subsolution of*

$$\tilde{H}(x,V,DV) \leq -(1-\eta)\delta < 0, \tag{2}$$

*Received by the editors April 21, 1997; accepted for publication (in revised form) June 3, 1997; published electronically May 28, 1998.
http://www.siam.org/journals/sicon/36-4/32030.html
†Tokyo-Mitsubishi International Plc, 6 Broadgate, London EC2M 2AA, UK (mark.davis@t-mi.com).
‡Control Engineering Research Center, City University, London EC1V 0HB, UK (m.farid@city.ac.uk).

where $w(x)$ is a strict subsolution and $\eta \in (0,1)$.

Proof. v_1 is a viscosity subsolution of $\tilde{H}(x,V,DV)=0$, so for all $\varphi \in C^1(\mathbb{R}^n)$, if $v_1-\varphi$ attains a local maximum at $x_0 \in \mathbb{R}^n$, then

$$\tilde{H}(x_0, v_1(x_0), D\varphi(x_0)) \le 0.$$

When $v_1-\varphi$ attains a local maximum at x_0, so does $\eta v_1+(1-\eta)w-(\eta\varphi+(1-\eta)w)$ for $\eta \in (0,1)$. Here $w \in C^1(\mathbb{R}^n)$ is a strict subsolution of $\tilde{H}$. Now we show that $\eta v_1+(1-\eta)w$ is a strict viscosity subsolution of $\tilde{H}$:

$$\begin{aligned}
&\tilde{H}(x_0,\ \eta v_1(x_0)+(1-\eta)w(x_0),\ \eta D\varphi(x_0)+(1-\eta)Dw(x_0))\\
&\quad = \max\{H(x_0,\eta D\varphi(x_0)+(1-\eta)Dw(x_0)), \eta v_1(x_0)+(1-\eta)w(x_0)-\psi(x_0)\}\\
&\quad \le \max\{\eta H(x_0,D\varphi(x_0))+(1-\eta)H(x_0,Dw(x_0)), \eta v_1(x_0)+(1-\eta)w(x_0)-\psi(x_0)\}\\
&\quad \le \eta\max\{H(x_0,D\varphi(x_0)),\ v_1(x_0)-\psi(x_0)\}\\
&\qquad +(1-\eta)\max\{H(x_0,Dw(x_0)),\ w(x_0)-\psi(x_0)\}\\
&\quad \le -(1-\eta)\delta < 0,
\end{aligned}$$

because v_1 is a viscosity subsolution and w is a strict subsolution. ∎

Proof of Theorem 4.6. Consider the following auxiliary test function:

$$\Phi_{\epsilon,\alpha}(x,y) = v_\eta(x)-v_2(y)-\frac{|\,x-y\,|^2}{2\epsilon}-\frac{\alpha}{2}(|\,x\,|^2+|\,y\,|^2), \qquad \epsilon,\alpha>0,$$

where v_η and v_2 are a viscosity subsolution and a viscosity supersolution of (2) and (1), respectively. By definition, $v_\eta, v_2 \in BUC(\mathbb{R}^n)$. Let us define $M_\eta = \sup_{\mathbb{R}^n}(v_\eta - v_2)$; then if $M_{\epsilon,\alpha}$ denotes the maximum of $\Phi_{\epsilon,\alpha}$, one can prove the following properties (refer to Theorem 2.11 in [Bar94]).

- $M_{\epsilon,\alpha} \to M_\eta$ as $(\epsilon,\alpha)\to 0$.
- If $(x_{\epsilon,\alpha}, y_{\epsilon,\alpha})$ denotes the maximum point of $\Phi_{\epsilon,\alpha}$ then
  (a) $v_\eta(x_{\epsilon,\alpha})-v_2(y_{\epsilon,\alpha}) \to M_\eta$ when $(\epsilon,\alpha)\to 0$,
  (b) $\frac{|x_{\epsilon,\alpha}-y_{\epsilon,\alpha}|^2}{\epsilon} \to 0$ when $(\epsilon,\alpha)\to 0$,
  (c) $\alpha(|\,x_{\epsilon,\alpha}\,|^2+|\,y_{\epsilon,\alpha}\,|^2)\to 0$ when $(\epsilon,\alpha)\to 0$.

Thus one can conclude that $|\,x_{\epsilon,\alpha}-y_{\epsilon,\alpha}\,| \to 0$ and $\alpha x_{\epsilon,\alpha}, \alpha y_{\epsilon,\alpha} \to 0$ when $(\epsilon,\alpha)\to 0$. Define

$$w_1(x) = v_2(y_{\epsilon,\alpha})+\frac{|\,x-y_{\epsilon,\alpha}\,|^2}{2\epsilon}+\frac{\alpha}{2}(|\,x\,|^2+|\,y_{\epsilon,\alpha}\,|^2);$$

then obviously $w_1 \in C^\infty(\mathbb{R}^n)$ and $v_\eta - w_1$ attains its maximum at $x_{\epsilon,\alpha}$. According to Lemma 4.9 we have

$$\max\{H(x_{\epsilon,\alpha}, p_{\epsilon,\alpha}+\alpha x_{\epsilon,\alpha}),\ v_\eta(x_{\epsilon,\alpha})-\psi(x_{\epsilon,\alpha})\} \le -\delta_\eta < 0, \tag{3}$$

where $p_{\epsilon,\alpha} := (x_{\epsilon,\alpha}-y_{\epsilon,\alpha})/\epsilon$ and $\delta_\eta := (1-\eta)\delta$. Now define

$$w_2(y) = v_\eta(x_{\epsilon,\alpha})-\frac{|\,x_{\epsilon,\alpha}-y\,|^2}{2\epsilon}-\frac{\alpha}{2}(|\,x_{\epsilon,\alpha}\,|^2+|\,y\,|^2),$$

where $w_2 \in C^\infty(\mathbb{R}^n)$ and v_2-w_2 attains its minimum at $y_{\epsilon,\alpha}$, thus

$$\max\{H(y_{\epsilon,\alpha}, p_{\epsilon,\alpha}-\alpha y_{\epsilon,\alpha}),\ v_2(y_{\epsilon,\alpha})-\psi(y_{\epsilon,\alpha})\} \ge 0. \tag{4}$$

The inequality in (3) implies that

$$\begin{cases} H(x_{\epsilon,\alpha}, p_{\epsilon,\alpha} + \alpha x_{\epsilon,\alpha}) \le -\delta_\eta < 0 \\ v_\eta(x_{\epsilon,\alpha}) - \psi(x_{\epsilon,\alpha}) \le -\delta_\eta < 0 \end{cases} \qquad \forall \epsilon, \alpha > 0. \tag{5}$$

One can show that

$$\begin{aligned} |H(x_{\epsilon,\alpha},\ p_{\epsilon,\alpha} + \alpha x_{\epsilon,\alpha}) - H(y_{\epsilon,\alpha},\ p_{\epsilon,\alpha} - \alpha y_{\epsilon,\alpha})| \le\ & m_\ell(|x_{\epsilon,\alpha} - y_{\epsilon,\alpha}|) \\ & + K|x_{\epsilon,\alpha} - y_{\epsilon,\alpha}||p_{\epsilon,\alpha} + \alpha x_{\epsilon,\alpha}| \\ & + C\alpha|x_{\epsilon,\alpha} + y_{\epsilon,\alpha}|, \end{aligned}$$

where $m_\ell \in C([0,\infty))$ with $m_\ell(0) = 0$ (i.e., m_ℓ is the modulus of continuity for ℓ), K is the Lipschitz constant of g, and $C = \sup_{(x,u)} |g(x,u)|$. Finally taking the limit when $(\epsilon, \alpha) \to 0$ gives

$$\lim_{(\epsilon,\alpha)\to 0} |H(x_{\epsilon,\alpha}, p_{\epsilon,\alpha} + \alpha x_{\epsilon,\alpha}) - H(y_{\epsilon,\alpha}, p_{\epsilon,\alpha} - \alpha y_{\epsilon,\alpha})| = 0;$$

thus there exists $\epsilon_0, \alpha_0 > 0$ such that for all $0 < \epsilon < \epsilon_0$ and $0 < \alpha < \alpha_0$ we have (see the first inequality in (5))

$$H(y_{\epsilon,\alpha}, p_{\epsilon,\alpha} - \alpha y_{\epsilon,\alpha}) \le -\delta_\eta/2 < 0. \tag{6}$$

Using (4) and (6) one can easily conclude that

$$v_2(y_{\epsilon,\alpha}) - \psi(y_{\epsilon,\alpha}) \ge 0 \quad \forall\ 0 < \epsilon < \epsilon_0 \text{ and } 0 < \alpha < \alpha_0.$$

Since for all $x \in \mathbb{R}^n$, $\Phi_{\epsilon,\alpha}(x,x) \le \Phi_{\epsilon,\alpha}(x_{\epsilon,\alpha}, y_{\epsilon,\alpha})$, one can write

$$\begin{aligned} v_\eta(x) - v_2(x) - \alpha \mid x \mid^2 &\le v_\eta(x_{\epsilon,\alpha}) - v_2(y_{\epsilon,\alpha}) - \frac{|x_{\epsilon,\alpha} - y_{\epsilon,\alpha}|^2}{2\epsilon} \\ &\quad - \frac{\alpha}{2}(|x_{\epsilon,\alpha}|^2 + |y_{\epsilon,\alpha}|^2), \\ &= (v_\eta(x_{\epsilon,\alpha}) - \psi(x_{\epsilon,\alpha})) - (v_2(y_{\epsilon,\alpha}) - \psi(y_{\epsilon,\alpha})) \\ &\quad + (\psi(x_{\epsilon,\alpha}) - \psi(y_{\epsilon,\alpha})) - \frac{|x_{\epsilon,\alpha} - y_{\epsilon,\alpha}|^2}{2\epsilon} \\ &\quad - \frac{\alpha}{2}(|x_{\epsilon,\alpha}|^2 + |y_{\epsilon,\alpha}|^2), \\ &\le (\psi(x_{\epsilon,\alpha}) - \psi(y_{\epsilon,\alpha})) - \frac{|x_{\epsilon,\alpha} - y_{\epsilon,\alpha}|^2}{2\epsilon} \\ &\quad - \frac{\alpha}{2}(|x_{\epsilon,\alpha}|^2 + |y_{\epsilon,\alpha}|^2), \end{aligned}$$

where in the last inequality we have $0 < \epsilon < \epsilon_0$ and $0 < \alpha < \alpha_0$, so when $(\epsilon, \alpha) \to 0$ one gets $v_\eta(x) - v_2(x) \le 0$ or

$$v_\eta(x) \le v_2(x).$$

Now let $v_\eta(x) = \eta v_1(x) + (1-\eta)w(x)$, where v_1 is a viscosity subsolution and w is a strict subsolution (see Lemma 4.9); then we have

$$\eta v_1(x) + (1-\eta)w(x) \le v_2(x) \quad \forall x \in \mathbb{R}^n \text{ and } \eta \in (0,1).$$

Let $\eta \to 1$ and one finally gets

$$v_1(x) \leq v_2(x). \tag{7}$$

The inequality in (7) simply says that any viscosity subsolution is less than or equal to any viscosity supersolution. A viscosity solution is both a viscosity subsolution and a viscosity supersolution, so if V_1 and V_2 are two different viscosity solutions for (1), then by (7) we must have $V_1 \leq V_2$ and $V_1 \geq V_2$, which implies that $V_1 = V_2$. So there is at most one viscosity solution of (1). □

REFERENCES

[Bar94] G. Barles, *Solutions de Viscosité des Équations de Hamilton-Jacobi*, Math. Appl. 17, Springer-Verlag, Paris, 1994.

[BP88] G. Barles and B. Perthame, *Exit time problems in optimal control and vanishing viscosity method*, SIAM J. Control Optim., 26 (1988), pp. 1133–1148.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 5, pp. 1485–1503, September 1998

001

ASYMPTOTIC CONTROLLABILITY AND EXPONENTIAL STABILIZATION OF NONLINEAR CONTROL SYSTEMS AT SINGULAR POINTS*

LARS GRÜNE†

Abstract. We discuss the relation between exponential stabilization and asymptotic controllability of nonlinear control systems with constrained control range at singular points. Using a discounted optimal control approach, we construct discrete feedback laws minimizing the Lyapunov exponent of the linearization. Thus we obtain an equivalence result between uniform exponential controllability and uniform exponential stabilizability by means of a discrete feedback law.

Key words. stabilization, nonlinear control systems, singular points, Lyapunov exponents, discounted optimal control problems, discrete feedback control

AMS subject classifications. 93D15, 93D22

PII. S0363012997315919

1. Introduction.

In this paper we will present a technique for the exponential stabilization of nonlinear control systems with constrained control range at singular points. In particular we address the relation between asymptotic controllability and exponential stabilization and will derive an equivalence theorem. In our context a singular point is a fixed point for each admissible control value of the control system. Such singular situations do typically occur if the control enters in the parameters of an uncontrolled system at a fixed point, for instance, when the restoring force of a nonlinear oscillator is controlled. One example to which our results can be applied is the stabilization problem of an inverted pendulum for which the suspension point is moved up and down periodically and the period of this motion can be controlled; cf. [14]. The main tool used throughout this paper is the linearization of the nonlinear system which forms a semilinear system. For two-dimensional control affine systems this linearization approach has been carried out in [4], giving a characterization of feedback stabilizability by algebraic methods.

The approach we follow here is based on optimal control techniques. More precisely, we consider the Lyapunov exponents of the linearization and formulate a discounted optimal control problem in order to minimize these exponents—an idea that was first presented in [12]. Lyapunov exponents have recently turned out to be a suitable tool for the stability analysis of semilinear systems, see, e.g., [7] and [8], and also for their stabilization [11]. However, due to the fact that for discounted optimal control problems optimal feedback laws are in general not available, we modify the feedback concept and introduce *discrete feedback laws* that are based on a discrete time sampled approximation of the given continuous time system. Using this approach it could be shown in [11] that, for semilinear systems satisfying an accessibility condition, exponential null controllability is equivalent to exponential stabilizability by discrete feedback. Using a similar feedback concept, a result on the relation between

*Received by the editors February 3, 1997; accepted for publication (in revised form) September 24, 1997; published electronically June 2, 1998. This research was partially supported by DFG grant Co 124/12-2.

http://www.siam.org/journals/sicon/36-5/31591.html

†Fachbereich Mathematik, AG 1.1, J.-W.-Goethe Universität, Postfach 11 19 32, 60054 Frankfurt a.M., Germany (gruene@math.uni-frankfurt.de).

asymptotic null controllability and practical stabilization for nonlinear systems has been developed in [5] using Lyapunov functions.

This paper is organized into two parts. In the first part we will focus on semilinear systems and extend the results from [11] and [12]. In particular in section 3 we will discuss different null controllability concepts for semilinear systems and extend the approximation results from [12] to general semilinear systems without any accessibility assumptions. Then in section 4 we will use this result in order to construct a stabilizing discrete feedback law following the outline of [11].

In the second part we will apply this discrete feedback to a general nonlinear system at a singular point. For this purpose we will first prove a robustness property of the discrete feedback in section 5. Using this result we will present the main theorem in section 6, stating that (local) uniform exponential null controllability is equivalent to (local) exponential stabilizability by means of a discrete feedback.

2. Preliminaries. We are interested in the stabilization of nonlinear control systems on $\mathbb{R}^d \times M$ given by

$$\begin{aligned} \dot{x}(t) &= f(x(t), y(t), u(t)), \\ \dot{y}(t) &= g(y(t), u(t)), \end{aligned} \tag{2.1}$$

where $x \in \mathbb{R}^d$ and $y \in M$, M is some Riemannian manifold and f and g are vector fields which are C^2 in x, Lipschitz in y, and continuous in u. The control function $u(\cdot)$ may be chosen from the set $\mathcal{U} := \{u : \mathbb{R} \to U \,|\, u(\cdot) \text{ measurable}\}$, where $U \subset \mathbb{R}^m$ is compact, i.e., we have a constrained set of control values.

For each pair (x_0, y_0) of initial values, the trajectories of (2.1) will be denoted by the pair $(x(t, x_0, y_0, u(\cdot)),\ y(t, y_0, u(\cdot)))$ and we assume them to exist uniquely for all times.

Our interest lies in the stabilization of the x-component at a *singular point* x^*, i.e., a point where $f(x^*, y, u) = 0$ for all $(y, u) \in M \times U$. Throughout the paper we will assume $x^* = 0$.

Note that our general setup covers several models: the additional equation for y allows us to model systems where time varying parametric excitations governed by an additional (nonlinear) control or dynamical system enter the system to be stabilized. The case in which the control u does not enter explicitly in the function f and the case in which f does not depend on y occur as special situations in this setup; hence they are also covered.

Our main tool for the stabilization is the linearization of (2.1) at the singular point which is given by

$$\begin{aligned} \dot{z}(t) &= A(y(t), u(t))z(t), \\ \dot{y}(t) &= g(y(t), u(t)). \end{aligned} \tag{2.2}$$

Here $A(y, u) := \frac{\partial}{\partial x} f(x^*, y, u) \in \mathbb{R}^{d \times d}$ and $f(x, y, u) = A(y, u)x + \tilde{f}(x, y, u)$. Then for any given compact subset $K \subset M$ the differentiability assumption on f implies the inequality

$$\|\tilde{f}(x, y, u)\| \le C_f \|x\|^2 \tag{2.3}$$

which holds for some constant C_f for all $y \in K$ and all x in a neighborhood of x^*.

As above we denote the trajectories of (2.2) by $(z(t, z_0, y_0, u(\cdot)),\ y(t, y_0, u(\cdot)))$ for the pair of initial values $(z_0,\ y_0)$.

The first step is now to analyze and characterize the null controllability of (2.2).

3. Lyapunov exponents and their approximation. This section is concerned with the asymptotic null controllability of the semilinear system (2.2). From [7] it is known for bilinear systems that exponential null controllability of (2.2) can be characterized by certain Lyapunov exponents, provided an accessibility condition holds and the matrix A does not depend on y. These conditions will be dropped here and in addition we will show that the characterization is also valid if we replace *exponential* null controllability with *asymptotic* null controllability.

We will first introduce some concepts that will help us characterize the properties of (2.2); see [6] and [7] for more details. Afterwards we will show the relation between different concepts of null controllability and then use these results in order to extend the approximation results from [12].

In order to measure the exponential null controllability we define the Lyapunov exponent of a trajectory of (2.2) by

$$\lambda(z_0, y_0, u(\cdot)) := \limsup_{t\to\infty} \frac{1}{t} \ln \|z(t, z_0, y_0, u(\cdot)\|.$$

Clearly $\lambda(z_0, y_0, u(\cdot)) < 0$ iff the corresponding trajectory converges to the origin exponentially fast. For each pair of initial values we define the infimal Lyapunov exponent by

$$\lambda^*(z_0, y_0) := \inf_{u(\cdot)\in\mathcal{U}} \lambda(z_0, y_0, u(\cdot)).$$

From the linearity of (2.2) it follows that $\lambda(z_0, y_0, u(\cdot)) = \lambda(\alpha z_0, y_0, u(\cdot))$ for all $\alpha \in \mathbb{R} \setminus \{0\}$. Hence we can use the projection of the z component to the unit sphere $\mathbb{S}^{d-1}$ which is given by

$$\begin{aligned} \dot{s}(t) &= h(s(t), y(t), u(t)), \\ \dot{y}(t) &= g(y(t), u(t)), \end{aligned} \tag{3.1}$$

where $h(s, y, u) = [A(y, u) - s^T A(y, u)s, \mathrm{Id}]s$, where Id denotes the $d \times d$ identity matrix. Denoting the projected trajectory by $s(t, s_0, y_0, u(\cdot))$, it follows from the chain rule that for $s_0 = \frac{z_0}{\|z_0\|}$ the Lyapunov exponent can be written as

$$\lambda(s_0, y_0, u(\cdot)) = \limsup_{t\to\infty} \frac{1}{t} \int_0^t q(s(\tau, s_0, y_0, u(\cdot)),\, y(\tau, y_0, u(\cdot)), u(\tau))d\tau, \tag{3.2}$$

where $q(s, y, u) := s^T A(y, u)s$. This integral is also referred to as an *averaged functional*.

By defining the exponential growth rate in finite time t,

$$\lambda^t(z_0, y_0, u(\cdot)) := \frac{1}{t} \ln \frac{\|z(t, z_0, y_0, u(\cdot)\|}{\|z_0\|},$$

it is easily seen that

$$\|z(t, z_0, y_0, u(\cdot))\| = e^{t\lambda^t(z_0, y_0, u(\cdot))}\|z_0\|. \tag{3.3}$$

As above, this expression can be written in integral form using the projected system, i.e., for $s_0 = \frac{z_0}{\|z_0\|}$ we obtain

$$\lambda^t(s_0, y_0, u(\cdot)) = \frac{1}{t} \int_0^t q(s(\tau, s_0, y_0, u(\cdot)),\, y(\tau, y_0, u(\cdot)), u(\tau))d\tau.$$

In our definitions of null controllability we need the notion of a positively invariant set for the subsystem on M.

DEFINITION 3.1. *A subset $K \subseteq M$ is called positively invariant for the subsystem of* (2.2) *on M if for all $y_0 \in K$ and all control functions $u(\cdot) \in \mathcal{U}$ the corresponding trajectory satisfies $y(t, y_0, u(\cdot)) \in K$ for all $t > 0$.*

Now we can define the concepts of null controllability; cf. also the stability concepts in [15].

DEFINITION 3.2. *Let $K \subseteq M$ be a compact positively invariant set for the subsystem of* (2.2) *on M.*

(i) *The system* (2.2) *is called* asymptotically null controllable *over K if for any pair of initial values $(z_0, y_0) \in \mathbb{R}^d \times K$ there exists a control function $u(\cdot) \in \mathcal{U}$ such that*

$$\lim_{t \to 0} \|z(t, z_0, y_0, u(\cdot))\| = 0.$$

(ii) *The system* (2.2) *is called* exponentially null controllable *over K if λ^* satisfies* $\sup_{(z_0,y_0)\in\mathbb{R}^d\times K} \lambda^*(z_0, y_0) < 0$.

(iii) *The system* (2.2) *is called* uniformly exponentially null controllable *over K if there exist constants C, $\alpha > 0$, such that for any pair of initial values $(z_0, y_0) \in \mathbb{R}^d \times K$ there exists a control function $u_{(z_0,y_0)}(\cdot) \in \mathcal{U}$ with*

$$\|z(t, z_0, y_0, u_{(z_0,y_0)}(\cdot))\| \le Ce^{-\alpha t}\|z_0\|.$$

An immediate consequence from (3.3) is that (2.2) is uniformly exponentially null controllable over K iff there exists a time $T > 0$ and a constant $\sigma < 0$ such that for any pair of initial values $(z_0, y_0) \in \mathbb{R}^d \times K$ there exists a control function $u_{(z_0,y_0)}(\cdot) \in \mathcal{U}$ with

$$\lambda^t(z_0, y_0, u_{(z_0,y_0)}(\cdot)) \le \sigma < 0$$

for all $t \ge T$.

It is easily seen from this definition that (iii) $\Rightarrow$ (ii) $\Rightarrow$ (i). In fact the converse is also true, i.e., the definitions are equivalent as the following proposition shows.

PROPOSITION 3.3. *Let $K \subseteq M$ be a compact positively invariant set for the subsystem of* (2.2) *on M. Then for the system* (2.2)*, asymptotic null controllability over K implies uniform exponential null controllability over K.*

Proof. We will first show the following property: there exist $T > 0$ and $\sigma < 0$ such that for each $(z, y) \in \mathbb{R}^d \times K$ there exists a control function $u_{(z,y)}(\cdot) \in \mathcal{U}$ and a time $t_{(z,y)} \le T$ such that $\lambda^{t_{(z,y)}}(z, y, u_{(z,y)}(\cdot)) < \sigma$.

The asymptotic null controllability implies that for each $(\tilde{z}_0, \tilde{y}_0) \in \mathbb{R}^d \times K$ there exists a time $\tilde{t}_{(\tilde{z}_0,\tilde{y}_0)}$ and a control function $\tilde{u}_{(\tilde{z}_0,\tilde{y}_0)}(\cdot)$, such that

$$\|z(\tilde{t}_{(\tilde{z}_0,\tilde{y}_0)}, \tilde{z}_0, \tilde{y}_0, \tilde{u}_{(\tilde{z}_0,\tilde{y}_0)}(\cdot))\| < \frac{1}{3}\|\tilde{z}_0\|.$$

Considering only those $\tilde{z}_0$ with $\|\tilde{z}_0\| = 1$ (i.e., $\tilde{z}_0 \in \mathbb{S}^{d-1}$) and using the continuous dependence on the initial value, we find a neighborhood $U(\tilde{z}_0, \tilde{y}_0)$ in $\mathbb{S}^{d-1} \times K$ such that for each $(z, y) \in U(\tilde{z}_0, \tilde{y}_0)$ it holds that

$$\|z(\tilde{t}_{(\tilde{z}_0,\tilde{y}_0)}, z, y, \tilde{u}_{(\tilde{z}_0,\tilde{y}_0)}(\cdot))\| < \frac{1}{2}\|z\|.$$

Hence it follows that $\lambda^{\tilde{t}_{(\tilde{z}_0,\tilde{y}_0)}}(z,y,\tilde{u}_{(\tilde{z}_0,\tilde{y}_0)}(\cdot)) < \gamma_{(\tilde{z}_0,\tilde{y}_0)} < 0$, where

$$\gamma_{(\tilde{z}_0,\tilde{y}_0)} = \frac{\ln \frac{1}{2}}{\tilde{t}_{(\tilde{z}_0,\tilde{y}_0)}}.$$

By the compactness of $\mathbb{S}^{d-1} \times K$ we may pick a finite number of pairs $(\tilde{z}_0, \tilde{y}_0)$ such that the neighborhoods $U(\tilde{z}_0, \tilde{y}_0)$ cover $\mathbb{S}^{d-1} \times K$. Now the independence of λ^t from the norm of z yields the asserted property, where T is the maximum over all $\tilde{t}_{(\tilde{z}_0,\tilde{y}_0)}$ and $\sigma < 0$ the maximum over all $\gamma_{(\tilde{z}_0,\tilde{y}_0)}$.

Now pick an arbitrary pair (z_0, y_0) of initial values. We use the control $u_0(\cdot) := u_{(z_0,y_0)}(\cdot)$ from above up to the time $t_1 := t_{(z_0,y_0)} < T$ from above and end up at the point $(z_1, y_1) = (z(t_1, z_0, y_0, u_0(\cdot)), y(t_0, y_0, u_0(\cdot)))$. We continue iteratively by defining $t_{i+1} := t_i + t_{(z_i,y_i)}$ and $u_i(\cdot) := u_{(z_i,y_i)}(\cdot)$ and define a control function $u : \mathbb{R}^+ \to U$ by

$$u(t) := u_i(t - t_i),\ t \in [t_i, t_{i+1}]$$

for $i \in \mathbb{N}_0$, where $t_0 := 0$.

This yields $\lambda^{t_i}(z_0, y_0, u(\cdot)) < \sigma$ for all t_i, $i \in \mathbb{N}_0$, and since $t_i - t_{i-1} < T$, it follows that for any $t > 0$ there exists $t_i =: t_i(t)$ with $0 \le t - t_i(t) < T$. By the definition of λ^t we obtain

$$\lambda^t(z_0, y_0, u(\cdot)) = \frac{t_i(t)}{t}\lambda^{t_i(t)}(z_0, y_0, u(\cdot)) + \frac{t - t_i(t)}{t}\lambda^{t-t_i(t)}(z_i, y_i, u_i(\cdot))$$

which yields

$$\lambda^t(z_0, y_0, u(\cdot)) < \sigma + \varepsilon(t),$$

where

$$\varepsilon(t) = \frac{t - t_i(t)}{t}\big(\lambda^{t-t_i(t)}(z_i, y_i, u_i(\cdot)) - \lambda^{t_i(t)}(z_0, y_0, u(\cdot))\big),$$

implying $\varepsilon(t) \to 0$ for $t \to \infty$ independently from (z_0, y_0) since λ^t is uniformly bounded for all $t > 0$ and all $(z, y) \in \mathbb{R}^d \times K$. Hence there exists $\varepsilon > 0$ and a time $T > 0$ such that $\lambda^t(z_0, y_0, u(\cdot)) < \sigma + \varepsilon < 0$ for all $t \ge T$, and the assertion follows. □

Using essentially the same arguments as in the previous proof, we can also determine the uniform upper bound for the values of the λ^t.

PROPOSITION 3.4. *Let $K \subseteq M$ be a compact positively invariant set for the subsystem of (2.2) on M. Let $\sigma := \sup_{(z_0,y_0)\in\mathbb{R}^d\times K} \lambda^*(z_0, y_0)$. Then for each $\varepsilon > 0$ there exists a $T > 0$ such that for any $(z_0, y_0) \in \mathbb{R}^d \times K$ there exists a control function $u(\cdot) \in \mathcal{U}$ satisfying*

$$\lambda^t(z_0, y_0, u(\cdot)) < \sigma + \varepsilon$$

for all $t \ge T$.

Proof. For any pair $(\tilde{z}_0, \tilde{y}_0) \in \mathbb{R}^d \times K$ there exists a control function $\tilde{u}_{(\tilde{z}_0,\tilde{y}_0)}(\cdot)$ and a time $\tilde{t}_{(\tilde{z}_0,\tilde{y}_0)}$ such that

$$\lambda^{\tilde{t}_{(\tilde{z}_0,\tilde{y}_0)}}(\tilde{z}_0, \tilde{y}_0, \tilde{u}_{(\tilde{z}_0,\tilde{y}_0)}(\cdot)) < \sigma + \frac{\varepsilon}{3}.$$

As in the previous proof, continuous dependence and compactness implies that for any pair (z, y) there exist $t_{(z,y)}$ bounded by some $\tilde{T}$ and control functions $u_{(z,y)} \in \mathcal{U}$, such that

$$\lambda^{t_{(z,y)}}(z, y, u_{(z,y)}(\cdot)) < \sigma + \frac{\varepsilon}{2}.$$

Following the previous proof we can iteratively construct control functions satisfying

$$\lambda^{t}(z_0, y_0, u(\cdot)) < \sigma + \frac{\varepsilon}{2} + \varepsilon(t).$$

Again $\varepsilon(t)$ can be chosen independently from (z_0, y_0) and $\varepsilon(t) \to 0$ as $t \to \infty$; hence the assertion follows by choosing T such that $\varepsilon(t) < \frac{\varepsilon}{2}$ for all $t \geq T$. □

This result implies that the α in Definition 3.2 (iii) can be chosen arbitrarily close to the sup-inf Lyapunov exponent σ as defined in Proposition 3.4. This Lyapunov exponent therefore gives the characteristic value for the null controllability of (2.2).

The construction of the stabilizing discrete feedback in the next section — following the outline of [11] — is based on the minimization of the Lyapunov exponent. This is related to minimizing (3.2) which forms an average time optimal control problem, for which the construction of optimal feedback controls is still an unsolved problem.

Hence we will not approach this problem directly but will use the approximation of (3.2) by a *discounted functional* with *discount rate* $\delta > 0$ defined by

$$J_\delta(s_0, y_0, u(\cdot)) := \int_0^\infty e^{-\delta\tau} q(s(\tau, s_0, y_0, u(\cdot)),\, y(\tau, y_0, u(\cdot)), u(\tau)) d\tau. \tag{3.4}$$

The function

$$v_\delta(s_0, y_0) := \inf_{u(\cdot)\in\mathcal{U}} J_\delta(s_0, y_0, u(\cdot)) \tag{3.5}$$

is called the *optimal value function* of this discounted optimal control problem.

The relation between this problem and the minimization of (3.2) has been discussed in [12] for the case where (3.1) is locally accessible, exploiting the controllability properties of (3.1). Here we will use Proposition 3.4 in combination with a stronger version of the approximation theorems from [12] in order to show this relation without assuming local accessibility.

LEMMA 3.5 (approximation theorems). *Let $q : \mathbb{R} \to \mathbb{R}$ be a measurable function satisfying $|q(s)| < M_q$ for almost all $s \in \mathbb{R}$.*

(i) *Assume there exists a time $T > 0$ such that*

$$\frac{1}{t}\int_0^t q(\tau)d\tau < \sigma \text{ for all } t \geq T.$$

Then for any $\varepsilon > 0$ and all $0 < \delta < \frac{\varepsilon}{(M_q+\sigma+\varepsilon)T}$ the following inequality holds:

$$\delta \int_0^\infty e^{-\delta\tau} q(\tau)d\tau \leq \sigma + \varepsilon.$$

(ii) *Let $\delta > 0$ be arbitrary and let*

$$\delta \int_0^\infty e^{-\delta\tau} q(\tau)d\tau =: \sigma.$$

Then for any $\varepsilon > 0$ there exists a $T \in [\frac{\varepsilon}{(4M_q+4\sigma+\varepsilon)\delta}, -\frac{1}{\delta}\ln\frac{\varepsilon}{4M_q}]$ satisfying

$$\frac{1}{T}\int_0^T q(\tau)d\tau \le \sigma + \varepsilon.$$

(iii) *Let $\delta > 0$ be arbitrary and let $\sigma \in \mathbb{R}$ such that*

$$\delta \int_0^\infty e^{-\delta\tau} q(t+\tau)d\tau \le \sigma \text{ for all } t \ge 0.$$

Then

$$\limsup_{T\to\infty} \frac{1}{T}\int_0^T q(\tau)d\tau \le \sigma$$

Proof. The rather technical proof can be found in the appendix. ∎

Next we can formulate the consequence for the optimal value function.

THEOREM 3.6. *Let $K \subseteq M$ be a compact positively invariant set for the subsystem on M of (2.2). Then*

$$\lim_{\delta\to 0} \sup_{(s,y)\in\mathbb{S}^{d-1}\times K} \delta v_\delta(s,y) = \sup_{(s,y)\in\mathbb{S}^{d-1}\times K} \lambda^*(s,y).$$

Proof. Let $\sigma := \sup_{(s,y)\in\mathbb{S}^{d-1}\times K} \lambda^*(s,y)$ and $\varepsilon > 0$. By Proposition 3.4 there exists a time $T > 0$ such that for each pair $(s,y) \in \mathbb{S}^{d-1} \times K$ there exists a control function $u(\cdot) \in \mathcal{U}$ such that

$$\lambda^t(s,y,u(\cdot)) < \sigma + \frac{\varepsilon}{2}.$$

By Lemma 3.5 (i) this implies

$$\delta J_\delta(s,y,u(\cdot)) < \sigma + \varepsilon$$

for all sufficiently small $\delta > 0$. Since $\varepsilon > 0$ was arbitrary this implies

$$\limsup_{\delta\to 0} \sup_{(s,y)\in\mathbb{S}^{d-1}\times K} \delta v_\delta(s,y) \le \sigma.$$

Now assume $\liminf_{\delta\to 0} \sup_{(s,y)\in\mathbb{S}^{d-1}\times K} \delta v_\delta(s,y) = \gamma < \sigma$. Then there exists $\delta > 0$ such that by Bellman's optimality principle [19, Theorem 1.2] for each pair (s,y) there exists a control function $u(\cdot)$ satisfying

$$\delta J_\delta(s(t,s,y,u(\cdot)), y(t,y,u(\cdot)), u(t+\cdot)) < \tilde{\gamma} < \sigma$$

for all $t \ge 0$. Now by Lemma 3.5 (iii) it follows that $\lambda^*(s,y) \le \tilde{\gamma} < \sigma$ which contradicts the definition of σ. Hence the assertion follows. ∎

This theorem states that the Lyapunov exponent that gives the characteristic number for null controllability can be approximated by the value function of a discounted optimal control problem.

Since algorithms for the numerical computation of v_δ are known (cf., e.g., [9] and [13]) this theorem also lays the foundation for the numerical null controllability analysis of semilinear systems; see also [12]. This is of particular interest because the question of whether (2.2) is null controllable cannot in general be answered by analytical methods.

4. Construction of the discrete feedback. We will now present a feedback construction for the (approximately) optimal solution of the discounted optimal control problem defined by (3.4) and (3.5), which will then be stabilizing for (2.2).

In general the construction of optimal feedback laws for discounted optimal control problems is an unsolved problem. One of the main problems is that optimal feedbacks are typically discontinuous and hence properties such as the existence and uniqueness of the corresponding solutions are no longer guaranteed. Some effort has been made in order to take these difficulties into account, e.g., by using differential inclusions (see [10] and [1]). However, apart from the fact that this approach leads to a characterization of optimal trajectories rather than to a construction of a feedback law, from the stabilization (and application) point of view it seems desirable to preserve these properties. Furthermore we will need a certain robustness property, as discussed in section 5, in order to apply the feedback to the nonlinear system.

These considerations lead to a somewhat modified feedback concept which is based on an approximation of $\mathcal{U}$ as introduced in [11]. Theorem 3.6 yields the property needed for the construction of the stabilizing discrete feedback in sections 3 and 4 of [11] and our construction now follows this outline. We will therefore just give the idea of the construction and omit the proofs except for the concluding theorem.

We approximate $\mathcal{U}$ by

$$\mathcal{U}_h := \{u : \mathbb{R} \to U \mid u|_{[ih,(i+1)h)} \equiv u_i \text{ for all } i \in \mathbb{Z}\}$$

for some time step $h > 0$. This discretization for discounted optimal control problems bears some similarity to the discretization in [2] and [3]; in fact what we obtain is a discrete time system by the process of sampling (cf., [21, section 2.10]):

$$s_{i+1} = s(h, s_i, y_i, u_i),\ y_{i+1} = y(h, y_i, u_i), \tag{4.1}$$

where $(u_i)_{i\in\mathbb{Z}} \in U^{\mathbb{Z}}$.

Defining

$$v_\delta^h(s_0, y_0) := \inf_{u(\cdot)\in\mathcal{U}_h} J_\delta(s_0, y_0, u(\cdot)),$$

the approximation property

$$\|v_\delta - v_\delta^h\|_\infty \le Ch^{\frac{\gamma}{2}}$$

holds for $\gamma = \delta/L$ where L denotes the Lipschitz constant of (3.1); see [3].

Bellman's optimality principle [19, Theorem 1.2] yields

$$v_\delta^h(s_0, y_0) = \inf_{u\in U} \left\{ \int_0^h e^{-\delta\tau} q(s(\tau, s_0, y_0, u),\ y(\tau, y_0, u), u)d\tau + e^{-\delta h} v_\delta^h(s(h, s_0, y_0, u),\ y(h, y_0, u)) \right\}.$$

By the continuity of all functions involved and the compactness of U, we can now define a function $F : \mathbb{S}^{d-1} \times K \to U$ by choosing $F(s_0, y_0) := u \in U$ such that the infimum above is attained in u.

We may now apply F to (3.1) by

$$\begin{aligned} \dot{s}(t) &= h(s(t), y(t), F(s(\left[\tfrac{t}{h}\right] h), y(\left[\tfrac{t}{h}\right] h))), \\ \dot{y}(t) &= g(y(t), F(s(\left[\tfrac{t}{h}\right] h), y(\left[\tfrac{t}{h}\right] h))). \end{aligned} \tag{4.2}$$

We denote the solution trajectories of (4.2) by $(s_F(t, s_0, y_0), y_F(t, y_0))$.

Feedback laws of this kind can be found in the literature under the name of *modified feedback control* [16], [17], *sample-and-hold* control, or *sampled feedback* [20], [22], and *step-by-step control* [18]. Of particular interest in this context is the recent work [5] where a stabilization result using a "sampled feedback" control is presented. We will discuss the relation between this work and the present paper in section 6.

In our terminology we call F a "discrete" feedback control, a notion being motivated by the fact that F is indeed a feedback control for the discrete time system (4.1). From this interpretation the existence and uniqueness of the trajectories of (4.2) is immediately clear.

If we evaluate

$$J_\delta(s,y,F) := \int_0^\infty e^{-\delta\tau} q\left(s_F(\tau,s,y),\, y_F(\tau,y), F\left(s_F\left(\left[\frac{\tau}{h}\right]h, s, y\right),\, y_F\left(\left[\frac{\tau}{h}\right]h, y\right)\tau\right)\right) d\tau,$$

i.e., the discounted value along the trajectories of (4.2), it follows that $J_\delta(s,y,F) = v_\delta^h(s,y)$ for all initial values $(s,y) \in \mathbb{S} \times K$ ([11, Theorem 3.6]). Hence F forms an optimal discrete feedback for the discounted optimal control problem with respect to the discretized control functions from $\mathcal{U}_h$.

In the same way we define the averaged value along the trajectories by

$$\lambda^t(s,y,F) := \frac{1}{t}\int_0^t q\left(s_F(\tau,s,y),\, y_F(\tau,y), F\left(s_F\left(\left[\frac{\tau}{h}\right]h, s, y\right),\, y_F\left(\left[\frac{\tau}{h}\right]h, y\right)\tau\right)\right) d\tau.$$

By defining $F_{\mathbb{R}}(z,y) := F(z/\|z\|, y)$ we can apply $F_{\mathbb{R}}$ to the nonprojected system (2.2) by

$$\begin{aligned} \dot z(t) &= A(y(t), F_{\mathbb{R}}(s(\left[\tfrac{t}{h}\right]h), y(\left[\tfrac{t}{h}\right]h)))z(t), \\ \dot y(t) &= g(y(t), F_{\mathbb{R}}(z(\left[\tfrac{t}{h}\right]h), y(\left[\tfrac{t}{h}\right]h))). \end{aligned} \tag{4.3}$$

As above we denote the corresponding trajectories by $(x_{F_{\mathbb{R}}}(t, x_0, y_0), y_{F_{\mathbb{R}}}(t, y_0))$. Applying $F_{\mathbb{R}}$ this way we can state the following theorem.

THEOREM 4.1. *Let $K \subseteq M$ be a compact positively invariant set for the subsystem of* (2.2) *on M. Then* (2.2) *is asymptotically null controllable over K iff there exists a time step h and a discrete feedback law $F_{\mathbb{R}} : \mathbb{R} \times K \to U$ such that* (4.3) *is uniformly exponentially stable, i.e., there exists C, $\alpha > 0$ such that every trajectory of* (4.3) *satisfies the condition from Definition* 3.2 (iii).

Proof ("$\Rightarrow$"). Assume asymptotic null controllability of (2.2). By [11, Corollary 3.7], it follows that for any $\varepsilon > 0$ there exists $h > 0$ such that the discrete feedback as defined above satisfies

$$\delta J_\delta(s,y,F) < \delta v_\delta(s,y) + \varepsilon. \tag{4.4}$$

Choosing $\delta > 0$ sufficiently small, Proposition 3.3 implies that there exists $\sigma < 0$ such that $\delta J_\delta(s,y,F) < \sigma$, hence from Lemma 3.5 (ii) we can conclude that for any $\varepsilon > 0$ there exists a bounded time $t = t(\varepsilon) > 0$ such that $\lambda^t(s,y,F) < \sigma + \varepsilon$. Using [11, Lemma 4.1] we obtain estimate (4.4) for the next trajectory piece and can inductively obtain the assertion as in the proof of Proposition 3.3.

("$\Leftarrow$"). This direction is immediately clear. □

Note that this stabilizing discrete feedback law is numerically computable — at least for lower-dimensional systems — using the algorithm proposed in [11] and [13].

5. Robustness of the discrete feedback control. From the definition of the discrete feedback F and $F_{\mathbb{R}}$ it is obvious that these functions are typically discontinuous. Hence by applying this feedback law, continuous dependence of the trajectories on the initial value will in general not hold.

This gives rise to the question of the robustness of the optimal trajectories. More precisely: do optimal trajectories remain approximately optimal under small perturbations?

The answer is given in the following proposition and is essentially based on the Hölder continuity of v_δ^h which satisfies

$$|v_\delta^h(s,y) - v_\delta^h(\tilde{s},\tilde{y})| \le C(d_{\mathbb{S}}(s,\tilde{s}) + d_M(y,\tilde{y}))^\gamma,$$

where $\gamma = \delta/L$ and L is the Lipschitz constant of (3.1). For systems in $\mathbb{R}^n$ this immediately follows from [3, Lemma 4.1]; the proof is easily transferred to general manifolds. Here $d_{\mathbb{S}}$ and d_M denote some metrics on $\mathbb{S}$ and M, respectively.

In what follows we allow time varying perturbations of the following kind: assume that we have a time varying system on $\mathbb{S}^{d-1} \times K$ given by

$$\begin{aligned} \dot{s}(t) &= \tilde{h}(t, s(t), y(t), u(t)), \\ \dot{y}(t) &= \tilde{g}(t, y(t), u(t)), \end{aligned} \tag{5.1}$$

with trajectories $(\tilde{s}(t,t^*,s_0,y_0,u(\cdot)),\ \tilde{y}(t,t^*,y_0,u(\cdot)))$ using the initial time t^*. For some pair of initial values (s_0,y_0) and a discrete Feedback F with time step $h>0$, we denote the solution trajectories of (5.1) applying F with initial time $t^*=0$ by $(\tilde{s}_F(t,s_0,y_0),\tilde{y}_F(t,y_0))$. Using the abbreviations $t_i := ih$, $\tilde{s}_i := \tilde{s}_F(t_i,s_0,y_0)$, $\tilde{y}_i := \tilde{y}_F(t_i,y_0)$, and $u_i := F(\tilde{s}_i,\tilde{y}_i)$ we assume

$$d_{\mathbb{S}}(\tilde{s}(t,t_i,\tilde{s}_i,\tilde{y}_i,u_i), s(t,\tilde{s}_i,\tilde{y}_i,u_i)) + d_M(\tilde{y}(t,t_i,\tilde{y}_i,u_i), y(t,\tilde{y}_i,u_i)) < \varepsilon_i \tag{5.2}$$

for all $t\in[0,h]$, all $i\in\mathbb{N}$, and some sequence $(\varepsilon_i)_{i\in\mathbb{N}}$.

PROPOSITION 5.1. *Consider the system* (3.1)*, a time step h, the corresponding optimal value function v_δ^h, and the optimal discrete feedback F. Assume that a system* (5.1) *with the property* (5.2) *for some pair of initial values (s,y) is given and denote the trajectories of* (5.1) *with initial time $t^*=0$ and the discrete feedback F by $(\tilde{s}_F(t,s_0,y_0),\tilde{y}_F(t,y_0))$.*

Then for any $k\in\mathbb{N}$ the following inequality holds:

$$|v_\delta^h(s,y) - \tilde{J}_\delta(s,y,F)| < C\sum_{i=0}^{k-1} e^{-\delta h i}\varepsilon_i^\gamma + 2e^{-\delta hk}\frac{M_q}{\delta},$$

where

$$\tilde{J}_\delta(s,y,F) := \int_0^\infty e^{-\delta\tau} q\Big(\tilde{s}_F(\tau,s,y),\, \tilde{y}_F(\tau,y), F\Big(\tilde{s}_F\Big(\Big[\frac{\tau}{h}\Big]h,s,y\Big),\, \tilde{y}_F\Big(\Big[\frac{\tau}{h}\Big]h,y\Big)\tau\Big)\Big)\,d\tau$$

is the value along the discrete feedback controlled trajectory of (5.1) *and M_q is the bound of $|q|$ on $\mathbb{S}^{d-1}\times K$.*

Remark 5.2. Note that the right-hand side of this inequality becomes small if the ε_i are small for all sufficiently large $i\in\mathbb{N}$.

Proof. From the definition of F and the assumption (5.2) it follows that

$$\begin{aligned} v_\delta^h(s,y) &= \int_0^h q(s_F(\tau,s,y), y_F(\tau,y), F(s,y)) + e^{-\delta h} v_\delta^h(s_F(h,s,y), y_F(h,y)) \\ &= \int_0^h q(\tilde{s}_F(\tau,s,y), \tilde{y}_F(\tau,y), F(s,y)) + e^{-\delta h} v_\delta^h(\tilde{s}_F(h,s,y), \tilde{y}_F(h,y)) + \tilde{C}\varepsilon_0^\gamma, \end{aligned}$$

where $|\tilde{C}| < C$. On the other hand, we obtain

$$\tilde{J}_\delta(s,y,F) = \int_0^h q(\tilde{s}_F(\tau,s,y),\tilde{y}_F(\tau,y),F(s,y)) + e^{-\delta h}\tilde{J}_\delta(\tilde{s}_F(h,s,y),\tilde{y}_F(h,y),F).$$

This yields

$$\begin{aligned}&|v_\delta^h(s,y) - \tilde{J}_\delta(s,y,F)|\\ &\quad\le\quad e^{-\delta h}|v_\delta^h(\tilde{s}_F(h,s,y),\tilde{y}_F(h,y)) - \tilde{J}_\delta(\tilde{s}_F(h,s,y),\tilde{y}_F(h,y),F)| + C\varepsilon_0^\gamma.\end{aligned}$$

By observing that v_δ^h and $\tilde{J}_\delta$ are bounded by M_q/δ, the assertion follows by induction. □

This robustness property is the main tool for the linearization result in the next section.

6. Stabilization of the nonlinear system. We will now return to our original system (2.1). We recall the fact that $f(x,y,u) = A(y,u)x + \tilde{f}(x,y,u)$, where for y in a compact set $K \subset M$ the estimate $\|\tilde{f}(x,y,u)\| \le C_f\|x\|^2$ holds for all $x \in B_{\eta_f}(0)$, the ball with radius η_f around 0; cf. (2.3).

In analogy to Definition 3.2 we begin by defining the controllability concepts for system (2.1). Since we assume that the singular point x^* coincides with the origin we may again formulate these concepts in terms of null controllability. As in the semilinear case, we denote the exponential growth rates of a trajectory by

$$\lambda_f^t(x_0,y_0,u(\cdot)) := \frac{1}{t}\ln\frac{\|x(t,x_0,y_0,u(\cdot))\|}{\|x_0\|}$$

and

$$\lambda_f^*(x_0,y_0) := \inf_{u(\cdot)\in\mathcal{U}} \limsup_{t\to\infty} \lambda^t(x_0,y_0,u(\cdot)).$$

DEFINITION 6.1. *Let $K \subseteq M$ be a compact positively invariant set for the subsystem of* (2.1) *on M.*

(i) *The system* (2.1) *is called* (locally) asymptotically null controllable *over K if there exists a neighborhood $B(0)$ of 0 such that for any pair of initial values $(x_0,y_0) \in B(0)\times K$ there exists a control function $u(\cdot)\in\mathcal{U}$ with*

$$\lim_{t\to 0}\|x(t,z_0,y_0,u(\cdot))\| = 0.$$

(ii) *The system* (2.1) *is called* (locally) exponentially null controllable *over K if there exists a neighborhood $B(0)$ of 0 such that $\sup_{(x_0,y_0)\in B(0)\times K}\lambda_f^*(x_0,y_0) < 0$.*

(iii) *The system* (2.2) *is called* (locally) uniformly exponentially null controllable *over K if there exists a neighborhood $B(0)$ of 0 and constants C, $\alpha > 0$, such that for any pair of initial values $(x_0,y_0) \in B(0)\times K$ there exists a control function $u_{(x_0,y_0)}(\cdot)\in\mathcal{U}$ with*

$$\|z(t,x_0,y_0,u_{(x_0,y_0)}(\cdot))\| \le Ce^{-\alpha t}\|x_0\|.$$

As in the semilinear case, the implications (iii) ⇒ (ii) ⇒ (i) are obvious. However, for nonlinear systems the converse is not true, as the example below will show. Note

that frequently the notion of *exponential stability* already demands the uniformity as in (iii); cf., e.g., [23] or [24].

We will now first prove some a priori estimates for the solutions of (2.1) and (2.2).

LEMMA 6.2. *Abbreviate with* $(x(t), y(t))$ *and* $(z(t), y(t))$ *the solutions of the systems* (2.1) *and* (2.2) *for a pair of initial values* (x_0, y_0) *and a control function* $u(\cdot)$. *Let* $T > 0$ *be a given time.*

Then there exist constants α, β, $C > 0$, *and* $\eta(T) > 0$ *independent from* $u(\cdot)$ *such that for all* $t \in [0, T]$ *the following estimates hold:*

(i) $\|x(t)\| \in [e^{-\alpha t}\|x_0\|, e^{\alpha t}\|x_0\|]$ *for all* $x_0 \in B_{\eta(T)}(0)$,

(ii) $\|z(t)\| \in [e^{-\alpha t}\|x_0\|, e^{\alpha t}\|x_0\|]$ *for all* $x_0 \in \mathbb{R}^d$,

(iii) $\|x(t) - z(t)\| \le tCe^{\beta t}\|x_0\|^2$ *for all* $x_0 \in B_\eta(T)(0)$,

where $B_\eta(T)(0)$ *denotes the ball with radius* $\eta(T)$ *around the origin.*

Proof. (i) We show the estimate for the upper bound; the estimate for the lower bound follows from (ii) and (iii). From the linearization it follows that

$$x(t) = x_0 + \int_0^t A(y(\tau),\, u(\tau))x(\tau) + \tilde{f}(x(\tau), y(\tau), u(\tau))d\tau.$$

As long as $x(t) \in B_{\eta_f}(0)$, this implies

$$\|x(t)\| \le \|x_0\| + \int_0^t \alpha\|y(\tau)\|d\tau$$

for some constant $\alpha > 0$. This yields $\|x(t)\| \le e^{\alpha t}\|x_0\|$ as long as $e^{\alpha t}\|x_0\| \le \eta_f$ and hence the assertion follows with $\eta(T) = \eta_f/e^{\alpha T}$.

(ii) This is an easy consequence from the linearity of the system.

(iii) Define $m(t) := x(t) - z(t)$. From (i) and (ii) it follows that $\|m(t)\| \le e^{\alpha t}\|x_0\|$. Furthermore m is a solution of the differential equation

$$\dot{m}(t) = A(y(t),\, u(t))m(t) + \tilde{f}(y(s),\, m(t) + z(t),\, u(t)), \;\; z(0) = 0$$

and thus satisfies

$$\begin{aligned}
\|m(t)\| &\le \int_0^t \|A(y(\tau),\, u(\tau))m(\tau)\| + \|\tilde{f}(y(\tau),\, m(\tau) + z(\tau),\, u(\tau))\|d\tau \\
&\le \int_0^t \|A(y(\tau),\, u(\tau))m(\tau)\| + C_f(\|m(\tau)\|^2 + \|z(\tau)\|\|m(\tau)\| + \|z(\tau)\|^2)d\tau \\
&\le tC_f e^{2\alpha t}\|x_0\|^2 + \int_0^t \gamma\|m(s)\|ds
\end{aligned}$$

for some constant $\gamma > 0$. Now the Gronwall lemma yields

$$\|m(t)\| \le tC_f e^{2\alpha t}\|x_0\|^2 e^{\gamma t}$$

and thus the assertion. ∎

As in the semilinear case, we may now write the exponential growth rate in finite time in integral form

$$\lambda_f^t(x_0, y_0, u(\cdot)) = \frac{1}{t}\int_0^t q_f(x(\tau, x_0, y_0, u(\cdot)), y(\tau, y_0, u(\cdot)), u(\tau))d\tau,$$

where

$$q_f(x,y,u) = q\left(\frac{x}{\|x\|}, y, u\right) + \frac{x^t \tilde{f}(x,y,u)}{\|x\|^2}$$

which can be calculated using the chain rule.

A simple calculation shows that

$$\|x(t,x_0,y_0,u(\cdot))\| = \|x_0\| e^{t\lambda_f^t(x_0,y_0,u(\cdot))t}.$$

We can now apply the discrete feedback $F_{\mathbb{R}}$ from the previous sections to (2.1) by

$$\begin{aligned} \dot{x}(t) &= f(x(t),y(t),F_{\mathbb{R}}(x(\left[\tfrac{t}{h}\right]h), y(\left[\tfrac{t}{h}\right]h))), \\ \dot{y}(t) &= g(y(t),F_{\mathbb{R}}(x(\left[\tfrac{t}{h}\right]h), y(\left[\tfrac{t}{h}\right]h))), \end{aligned} \tag{6.1}$$

and denote the resulting trajectories by $(x_{F_{\mathbb{R}}}(t,x_0,y_0), y_{F_{\mathbb{R}}}(t,y_0))$.

Defining the growth rate of $\|x_{F_{\mathbb{R}}}(t,x_0,y_0)\|$ in finite time by

$$\lambda_f^t(x_0,y_0,F_{\mathbb{R}}) := \frac{1}{t}\ln\frac{\|x_{F_{\mathbb{R}}}(t,x_0,y_0)\|}{\|x_0\|},$$

we obtain the following estimate.

LEMMA 6.3. *Let δ, $h>0$ and let F be an optimal discrete feedback with respect to v_δ^h for the linearization* (2.2). *Let $\sigma := \sup_{(s,y)\in\mathbb{S}^{d-1}\times K}\delta v_\delta^h(x,y)$. Then for any $\varepsilon>0$ there exists an interval $[C_-(\varepsilon), C^-(\varepsilon)]$ and a constant $\eta(\varepsilon)>0$ such that for all pairs of initial values x_0, y_0 where $x_0 \in B_{\eta(\varepsilon)}(0)$ the estimate*

$$\lambda_f^t(x_0,y_0,F_{\mathbb{R}}) \le \sigma + \varepsilon$$

holds for some $t \in [C_-(\varepsilon), C^-(\varepsilon)]$.

Proof. For a fixed pair of initial values (x_0,y_0) and a control function $u(\cdot)\in\mathcal{U}$ we abbreviate $x(t) := x(t,x_0,y_0,u(\cdot))$ and define

$$\tilde{h}(t,s,y,u) = \frac{f(x(t),y,u)}{\|x(t)\|} - \left\langle \frac{f(x(t),y,u)}{\|x(t)\|}, s\right\rangle s$$

for $s\in\mathbb{S}^{d-1}$. With $s_0 := x_0/\|x_0\|$ and $\tilde{s}(t,s_0,y_0,u(\cdot)) := x(t)/\|x(t)\|$ it follows that

$$\dot{\tilde{s}}(t,s_0,y_0,u(\cdot)) = \tilde{h}(t,\tilde{s}(t,s_0,y_0,u(\cdot)), y(t,y_0,u(\cdot)), u(t));$$

hence the projection of the trajectory $x(t)$ onto $\mathbb{S}$ forms a solution trajectory of this time varying control system.

Now let $x_i := x_{F_{\mathbb{R}}}(ih,x_0,y_0)$. Using Lemma 6.2 we obtain

$$\left\| \frac{z(h,x_i,y_i,u)}{\|z(h,x_i,y_i,u)\|} - \frac{x(h,x_i,y_i,u)}{\|x(h,x_i,y_i,u)\|}\right\| \le hC_1\|x_i\|.$$

By Lemma 6.2, x_i can be made arbitrarily small for each fixed $i\in\mathbb{N}$ by choosing x_0 sufficiently small, and we can use Proposition 5.1 with $s = z/\|z\|$ and $\tilde{s} = x/\|x\|$ in order to obtain the estimate

$$\delta\tilde{J}_\delta(x_0,y_0,F) \le \sigma + \frac{\varepsilon}{4}$$

for all sufficiently small x_0.

From the linearization estimates we obtain

$$\left| \frac{x_i^t \tilde{f}(x_i, u)}{\|x_i\|^2} \right| \leq C_f \|x_i\| \leq C_f e^{\gamma t} \|x_0\|$$

for all sufficiently small $\|x_i\|$, i.e., all sufficiently small $\|x_0\|$.

Hence $\|q(x_{F_{\mathbb{R}}}(t, x_0, y_0), \cdot, \cdot)/\|x_{F_{\mathbb{R}}}(t, x_0, y_0)\| - q_f(x_{F_{\mathbb{R}}}(t, x_0, y_0), \cdot, \cdot)\|$ can be made arbitrarily small on each bounded time interval by choosing x_0 sufficiently close to the origin, and using [3, Lemma 4.1] we can conclude

$$\delta \int_0^\infty e^{-\delta\tau} q_f \left(x_{F_{\mathbb{R}}}(\tau, x, y),\, y_{F_{\mathbb{R}}}(\tau, y), F_{\mathbb{R}} \left(x_{F_{\mathbb{R}}} \left(\left[\frac{\tau}{h} \right] h, x, y \right),\, y_{F_{\mathbb{R}}} \left(\left[\frac{\tau}{h} \right] h, y \right) \right) \right) d\tau \leq \sigma + \frac{\varepsilon}{2}$$

for all sufficiently small $\|x_0\|$.

Now Lemma 3.5 (ii) yields the assertion. □

In order prove the stability of (6.1) the last thing that remains to do is putting together the trajectory pieces.

PROPOSITION 6.4. *Consider system* (2.1). *Let $K \subseteq M$ be a compact positively invariant set for the subsystem of* (2.1) *on M. Assume that the linearization* (2.2) *is asymptotically null controllable over K. Then there is $\delta > 0$ and $h > 0$ such that the system* (6.1) *with the discrete feedback $F_{\mathbb{R}}$ is uniformly exponentially stable in some neighborhood of the origin.*

Proof. From the assumptions on the linearization, Lemma 6.3 can be applied with $\sigma < 0$.

Hence for all sufficiently small initial values $\|x_0\|$ there exists a $t \in [C_-(\varepsilon), C^-(\varepsilon)]$ such that

$$\frac{1}{t} \ln \frac{\|x_{F_{\mathbb{R}}}(t, x_0, y_0)\|}{\|x_0\|} \leq \sigma + \varepsilon < 0.$$

Abbreviating $x_1 := x_{F_{\mathbb{R}}}(t, x_0, y_0)$, it holds that $\|x_1\| < \|x_0\|$. Thus we can proceed inductively as in the proof of Proposition 3.3 and the assertion follows. □

This proposition gives a characterization of exponential discrete feedback stabilizability by looking at its linearization. However, we would also like to have a characterization in terms of the nonlinear system itself. Clearly, since we are dealing with linearizations, asymptotic null controllability of the nonlinear system is not sufficient; see, e.g., [4, Example 15].

In fact, even exponential null controllability is not sufficient, as the following example shows. Consider

$$\dot{x} = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} x + u_1 \begin{pmatrix} -\frac{1}{2} & 0 \\ -\frac{1}{3} & \frac{1}{2} \end{pmatrix} x + u_2 \begin{pmatrix} -\frac{1}{2} & 0 \\ \frac{1}{3} & \frac{1}{2} \end{pmatrix} x + u_3 \begin{pmatrix} x_2^2 \\ 0 \end{pmatrix},$$

where $U = [-1, 1]^3$.

We claim that the linearized system is not asymptotically null controllable: looking at the initial values $z_0 = (0,\, z_2)^T$, $z_2 > 0$, it is easily seen that

$$A(u) z_0 = \begin{pmatrix} 0 \\ (1 + \frac{1}{2}u_1 + \frac{1}{2}u_2) z_2 \end{pmatrix}.$$

Denoting the solution by $z(t, x_0, u(\cdot)) = (z_1(t, z_0, u(\cdot)), z_2(t, z_0, u(\cdot)))^T$, we obtain $z_1(t, z_0, u(\cdot)) \equiv 0$ and $z_2(t, z_0, u(\cdot)) \geq z_2$ since $(1 + \frac{1}{2}u_1(t) + \frac{1}{2}u_2(t)) \geq 0$ for all $u(\cdot) \in U$ and all $t \geq 0$. Thus we can conclude that $\|z(t, z_0, u(\cdot))\| \geq \|z_2\|$ for all $u(\cdot) \in \mathcal{U}$, meaning that for all initial values z_0 of the considered form no possible trajectory converges to the origin, which implies our claim.

However, for any $x = (x_1, x_2)^T \in \mathcal{C}$, where $\mathcal{C}$ is the cone defined by

$$\mathcal{C} := \left\{ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2 \;\middle|\; \left|\frac{x_2}{x_1}\right| < \frac{1}{10} \right\},$$

we can choose the control $u_x = (u_{x1}, u_{x2}, u_{x3}) := (\frac{2x_2}{x_1/3 - x_2}, 0, 0) \in U$. Then a simple computation yields

$$f(x, u_x) = \left(-1 - \frac{1}{2}u_{x1}\right) x,$$

and since $u_{\alpha x} = u_x$ for all $\alpha \in \mathbb{R} \setminus \{0\}$ the solution for $u_x(\cdot) \equiv u_x$ satisfies

$$x(t, x, u_x(\cdot)) = e^{(-1 - \frac{1}{2}u_{x1})t} x.$$

Hence the corresponding trajectory satisfies $\|x(t, x, u_x)\| \leq e^{-\frac{1}{2}t}\|x\|$ and thus converges to the origin exponentially fast.

For all initial values $x_0 \in \mathbb{R}^2 \setminus \mathcal{C}$ we choose $u(t) \equiv (-1, -1, \operatorname{sgn}(x_1))$ (with the convention $\operatorname{sgn}(0) = 1$) as long as the corresponding trajectory stays outside $\mathcal{C}$ and switch to u_x from above once the trajectory reaches a point $x \in \mathcal{C}$.

Using this control function, any trajectory will enter the cone $\mathcal{C}$ in some finite time and then converge to the origin exponentially fast; thus the overall trajectory also converges to the origin exponentially fast. Hence the nonlinear system is exponentially null controllable, although the semilinear system is not even asymptotically null controllable. Thus exponential null controllability of the nonlinear system does not imply asymptotic null controllability of the linearized system.

In order to formulate the desired result we therefore need the notion of uniform exponential null controllability.

THEOREM 6.5. *Consider system* (2.1). *Let $K \subseteq M$ be a compact positively invariant set for the subsystem of* (2.1) *on M. Then the following properties are equivalent:*

(i) (2.1) *is (locally) uniformly exponentially null controllable over K.*
(ii) (2.2) *is asymptotically null controllable over K.*
(iii) *There is $h > 0$ and a discrete feedback that (locally) stabilizes* (2.1) *uniformly exponentially over K.*

Proof. "(ii)⇒(iii)" is Proposition 6.4; "(iii)⇒(i)" is immediately clear. It remains to show "(i)⇒(ii)":

Let $B(0)$ be the neighborhood in which uniform exponential null controllability holds. From (i) it follows that for any $\varepsilon > 0$ there exists a $T > 0$ such that for all $(x_0, y_0) \in B(0) \times K$ there exists a control function $u_{(x_0,y_0)}(\cdot) \in \mathcal{U}$ with

$$\lambda_f^T(x_0, y_0, u_{(x_0,y_0)}(\cdot)) < -\alpha + \varepsilon.$$

Using the estimates from the proof of Lemma 6.3, we obtain for the growth rate of (2.2),

$$\lambda^T(x_0, y_0, u_{(x_0,y_0)}(\cdot)) < -\alpha + 2\varepsilon$$

for x_0 sufficiently close to the origin. Due to the linearity of (2.2) this estimate holds for all $x_0 \in \mathbb{R}^d$. Now by induction we obtain the assertion as in the proof of Proposition 3.3. □

This theorem shows in particular that any attempt to stabilize (2.1) at a singular point by using its linearization must fail if uniform exponential controllability is not satisfied, because the linearized system will not even be asymptotically null controllable. Conversely, exponential discrete feedback stabilization is always possible under this condition. We have therefore obtained the strongest result possible within the linearization approach.

A related result has been developed in [5] using Lyapunov functions: it is shown that for nonlinear systems, asymptotic controllability to a (not necessarily singular) point x implies stabilizability by means of a discrete feedback, where in order to reach x the step size h must tend to 0. The result can therefore be interpreted as a kind of practical stabilization. In contrast to this practical stability here we obtain *exponential* stability using a discrete feedback with a *fixed* step size.

7. Conclusions. In this paper we developed results on the relation between null controllability and exponential stabilization by using a discrete feedback law for nonlinear systems at singular points. The construction of the feedback is obtained by minimizing the Lyapunov exponent of the linearized system, which forms a semilinear system. For semilinear systems, asymptotic null controllability and exponential stabilizability by a discrete feedback turned out to be equivalent. For general nonlinear systems the equivalence between uniform exponential controllability and uniform exponential stabilizability has been shown. An example illustrated that uniform exponential controllability is in fact a necessary condition for the applicability of linearization techniques.

8. Appendix: Proof of Lemma 3.5. (i) Fix $\varepsilon > 0$. We may assume $\sigma = -\varepsilon$; otherwise we use $q - \sigma - \varepsilon$ and $M_q + \sigma + \varepsilon$ instead of q and M_q. Hence there exists $0 \le T_0 < T$ such that

$$\int_0^{T_0} q(\tau)\,d\tau = -T_0\varepsilon \quad \text{and} \quad \int_0^{t} q(\tau)\,d\tau < -t\varepsilon \quad \text{for all } t > T_0. \tag{8.1}$$

This yields

$$\int_{T_0}^{t} q(\tau)\,d\tau < (t - T_0)(-\varepsilon) \quad \text{for all } t > T_0. \tag{8.2}$$

Since for all $y \in [0,1)$ the inequality $\ln(1-y) \le -y$ and hence $e^{-y} \ge 1 - y$ holds, we obtain

$$\begin{aligned}\left| \int_0^{T_0} q(\tau)\,d\tau - \int_0^{T_0} e^{-\delta t} q(\tau)\,d\tau \right| &\le T_0(1 - e^{-\delta T_0})M \\ &\le T(1 - e^{-\delta T})M \\ &\le \delta T^2 M.\end{aligned}$$

Thus the inequality in (8.1) implies for $\delta < \frac{\varepsilon}{MT} < \frac{\varepsilon}{MT_0}$

$$\delta \int_0^{T_0} e^{-\delta\tau} q(\tau)\,d\tau < \delta(-T_0\varepsilon + \delta T^2 M) < 0. \tag{8.3}$$

Now fix $\tilde{\varepsilon} > 0$. Since q is bounded there exists $T_1 > T_0$ such that

$$\left| \int_{T_0}^{\infty} e^{-\delta\tau} q(\tau)\, d\tau - \int_{T_0}^{T_1} e^{-\delta\tau} q(\tau)\, d\tau \right| \leq \tilde{\varepsilon}. \tag{8.4}$$

From estimate (8.2) choose $\gamma > 0$ maximal with the property

$$\int_{T_0}^{T_1} q^+(\tau)\, d\tau - \int_{T_0}^{T_1-\gamma} q^-(\tau)\, d\tau = 0, \tag{8.5}$$

where q^+ and q^- denote the positive and negative parts of q, respectively. Now we can define a monotonically decreasing sequence τ_i, $i \in \mathbb{N}$ by $\tau_1 := T_1,\ \tau_2 := T_1 - \gamma$, and

$$\tau_{i+1} := \min \left\{ t \in [T_0, \tau_i] \mid -\int_t^{\tau_i} q^-(\tau)\, d\tau + \int_{\tau_i}^{\tau_{i-1}} q^+(\tau)\, d\tau = 0 \right\}.$$

This sequence is well defined: assume that there exists τ_i for some $i \geq 2$. In the case $i > 2$ for all j with $i \geq j \geq 3$, the equality

$$-\int_{T_0}^{\tau_j} q^-(s)\, ds + \int_{\tau_j}^{\tau_{j-1}} q^+(s)\, ds = -\int_{T_0}^{\tau_{j-1}} q^-(s)\, ds + \int_{\tau_{j-1}}^{\tau_{j-2}} q^+(s)\, ds$$

holds, and by induction and the choice of γ in (8.5) it follows

$$-\int_{T_0}^{\tau_i} q^-(\tau)\, d\tau + \int_{\tau_i}^{\tau_{i-1}} q^+(\tau)\, d\tau = -\int_{T_0}^{T_1-\gamma} q^-(\tau)\, d\tau + \int_{T_1-\gamma}^{T_1} q^+(\tau)\, d\tau \ \leq 0.$$

This guarantees the existence of τ_{i+1}. Since (τ_i) is monotone and bounded, the sequence converges to some $\tilde{\tau} \geq T_0$. We claim $\tilde{\tau} = T_0$.

By the definition of τ_i it follows that $-\int_{\tau_{i+1}}^{T_1-\gamma} q^-(\tau)\, d\tau + \int_{\tau_i}^{T_1} q^+(\tau)\, d\tau = 0$. The convergence $\tau_i \to \tilde{\tau}$ yields the equality

$$-\int_{\tilde{\tau}}^{T_1-\gamma} q^-(\tau)\, d\tau + \int_{\tilde{\tau}}^{T_1} q^+(\tau)\, d\tau = 0.$$

This implies $\int_{T_0}^{\tilde{\tau}} q(\tau)\, d\tau = 0$, which shows the asserted equality using (8.2).

Hence we can choose $k \in \mathbb{N}$ such that $|\tau_{k-1} - T_0| \leq \tilde{\varepsilon}$ and replace τ_k by $\tau_k = T_0$. Thus we can estimate

$$\begin{aligned}
\int_{T_0}^{T_1} e^{-\delta\tau} q(\tau)\, d\tau &\leq \sum_{i=2}^{k-1} \left(-\int_{\tau_{i+1}}^{\tau_i} e^{-\delta\tau} q^-(\tau)\, d\tau + \int_{\tau_i}^{\tau_{i-1}} e^{-\delta\tau} q^+(\tau)\, d\tau \right) + M\tilde{\varepsilon} \\
&\leq \sum_{i=2}^{k-1} \underbrace{\left(-\int_{\tau_{i+1}}^{\tau_i} e^{-\delta\tau_i} q^-(\tau) d\tau + \int_{\tau_i}^{\tau_{i-1}} e^{-\delta\tau_i} q^+(\tau) d\tau \right)}_{=0} + M\tilde{\varepsilon} \\
&= M\tilde{\varepsilon}.
\end{aligned}$$

In connection with (8.3) and (8.4) this yields

$$\int_0^{\infty} e^{-\delta\tau} q(\tau)\, d\tau = \int_0^{T_0} e^{-\delta\tau} q(\tau)\, d\tau + \int_{T_0}^{T_1} e^{-\delta\tau} q(\tau)\, d\tau + \int_{T_1}^{\infty} e^{-\delta\tau} q(\tau)\, d\tau < 0 + M\tilde{\varepsilon} + \tilde{\varepsilon}.$$

Since $\tilde{\varepsilon} > 0$ was arbitrary this proves (i).

(ii) Assume the opposite: let

$$\frac{1}{t}\int_0^t q(\tau)d\tau > \sigma + \varepsilon \text{ for all } t \in \left[\frac{\varepsilon}{(4M+4\sigma+\varepsilon)\delta}, -\frac{\ln\frac{\varepsilon}{4M}}{\delta}\right].$$

We define $\tilde{q}$ via

$$\tilde{q}(\tau) := \begin{cases} q(\tau), & \tau \le -\frac{\ln\frac{\varepsilon}{4M}}{\delta}, \\ \sigma+\varepsilon, & \tau > -\frac{\ln\frac{\varepsilon}{4M}}{\delta}. \end{cases}$$

This yields

$$\frac{1}{t}\int_0^t \tilde{q}(\tau)d\tau > \sigma + \varepsilon \text{ for all } t \ge \frac{\varepsilon}{(4M+4\sigma+\varepsilon)\delta},$$

and by (i) (with opposite signs and inequalities) we obtain

$$\delta\int_0^\infty e^{-\delta\tau}\tilde{q}(\tau)d\tau \ge \sigma + \frac{3}{4}\varepsilon.$$

Hence

$$\begin{aligned}\delta\int_0^\infty e^{-\delta\tau}q(\tau)d\tau &= \delta\int_0^\infty e^{-\delta\tau}\tilde{q}(\tau)d\tau - \delta\int_{-\frac{\ln\frac{\varepsilon}{4M}}{\delta}}^\infty \tilde{q}(\tau) - q(\tau)d\tau \\ &\ge \sigma + \frac{3}{4}\varepsilon - \delta\int_{-\frac{\ln\frac{\varepsilon}{4M}}{\delta}}^\infty e^{-\delta\tau}2Md\tau \quad = \quad \sigma + \frac{1}{4}\varepsilon,\end{aligned}$$

which contradicts the assumption on this discounted integral.

(iii) By (ii) for any $\varepsilon > 0$ there exist times $\tau(t)$ bounded from below and above such that

$$\frac{1}{\tau(t)}\int_0^{\tau(t)} q(t+\tau)d\tau < \sigma + \varepsilon.$$

Hence $\tau_0 := 0$, $\tau_{i+1} = \tau_i + \tau(\tau_i)$ defines a monotonically increasing sequence diverging to infinity for which there exists $a \in \mathbb{R}$ such that $\tau_{i+1} - \tau_i < a$ for all $i \in \mathbb{N}$. For arbitrary $T > 0$ let $\tau_{i(T)}$ be the maximal element of this sequence satisfying $\tau_{i(T)} \le T$. Thus we obtain

$$\begin{aligned}\limsup_{T\to\infty}\frac{1}{T}\int_0^T q(\tau)d\tau &= \limsup_{T\to\infty}\frac{1}{T}\left(\sum_{j=0}^{i(T)}\int_{\tau_{j-1}}^{\tau_j} q(\tau)d\tau + \int_{\tau_{i(T)}}^T q(\tau)d\tau\right) \\ &\le \limsup_{T\to\infty}\left(\sigma+\varepsilon+\frac{aM}{T}\right) \quad = \quad \sigma+\varepsilon.\end{aligned}$$

Since $\varepsilon > 0$ was arbitrary the assertion follows. □

REFERENCES

[1] P. Cannarsa and H. Frankowska, *Some characterizations of optimal trajectories in control theory*, SIAM J. Control Optim., 29 (1991), pp. 1322–1347.

[2] I. Capuzzo Dolcetta, *On a discrete approximation of the Hamilton-Jacobi equation of dynamic programming*, Appl. Math. Optim., 10 (1983), pp. 367–377.
[3] I. Capuzzo Dolcetta and H. Ishii, *Approximate solutions of the Bellman equation of deterministic control theory*, Appl. Math. Optim., 11 (1984), pp. 161–181.
[4] R. Chabour, G. Sallet, and J. Vivalda, *Stabilization of nonlinear systems: A bilinear approach*, Math. Control Signals Systems, 6 (1993), pp. 224–246.
[5] F. Clarke, Y. Ledyaev, E. Sontag, and A. Subbotin, *Asymptotic controllability implies feedback stabilization*, IEEE Trans. Automat. Control, 42(1997), pp. 1394–1407.
[6] F. Colonius and W. Kliemann, *Linear control semigroups acting on projective space*, J. Dyn. Differential Equations, 5 (1993), pp. 495–528.
[7] F. Colonius and W. Kliemann, *Maximal and minimal Lyapunov exponents of bilinear control systems*, J. Differential Equations, 101 (1993), pp. 232–275.
[8] F. Colonius and W. Kliemann, *Asymptotic null controllability of bilinear systems*, in Geometry in Nonlinear Control and Differential Inclusions, B. Jakubczyk and W. Respondek, eds., Banach Center Publications, Vol. 32, Warsaw, 1995, pp. 139–148.
[9] M. Falcone, *A numerical approach to the infinite horizon problem of deterministic control theory*, Appl. Math. Optim., 15 (1987), pp. 1–13. *Corrigenda*, ibid., 23 (1991), pp. 213–214.
[10] H. Frankowska, *Optimal trajectories associated with a solution of the contingent Hamilton-Jacobi Equation*, Appl. Math. Optim., 19 (1989), pp. 291–311.
[11] L. Grüne, *Discrete feedback stabilization of semilinear control systems*, ESAIM *Contrôle*, Optim. and Calc. of Var., 1 (1995/1996), pp. 207–224.
[12] L. Grüne, *Numerical stabilization of bilinear control systems*, SIAM J. Control Optim., 34 (1996), pp. 2024–2050.
[13] L. Grüne, *An adaptive grid scheme for the discrete Hamilton-Jacobi-Bellman equation*, Numer. Math., 75 (1997), pp. 319–337.
[14] L. Grüne, *Discrete feedback stabilization of nonlinear control systems at a singular point*, in Proceedings of the 4th European Control Conference, Brussels, Belgium, 1997, CD-ROM; European Union Control Association (EUCA), 1997, http://www.auto.ucl.ac.be/INMA/ECC97.html.
[15] W. Hahn, *Stability of Motion*, Springer-Verlag, Heidelberg, 1967.
[16] H. Hermes, *On stabilizing feedback attitude control*, J. Optim. Theory Appl., 31 (1980), pp. 373–384.
[17] H. Hermes, *On the synthesis of stabilizing feedback control via Lie algebraic methods*, SIAM J. Control Optim., 18 (1980), pp. 352–361.
[18] N. N. Krasovskiĭ and A. I. Subbotin, *Game-Theoretical Control Problems*, Springer-Verlag, New York, 1988.
[19] P. L. Lions, *Generalized Solutions of Hamilton-Jacobi Equations*, Pitman, London, 1982.
[20] E. D. Sontag, *Nonlinear regulation: The piecewise linear approach*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 346–358.
[21] E. D. Sontag, *Mathematical Control Theory*, Springer-Verlag, New York, 1990.
[22] E. D. SontagT, *Feedback stabilization using two-hidden-layer nets*, IEEE Trans. Neural Networks, 3 (1992), pp. 981–990.
[23] M. Vidyasagar, *Nonlinear System Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1993.
[24] J. Zabczyk, *Some comments on stabilizability*, Appl. Math. Optim., 19 (1989), pp. 1–9.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 5, pp. 1504–1538, September 1998

002

STOCHASTIC H^∞*

D. HINRICHSEN† AND A. J. PRITCHARD‡

Abstract. We consider stochastic linear plants which are controlled by dynamic output feedback and subjected to both deterministic and stochastic perturbations. Our objective is to develop an H^∞-type theory for such systems. We prove a bounded real lemma for stochastic systems with deterministic and stochastic perturbations. This enables us to obtain necessary and sufficient conditions for the existence of a stabilizing compensator which keeps the effect of the perturbations on the to-be-controlled output below a given threshhold $\gamma > 0$. In the deterministic case, the analogous conditions involve two uncoupled linear matrix inequalities, but in the stochastic setting we obtain coupled nonlinear matrix inequalities instead. The connection between H^∞ theory and stability radii is discussed and leads to a lower bound for the radii, which is shown to be tight in some special cases.

Key words. stochastic systems, state dependent noise, H^∞ control, bounded real lemma, matrix inequalities

AMS subject classifications. 93C55, 93D09, 93E15

PII. S0363012996301336

1. Introduction.

The objective of this paper is to develop an H^∞-type theory over infinite time horizons for the disturbance attenuation of stochastic systems by dynamic output feedback. We consider systems Σ described by Ito stochastic differential equations of the form

$$
\begin{aligned}
(1) \qquad dx(t) &= Ax(t)dt + A_0x(t)dw_1(t) + B_0v(t)dw_2(t) + B_1v(t)dt + B_2u(t)dt,\\
z(t) &= C_1x(t) + D_{11}v(t) + D_{12}u(t),\\
y(t) &= C_2x(t) + D_{21}v(t),
\end{aligned}
$$

where w_i, $i = 1, 2$ are zero mean scalar Wiener processes, not necessarily independent. In applications, such models are often obtained by linearization, and then $x(t)$, $z(t)$, and $u(t)$ represent deviations from desired fixed values of the state, the output, and the control (for instance, in a tracking problem; see [30]). We view v as an *unknown* finite energy stochastic disturbance which adversely affects the to-be-controlled output z (whose desired value is represented by 0). The disturbing effect is to be ameliorated via control action u based on dynamic feedback from the measured output y. A feedback controller $K : y \mapsto u$ has to be chosen in such a way that the closed loop system Σ_{cl} is stabilized. The effect of the disturbances on the to-be-controlled output z of Σ_{cl} is then described by the perturbation operator $\mathbb{L}_{cl} : v \mapsto z$ of Σ_{cl} which (for zero initial state) maps finite energy disturbance signals v into the corresponding finite energy output signals z of the closed loop system. The size of this linear operator is measured by the induced norm. The larger that this norm is, the larger is the effect of the unknown disturbance v on the to-be-controlled output z *in the worst case.*

*Received by the editors April 1, 1996; accepted for publication (in revised form) September 5, 1997; published electronically June 2, 1998.
http://www.siam.org/journals/sicon/36-5/30133.html
†Institut für Dynamische Systeme, Universität Bremen, D-28334 Bremen, Germany, (dh@mathematik.uni-bremen.de).
‡Mathematics Institute, University of Warwick, Coventry CV4 7AL, UK (ajp@maths.warwick.ac.uk).

The problem is to determine whether or not for each $\gamma > 0$ there exists a stabilizing controller K achieving $\|\mathbb{L}_{cl}\| < \gamma$. Moreover, we want to know how such controllers, if they exist, can be constructed.

In the deterministic case, the norm $\|\mathbb{L}_{cl}\|$ is given by the H^∞-norm of the associated rational transfer matrix, and so the theory dealing with the above problem is known as H^∞-control theory. In the present stochastic context the term H^∞ control may be a misnomer, but we use it nevertheless to refer, in a succinct and suggestive way, to the above disturbance attenuation problem.

System (1) may be regarded as a perturbed version of the stochastic system

$$\begin{aligned} \dot{x}(t) &= (A + A_0\dot{w}_1(t))x(t) + B_2u(t), \\ z(t) &= C_1x(t) + D_{12}u(t), \\ y(t) &= C_2x(t), \end{aligned} \tag{2}$$

representing a linear time-invariant system with multiplicative white noise. Such systems are widely considered in the stochastic literature, especially in stochastic stability analysis; see [10, ch. 6], [1, ch. 11], and [4, ch. 11]. By adding unknown disturbances to this equation we lay the groundwork for an analysis of *robust stability* of these systems; see section 5.

The disturbance of the state equation in our model (1) is composed of two parts, $B_1v(t)dt$ and $B_0v(t)dw_2(t)$. Although v is in general a stochastic vector, we view the first term as the *deterministic* and the second as the *stochastic component* of the disturbance. To motivate this terminology, let $v = \Delta z$, where Δ is an unknown matrix, and assume that the two Wiener processes in (1) are equal: $w_1 = w_2 = w$. Then the state equation in (1) reads

$$dx(t) = (A + B_1\Delta C_1)x(t)dt + (A_0 + B_0\Delta C_1)x(t)dw(t) + B_2u(t)dt. \tag{3}$$

So the deterministic disturbance term $B_1\Delta C_1$ represents a perturbation of A, i.e., of the deterministic parameters, and the stochastic disturbance term $B_0\Delta C_1$ represents a perturbation of the stochastic parameters of the system. The presence of both types of disturbances in (1) is essential to obtain a full generalization of the H^∞-control problem to the stochastic context. As a special case, it contains, on the one hand, $(A_0 = 0, B_0 = 0)$, the general deterministic H^∞-control problem (without regularity assumptions) as stated, e.g., in [8], [9], [17], [25], and [26]. On the other hand, it also includes the "purely stochastic" case where $B_1 = 0$, see [16].

It may seem odd that we use the same disturbance vector v in both the deterministic and stochastic disturbance terms. But this is in fact more general since distinct disturbance vectors v_0 and v_1 can be accounted for by setting $B_0 = [B_0^0\ 0]$, $B_1 = [0\ B_1^1]$, and $v = \left[\begin{smallmatrix} v_0 \\ v_1 \end{smallmatrix}\right]$. Similarly, it would simplify the situation if we assumed that the two Wiener processes w_1 and w_2 are independent. But we avoid this assumption in order to derive formulae which are equally applicable to the case where, e.g., $w_1 = w_2 = w$; see (3).

In order to keep the notational burden as low as possible, we do not deal with more general models or more general multiperturbation structures, where the single stochastic terms in (1) are replaced by sums of similar terms. In section 6 we make some comments about the extension of our results to such systems.

The main complication in the H^∞-control problem studied here is due to the presence of both deterministic and stochastic perturbations terms in (1). In the case of purely stochastic disturbances, the problem has been solved in [16]. The key result

on which the general solution will be based is the Stochastic Bounded Real Lemma which will be proved in the next section. This result states necessary and sufficient conditions for a given stochastic system to be stable with $\|\mathbb{L}\| < \gamma$. It is of independent interest, because it allows one to determine $\|\mathbb{L}\|$ which measures the influence of the disturbances in the worst case scenario. The centerpiece of our conditions is no longer a Riccati-type equation or the corresponding matrix inequality (as in the deterministic case) but is a rational matrix equation which appears to be new. For the associated matrix inequality we could not find existence results in the literature. Our proof proceeds—as in the theory of algebraic Riccati equations—via the study of a finite time optimization problem which, due to the structure of our disturbance model, has a number of subtleties.

While section 2 deals with a problem of system analysis (under which conditions does the input output operator $\mathbb{L}$ of a stable stochastic system have a norm $\|\mathbb{L}\| < \gamma$?) the synthesis problem of H^∞-control will be treated in section 3. Here we follow the linear matrix inequalities (LMI) approach developed for deterministic systems; see [8], [17]. The idea is to apply the Stochastic Bounded Real Lemma to the compensated system Σ_{cl} in such a way that the matrices of compensator parameters, which achieve stability and $\|\mathbb{L}_{cl}\| < \gamma$, are characterized by a linear matrix inequality. Then, applying the projection lemma [8] we are able to obtain necessary and sufficient conditions in terms of the given data. The result is a characterization in terms of a pair of matrix inequalities. But in contrast to the deterministic case the two matrix inequalities are coupled and nonlinear. Specializing to the case $A_0 = B_0 = 0$, however, the inequalities are linear and decoupled, and we regain the deterministic results as given in [8].

In section 4 we deal with the so-called regular case. The matrix inequalities—although still coupled and nonlinear—are greatly simplified via the regularity assumptions, and it is possible to derive explicit formulae for full order suboptimal compensators. In the deterministic context, this was the case for which the H^∞-control problem was first resolved via a pair of Riccati equations [5]. In the stochastic context, we do not obtain an analogous pair of rational matrix *equations*. In fact it has been shown for the special case $A_0 = B_1 = 0$ that, in general, it is not possible to replace the two inequalities by equalities; see [16]. Therefore, even for regular data, matrix *in*equalities seem to be indispensable tools of an H^∞-type optimal control theory in the stochastic context. For the deterministic case $A_0 = B_0 = 0$, however, our conditions reduce to the well-known Riccati inequalities from which the results in [5] follow via standard theorems about the relationship between Riccati inequalities and equations.

In section 5, we consider stability radii for a nominal system where all direct input output couplings are zero (a singular problem). If (2) is stable, $u = 0$, and $v = \Delta z$, the associated stability radius is the maximum ρ such that all the perturbed systems with $\|\Delta\| < \rho$ are stable. We will use the results of section 2 to obtain a lower bound for the radius. In the deterministic case there is a close relationship between the singular H^∞-control problem and the problem of maximizing the stability radius of a given system by state or dynamic output feedback; see [13]. We will use this relationship to show how the radius may be enhanced by feedback. However, we do not determine precise formulae for the stability radius or for the supremal stability radius achievable by feedback. The general problem of characterizing and maximizing stability radii of stochastic systems of the form (1) is still open.

In some concluding remarks (section 6) we will comment on further open problems and possible extensions of our work.

The stabilization of stochastic systems with multiplicative noise has been studied since the late sixties, particularly in the context of linear quadratic optimal control; see, e.g., [20], [27], and [29]. The subject of robust stabilization is of more recent vintage. An early reference is [28], where the problem is considered in an almost disturbance decoupling framework. More recently, a number of papers have been published which deal with robust stability and robust stabilization problems in the spirit of H^∞-control or the stability radius approach. In [6], El Ghaoui describes how the maximization of an estimate for the stability radius by state feedback can be formulated—for general multi noise structures—as a convex optimization problem over linear matrix inequalities. In [22], the norm of the perturbation operator of a time-varying stable linear system with state dependent noise ($B_0 = 0, B_1 = 0$) is related to a parametrized Riccati equation. Formulae for stability radii and supremal stability radii of stochastic systems have been obtained for various special cases. The first formula was derived for the case $A_0 = 0$, $B_1 = 0$ in [2]. Morozan [21] extended this formula to the case where $B_1 = 0$ and the nominal system contains a sum of white noise terms. The maximization of stability radii via state feedback was first considered in [13]. A full H^∞ and stability radius theory for stochastic multiperturbations of a deterministic system was developed in [16]. *However, all these characterizations of stability radii apply only to the special case of purely stochastic parameter disturbances* ($B_1 = 0$). First results concerning stability radii subjected to simultaneous deterministic and stochastic parameter perturbations have been presented in [15]. These are improved in the present paper. However, our main intention is to develop a counterpart of H^∞-control for linear stochastic systems. The conceptual difference between H^∞-control theory and the theory of stability radii, which has been blurred by the fact that they yield similar results for deterministic time-invariant linear systems, stands out more clearly in the stochastic context.

2. A stochastic version of the bounded real lemma. The main tool that we will use in our analysis of the stochastic disturbance attenuation problem is an extension of the bounded real lemma to stochastic systems. This result is of independent interest and, in fact, we regard it as the main result of this paper. In order to describe it we consider the following system

$$\begin{aligned} dx(t) &= Ax(t)dt + A_0x(t)dw_1(t) + B_0v(t)dw_2(t) + Bv(t)dt, \\ z(t) &= Cx(t) + Dv(t), \end{aligned} \tag{4}$$

where

(5)

$$(A, A_0, B_0, B, C, D) \in \mathbb{K}^{n\times n} \times \mathbb{K}^{n\times n} \times \mathbb{K}^{n\times \ell} \times \mathbb{K}^{n\times \ell} \times \mathbb{K}^{q\times n} \times \mathbb{K}^{q\times \ell}, \qquad \mathbb{K} = \mathbb{R} \text{ or } \mathbb{C}.$$

w_1, w_2 are zero mean real scalar Wiener processes on a probability space $(\Omega, \mathcal{F}, \mu)$ relative to an increasing family $(\mathcal{F}_t)_{t\in\mathbb{R}_+}$ of σ-algebras $\mathcal{F}_t \subset \mathcal{F}$. We assume that

$$\mathcal{E}((w_i(t) - w_i(s))(w_j(t) - w_j(s))) = q_{ij}(t - s), \quad i, j = 1, 2, \quad t, s \in \mathbb{R}_+, \quad t > s.$$

So $Q = (q_{ij})$ is the incremental covariance matrix of the two-dimensional Wiener process

$$\begin{bmatrix} w_1(t) \\ w_2(t) \end{bmatrix}.$$

In (4), the input process $v(t)$ is viewed as a stochastic disturbance and the output process $z(t)$ is viewed as a vector of the to-be-controlled variables. The system equation contains multiplicative state and input dependent noise terms which may be interpreted as white noise parameter perturbations of the following matrices A and B:

$$dx(t) = (A + A_0\dot{w}_1(t))x(t)dt + (B + B_0\dot{w}_2(t))v(t)dt.$$

In this paper, we provide all spaces $\mathbb{K}^k$, $k \geq 1$ with the usual inner product $\langle \cdot, \cdot \rangle$ and the corresponding 2-norm $\| \cdot \|$. Let $L^2(\Omega, \mathbb{K}^k)$ denote the space of square-integrable $\mathbb{K}^k$-valued functions (modulo equivalence) on the probability space $(\Omega, \mathcal{F}, \mu)$. For any $0 < T \leq \infty$, we write $[0, T]$ for the closure of the open interval $(0, T)$ in $\mathbb{R}$ and denote by $L^2_w([0, T]; L^2(\Omega, \mathbb{K}^k))$ the space of nonanticipative stochastic processes $y(\cdot) = (y(t))_{t\in[0,T]}$ with respect to $(\mathcal{F}_t)_{t\in[0,T]}$ (see, e.g., [7]) satisfying

$$\|y(\cdot)\|^2_{L^2_w} = \mathcal{E}\left(\int_0^T \|y(t)\|^2 dt\right) = \int_0^T \mathcal{E}(\|y(t)\|^2)dt < \infty. \tag{6}$$

For arbitrary $0 < T < \infty$ and $(v, x^0) \in L^2_w([0, T]; L^2(\Omega, \mathbb{K}^\ell)) \times \mathbb{K}^n$, there exists a unique solution $x(\cdot) = x(\cdot, v, x^0) \in L^2_w([0, T]; L^2(\Omega, \mathbb{K}^n))$ of (4) with $x(0) = x^0$ [18], i.e., $x(\cdot)$ is a continuous nonanticipative stochastic process satisfying the Ito integral equation

(7)

$$x(t) = x^0 + \int_0^t (Ax(s) + Bv(s))ds + \int_0^t [A_0x(s)\ B_0v(s)]d\begin{bmatrix} w_1(s) \\ w_2(s) \end{bmatrix}, \quad t \in [0, T].$$

Moreover, $x(\cdot)$ has bounded second moments on $[0, T]$.

Definition 2.1. *The system* (4) *is called* internally stable *if there exists a constant $c > 0$ such that*

$$\mathcal{E}\int_0^\infty \|x(t)\|^2 dt \leq c\|x^0\|^2, \qquad x^0 \in \mathbb{K}^n,$$

where $x(\cdot) = x(\cdot; 0, x^0)$ is the free *trajectory of* (7) *starting at x^0 (i.e., $v = 0$).*

It has been shown (see [4]) that an equivalent condition is that there exist constants $M \geq 1$, $\omega > 0$ such that

$$\mathcal{E}\|x(t; 0, x^0)\|^2 \leq Me^{-\omega t}\|x^0\|^2 \quad \text{for all } x^0 \in \mathbb{K}^n, \quad t \geq 0.$$

Let $\mathcal{H}_n(\mathbb{K})$ denote the set of Hermitian matrices in $\mathbb{K}^{n\times n}$. It is known [4] that (4) is stable in the above sense if and only if there exists $P \in \mathcal{H}_n(\mathbb{K})$, $P \prec 0$ such that

$$PA + A^*P + q_{11}A_0^*PA_0 = I_n. \tag{8}$$

It is easily seen that in this stability criterion the identity matrix (on the right-hand side of (8)) may be replaced by any other positive definite matrix $Q_0 \in \mathcal{H}_n(\mathbb{K})$.

The following definition generalizes the concept of finite gain L^2 stability from deterministic input output systems to stochastic systems of the form (4).

Definition 2.2. *The system* (4) *is said to be* externally stable *or* L^2 input-output stable *if, for every $v(\cdot) \in L^2_w(\mathbb{R}_+; L^2(\Omega, \mathbb{K}^\ell))$,*

$$z(\cdot) = Cx(\cdot, v, 0) + Dv(\cdot) \in L^2_w(\mathbb{R}_+; L^2(\Omega, \mathbb{K}^q)),$$

and there exists a constant $\gamma \geq 0$ *such that*

$$\|z(\cdot)\|_{L_w^2(\mathbb{R}_+;L^2(\Omega,\mathbb{K}^q))} \leq \gamma \|v(\cdot)\|_{L_w^2(\mathbb{R}_+;L^2(\Omega,\mathbb{K}^\ell))}, \qquad v \in L_w^2(\mathbb{R}_+;L^2(\Omega,\mathbb{K}^\ell)). \tag{9}$$

DEFINITION 2.3. *Suppose that* (4) *is externally stable. The operator*

$$\mathbb{L} : L_w^2(\mathbb{R}_+;L^2(\Omega,\mathbb{K}^\ell)) \to L_w^2(\mathbb{R}_+;L^2(\Omega,\mathbb{K}^q)),$$

defined by

$$(\mathbb{L}v)(t) = Cx(t,v,0) + Dv(t), \quad t \geq 0, \quad v(\cdot) \in L_w^2(\mathbb{R}_+;L^2(\Omega,\mathbb{K}^\ell)), \tag{10}$$

is called the perturbation operator *of* (4). *Its norm is defined as the minimal* $\gamma \geq 0$ *such that* (9) *is satisfied, i.e.,*

$$\|\mathbb{L}\| = \sup_{v \in L_w^2(\mathbb{R}_+;L^2(\Omega,\mathbb{K}^\ell)), v \neq 0} \frac{\|Cx(\cdot,v,0) + Dv(\cdot)\|_{L_w^2(\mathbb{R}_+;L^2(\Omega,\mathbb{K}^q))}}{\|v(\cdot)\|_{L_w^2(\mathbb{R}_+;L^2(\Omega,\mathbb{K}^\ell))}}. \tag{11}$$

$\|\mathbb{L}\|$ *is a measure of the worst effect the stochastic disturbance* $v(\cdot)$ *may have on the to-be-controlled output* $z(\cdot)$ *of the system. Therefore it is important to find a way of determining the norm* $\|L\|$. *The stochastic bounded real lemma which we will derive in this section provides a method for computing* $\|\mathbb{L}\|$.

We proceed by associating a finite time quadratic cost functional with the problem

$$J_T^{\gamma^2}(x^0,v) = \int_0^T \mathcal{E}[\gamma^2\|v(t)\|^2 - \|z(t)\|^2]dt = \int_0^T \mathcal{E}[\gamma^2\|v(t)\|^2 - \|Cx(t) + Dv(t)\|^2]dt, \tag{12}$$

where $x(\cdot) = x(\cdot,v,x^0)$ denotes the solution of (4) with $x(0) = x^0$ and $v(\cdot) \in L_w^2 = L_w^2([0,T];L^2(\Omega,\mathbb{K}^\ell))$, and $z(\cdot) = z(\cdot,v,x^0)$ is the corresponding output. We will see that the problem of minimizing this functional will lead us to a solution of the supremum problem on the right-hand side of (11). Formally, the problem of minimizing $J_T^{\gamma^2}(x^0,v)$ has the form of an optimal control problem and so in our development in this section we will refer to the disturbance v as a "control." Our first step is to show that an internally stable system (4) is also externally stable. For every $P \in \mathcal{H}_n(\mathbb{K})$, we set

$$M(P) = \begin{bmatrix} PA + A^*P + q_{11}A_0^*PA_0 - C^*C & PB + q_{12}A_0^*PB_0 - C^*D \\ B^*P + q_{12}B_0^*PA_0 - D^*C & \gamma^2 I_\ell + q_{22}B_0^*PB_0 - D^*D \end{bmatrix}. \tag{13}$$

LEMMA 2.4. *Suppose* $P(\cdot) : [0,T] \mapsto \mathcal{H}_n(\mathbb{K})$ *is continuously differentiable,* $T > 0$. *Then for every* $x^0 \in \mathbb{K}^n, v(\cdot) \in L_w^2$,

$$\begin{aligned} J_T^{\gamma^2}(x^0,v) &= \langle x^0, P(0)x^0\rangle - \mathcal{E}\langle x(T), P(T)x(T)\rangle \\ &\quad + \int_0^T \mathcal{E}\left(\langle x(t), \dot{P}(t)x(t)\rangle + \left\langle \begin{bmatrix} x(t) \\ v(t) \end{bmatrix}, M(P(t)) \begin{bmatrix} x(t) \\ v(t) \end{bmatrix} \right\rangle\right) dt, \end{aligned} \tag{14}$$

where $M(P)$ *is defined by* (13) *and* $x(\cdot) = x(\cdot,v,x^0)$.

Proof. Let $x^0 \in \mathbb{K}^n$, $v(\cdot) \in L^2_w$, and let $x(\cdot) = x(\cdot, v, x^0)$ denote the corresponding solution of (4). Then the vector function $\varphi(s) = (Ax(s) + Bv(s))$ and the $n \times 2$ matrix function $\Phi(s) = [A_0 x(s)\ B_0 v(s)]$ satisfy the conditions of Ito's lemma (see [4, ch. 4, section 5]) and, by (7),

$$x(t) = x^0 + \int_0^t \varphi(s)ds + \int_0^t \Phi(s)d \begin{bmatrix} w_1(s) \\ w_2(s) \end{bmatrix}, \qquad t \in [0, T].$$

Applying Ito's formula to $F(t, x(t)) = \langle x(t),\ P(t)x(t)\rangle$ and taking expectations we obtain, for every $T > 0$.

$$\begin{aligned}
\mathcal{E}\langle x(T), P(T)x(T)\rangle - \langle x^0, P(0)x^0\rangle &= \mathcal{E}\int_0^T \langle x(t), \dot{P}(t)x(t)\rangle dt \\
&+ \mathcal{E}\int_0^T 2\mathrm{Re}\left\langle P(t)x(t), \Phi(t)d \begin{bmatrix} w_1(s) \\ w_2(s) \end{bmatrix} \right\rangle \\
&+ \mathcal{E}\int_0^T 2\mathrm{Re}\langle P(t)x(t), Ax(t) + Bv(t)\rangle dt \\
&+ \mathcal{E}\int_0^T \mathrm{tr}\{P(t)[A_0x(t)\ B_0v(t)]Q[A_0x(t)\ B_0v(t)]^*\}dt,
\end{aligned}$$

where tr denotes the trace. We first prove (14) under the condition that $v(\cdot) \in L^2_w$ is bounded, i.e.,

$$\exists c > 0 : \qquad \|v(t, \omega)\|_{\mathbb{K}^\ell} \le c, \quad (t, \omega) \in [0, T] \times \Omega.$$

Applying an estimate for the moments of $x(\cdot)$ (see [19, p. 81, Corollary 6]) there exist constants $c_0, c_1 > 0$ such that

$$\mathcal{E}\|x(t, v, x^0)\|^2_{\mathbb{K}^n} \le c_0\|x^0\|^2_{\mathbb{K}^n} + c_1\mathcal{E}\int_0^t \|v(s)\|^2_{\mathbb{K}^\ell} ds. \tag{15}$$

Hence, $\mathcal{E}\|x(t, v, x^0)\|^2_{\mathbb{K}^n}$ is bounded on $[0, T]$ and therefore,

$$\begin{aligned}
\mathcal{E}\int_0^T \left\langle P(t)x(t), \Phi(t)d \begin{bmatrix} w_1(s) \\ w_2(s) \end{bmatrix} \right\rangle &= \mathcal{E}\int_0^T \langle P(t)x(t), A_0x(t)\rangle dw_1(t) \\
&+ \mathcal{E}\int_0^T \langle P(t)x(t), B_0v(t)\rangle dw_2(t) = 0.
\end{aligned}$$

Now

$$\begin{aligned}
&\mathrm{tr}\{P(t)[A_0x(t)\ B_0v(t)]Q[A_0x(t)\ B_0v(t)]^*\} \\
&= \mathrm{tr}\left\{ \begin{bmatrix} x(t)^*A_0^* \\ v(t)^*B_0^* \end{bmatrix} P(t)[A_0x(t)\ B_0v(t)]Q \right\} \\
&= \mathrm{tr}\left\{ \begin{bmatrix} x(t)^*A_0^*P(t)A_0x(t) & x(t)^*A_0^*P(t)B_0v(t) \\ v(t)^*B_0^*P(t)A_0x(t) & v(t)^*B_0^*P(t)B_0v(t) \end{bmatrix} \begin{bmatrix} q_{11} & q_{12} \\ q_{12} & q_{22} \end{bmatrix} \right\} \\
&= \left\langle \begin{bmatrix} x(t) \\ v(t) \end{bmatrix}, \begin{bmatrix} q_{11}A_0^*P(t)A_0 & q_{12}A_0^*P(t)B_0 \\ q_{12}B_0^*P(t)A_0 & q_{22}B_0^*P(t)B_0 \end{bmatrix} \begin{bmatrix} x(t) \\ v(t) \end{bmatrix} \right\rangle.
\end{aligned}$$

Hence,

$$
\begin{aligned}
&J_T^{\gamma^2}(x^0,v)+\mathcal{E}\langle x(T),P(T)x(T)\rangle-\langle x^0,P(0)x^0\rangle\\
&\quad=\ \mathcal{E}\int_0^T\Bigg\{\langle x(t),\dot P(t)x(t)\rangle+\gamma^2\|v(t)\|^2-\|Cx(t)+Dv(t)\|^2\\
&\qquad+\langle P(t)x(t),Ax(t)+Bv(t)\rangle+\langle Ax(t)+Bv(t),P(t)x(t)\rangle\\
&\qquad+\left\langle\begin{bmatrix}x(t)\\v(t)\end{bmatrix},\begin{bmatrix}q_{11}A_0^*P(t)A_0 & q_{12}A_0^*P(t)B_0\\ q_{12}B_0^*P(t)A_0 & q_{22}B_0^*P(t)B_0\end{bmatrix}\begin{bmatrix}x(t)\\v(t)\end{bmatrix}\right\rangle\Bigg\}dt\\
&\quad=\ \mathcal{E}\int_0^T\Bigg\{\langle x(t),\dot P(t)x(t)\rangle\\
&\qquad+\left\langle\begin{bmatrix}x(t)\\v(t)\end{bmatrix},\begin{bmatrix}P(t)A+A^*P(t)-C^*C & P(t)B-C^*D\\ B^*P(t)-D^*C & \gamma^2I-D^*D\end{bmatrix}\begin{bmatrix}x(t)\\v(t)\end{bmatrix}\right\rangle\\
&\qquad+\left\langle\begin{bmatrix}x(t)\\v(t)\end{bmatrix},\begin{bmatrix}q_{11}A_0^*P(t)A_0 & q_{12}A_0^*P(t)B_0\\ q_{12}B_0^*P(t)A_0 & q_{22}B_0^*P(t)B_0\end{bmatrix}\begin{bmatrix}x(t)\\v(t)\end{bmatrix}\right\rangle\Bigg\}dt\\
&\quad=\ \int_0^T\mathcal{E}\left\{\langle x(t),\dot P(t)x(t)\rangle+\left\langle\begin{bmatrix}x(t)\\v(t)\end{bmatrix},M(P(t))\begin{bmatrix}x(t)\\v(t)\end{bmatrix}\right\rangle\right\}dt.
\end{aligned}
$$

This proves (14) for all bounded $v(\cdot)\in L_w^2$ and $x^0\in\mathbb{K}^n$. Now consider the linear maps

$$
\begin{aligned}
L_w^2\times\mathbb{K}^n\to L_w^2([0,T];L^2(\Omega,\mathbb{K}^n)),\qquad (v,x^0)\mapsto x(\cdot,v,x^0),\\
L_w^2\times\mathbb{K}^n\to L^2(\Omega,\mathbb{K}^n),\qquad (v,x^0)\mapsto x(T,v,x^0),
\end{aligned}
$$

where we endow $L_w^2\times\mathbb{K}^n$ with the norm $\|(v,x^0)\|=(\|x^0\|_{\mathbb{K}^n}^2+\|v(\cdot)\|_{L_w^2}^2)^{1/2}$. By the estimate (15), these are bounded linear operators. As a consequence, for any fixed $x^0\in\mathbb{K}^n$ the left- and right-hand sides of (12) depend continuously on $v\in L_w^2$. They coincide on the linear subspace L_b^2 of bounded $v\in L_w^2$ which is dense in L_w^2. Therefore (12) holds for all $v\in L_w^2$, $x^0\in\mathbb{K}^n$. ∎

PROPOSITION 2.5. *Suppose* (4) *is internally stable. Then* (4) *is externally stable. Moreover, there exist* $\gamma>0$ *and* $P\in\mathcal{H}_n(\mathbb{K})$, $P\prec 0$ *such that*

(16)
$$
M(P)=\begin{bmatrix}PA+A^*P+q_{11}A_0^*PA_0-C^*C & PB+q_{12}A_0^*PB_0-C^*D\\ B^*P+q_{12}B_0^*PA_0-D^*C & \gamma^2I_\ell+q_{22}B_0^*PB_0-D^*D\end{bmatrix}\succ 0,
$$

and, for each pair $(\gamma,P)\in(0,\infty)\times\mathcal{H}_n(\mathbb{K})$ *satisfying* (16) *and* $P\prec 0$, *we have* $\|\mathbb{L}\|<\gamma$.

Proof. Since (4) is internally stable, there exists $P\in\mathcal{H}_n(\mathbb{K})$, $P\prec 0$ such that

$$
PA+A^*P+q_{11}A_0^*PA_0-C^*C\succ 0. \tag{17}
$$

For any Hermitian block matrix $M=\left[\begin{smallmatrix}M_{11} & M_{12}\\ M_{21} & M_{22}\end{smallmatrix}\right]\in\mathbb{K}^{(n+\ell)\times(n+\ell)}$, we have the well-known definiteness criterion

$$
M\succ 0\quad\Leftrightarrow\quad(M_{22}\succ 0\text{ and }M_{11}-M_{12}M_{22}^{-1}M_{21}\succ 0). \tag{18}
$$

Applying this criterion to $M(P)$ (with the above P) we see that, for γ sufficiently large, $M(P) \succ 0$. Hence, there exists a pair $(\gamma, P) \in (0,\infty) \times \mathcal{H}_n(\mathbb{K})$ satisfying (16) and $P \prec 0$.

Now assume that $(\gamma, P) \in (0,\infty) \times \mathcal{H}_n(\mathbb{K})$ is any pair satisfying (16) and $P \prec 0$. Choose $\varepsilon > 0$ sufficiently small such that $M(P) \succeq \varepsilon^2 I$. Then, setting $P(t) = P$ and $x^0 = 0$ in (14), we obtain for all $v(\cdot) \in L^2_w(\mathbb{R}_+; L^2(\Omega, \mathbb{K}^\ell))$ and all $T > 0$,

$$J_T^{\gamma^2}(0, v) = \int_0^T \mathcal{E}[\gamma^2 \|v(t)\|^2 - \|z(t)\|^2] dt \geq \varepsilon^2 \int_0^T \mathcal{E}\|v(t)\|^2 dt, \tag{19}$$

since $P \prec 0$. It follows that $z(\cdot) = z(\cdot, v, 0) \in L^2_w(\mathbb{R}_+; L^2(\Omega, \mathbb{K}^q))$ and

$$\|\mathbb{L}v\|^2_{L^2_w(\mathbb{R}_+; L^2(\Omega,\mathbb{K}^q))} = \int_0^\infty \mathcal{E}\|z(t)\|^2 dt \leq (\gamma^2 - \varepsilon^2) \int_0^\infty \mathcal{E}\|v(t)\|^2 dt$$

for all $v(\cdot) \in L^2_w(\mathbb{R}_+; L^2(\Omega, \mathbb{K}^\ell))$. This concludes the proof. ∎

Remark 2.6. (i) By setting $C = I_n$ and $D = 0$, we see that $x(\cdot, v, 0) \in L^2_w(\mathbb{R}_+; L^2(\Omega, \mathbb{K}^n))$ for $v(\cdot) \in L^2_w(\mathbb{R}_+; L^2(\Omega, \mathbb{K}^\ell))$, and since $x(t, v, x^0) = x(t, 0, x^0) + x(t, v, 0)$, we conclude that $x(\cdot, v, x^0) \in L^2_w(\mathbb{R}_+; L^2(\Omega, \mathbb{K}^n))$ for all $(v(\cdot), x^0) \in L^2_w \times \mathbb{K}^n$.

(ii) Suppose that $M(P) \succ 0$ for some $\gamma > 0$ and $P \in \mathcal{H}_n(\mathbb{K})$ with $P \prec 0$ (as in Proposition 2.5). Then there exists $\delta > 0$ such that $P \preceq -\delta^2 I$, and by (14),

$$\begin{aligned}\gamma^2 \int_0^\infty \mathcal{E}\|v(t)\|^2 dt \geq J_T^{\gamma^2}(x^0, v) &\geq \langle x^0, Px^0\rangle - \mathcal{E}\langle x(T), Px(T)\rangle \\ &\geq \langle x^0, Px^0\rangle + \delta^2 \mathcal{E}\|x(T)\|^2, \qquad T > 0.\end{aligned}$$

It follows that $\mathcal{E}\|x(t, v, x^0)\|^2$ is bounded in $t \in \mathbb{R}_+$, for all $(v(\cdot), x^0) \in L^2_w \times \mathbb{K}^n$.

COROLLARY 2.7. *Suppose that* (16) *holds for some pair* $(\gamma, P) \in (0,\infty) \times \mathcal{H}_n(\mathbb{K})$ *with* $P \prec 0$. *Then* (4) *is internally stable and* $\|\mathbb{L}\| < \gamma$.

Proof. Suppose that (16) holds. Since (16) implies (17), system (4) is internally stable. Hence, $\|\mathbb{L}\| < \gamma$ follows from the previous proposition. ∎

We will now show the converse of Corollary 2.7, i.e., we will prove the following characterization of the norm of the perturbation operator which can be viewed as a stochastic version of the bounded real lemma.

THEOREM 2.8 (stochastic bounded real lemma). *For any set of data* (5) *and any positive real number* γ, *the following statements are equivalent:*

(i) *The system* (4) *is internally stable and* $\|\mathbb{L}\| < \gamma$.

(ii) *There exists* $P \in \mathcal{H}_n(\mathbb{K})$ *such that* (16) *is satisfied.*

It remains to prove that (i) implies (ii). In order to do this we need a number of lemmata. Using the notation

$$H^{\gamma^2}(P) = \gamma^2 I_\ell + q_{22} B_0^* P B_0 - D^* D \in \mathcal{H}_\ell(\mathbb{K}), \quad K(P) = PB + q_{12} A_0^* P B_0 - C^* D \in \mathbb{K}^{n\times\ell},$$

we can write $M(P)$, defined by (13), in the following way:

$$M(P) = \begin{bmatrix} PA + A^*P + q_{11}A_0^*PA_0 - C^*C & K(P) \\ K(P)^* & H^{\gamma^2}(P) \end{bmatrix} \succ 0. \tag{20}$$

LEMMA 2.9. *Suppose* $F(\cdot) \in C([0,T], \mathbb{K}^{\ell\times n})$ *and* $P_F^{\gamma^2}(\cdot)$ *satisfies the linear differential matrix equation*

$$\begin{aligned}&\dot X(t) + X(t)(A + BF(t)) + (A + BF(t))^* X(t) + q_{11}A_0^* X(t) A_0 \\ &\quad + q_{22} F(t)^* B_0^* X(t) B_0 F(t) + q_{12} A_0^* X(t) B_0 F(t) \\ &\quad + q_{12} F(t)^* B_0^* X(t) A_0 + \gamma^2 F(t)^* F(t) - (C + DF(t))^*(C + DF(t)) = 0,\end{aligned} \tag{21}$$

with $P_F^{\gamma^2}(T) = 0$. *Then if* $v(\cdot) \in L_w^2([0,T]; L^2(\Omega, \mathbb{K}^\ell))$, *we have*

$$\begin{aligned} J_T^{\gamma^2}(x^0, v + Fx_F) &= \langle x^0, P_F^{\gamma^2}(0)x^0\rangle \\ &\quad + \int_0^T \mathcal{E}[\langle v, Nx_F\rangle + \langle Nx_F, v\rangle + \langle v, H^{\gamma^2}(P_F^{\gamma^2})v\rangle]dt, \end{aligned} \tag{22}$$

where $x_F(\cdot) = x_F(\cdot, v(\cdot), x^0) = x(\cdot, F(\cdot)x_F(\cdot) + v(\cdot), x^0)$ *is the solution of*

$$\begin{aligned} dx_F(t) &= (A + BF(t))x_F(t)dt + A_0x_F(t)dw_1(t) + B_0F(t)x_F(t)dw_2(t) \\ &\quad + B_0v(t)dw_2(t) + Bv(t)dt, \end{aligned} \tag{23}$$

with $x_F(0) = x^0$ *and* $N(t) = K(P_F^{\gamma^2}(t))^* + H^{\gamma^2}(P_F^{\gamma^2}(t))F(t)$. *In particular, if* $v = 0$, *then*

$$J_T^{\gamma^2}(x^0, Fx_F) = \langle x^0, P_F^{\gamma^2}(0)x^0\rangle. \tag{24}$$

Proof. The left-hand side of (21) can be written as

$$\dot{X}(t) + \begin{bmatrix} I \\ F(t) \end{bmatrix}^* \begin{bmatrix} X(t)A + A^*X(t) + q_{11}A_0^*X(t)A_0 - C^*C & X(t)B + q_{12}A_0^*X(t)B_0 - C^*D \\ B^*X(t) + q_{12}B_0^*X(t)A_0 - D^*C & \gamma^2 I + q_{22}B_0^*X(t)B_0 - D^*D \end{bmatrix} \begin{bmatrix} I \\ F(t) \end{bmatrix}.$$

Hence, $P_F^{\gamma^2}(t)$ satisfies

$$\dot{X}(t) + [I \quad F^*(t)]M(X(t)) \begin{bmatrix} I \\ F(t) \end{bmatrix} = 0, \qquad X(T) = 0. \tag{25}$$

Therefore, applying Lemma 2.4 with $P(\cdot) = P_F^{\gamma^2}(\cdot)$ and $F(\cdot)x_F(\cdot) + v(\cdot)$ for $v(\cdot)$, we obtain that $J_T^{\gamma^2}(x^0, Fx_F + v)$ is equal to

$$\begin{aligned} &\langle x^0, P_F^{\gamma^2}(0)x^0\rangle + \mathcal{E}\int_0^T \left\{ \langle x_F(t), \dot{P}_F^{\gamma^2}(t)x_F(t)\rangle \right. \\ &\qquad \left. + \begin{bmatrix} x_F(t) \\ F(t)x_F(t) + v(t) \end{bmatrix}^* M(P_F^{\gamma^2}(t)) \begin{bmatrix} x_F(t) \\ F(t)x_F(t) + v(t) \end{bmatrix} \right\} dt \\ &= \langle x^0, P_F^{\gamma^2}(0)x^0\rangle + \mathcal{E}\int_0^T [\langle v(t), N(t)x_F(t)\rangle \\ &\qquad + \langle N(t)x_F(t), v(t)\rangle + \langle v(t), H^{\gamma^2}(P_F^{\gamma^2}(t))v(t)\rangle]dt. \end{aligned}$$

Hence, (22) holds. Setting $v = 0$ in (22), we obtain (24). ∎

LEMMA 2.10. *Suppose* (4) *is internally stable and* $\|\mathbb{L}\| < \gamma$. *Then there exists* $c > 0$ *such that*

$$J_T^{\gamma^2}(x^0, v) \geq -c\|x^0\|^2, \quad x^0 \in \mathbb{K}^n, \quad v(\cdot) \in L^2_w([0,T]; L^2(\Omega, \mathbb{K}^\ell)), \quad T > 0. \tag{26}$$

Proof. Denote by $X_T(t)$ the solution of (21) with $F(t) \equiv 0$ and final value $X_T(T) = 0$, i.e., $X_T(t)$ solves

$$\dot{X}(t) + X(t)A + A^*X(t) + q_{11}A_0^*X(t)A_0 - C^*C = 0, \qquad X(T) = 0.$$

By time invariance, $X_T(t) = X_{T-t}(0)$. By linearity, we have $x(t, v, x^0) = x(t, 0, x^0) + x(t, v, 0)$. Applying (22) with $F(t) \equiv 0$, we get

$$\begin{aligned} J_T^{\gamma^2}(x^0, v) - J_T^{\gamma^2}(0, v) &= \langle x^0, X_T(0)x^0\rangle \\ &\quad + \mathcal{E}\int_0^T [\langle v(t), N_T(t)x(t,0,x^0)\rangle + \langle N_T(t)x(t,0,x^0), v(t)\rangle]dt, \end{aligned}$$

where $N_T(t) = K(X_T(t))^*$. Let $0 < \varepsilon^2 < \gamma^2 - \|\mathbb{L}\|^2$. Then,

$$\begin{aligned} J_T^{\gamma^2}(0, v) &\geq \gamma^2\|\bar{v}\|^2_{L^2_w([0,\infty];L^2(\Omega,\mathbb{K}^\ell))} - \|(\mathbb{L}\bar{v}\|^2_{L^2_w([0,\infty];L^2(\Omega,\mathbb{K}^\ell))}, \\ &\geq \varepsilon^2\|\bar{v}\|^2_{L^2_w([0,\infty];L^2(\Omega,\mathbb{K}^\ell))} = \varepsilon^2\|v\|^2_{L^2_w([0,T];L^2(\Omega,\mathbb{K}^\ell))}, \end{aligned}$$

where $\bar{v}$ denotes the extension of v from $[0, T]$ to $\mathbb{R}_+$ by 0. Hence,

$$\begin{aligned} J_T^{\gamma^2}(x^0, v) &\geq \langle x^0, X_T(0)x^0\rangle + \int_0^T \mathcal{E}[\varepsilon^2\langle v(t), v(t)\rangle + \langle v(t), N_T(t)x(t,0,x^0)\rangle \\ &\qquad + \langle N_T(t)x(t,0,x^0), v(t)\rangle]dt \\ &= \langle x^0, X_T(0)x^0\rangle + \int_0^T \mathcal{E}[\|\varepsilon v(t) + \varepsilon^{-1}N_T(t)x(t,0,x^0)\|^2 \\ &\qquad - \|\varepsilon^{-1}N_T(t)x(t,0,x^0)\|^2]dt \\ &\geq \langle x^0, X_T(0)x^0\rangle - \int_0^T \mathcal{E}\|\varepsilon^{-1}N_T(t)x(t,0,x^0)\|^2 dt. \end{aligned} \tag{27}$$

Since (4) is stable, there exists $c_0 > 0$ such that

$$\int_0^\infty \mathcal{E}\|x(t,0,x^0)\|^2 dt \leq c_0\|x^0\|^2.$$

Hence, by (24) there exist constants $c_1, c_2 > 0$ independent on T such that

$$\begin{aligned} 0 \geq \langle x^0, X_T(t)x^0\rangle &= \langle x^0, X_{T-t}(0)x^0\rangle = J^{\gamma^2}_{T-t}(x^0, 0) \\ &\geq -\int_0^\infty \mathcal{E}\|Cx(s,0,x^0)\|^2 ds \geq -c_1\|x^0\|^2 \end{aligned}$$

and

$$\|N_T(t)\| = \|X_T(t)B + q_{12}A_0^*X_T(t)B_0 - C^*D\| \leq c_2, \quad t \in [0,T], \quad T > 0.$$

Thus, by (27)

$$J_T^{\gamma^2}(x^0, v) \geq -c_1\|x^0\|^2 - c_2^2\varepsilon^{-2}c_0\|x^0\|^2, \qquad T > 0.$$

This concludes the proof. □

LEMMA 2.11. *Suppose* (4) *is internally stable,* $\|\mathbb{L}\| < \gamma$, $F(\cdot) \in C([0,T], \mathbb{K}^{\ell\times n})$, $T > 0$, *and* $P_F^{\gamma^2}(\cdot)$ *satisfies* (21) *with* $P_F^{\gamma^2}(T) = 0$. *Then,*

$$\gamma^2 I - D^*D \succ 0 \quad \textit{and} \quad H^{\gamma^2}(P_F^{\gamma^2}(t)) \succeq (\gamma^2 - \|\mathbb{L}\|^2)I_\ell, \qquad t \in [0,T]. \tag{28}$$

Proof. We will first prove that $H^{\gamma^2}(P_F^{\gamma^2}(t)) \succeq 0$. Suppose this is false and there exists $\hat{t} \in [0,T]$, $u \in \mathbb{K}^\ell$, $\|u\| = 1$ such that $\langle u, H^{\gamma^2}(P_F^{\gamma^2}(\hat{t}\,))u\rangle \leq -\eta$ for some $\eta > 0$. Assume $\hat{t} < T$. Then, for $\delta > 0$ sufficiently small,

$$\langle u, H^{\gamma^2}(P_F^{\gamma^2}(t))u\rangle \leq -\eta/2, \qquad t \in [\hat{t}, \hat{t}+\delta] \subset [0,T].$$

Define

$$v(t) = \begin{cases} 0 & \text{if} \quad t \in [0,\hat{t}\,) \cup (\hat{t}+\delta, \infty), \\ u & \text{if} \quad t \in [\hat{t}, \hat{t}+\delta]. \end{cases}$$

Now apply Lemma 2.9 to this $v(\cdot)$ and $x^0 = 0$. Then, $x_F(t) = x_F(t, v(\cdot), 0) = 0$ for $t \in [0,t]$, and

$$\begin{aligned}
\mathcal{E}\int_0^\infty \left[\gamma^2\|v(t)\|^2 - \|Cx_F(t) + Dv(t)\|^2\right] dt &\leq \mathcal{E}\int_0^T \left[\gamma^2\|v(t)\|^2 - \|Cx_F(t) + Dv(t)\|^2\right] dt \\
&= \int_0^T \mathcal{E}[\langle v(t), N(t)x_F(t)\rangle + \langle N(t)x_F(t), v(t)\rangle + \langle v(t), H^{\gamma^2}(P_F^{\gamma^2}(t))v(t)\rangle]dt \\
&\leq \int_{\hat{t}}^{\hat{t}+\delta} (2\|N(t)^*u\|\,\|\mathcal{E}x_F(t)\| - \eta/2)dt.
\end{aligned}$$

Choosing $\delta > 0$ sufficiently small, the integrand becomes negative, since $\mathcal{E}x_F(t)$ is continuous and $\mathcal{E}x_F(\hat{t}\,) = 0$. This yields a contradiction whence $H^{\gamma^2}(P_F^{\gamma^2}(t)) \succeq 0$. If $\hat{t} = T$, a similar proof applies, replacing the interval $[\hat{t}, \hat{t}+\delta]$ by $[T-\delta, T]$.

Now let ε be any positive number such that $\|\mathbb{L}\|^2 < \gamma^2 - \varepsilon^2$. Applying the previous step with $\tilde{\gamma} = (\gamma^2 - \varepsilon^2)^{1/2}$ instead of γ we obtain, for the corresponding solution $P_F^{\tilde{\gamma}^2}(t)$ of (21) (with $\tilde{\gamma}$ instead of γ), $H^{\tilde{\gamma}^2}(P_F^{\tilde{\gamma}^2}(t)) \succeq 0$. For any $t_0 \in [0,T)$, define $F_{t_0}(t) = F(t+t_0)$, $t \in [0, T-t_0]$. Let $P_{F_{t_0}}^{\tilde{\gamma}^2}(t)$ be the solution of (21) with γ replaced by $\tilde{\gamma}$ and F replaced by F_{t_0} on the interval $[0, T-t_0]$ such that $P_{F_{t_0}}^{\tilde{\gamma}^2}(T-t_0) = 0$. Then,

$$P_{F_{t_0}}^{\tilde{\gamma}^2}(t) = P_F^{\tilde{\gamma}^2}(t+t_0), \qquad t \in [0, T-t_0].$$

Hence by (24), for any $t_0 \in [0,T)$, $x^0 \in \mathbb{K}^n$,

$$\begin{aligned}
\langle x^0, P_F^{\tilde{\gamma}^2}(t_0)x^0\rangle = \langle x^0, P_{F_{t_0}}^{\tilde{\gamma}^2}(0)x^0\rangle &= J_{T-t_0}^{\tilde{\gamma}^2}(x^0, F_{t_0}x_{F_{t_0}}) \\
&\leq J_{T-t_0}^{\gamma^2}(x^0, F_{t_0}x_{F_{t_0}}) = \langle x^0, P_F^{\gamma^2}(t_0)x^0\rangle,
\end{aligned}$$

and so $H^{\gamma^2-\varepsilon^2}(P_F^{\gamma^2}(t_0)) \succeq H^{\gamma^2-\varepsilon^2}(P_F^{\tilde{\gamma}^2}(t_0)) \succeq 0$, i.e., $H^{\gamma^2}(P_F^{\gamma^2}(t)) \succeq \varepsilon^2 I$ for all $t \in [0,T]$ (by continuity). Since this holds for arbitrary $\varepsilon^2 < \gamma^2 - \|\mathbb{L}\|^2$, (28) follows, and

$$\gamma^2 I - D^*D = H^{\gamma^2}(P_F^{\gamma^2}(T)) \succ 0.$$

This completes the proof. □

We will now study the matrix differential equation

(29)

$$\dot{X} + XA + A^*X + q_{11}A_0^*XA_0 - C^*C - K(X)H^{\gamma^2}(X)^{-1}K(X)^* = 0, \qquad X(T) = 0.$$

The function

$$f(X) = XA + A^*X + q_{11}A_0^*XA_0 - C^*C - K(X)H^{\gamma^2}(X)^{-1}K(X)^*$$

is continuously differentiable on its domain of definition $D_f = \{X \in \mathcal{H}_n(\mathbb{K}); \det(H^{\gamma^2}(X)) \neq 0\}$ in the real vector space $\mathcal{H}_n(\mathbb{K})$. For every $T > 0$, there exists a (unique) solution of (29) backwards in time on a maximal interval $(t_-(T), T]$. The following proposition shows, in particular, that $t_-(T) < 0$ for all $T > 0$.

PROPOSITION 2.12. *Suppose* (4) *is internally stable and* $\|\mathbb{L}\| < \gamma$. *Then* (29) *has a unique solution* $P_T(\cdot)$ *on* $[0,T]$ *for every* $T > 0$. *Moreover, the feedback control*

$$v_T(t) = F_T(t)x_{F_T}(t), \qquad F_T(t) = -H^{\gamma^2}(P_T(t))^{-1}K(P_T(t))^*, \tag{30}$$

where $x_{F_T}(\cdot)$ *satisfies*

$$\begin{aligned} dx_{F_T}(t) = {} & (A + BF_T(t))x_{F_T}(t)dt + A_0x_{F_T}(t)dw_1(t) \\ & + B_0F_T(t)x_{F_T}(t)dw_2(t), \qquad x_{F_T}(0) = x^0, \end{aligned}$$

minimizes $J_T^{\gamma^2}(x^0, v)$, *and the optimal cost is*

$$\min_{v \in L_w^2} J_T^{\gamma^2}(x^0, v) = \langle x^0, P_T(0)x^0 \rangle. \tag{31}$$

Proof. Since (29) is time invariant, we have

$$P_T(t) = P_{T-t}(0), \quad t \in (t_-(T), T] \quad \text{and} \quad t_-(T-\tau) = t_-(T) - \tau, \quad \tau \in \mathbb{R}. \tag{32}$$

Let $\tilde{T} = \inf\{T \geq 0; t_-(T) \geq 0\}$. Then $\tilde{T} > 0$, and $t_-(\tilde{T}) = 0$ if $\tilde{T} < \infty$. For every $T < \tilde{T}$, we have $t_-(T) < 0$, and $P_T(\cdot)$ is continuously differentiable on $[0,T]$. Setting $F(t) = F_T(t)$, $t \in [0,T]$ in (25), we get from (20) and (30)

$$\begin{aligned} & \dot{P}_T + \begin{bmatrix} I \\ F_T \end{bmatrix}^* M(P_T) \begin{bmatrix} I \\ F_T \end{bmatrix} \\ & = \dot{P}_T + \begin{bmatrix} I \\ F_T \end{bmatrix}^* \begin{bmatrix} P_TA + A^*P_T + q_{11}A_0^*P_TA_0 - C^*C & K(P_T) \\ K(P_T)^* & H^{\gamma^2}(P_T) \end{bmatrix} \begin{bmatrix} I \\ F_T \end{bmatrix} \\ & = \dot{P}_T + P_TA + A^*P_T + q_{11}A_0^*P_TA_0 - C^*C - K(P_T)H^{\gamma^2}(P_T)^{-1}K(P_T)^* = 0. \end{aligned}$$

Hence $P_T(\cdot)$ satisfies (25), or equivalently (21), with $F(t) = F_T(t)$ on $[0,T]$ for all $T < \tilde{T}$, i.e.,

$$P^{\gamma^2}_{F_T}(t) = P_T(t), \qquad t \in [0,T].$$

Moreover, with this choice of $F(t)$,

$$N(t) = K(P_T(t))^* + H^{\gamma^2}(P_T(t))F_T(t) = 0,$$

and so Lemma 2.9 implies that

$$J^{\gamma^2}_T(x^0, v + F_T x) = \langle x^0, P_T(0)x^0\rangle + \int_0^T \mathcal{E}[\langle v(t), H^{\gamma^2}(P_T(t))v(t)\rangle]dt.$$

But by Lemma 2.11,

$$H^{\gamma^2}(P_T(t)) = H^{\gamma^2}(P^{\gamma^2}_{F_T}(t)) \succeq (\gamma^2 - \|\mathbb{L}\|^2)I_\ell \succ 0, \qquad t \in [0,T]. \tag{33}$$

Hence, the control $v_T(t) = F_T(t)x(t)$ minimizes $J^{\gamma^2}_T(x^0, v)$ and the optimal costs are given by (31), for all $T < \tilde{T}$. As a consequence, we obtain

$$\langle x^0, P_T(\tau)x^0\rangle = \langle x^0, P_{T-\tau}(0)x^0\rangle = J^{\gamma^2}_{T-\tau}(x^0, v_{T-\tau}) \le J^{\gamma^2}_{T-\tau}(x^0, 0) \le 0, \qquad \tau \in [0,T].$$

On the other hand,

$$\langle x^0, P_T(\tau)x^0\rangle = J^{\gamma^2}_{T-\tau}(x^0, v_{T-\tau}) \ge -c\|x^0\|^2, \quad x^0 \in \mathbb{K}^n, \quad \tau \in [0,T],$$

for all $T < \tilde{T}$ by Lemma 2.10. Hence,

$$-cI_n \preceq P_T(t) \preceq 0, \qquad t \in [0,T], \quad T < \tilde{T}. \tag{34}$$

Now, suppose $\tilde{T} < \infty$ so that $t_-(\tilde{T}) = 0$. Then $-cI \preceq P_{\tilde{T}}(t) \preceq 0$ for all $t \in (0, \tilde{T}]$ and hence, the solution $P_{\tilde{T}}(t)$ of (29) (with $T = \tilde{T}$) cannot escape to ∞ as $t \downarrow 0$. It follows that there exists a boundary point $P^0 \in \mathcal{H}_n(\mathbb{K})$, $\det(H^{\gamma^2}(P^0)) = 0$ of the domain D_f which is a limit point of $P_{\tilde{T}}(t)$ as $t \downarrow 0$. But this contradicts the fact that by (33), $H^{\gamma^2}(P_{\tilde{T}}(t)) = H^{\gamma^2}(P_{\tilde{T}-t}(0)) \succeq (\gamma^2 - \|\mathbb{L}\|^2)I_\ell$ for all $t \in (0, \tilde{T})$. Thus, $\tilde{T} = \infty$ and the proposition is proved. □

Now we examine what happens as $T \to \infty$.

LEMMA 2.13. *Suppose* (4) *is internally stable and* $\|\mathbb{L}\| < \gamma$. *Then* $P_T(t)$ *decreases as* T *increases for each* $t \in [0,T]$.

Proof. Suppose $T' > T, t \in [0,T]$, and $x^0 \in \mathbb{K}^n$. Let v_{T-t} be optimal for x^0 on $[0, T-t]$, and set $v(\tau) = v_{T-t}(\tau)$ for $\tau \in [0, T-t]$ and $v(\tau) = 0$ for $\tau \in (T-t, T'-t]$. Then,

$$\begin{aligned}\langle x^0, P_{T'}(t)x^0\rangle &\le J_{T'-t}(x^0, v) = J_{T-t}(x^0, v_{T-t}) - \int_{T-t}^{T'-t} \mathcal{E}\|z(s)\|^2 ds \\ &\le J_{T-t}(x^0, v_{T-t}) = \langle x^0, P_T(t)x^0\rangle. \qquad \square\end{aligned}$$

We are now in a position to prove Theorem 2.8.

Proof of Theorem 2.8. By Corollary 2.7, it only remains to prove that (i) implies (ii). Assume (i), i.e., (4) is internally stable and $\|\mathbb{L}\| < \gamma$. Using (34), it follows from

Lemma 2.13 that $P_T(t)$ converges as $T \to \infty$ for any $t \geq 0$. But $P_T(t) = P_{T-t}(0)$, so the limit $\lim_{T\to\infty} P_T(t) = \lim_{T\to\infty} P_T(0) = P$ is constant. It follows from (34) and (33) that P satisfies

$$P \preceq 0 \quad \text{and} \quad H^{\gamma^2}(P) \succ 0 \tag{35}$$

and is a solution of the rational matrix equation

$$PA + A^*P + q_{11}A_0^*PA_0 - C^*C - K(P)H^{\gamma^2}(P)^{-1}K(P)^* = 0. \tag{36}$$

Now replace C by $C_\delta = \left[\begin{smallmatrix} C \\ \delta I \end{smallmatrix}\right]$ and D by $D_\delta = \left[\begin{smallmatrix} D \\ 0 \end{smallmatrix}\right]$ in Definition 2.3 to obtain the perturbation operator $\mathbb{L}_\delta$ for the modified data. Then $\|\mathbb{L}_\delta\| < \gamma$ for sufficiently small $\delta > 0$, and so applying the above result to the modified data we find that there exists $P_\delta \in \mathcal{H}_n(\mathbb{K})$, $P_\delta \preceq 0$ satisfying

(37)

$$P_\delta A + A^*P_\delta + q_{11}A_0^*P_\delta A_0 - C^*C - \delta^2 I - K(P_\delta)H^{\gamma^2}(P_\delta)^{-1}K(P_\delta)^* = 0, \qquad H^{\gamma^2}(P_\delta) \succ 0.$$

By stability, $P_\delta \prec 0$, and the above equation implies

$$P_\delta A + A^*P_\delta + q_{11}A_0^*P_\delta A_0 - C^*C - K(P_\delta)H^{\gamma^2}(P_\delta)^{-1}K(P_\delta)^* \succ 0, \qquad H^{\gamma^2}(P_\delta) \succ 0.$$

Applying the definiteness criterion (18), we get that $M(P_\delta) \succ 0$, and (ii) is proved. ∎

For later use we add the following consequence of Theorem 2.8.

COROLLARY 2.14. *For any set of data* (5)*, the following conditions are equivalent:*

(a) *The system* (4) *is internally stable and* $\|\mathbb{L}\| < \gamma$.

(b) *There exist* $\delta > 0$ *and* $P_\delta \prec 0$ *satisfying* (37).

(c) *There exist* $P \in \mathcal{H}_n(\mathbb{K}), P \prec 0$ *such that*

$$\begin{bmatrix} PA + A^*P + q_{11}A_0^*PA_0 & PB + q_{12}A_0^*PB_0 & C^* \\ B^*P + q_{12}B_0^*PA_0 & \gamma^2 I + q_{22}B_0^*PB_0 & D^* \\ C & D & I \end{bmatrix} \succ 0. \tag{38}$$

Proof. In the above proof, we have shown that (a) implies (b) and (b) implies condition (ii) of Theorem 2.8 (hence (a)). So it remains to show the equivalence of conditions (ii) and (c). This follows from the equality

$$\begin{bmatrix} I & 0 & -C^* \\ 0 & I & -D^* \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} PA + A^*P + q_{11}A_0^*PA_0 & PB + q_{12}A_0^*PB_0 & C^* \\ B^*P + q_{12}B_0^*PA_0 & \gamma^2 I + q_{22}B_0^*PB_0 & D^* \\ C & D & I \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ -C & -D & I \end{bmatrix}$$

$$= \begin{bmatrix} PA + A^*P + q_{11}A_0^*PA_0 - C^*C & PB + q_{12}A_0^*PB_0 - C^*D & 0 \\ B^*P + q_{12}B_0^*PA_0 - D^*C & \gamma^2 I + q_{22}B_0^*PB_0 - D^*D & 0 \\ 0 & 0 & I \end{bmatrix}$$

$$= \begin{bmatrix} M(P) & 0 \\ 0 & I \end{bmatrix}. \qquad \blacksquare$$

The following scalar example illustrates the above results.

Example 2.15. Consider the following system of the form (4), with $n = 1$, $\mathbb{K} = \mathbb{R}$, and $D = 0$:

$$dx(t) = ax(t)dt + a_0x(t)dw_1(t) + b_0v(t)dw_2(t) + bv(t)dt, \qquad y(t) = cx(t), \tag{39}$$

where $a, a_0, b, b_0, c \in \mathbb{R}$ and $w_1(t)$, $w_2(t)$ are Wiener processes, as before. Equation (16) is equivalent to

$$2pa + q_{11}a_0^2p - c^2 - (b + q_{12}a_0b_0)^2p^2/(\gamma^2 + q_{22}b_0^2p) > 0, \qquad \gamma^2 + q_{22}b_0^2p > 0. \tag{40}$$

Suppose $a = -1$, $a_0 = b = b_0 = c = 1$, and assume first that $w_1 = w_2 = q_{11} = q_{12} = q_{22} = 1$. The inequalities (40) become

$$-p - 1 - 4p^2/(\gamma^2 + p) > 0, \qquad \gamma^2 + p > 0,$$

and these in turn are equivalent to

$$0 > 5p^2 + (1 + \gamma^2)p + \gamma^2 \quad \text{and} \quad \gamma^2 + p > 0. \tag{41}$$

The first inequality holds if and only if

$$(9 + \sqrt{80} < \gamma^2 \text{or } \gamma^2 < 9 - \sqrt{80}\,) \quad \text{and} \quad 10p < -(1 + \gamma^2) + (\gamma^4 - 18\gamma^2 + 1)^{1/2}.$$

So

$$10(\gamma^2 + p) < (\gamma^4 - 18\gamma^2 + 1)^{1/2} + 9\gamma^2 - 1.$$

Hence, the constraint $\gamma^2 + p > 0$ requires $\gamma^2 \geq 1/9 > 9 - \sqrt{80}$ which excludes the alternative $\gamma^2 < 9 - \sqrt{80}$. Therefore, $9 + \sqrt{80} < \gamma^2$ is a necessary and sufficient condition for (41) to have a joint negative solution p. Thus, $\|\mathbb{L}\|^2 = 9 + \sqrt{80}$.

We will now analyze what happens if the incremental covariance matrix of the system is changed. Let $q_{11} = 1$, $q_{12} = q_{22} = 0$ so that the stochastic perturbation term $v(t)dw_2(t)$ is absent from (39). In this case, the inequalities (40) reduce to

$$-p - 1 - p^2/\gamma^2 > 0.$$

This inequality has a negative solution p if and only if $\gamma^2 > 4$. Hence $\|\mathbb{L}\| = 2$. □

3. Resolution of the general disturbance attenuation problem. We will study the H^∞-type disturbance attenuation problem for stochastic systems of the form

$$\Sigma: \begin{aligned} dx(t) &= Ax(t)dt + A_0x(t)dw_1(t) + B_0v(t)dw_2(t) + B_1v(t)dt + B_2u(t)dt, \\ z(t) &= C_1x(t) + D_{11}v(t) + D_{12}u(t), \\ y(t) &= C_2x(t) + D_{21}v(t), \end{aligned} \tag{42}$$

where

$$(A, A_0, B_0, B_1, B_2) \in \mathbb{K}^{n\times n} \times \mathbb{K}^{n\times n} \times \mathbb{K}^{n\times \ell} \times \mathbb{K}^{n\times \ell} \times \mathbb{K}^{n\times m},$$
$$(C_1, C_2, D_{11}, D_{12}, D_{21}) \in \mathbb{K}^{q\times n} \times \mathbb{K}^{p\times n} \times \mathbb{K}^{q\times l} \times \mathbb{K}^{q\times m} \times \mathbb{K}^{p\times \ell},$$

and w_1, w_2 are as in the previous section. There are two vector valued input variables u, v and two vector valued output variables y, z. v represents an unknown stochastic disturbance signal, u the control, z the vector of the to-be-controlled variables,

and y the measurements. As compensator we choose—as usual in H^∞-theory—a finite-dimensional time-invariant deterministic linear system which is driven by the measurement process $y(\cdot)$ of Σ and produces the (random) control values $u(t)$:

$$\Sigma_K : d\hat{x}(t) = A_K\hat{x}(t)dt + B_K y(t)dt, \qquad u(t) = C_K\hat{x}(t) + D_K y(t), \tag{43}$$

where $(A_K, B_K, C_K, D_K) \in \mathbb{K}^{\hat{n}\times\hat{n}} \times \mathbb{K}^{\hat{n}\times p} \times \mathbb{K}^{m\times\hat{n}} \times \mathbb{K}^{m\times p}$ and the dimension $\hat{n} \geq 0$ is arbitrary. If $\hat{n} = 0$, the state equation of Σ_K vanishes and we obtain $u(t) = D_K y(t)$, i.e., a static linear output feedback control law. For arbitrary $\hat{n} \geq 0$, let

$$M_K = \begin{bmatrix} A_K & B_K \\ C_K & D_K \end{bmatrix}.$$

The resulting closed loop system is

$$\Sigma_{cl} \ : \ \begin{aligned} d\bar{x}(t) &= A_{cl}\bar{x}(t)dt + A^0_{cl}\bar{x}(t)dw_1(t) + B^0_{cl}v(t)dw_2(t) + B_{cl}v(t)dt, \\ z(t) &= C_{cl}\bar{x}(t) + D_{cl}v(t), \end{aligned} \tag{44}$$

where

$$\bar{x} = \begin{bmatrix} x \\ \hat{x} \end{bmatrix}, \quad A_{cl} = \begin{bmatrix} A + B_2D_KC_2 & B_2C_K \\ B_KC_2 & A_K \end{bmatrix}, \quad A^0_{cl} = \begin{bmatrix} A_0 & 0 \\ 0 & 0 \end{bmatrix},$$

$$B^0_{cl} = \begin{bmatrix} B_0 \\ 0 \end{bmatrix} \quad B_{cl} = \begin{bmatrix} B_1 + B_2D_KD_{21} \\ B_KD_{21} \end{bmatrix}, \tag{45}$$

$$C_{cl} = [C_1 + D_{12}D_KC_2, \quad D_{12}C_K], \quad D_{cl} = D_{11} + D_{12}D_KD_{21}.$$

Suppose (44) is internally stable in the sense of the previous section, and the linear operator

$$\mathbb{L}_{cl} : L^2_w(\mathbb{R}_+; L^2(\Omega, \mathbb{K}^\ell)) \to L^2_w(\mathbb{R}_+; L^2(\Omega, \mathbb{K}^q)),$$

is defined by

$$(\mathbb{L}_{cl}v)(t) = C_{cl}\bar{x}(t, v, 0) + D_{cl}v(t), \qquad t \geq 0, \qquad v(\cdot) \in L^2_w(\mathbb{R}_+; L^2(\Omega, \mathbb{K}^\ell)), \tag{46}$$

where $\bar{x}(t, v, \bar{x}^0)$ is the solution of (44) with $\bar{x}(0) = \bar{x}^0$, for every $v(\cdot) \in L^2_w(\mathbb{R}_+; L^2(\Omega, \mathbb{K}^\ell))$. $\mathbb{L}_{cl}$ describes the effect of the disturbance signal $v(\cdot)$ on the to-be-controlled output vector $z(\cdot)$ of the closed loop system. Given $\gamma > 0$, our aim is to determine whether or not there is a compensator (43) which stabilizes system (42) internally and achieves $\|\mathbb{L}_{cl}\| < \gamma$. Such controllers will be called *suboptimal* of level γ. In case of existence, we want to know how these compensators M_K can be constructed.

Remark 3.1. In the linear quadratic Gaussian control problem, additive noise terms are present in the basic model and one may ask why we have excluded them here. Consider the following model with additive white noise in the state and the measurement equations:

$$\begin{aligned} dx(t) &= Ax(t)dt + A_0x(t)dw_1(t) + B_0v(t)dw_2(t) + B_1v(t)dt + B_2u(t)dt + E_1dw_3(t), \\ z(t) &= C_1x(t) + D_{11}v(t) + D_{12}u(t), \\ dy(t) &= C_2x(t)dt + D_{21}v(t)dt + E_2dw_3(t), \end{aligned}$$

where $(E_1, E_2) \in \mathbb{K}^{n\times r} \times \mathbb{K}^{p\times r}$, and w_3 is a vector of r scalar Wiener processes. Suppose the compensator has the form

$$d\hat{x}(t) = A_K\hat{x}(t)dt + B_K dy(t), \qquad u(t) = C_K\hat{x}(t).$$

Then the closed loop system is

$$\begin{aligned} d\bar{x}(t) &= A_{cl}\bar{x}(t)dt + A^0_{cl}\bar{x}(t)dw_1(t) + B^0_{cl}v(t)dw_2(t) + B_{cl}v(t)dt + E_{cl}dw_3(t),\\ z(t) &= C_{cl}\bar{x}(t) + D_{cl}v(t), \end{aligned}$$

where $\bar{x} = \left[\begin{smallmatrix} x \\ \hat{x}\end{smallmatrix}\right]$, $A_{cl}, A^0_{cl}, B^0_{cl}, C_{cl}, D_{cl}$ are as in (45) with $D_K = 0$ and $E_{cl} = \left[\begin{smallmatrix} E_1 \\ B_K E_2\end{smallmatrix}\right]$. In order for the H^∞ problem to make sense over infinite time horizons, the map from v to z (with initial state $\bar{x}(0) = 0$) must be linear or must at least map the zero input onto the zero output. For the preceding closed loop system, this will only be the case if the additive noise term w_3 is completely decoupled from z. In the case where the diffusion term is absent in the nominal system equation ($A_0 = 0$), this would require that the kernel of C_{cl} must contain the smallest A_{cl}-invariant subspace generated by the range of E_{cl}. Thus, the presence of additive white noise would impose an additional, very restrictive, condition on the controllers. In fact, it is our opinion that adding a specific white noise term is not really appropriate in an H^∞-type disturbance attenuation problem. In this framework, measurement and state disturbances are modelled by the *unknown* random process v (so that, e.g., measurement noise is represented by the term $D_{21}v(t)$ in the second output equation. □

We will show that the above disturbance attenuation problem can be solved via the resolution of matrix inequalities. Our approach follows the one developed by Gahinet and Apkarian for the deterministic case [8]. The key tool which makes this possible is the stochastic version of the bounded real lemma derived in the previous section. From deterministic H^∞-control theory, we will need the following so-called projection lemma. A proof of this lemma can be found in [8].

LEMMA 3.2 (projection lemma). *Suppose* $N \in \mathbb{K}^{\ell\times m}, M \in \mathbb{K}^{n\times m}$, *and* $H \in \mathcal{H}_m(\mathbb{K})$. *Then the linear matrix inequality*

$$H + N^*X^*M + M^*XN \succ 0$$

has a solution $X \in \mathbb{K}^{n\times\ell}$ *if and only if* H *is positive definite on* $\ker N$ *and* $\ker M$.

To simplify the presentation, the following notations will be used:

$$A^0 = \begin{bmatrix} A & 0 \\ 0 & 0_{\hat{n}\times\hat{n}}\end{bmatrix}, \quad B^0 = \begin{bmatrix} B_1 \\ 0_{\hat{n}\times\ell}\end{bmatrix}, \quad C^0 = [C_1, 0_{q\times\hat{n}}], \quad D^0_{12} = [0_{q\times\hat{n}}, D_{12}],$$

$$B^I = \begin{bmatrix} 0 & B_2 \\ I_{\hat{n}} & 0\end{bmatrix}, \quad C^I = \begin{bmatrix} 0 & I_{\hat{n}} \\ C_2 & 0\end{bmatrix}, \quad D^0_{21} = \begin{bmatrix} 0_{\hat{n}\times\ell} \\ D_{21}\end{bmatrix}.$$

Then the closed loop matrices can be written as

$$\begin{aligned} &A_{cl} = A^0 + B^I M_K C^I, \quad B_{cl} = B^0 + B^I M_K D^0_{21}, \quad C_{cl} = C^0 + D^0_{12} M_K C^I,\\ &D_{cl} = D_{11} + D^0_{12} M_K D^0_{21}, \qquad A^0_{cl}, B^0_{cl} \quad \text{as in (45)}. \end{aligned} \tag{47}$$

In order to save space we will not write out the upper triangle of large Hermitian matrices but will use a $\star$ notation.

THEOREM 3.3. *For any system of the form* (42) *and* $\gamma > 0$, *the following conditions are equivalent:*

(i) *There exists a compensator* (43) *of dimension* $\hat{n}$ *such that the resulting closed loop system* (44) *is internally stable and* $\|\mathbb{L}_{cl}\| < \gamma$.

(ii) *There exists a* $P_{cl} \in \mathcal{H}_{n+\hat{n}}(\mathbb{K})$, $P_{cl} \prec 0$ *such that the matrix* $\Phi_{P_{cl}} = \Phi_{P_{cl}}^*$ *is positive definite on* $\ker U$ *and* $\Psi_{P_{cl}} = \Phi_{P_{cl}}^*$ *is positive definite on* $\ker V$, *where*

$$\Psi_{P_{cl}} = \begin{bmatrix} (A^0)^* P_{cl} + P_{cl} A^0 + q_{11}(A_{cl}^0)^* P_{cl} A_{cl}^0 & \star & \star \\ (B^0)^* P_{cl} + q_{12}(B_{cl}^0)^* P_{cl} A_{cl}^0 & \gamma I_\ell + q_{22}(B_{cl}^0)^* P_{cl} B_{cl}^0 & \star \\ C^0 & D_{11} & I_q \end{bmatrix},$$

$$\Phi_{P_{cl}} = \begin{bmatrix} P_{cl}^{-1} & 0 & 0 \\ 0 & I_\ell & 0 \\ 0 & 0 & I_q \end{bmatrix} \Psi_{P_{cl}} \begin{bmatrix} P_{cl}^{-1} & 0 & 0 \\ 0 & I_\ell & 0 \\ 0 & 0 & I_q \end{bmatrix},$$

and

$$U = [(B^I)^*, 0_{(\hat{n}+m)\times\ell}, (D_{12}^0)^*], \qquad V = [C^I, D_{21}^0, 0_{(\hat{n}+p)\times q}]. \tag{48}$$

Proof. Applying Corollary 2.14 with $A = A_{cl}$, $A_0 = A_{cl}^0$, etc. we see that (i) is equivalent to the existence of $P_{cl} \prec 0$ such that

$$\begin{bmatrix} (A_{cl})^* P_{cl} + P_{cl} A_{cl} + q_{11}(A_{cl}^0)^* P_{cl} A_{cl}^0 & \star & \star \\ (B_{cl})^* P_{cl} + q_{12}(B_{cl}^0)^* P_{cl} A_{cl}^0 & \gamma^2 I_\ell + q_{22}(B_{cl}^0)^* P_{cl} B_{cl}^0 & \star \\ C_{cl} & D_{cl} & I_q \end{bmatrix} \succ 0.$$

Substituting for A_{cl}, B_{cl}^0, etc. the expressions in (47), we obtain that this is equivalent to

$$\begin{bmatrix} (A^0 + B^I M_K C^I)^* P_{cl} + P_{cl}(A^0 + B^I M_K C^I) + q_{11}(A_{cl}^0)^* P_{cl} A_{cl}^0 & \star & \star \\ ((B^0)^* + (D_{21}^0)^* M_K^* (B^I)^*) P_{cl} + q_{12}(B_{cl}^0)^* P_{cl} A_{cl}^0 & \gamma^2 I_\ell + q_{22}(B_{cl}^0)^* P_{cl} B_{cl}^0 & \star \\ C^0 + D_{12}^0 M_K C^I & D_{11} + D_{12}^0 M_K D_{21}^0 & I_q \end{bmatrix} \succ 0.$$

Or, separating the data and the design parameters,

$$\Psi_{P_{cl}} + \begin{bmatrix} P_{cl} B^I \\ 0_{\ell\times(\hat{n}+\ell)} \\ D_{12}^0 \end{bmatrix} M_K [C^I, D_{21}^0, 0_{(\hat{n}+p)\times q}]$$

$$+ \begin{bmatrix} (C^I)^* \\ (D_{21}^0)^* \\ 0_{q\times(\hat{n}+p)} \end{bmatrix} M_K^* [(B^I)^* P_{cl},\ 0_{(\hat{n}+\ell)\times\ell}, (D_{12}^0)^*] \succ 0.$$

That is,

$$\Psi_{P_{cl}} + U_{P_{cl}}^* M_K V + V^* M_K^* U_{P_{cl}} \succ 0, \tag{49}$$

where $U_{P_{cl}} = [(B^I)^* P_{cl},\ 0_{(\hat{n}+m)\times\ell},\ (D_{12}^0)^*]$ and V is defined as in (48).

Applying the projection lemma, we conclude that (i) is equivalent to $\Psi_{P_{cl}}$ being positive definite on $\ker V$ and $\ker U_{P_{cl}}$. To complete the proof, note that

$$U_{P_{cl}} = U \begin{bmatrix} P_{cl} & 0 & 0 \\ 0 & I_\ell & 0 \\ 0 & 0 & I_q \end{bmatrix}. \qquad \square$$

The characterization in the above theorem is awkward since it involves both P_{cl} and its inverse. However, a simpler form can be obtained by partitioning P_{cl}. To achieve this, the following lemma will be useful; see Lemma 7.5 in [23].

LEMMA 3.4. *Let $n, \hat{n} \geq 1$. Suppose $P \in \mathcal{H}_{n+\hat{n}}(\mathbb{K})$ and its inverse P^{-1} are partitioned as follows:*

$$P = \begin{bmatrix} S & N \\ N^* & Q \end{bmatrix}, \qquad P^{-1} = \begin{bmatrix} R & M \\ M^* & T \end{bmatrix}, \qquad R, S \in \mathcal{H}_n(\mathbb{K}), \tag{50}$$

and $P \prec 0$, then

$$S \preceq R^{-1} \prec 0 \quad \text{and} \quad \operatorname{rank}[R^{-1} - S] \leq \hat{n}. \tag{51}$$

Conversely, if $R, S \in \mathcal{H}_n(\mathbb{K})$ are given such that (51) is satisfied, then there exists $P \in \mathcal{H}_{n+\hat{n}}(\mathbb{K})$, $P \prec 0$ such that P and its inverse can be partitioned as in (50) (with suitable N, Q, M, T).

THEOREM 3.5. *For any system of the form (42) and $\gamma > 0$, the following conditions are equivalent:*

(i) *There exists a stabilizing compensator (43) of dimension $\hat{n}$ such that $\|\mathbb{L}_{cl}\| < \gamma$.*

(ii) *There exists $(R, S) \in \mathcal{H}_n(\mathbb{K}) \times \mathcal{H}_n(\mathbb{K})$ such that*

$$S \preceq R^{-1} \prec 0, \quad \operatorname{rank}(R^{-1} - S) \leq \hat{n} \quad \text{and} \quad \Pi_\gamma(S) = \gamma^2 I_\ell + q_{22} B_0^* S B_0 \succ 0, \tag{52}$$

$$\begin{bmatrix} AR + RA^* + q_{11} RA_0^* S A_0 R & RC_1^* \\ C_1 R & I_\ell \end{bmatrix} - \begin{bmatrix} B_1 + q_{12} RA_0^* S B_0 \\ D_{11} \end{bmatrix} \Pi_\gamma(S)^{-1} \begin{bmatrix} B_1 + q_{12} RA_0^* S B_0 \\ D_{11} \end{bmatrix}^* \succ 0 \quad \text{on } \ker [B_2^* D_{12}^*] \tag{53}$$

and

$$\begin{bmatrix} SA + A^* S + q_{11} A_0^* S A_0 & SB_1 + q_{12} A_0^* S B_0 \\ B_1^* S + q_{12} B_0^* S A_0 & \Pi_\gamma(S) \end{bmatrix} - \begin{bmatrix} C_1^* \\ D_{11}^* \end{bmatrix} \begin{bmatrix} C_1^* \\ D_{11}^* \end{bmatrix}^* \succ 0 \quad \text{on } \ker [C_2 \ D_{21}]. \tag{54}$$

Proof. By Theorem 3.3, (i) is equivalent to the existence of $P_{cl} \in \mathcal{H}_{n+\hat{n}}(\mathbb{K})$, $P_{cl} \prec 0$ such that the matrix $\Phi_{P_{cl}}$ is positive definite on $\ker U$ and $\Psi_{P_{cl}}$ is positive definite on $\ker V$. If we partition

$$P_{cl} = \begin{bmatrix} S & N \\ N^* & Q \end{bmatrix}, \quad P_{cl}^{-1} = \begin{bmatrix} R & M \\ M^* & T \end{bmatrix}, \quad R, S \in \mathcal{H}_n(\mathbb{K}),$$

we obtain from the preceding lemma that

$$S \preceq R^{-1} \prec 0 \quad \text{and} \quad \text{rank } [R^{-1} - S] \leq \hat{n}.$$

Let us first consider the condition that $\Psi_{P_{cl}}$ is positive definite on $\ker V$. Since, by definition (48),

$$V = \begin{bmatrix} 0 & I_{\hat{n}} & 0 & 0_{\hat{n}\times q} \\ C_2 & 0 & D_{21} & 0_{p\times q} \end{bmatrix},$$

$\ker V$ can be represented as

$$\ker V = \text{Im} \begin{bmatrix} V_1 & 0 \\ 0 & 0 \\ V_2 & 0 \\ 0 & I_q \end{bmatrix},$$

where $\begin{bmatrix} V_1 \\ V_2 \end{bmatrix}$ is a basis matrix for $\ker [C_2, D_{21}]$. Partitioning $\Psi_{P_{cl}}$ accordingly, a straightforward calculation yields

$$\Psi_{P_{cl}} = \begin{bmatrix} SA + A^*S + q_{11}A_0^*SA_0 & \star & \star & \star \\ N^*A & 0 & \star & \star \\ B_1^*S + q_{12}B_0^*SA_0 & B_1^*N & \Pi_\gamma(S) & \star \\ C_1 & 0 & D_{11} & I_q \end{bmatrix}.$$

Now,

$$\begin{bmatrix} V_1^* & 0 & V_2^* & 0 \\ 0 & 0 & 0 & I_q \end{bmatrix} \begin{bmatrix} SA + A^*S + q_{11}A_0^*SA_0 & A^*N & SB_1 + q_{12}A_0^*SB_0 & C_1^* \\ N^*A & 0 & N^*B_1 & 0 \\ B_1^*S + q_{12}B_0^*SA_0 & B_1^*N & \Pi_\gamma(S) & D_{11}^* \\ C_1 & 0 & D_{11} & I_q \end{bmatrix} \begin{bmatrix} V_1 & 0 \\ 0 & 0 \\ V_2 & 0 \\ 0 & I_q \end{bmatrix}$$

$$= \begin{bmatrix} [V_1^* \;\; V_2^*] \begin{bmatrix} SA + A^*S + q_{11}A_0^*SA_0 & SB_1 + q_{12}A_0^*SB_0 \\ B_1^*S + q_{12}B_0^*SA_0 & \Pi_\gamma(S) \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} & [V_1^* \;\; V_2^*] \begin{bmatrix} C_1^* \\ D_{11}^* \end{bmatrix} \\ [C_1 \;\; D_{11}] \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} & I_q \end{bmatrix}.$$

Using (18), it follows therefore that $\Psi_{P_{cl}}$ is positive definite on $\ker V$ if and only if

$$\begin{bmatrix} SA + A^*S + q_{11}A_0^*SA_0 & SB_1 + q_{12}A_0^*SB_0 \\ B_1^*S + q_{12}B_0^*SA_0 & \Pi_\gamma(S) \end{bmatrix} - \begin{bmatrix} C_1^* \\ D_{11}^* \end{bmatrix} [C_1 \;\; D_{11}] \succ 0 \quad \text{on ker } [C_2 \;\; D_{21}],$$

i.e., if and only if (54) is satisfied.

The condition that the matrix $\Phi_{P_{cl}}$ is positive definite on $\ker U$ can be analyzed in a similar way. Since, by definition (48),

$$U = \begin{bmatrix} 0 & I_{\hat{n}} & 0_{\hat{n}\times\ell} & 0 \\ B_2^* & 0 & 0_{m\times\ell} & D_{12}^* \end{bmatrix},$$

$\ker U$ can be represented by

$$\ker U = \operatorname{Im} \begin{bmatrix} U_1 & 0 \\ 0 & 0 \\ 0 & I_\ell \\ U_2 & 0 \end{bmatrix},$$

where $\left[\begin{smallmatrix} U_1 \\ U_2 \end{smallmatrix}\right]$ is a basis for $\ker [B_2^* \ D_{12}^*]$. Partitioning $\Phi_{P_{cl}}$ accordingly, we obtain by a straightforward calculation that

$$\Phi_{P_{cl}} = \begin{bmatrix} AR + RA^* + q_{11}RA_0^*SA_0R & \star & \star & \star \\ (AM + q_{11}RA_0^*SA_0M)^* & q_{11}M^*A_0^*SA_0M & \star & \star \\ (B_1 + q_{12}RA_0^*SB_0)^* & q_{12}B_0^*SA_0M & \Pi_\gamma(S) & \star \\ C_1R & C_1M & D_{11} & I_q \end{bmatrix}.$$

Now,

$$\begin{bmatrix} U_1^* & 0 & 0 & U_2^* \\ 0 & 0 & I_\ell & 0 \end{bmatrix} \begin{bmatrix} AR + RA^* + q_{11}RA_0^*SA_0R & \star & \star & \star \\ (AM + q_{11}RA_0^*SA_0M)^* & q_{11}M^*A_0^*SA_0M & \star & \star \\ (B_1 + q_{12}RA_0^*SB_0)^* & q_{12}B_0^*SA_0M & \Pi_\gamma(S) & \star \\ C_1R & C_1M & D_{11} & I_q \end{bmatrix} \begin{bmatrix} U_1 & 0 \\ 0 & 0 \\ 0 & I_\ell \\ U_2 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} [U_1^* \ U_2^*] \begin{bmatrix} AR + RA^* + q_{11}RA_0^*SA_0R & RC_1^* \\ C_1R & I_q \end{bmatrix} \begin{bmatrix} U_1 \\ U_2 \end{bmatrix} & [U_1^* \ U_2^*] \begin{bmatrix} B_1 + q_{12}RA_0^*SB_0 \\ D_{11} \end{bmatrix} \\ \begin{bmatrix} B_1 + q_{12}RA_0^*SB_0 \\ D_{11} \end{bmatrix}^* \begin{bmatrix} U_1 \\ U_2 \end{bmatrix} & \Pi_\gamma(S) \end{bmatrix}.$$

Hence, again using (18), $\Phi_{P_{cl}}$ is positive definite on $\ker U$ if and only if $\Pi_\gamma(S) \succ 0$ and (53) hold. Altogether we see that (i) implies (ii).

Conversely, suppose $(R, S) \in \mathcal{H}_n(\mathbb{K}) \times \mathcal{H}_n(\mathbb{K})$, $R \prec 0$, $S \prec 0$ satisfy the conditions in (ii). Applying Lemma 3.4, we obtain that there exist N, $M \in \mathbb{K}^{n\times\ell}$, Q, $T \in \mathcal{H}_\ell(\mathbb{K})$ such that

$$P_{cl} := \begin{bmatrix} S & N \\ N^* & Q \end{bmatrix} \prec 0, \qquad P_{cl}^{-1} = \begin{bmatrix} R & M \\ M^* & T \end{bmatrix}.$$

Now define $\Psi_{P_{cl}}$, $\Phi_{P_{cl}}$ as in Theorem 3.3. We have just proved that (53) and (54) imply that the matrix $\Phi_{P_{cl}}$ is positive definite on $\ker U$ and $\Psi_{P_{cl}}$ is positive definite on $\ker V$. But this is equivalent to (i). □

Let γ_{opt} be the optimal value of our H^∞-control problem, i.e.,

(55)

$$\gamma_{\text{opt}} = \inf\{\gamma \geq 0; \exists \text{ compensator (43) s.t. (44) is internally stable and } \|\mathbb{L}_{cl}\| < \gamma\}.$$

By the previous theorem, γ_{opt} is the infimum of all $\gamma \geq 0$ for which there exist $(R, S) \in \mathcal{H}_n(\mathbb{K}) \times \mathcal{H}_n(\mathbb{K})$ such that (52), (53), and (54) are satisfied. Now the condition $\text{rk}(R^{-1} - S) \leq \hat{n}$ is automatically satisfied for $\hat{n} \geq n$ and conditions (53), (54) do not depend on $\hat{n}$. Therefore we obtain, as a consequence of the previous theorem, that for every $\gamma > \gamma_{\text{opt}}$ there exists a stabilizing controller (43) *of dimension* $\leq n$ such that $\|\mathbb{L}_{cl}\| < \gamma$.

Remark 3.6. (i) In the deterministic case ($A_0 = 0, B_0 = 0$), we have $\Pi_\gamma(S) = \gamma^2 I$ and so (53) and (54) become

$$\begin{bmatrix} AR + RA^* - B_1 B_1^*/\gamma^2 & RC_1^* - B_1 D_{11}^*/\gamma^2 \\ C_1 R - D_{11} B_1^*/\gamma^2 & I - D_{11} D_{11}^*/\gamma^2 \end{bmatrix} \succ 0, \quad \text{on ker } [B_2^* \;\; D_{12}^*], \tag{56}$$

$$\begin{bmatrix} SA + A^* S - C_1^* C_1 & SB_1 - C_1^* D_{11} \\ B_1^* S - D_{11}^* C_1 & \gamma^2 I - D_{11}^* D_{11} \end{bmatrix} \succ 0, \quad \text{on ker } [C_2 \;\; D_{21}]. \tag{57}$$

These, together with

$$S \preceq R^{-1} \prec 0, \qquad \text{rank } (R^{-1} - S) \leq \hat{n},$$

are precisely the LMI solvability conditions for the suboptimal H^∞ synthesis problem as stated in [8].

(ii) Suppose R, S satisfying the conditions in part (ii) of the previous theorem have been found for a given value of γ. Then, stabilizing compensators that achieve $\|\mathbb{L}_{cl}\| < \gamma$ can be constructed just as for deterministic systems [8]. First, one constructs P_{cl}, followed by $\Psi_{P_{cl}}$ and $\Phi_{P_{cl}}$, and then one solves (49) for M_K. An explicit construction will be given in the next section for the regular case.

(iii) Comparing (53), (54) with (56), (57), we see that the presence of multiplicative state and control dependent noise leads to nonlinear instead of linear matrix inequalities and to a one-sided coupling of "controller" and "observer" matrix inequalities. Note, however, that the "observer" inequality (54) is still linear and independent of (53) so that it can be considered separately. Its solutions then have to be fed into the controller inequality (53) which is a quadratic matrix inequality in R. □

We can replace inequalities (53), (54) by inequalities on the whole space at the sake of introducing scalar parameters. Namely, (53) is equivalent to the existence of $\alpha > 0$ such that

$$\begin{aligned} &\begin{bmatrix} AR + RA^* + q_{11} RA_0^* S A_0 R & RC_1^* \\ C_1 R & I \end{bmatrix} \\ &- \begin{pmatrix} B_1 + q_{12} RA_0^* S B_0 \\ D_{11} \end{pmatrix} \Pi_\gamma(S)^{-1} \begin{pmatrix} B_1 + q_{12} RA_0^* S B_0 \\ D_{11} \end{pmatrix}^* \\ &+ \alpha^2 \begin{bmatrix} B_2 \\ D_{12} \end{bmatrix} [B_2^* \; D_{12}^*] \succ 0, \end{aligned} \tag{58}$$

and (54) is equivalent to the existence of $\beta > 0$ such that

$$\begin{aligned}&\begin{bmatrix} SA + A^*S + q_{11}A_0^*SA_0 & SB_1 + q_{12}A_0^*SB_0 \\ B_1^*S + q_{12}B_0^*SA_0 & \Pi_\gamma(S) \end{bmatrix} \\ &\quad - \begin{bmatrix} C_1^* \\ D_{11}^* \end{bmatrix} \begin{bmatrix} C_1^* \\ D_{11}^* \end{bmatrix}^* + \beta^2 \begin{bmatrix} C_2^* \\ D_{21}^* \end{bmatrix} [C_2 \ D_{21}] \succ 0.\end{aligned} \tag{59}$$

Furthermore, we can replace the above inequalities with lower-dimensional ones. In fact, applying (18), (58) is equivalent to $I + \alpha^2 D_{12}D_{12}^* - D_{11}\Pi_\gamma(S)^{-1}D_{11}^* \succ 0$ and

$$\begin{aligned}&AR + RA^* + q_{11}RA_0^*SA_0R \\ &\quad - (B_1 + q_{12}RA_0^*SB_0)\Pi_\gamma(S)^{-1}(B_1 + q_{12}RA_0^*SB_0)^* + \alpha^2 B_2B_2^* \\ &\quad - \left[RC_1^* - (B_1 + q_{12}RA_0^*SB_0)\Pi_\gamma(S)^{-1}D_{11}^* + \alpha^2 B_2D_{12}^*\right] \\ &\qquad \left[I + \alpha^2 D_{12}D_{12}^* - D_{11}\Pi_\gamma(S)^{-1}D_{11}^*\right]^{-1} \\ &\quad \times \left[RC_1^* - (B_1 + q_{12}RA_0^*SB_0)\Pi_\gamma(S)^{-1}D_{11}^* + \alpha^2 B_2D_{12}^*\right]^* \succ 0.\end{aligned} \tag{60}$$

Similarly, we obtain that (59) is equivalent to $\Pi_\gamma(S) + \beta^2 D_{21}^*D_{21} - D_{11}^*D_{11} \succ 0$ and

(61)

$$\begin{aligned}&SA + A^*S + q_{11}A_0^*SA_0 - C_1^*C_1 + \beta^2 C_2^*C_2 \\ &\quad - \left[SB_1 + q_{12}A_0^*SB_0 - C_1^*D_{11} + \beta^2 C_2^*D_{21}\right] \\ &\quad \times \left[\Pi_\gamma(S) + \beta^2 D_{21}^*D_{21} - D_{11}^*D_{11}\right]^{-1}\left[SB_1 + q_{12}A_0^*SB_0 - C_1^*D_{11} + \beta^2 C_2^*D_{21}\right]^* \succ 0.\end{aligned}$$

Thus, we have the following corollary.

COROLLARY 3.7. *For any system of the form* (42) *and any* $\gamma > 0$*, the following conditions are equivalent:*

(i) *There exists a stabilizing compensator* (43) *of dimension* $\hat{n}$ *such that* $\|\mathbb{L}_{cl}\| < \gamma$.

(ii) *There exist* $(R, S) \in \mathcal{H}_n(\mathbb{K}) \times \mathcal{H}_n(\mathbb{K}), \alpha > 0, \beta > 0$*, such that*

$$\begin{aligned}&S \preceq R^{-1} \prec 0, \quad \operatorname{rk}(R^{-1} - S) \le \hat{n}, \quad \Pi_\gamma(S) = \gamma^2 I + q_{22}B_0^*SB_0 \succ 0, \\ &I + \alpha^2 D_{12}D_{12}^* - D_{11}\Pi_\gamma(S)^{-1}D_{11}^* \succ 0, \qquad \Pi_\gamma(S) + \beta^2 D_{21}^*D_{21} - D_{11}^*D_{11} \succ 0,\end{aligned}$$

and (60), (61) *hold.*

We conclude this section with a brief discussion of the *state feedback case*, where $C_2 = I_n$ and $D_{21} = 0$. Then (54) is equivalent to

$$\Pi_\gamma(S) - D_{11}^*D_{11} = \gamma^2 I + q_{22}B_0^*SB_0 - D_{11}^*D_{11} \succ 0.$$

The following corollary determines what can be achieved by static state feedback:

$$u(t) = Fx(t), \qquad F \in \mathbb{K}^{m\times n}. \tag{62}$$

COROLLARY 3.8. *There exists a stabilizing static state feedback controlled* (62) *such that* $\|\mathbb{L}_{cl}\| < \gamma$ *if and only if there exists* $R \in \mathcal{H}_n(\mathbb{K})$, $R \prec 0$ *satisfying*

$$\Pi_\gamma(R^{-1}) - D_{11}^*D_{11} = \gamma^2 I + q_{22}B_0^*R^{-1}B_0 - D_{11}^*D_{11} \succ 0 \tag{63}$$

and

$$\begin{bmatrix} AR + RA^* + q_{11}RA_0^*R^{-1}A_0R & RC_1^* \\ C_1R & I \end{bmatrix}$$

$$- \begin{bmatrix} B_1 + q_{12}RA_0^*R^{-1}B_0 \\ D_{11} \end{bmatrix} \Pi_\gamma(R^{-1})^{-1} \begin{bmatrix} B_1 + q_{12}RA_0^*R^{-1}B_0 \\ D_{11} \end{bmatrix}^* \succ 0 \tag{64}$$

$$\text{on } \ker \begin{bmatrix} B_2^* & D_{12}^* \end{bmatrix}.$$

Proof. In the static state feedback case, we have $\hat{n} = 0$ and hence (52) implies $S = R^{-1}$. But with $S = R^{-1}$, (53) and (54) are equivalent to (64) and (63), respectively. Thus the statement follows from Theorem 3.5. ∎

An interesting question is whether or not lower levels of γ can be achieved by employing *dynamic state feedback*. This has been answered in the negative for the deterministic case $A_0 = 0$, $B_0 = 0$, (see, e.g., [26]) and for the special stochastic case $A_0 = 0$, $B_1 = 0$, see Proposition 5.6 in [16]. The following corollary generalizes these results to the general stochastic case where the nominal system's noise $w_1(t)$ and the perturbation noise $w_2(t)$ are independent.

COROLLARY 3.9. *For any system of the form* (42) *with* $q_{12} = 0$ *and any* $\gamma > 0$, *the following conditions are equivalent:*

(i) *There exists a stabilizing static state feedback controller* (62) *such that* $\|\mathbb{L}_{cl}\| < \gamma$.

(ii) *There exists a stabilizing dynamic state feedback controller of dimension* $\hat{n} \geq 0$ (*i.e.,* (43) *with* $y(t) = x(t)$) *such that* $\|\mathbb{L}_{cl}\| < \gamma$.

Proof. Only the implication (ii) $\Rightarrow$ (i) needs to be proved. Assume (ii); then by Theorem 3.5 there exists $(R, S) \in \mathcal{H}_n(\mathbb{K}) \times \mathcal{H}_n(\mathbb{K})$, such that $S \preceq R^{-1} \prec 0$, and the following inequalities are satisfied:

$$\begin{bmatrix} AR + RA^* + q_{11}RA_0^*SA_0R & RC_1^* \\ C_1R & I \end{bmatrix}$$

$$- \begin{bmatrix} B_1 \\ D_{11} \end{bmatrix} \Pi_\gamma(S)^{-1} \begin{bmatrix} B_1 \\ D_{11} \end{bmatrix}^* \succ 0 \quad \text{on } \ker \begin{bmatrix} B_2^* & D_{12}^* \end{bmatrix},$$

$$\Pi_\gamma(S) - D_{11}^*D_{11} = \gamma^2 I + q_{22}B_0^*SB_0 - D_{11}^*D_{11} \succ 0.$$

But since $S \preceq R^{-1} \prec 0$, it follows that $0 \prec \Pi_\gamma(S) \preceq \Pi_\gamma(R^{-1})$ and $q_{11}RA_0^*SA_0R \preceq q_{11}RA_0^*R^{-1}A_0R$. Hence, the previous two inequalities hold with S replaced by R^{-1}. Therefore (i) follows from Corollary 3.8. ∎

4. The regular case. In this section we consider the so-called regular case and make the following usual assumptions [5]:

$$D_{11} = 0, D_{12}^*D_{12} = I, \quad D_{21}D_{21}^* = I, \quad D_{12}^*C_1 = 0, \quad D_{21}B_1^* = 0. \tag{65}$$

Then (60) becomes

$$\begin{aligned} &AR + RA^* + q_{11}RA_0^*SA_0R \\ &\quad - (B_1 + q_{12}RA_0^*SB_0)\Pi_\gamma(S)^{-1}(B_1 + q_{12}RA_0^*SB_0)^* + \alpha^2 B_2B_2^* \\ &\quad - (RC_1^* + \alpha^2 B_2D_{12}^*)(I + \alpha^2 D_{12}D_{12}^*)^{-1}(RC_1^* + \alpha^2 B_2D_{12}^*)^* \succ 0. \end{aligned}$$

By (65),

$$(I+\alpha^2 D_{12}D_{12}^*)^{-1}D_{12} = (1+\alpha^2)^{-1}D_{12}, \qquad (I+\alpha^2 D_{12}D_{12}^*)^{-1}C_1 = C_1,$$

and hence

$$RC_1^*(I+\alpha^2 D_{12}D_{12}^*)^{-1}(RC_1^*+\alpha^2 B_2 D_{12}^*)^* = RC_1^*C_1R.$$

Therefore, (60) is equivalent to

$$\begin{aligned} AR+RA^*+q_{11}RA_0^*SA_0R-(B_1+q_{12}RA_0^*SB_0)\Pi_\gamma(S)^{-1}(B_1+q_{12}RA_0^*SB_0)^* \\ +\alpha^2 B_2B_2^*-RC_1^*C_1R-\alpha^4 B_2D_{12}^*(I+\alpha^2D_{12}D_{12}^*)^{-1}D_{12}B_2^* \succ 0. \end{aligned}$$

Now $D_{12}^*(I+\alpha^2 D_{12}D_{12}^*)^{-1}D_{12} = (1+\alpha^2)^{-1}I$, so (60) is equivalent to

(66)
$$\begin{aligned} AR+RA^*+q_{11}RA_0^*SA_0R-(B_1+q_{12}RA_0^*SB_0)\Pi_\gamma(S)^{-1}(B_1+q_{12}RA_0^*SB_0)^* \\ +\alpha^2(1+\alpha^2)^{-1}B_2B_2^*-RC_1^*C_1R \succ 0. \end{aligned}$$

Equation (61) becomes

$$\begin{aligned} &SA+A^*S+q_{11}A_0^*SA_0-C_1^*C_1+\beta^2C_2^*C_2 \\ &-(SB_1+q_{12}A_0^*SB_0+\beta^2C_2^*D_{21})(\Pi_\gamma(S)+\beta^2D_{21}^*D_{21})^{-1} \\ &(SB_1+q_{12}A_0^*SB_0+\beta^2C_2^*D_{21})^* \succ 0. \end{aligned}$$

Suppose, in addition, that $D_{21}B_0^* = 0$, then after similar calculation to the ones above we obtain the equivalent inequality

(67)
$$\begin{aligned} SA+A^*S+q_{11}A_0^*SA_0-C_1^*C_1+\beta^2\gamma^2(\beta^2+\gamma^2)^{-1}C_2^*C_2 \\ -(SB_1+q_{12}A_0^*SB_0)\Pi_\gamma(S)^{-1}(SB_1+q_{12}A_0^*SB_0)^* \succ 0. \end{aligned}$$

If (66), (67) are satisfied for some given α, $\beta > 0$, they are also satisfied for all larger values. Taking limits as $\alpha \to \infty$, $\beta \to \infty$, we see that (66) and (67) are equivalent, respectively, to

$$\begin{aligned} &AR+RA^*+q_{11}RA_0^*SA_0R+B_2B_2^*-RC_1^*C_1R \\ &\quad -(B_1+q_{12}RA_0^*SB_0)\Pi_\gamma(S)^{-1}(B_1+q_{12}RA_0^*SB_0)^* \succ 0, \\ &SA+A^*S+q_{11}A_0^*SA_0-C_1^*C_1+\gamma^2C_2^*C_2 \\ &\quad -(SB_1+q_{12}A_0^*SB_0)\Pi_\gamma(S)^{-1}(SB_1+q_{12}A_0^*SB_0)^* \succ 0. \end{aligned} \tag{68}$$

Setting $R^{-1} = P$, the first inequality is equivalent to

$$\begin{aligned} &PA+A^*P+q_{11}A_0^*SA_0+PB_2B_2^*P-C_1^*C_1 \\ &\quad -(PB_1+q_{12}A_0^*SB_0)\Pi_\gamma(S)^{-1}(PB_1+q_{12}A_0^*SB_0)* \succ 0. \end{aligned} \tag{69}$$

Altogether we have derived the following consequence of Corollary 3.7.

PROPOSITION 4.1. *Suppose the regularity conditions* (65) *and* $D_{21}B_0^* = 0$. *Then the following statements are equivalent:*

(i) *There exists a stabilizing compensator* (43) *of dimension* $\hat{n}$ *such that* $\|\mathbb{L}_{cl}\| < \gamma$.

(ii) *There exist* $P, S \in \mathcal{H}_n(\mathbb{K})$ *such that* $S \preceq P \prec 0$, $\operatorname{rank}(P-S) \le \hat{n}$, $\gamma^2 I + q_{22} B_0^* S B_0 \succ 0$ *and* (68), (69) *hold.*

We now show how to explicitly calculate a compensator in the special case that $\hat{n} = n$. Suppose that condition (ii) of the previous proposition is satisfied with $\hat{n} = n$. Then there exist $P, S \in \mathcal{H}_n(\mathbb{K})$ such that $S \prec P \prec 0$, $\gamma^2 I + q_{22} B_0^* S B_0 \succ 0$, and $\Pi_S \succ 0$, $\Pi_P \succ 0$, where

$$\begin{aligned}\Pi_S &= SA + A^*S + q_{11}A_0^*SA_0 - C_1^*C_1 + \gamma^2 C_2^*C_2 \\ &\quad -(SB_1 + q_{12}A_0^*SB_0)\Pi_\gamma(S)^{-1}(SB_1 + q_{12}A_0^*SB_0)^*, \\ \Pi_P &= PA + A^*P + q_{11}A_0^*SA_0 + PB_2B_2^*P - C_1^*C_1 \\ &\quad -(PB_1 + q_{12}A_0^*SB_0)\Pi_\gamma(S)^{-1}(PB_1 + q_{12}A_0^*SB_0)^*.\end{aligned}$$

Define

$$B_K = \gamma^2(P-S)^{-1}C_2^*, \quad C_K = B_2^*P, \quad D_K = 0, \tag{70}$$

then,

$$A_{cl} = \begin{bmatrix} A & B_2B_2^*P \\ \gamma^2(P-S)^{-1}C_2^*C_2 & A_K \end{bmatrix}, \quad A_{cl}^0 = \begin{bmatrix} A_0 & 0 \\ 0 & 0 \end{bmatrix}, \quad B_{cl}^0 = \begin{bmatrix} B_0 \\ 0 \end{bmatrix},$$

$$B_{cl} = \begin{bmatrix} B_1 \\ \gamma^2(P-S)^{-1}C_2^*D_{21} \end{bmatrix}, \quad C_{cl} = [C_1 \;\; D_{12}B_2^*P], \quad D_{cl} = 0.$$

So condition (ii) of Theorem 2.8 is equivalent to the existence of $P_{cl} \prec 0$ such that

(71)

$$\begin{aligned}\Pi &:= P_{cl}A_{cl} + A_{cl}^*P_{cl} + q_{11}A_{cl}^{0*}P_{cl}A_{cl}^0 - C_{cl}^*C_{cl} \\ &\quad -(P_{cl}B_{cl} + q_{12}A_{cl}^{0*}P_{cl}B_{cl}^0)(\gamma^2 I + q_{22}B_{cl}^{0*}P_{cl}B_{cl}^0)^{-1}(P_{cl}B_{cl} + q_{12}A_{cl}^{0*}P_{cl}B_{cl}^0)^* \succ 0.\end{aligned}$$

Choosing $P_{cl} = \left[\begin{smallmatrix} S & N \\ N^* & Q \end{smallmatrix}\right]$, with $N = -Q = (P-S)$ then $P_{cl} \prec 0$ and partitioning $\Pi = \left[\begin{smallmatrix} \Pi_{11} & \Pi_{12} \\ \Pi_{12}^* & \Pi_{22} \end{smallmatrix}\right]$, we obtain from (71),

$$\begin{aligned}\Pi_{11} &= SA + A^*S + q_{11}A_0^*SA_0 + 2\gamma^2 C_2^*C_2 - C_1^*C_1 \\ &\quad -(SB_1 + q_{12}A_0^*SB_0 + \gamma^2 C_2^*D_{21})\Pi_\gamma(S)^{-1}(SB_1 + q_{12}A_0^*SB_0 + \gamma^2 C_2^*D_{21})^*, \\ \Pi_{12} &= SB_2B_2^*P + NA_K + A^*N - \gamma^2 C_2^*C_2 \\ &\quad -(SB_1 + q_{12}A_0^*SB_0 + \gamma^2 C_2^*D_{21})\Pi_\gamma(S)^{-1}(NB_1 - \gamma^2 C_2^*D_{21})^*, \\ \Pi_{22} &= NB_2B_2^*P + PB_2B_2^*N - NA_K - A_K^*N - PB_2B_2^*P \\ &\quad -(NB_1 - \gamma^2 C_2^*D_{21})\Pi_\gamma(S)^{-1}(NB_1 - \gamma^2 C_2^*D_{21})^*.\end{aligned}$$

Now $\Pi_{11} = \Pi_S$, and Π_{12} simplifies to

$$\Pi_{12} = SB_2B_2^*P + NA_K + A^*N - (SB_1 + q_{12}A_0^*SB_0)\Pi_\gamma(S)^{-1}B_1^*N.$$

Thus, choosing

$$A_K = -N^{-1}[SB_2B_2^*P + A^*N - (SB_1 + q_{12}A_0^*SB_0)\Pi_\gamma(S)^{-1}B_1^*N + \Pi_S], \tag{72}$$

we get $\Pi_{12} = -\Pi_S$. Finally, Π_{22} simplifies to

$$\Pi_{22} = NB_2B_2^*P + PB_2B_2^*N - NA_K - A_K^*N - PB_2B_2^*P - \gamma^2 C_2^*C_2 - NB_1\Pi_\gamma(S)^{-1}B_1^*N.$$

Substituting for A_K, we get $\Pi_{22} = \Pi_P + \Pi_S$. Hence $\Pi = \left[\begin{smallmatrix} \Pi_S & -\Pi_S \\ -\Pi_S & \Pi_P+\Pi_S \end{smallmatrix}\right]$, and since $\Pi_S \succ 0$, $\Pi_P \succ 0$, it follows that $\Pi \succ 0$, i.e., the above P_{cl} satisfies (71). We conclude from Theorem 2.8 that the closed loop system is stable and $\|\mathbb{L}_{cl}\| < \gamma$.

Remark 4.2. (i) Using (68) and (69) one can show that

$$A_K = A - B_K C_2 + B_2 C_K - B_1\Pi_\gamma(S)^{-1}(B_1^*P + q_{12}B_0^*SA_0) - N^{-1}\Pi_P.$$

The first three terms are familiar from pole-placement based deterministic dynamic output feedback stabilization. If $B_0 = 0$, the fourth term is $\gamma^{-2}B_1B_1^*P$ and this is familiar from deterministic H^∞ control. In this case, the last term ($N^{-1}\Pi_P$) is zero for the so-called "central controller."

(ii) We have shown in [16] that in a stochastic setting it is not possible, in general, to replace both inequalities (68) and (69) by equalities in Proposition 4.1.

5. Stability radii. In this section we turn to a *singular* control problem and discuss the application of our general results to stability radii. First we show how the results of section 2 can be used to derive a lower bound for stability radii of stochastic systems. Then we use the results of section 3 to show how to increase the radii via feedback. The corresponding H^∞-type control problem is singular because all three feedthrough matrices D_{11}, D_{12}, D_{21} are zero in this case. For the analysis problem, we adopt the notation of section 2 and, for the synthesis problem, the notation of section 3. Suppose that a stable linear stochastic model

$$\Sigma_0 : dx(t) = Ax(t)dt + A_0x(t)dw_1(t) \tag{73}$$

is perturbed to

$$\Sigma_\Delta : dx(t) = (A + B\Delta C)x(t)dt + A_0x(t)dw_1(t) + B_0\Delta Cx(t)dw_2(t), \tag{74}$$

where

$$(A, A_0, B_0, B, C) \in \mathbb{K}^{n\times n} \times \mathbb{K}^{n\times n} \times \mathbb{K}^{n\times \ell} \times \mathbb{K}^{n\times \ell} \times \mathbb{K}^{q\times n}.$$

The Wiener processes w_i, $i = 1, 2$ are as in section 2, and $\Delta \in \mathcal{D}(\mathbb{K}) = \mathbb{K}^{\ell\times q}$ represents an unknown disturbance matrix. We view the term $B\Delta C$ in (74) as a parameter perturbation of the nominal system matrix A and view $B_0\Delta Cx(t)dw_2(t)$ as a stochastic perturbation (multiplicative noise). If $w_1 = w_2 = w$, we can interpret $B_0\Delta C$ as a parameter perturbation of A_0 and write (74) in the following way:

$$\dot{x}(t) = (A + B\Delta C)x(t) + (A_0 + B_0\Delta C)x(t)dw(t). \tag{75}$$

The fact that the same Δ is used in both perturbations is not really a restriction since, if we set $B_0 = [B_0^0 \quad 0]$, $B = [0 \quad B^1]$, and $\Delta = \left[\begin{smallmatrix} \Delta_0 \\ \Delta_1 \end{smallmatrix}\right]$, we obtain

$$\dot{x}(t) = (A + B^1\Delta_1 C)x(t) + (A_0 + B_0^0\Delta_0 C)x(t)dw(t).$$

We will take this up again in Example 5.8.

The size of each $\Delta \in \mathcal{D}(\mathbb{K})$ is measured by its operator norm (with respect to the Euclidean norms on $\mathbb{K}^q, \mathbb{K}^\ell$). Our aim is to determine which bounds $\rho > 0$ on

the size of the perturbations ensure the stability of the perturbed system (74). The maximum ρ for which all the perturbed systems (74) with $\|\Delta\| < \rho$ are stable is called the stability radius of (73).

DEFINITION 5.1. *The* stability radius *of the stochastic system* (73), *with respect to perturbations as in* (74), *is*

$$r_{\mathbb{K}}^w = r_{\mathbb{K}}^w(A, A_0; B, B_0, C) = \inf\{\|\Delta\|; \Delta \in \mathcal{D}(\mathbb{K}), (74) \text{ is not stable}\}. \tag{76}$$

In particular, $r_{\mathbb{K}}^w = \infty$ *if there does not exist* $\Delta \in \mathcal{D}(\mathbb{K})$ *such that* (74) *is not stable.*

The stability radius is a quantitative index of robust stability of the system Σ_0 (73) under perturbations of the form $\Sigma_0 \rightsquigarrow \Sigma_\Delta$ (74). Since robust stability is a basic requirement for every control system with uncertain parameters, it is of considerable interest to have computable formulae or good estimates for the stability radii of a given system.

To connect the stability radius problem with the analysis in section 2, observe that the perturbed system (74) is identical with the closed loop system obtained from

$$\begin{aligned} dx(t) &= Ax(t)dt + A_0x(t)dw_1(t) + B_0v(t)dw_2(t) + Bv(t)dt \\ z(t) &= Cx(t), \end{aligned} \tag{77}$$

by setting $v(t) = \Delta z(t)$. The open loop gain of this closed loop system is given by $\|\mathbb{L}\Delta\|$, where $\mathbb{L} : v(\cdot) \mapsto Cx(\cdot, v, 0)$ is the perturbation operator associated with (77) (see Definition 2.3). It is therefore reasonable to expect that the norm of the perturbation operator plays a crucial role in the determination of the stability radius.

Remark 5.2. (i) If the data A, A_0, B, B_0, C are real, two stability radii are obtained depending on whether one chooses $\mathbb{K} = \mathbb{C}$ (complex perturbations) or $\mathbb{K} = \mathbb{R}$ (real perturbations) in (76). In a deterministic framework, the real and the complex stability radii are, in general, distinct; see [14]. The complex stability radius is equal to $\|\mathbb{L}\|^{-1}$ [12], whereas the real stability radius is characterized via second order singular values [24].

(ii) A stability radius with respect to *time-varying* and/or *nonlinear* perturbations can be defined via (76) by extending the perturbation class $\mathcal{D}(\mathbb{K})$ appropriately. For example, let $\mathcal{D}_{tn}(\mathbb{K})$ denote the set of all Lebesgue measurable $\Delta : \mathbb{R}_+ \times \mathbb{K}^q \mapsto \mathbb{K}^\ell$ which are *Lipschitz bounded* and *linearly bounded* in y; that is, for all $T > 0$ there exists $L = L(T)$ such that

$$\|\Delta(t, y) - \Delta(t, \hat{y})\| \le L\|y - \hat{y}\| \quad \text{for all } y, \hat{y} \in \mathbb{K}^q, \qquad t \in [0, T],$$

and there exists $K > 0$ such that

$$\|\Delta(t, y)\| \le K\|y\| \quad \text{for all } t \in \mathbb{R}_+, \qquad y \in \mathbb{K}^q. \tag{78}$$

The size of $\Delta \in \mathcal{D}_{tn}(\mathbb{K})$ is measured by the smallest K for which (78) holds. The stability radius of (73) with respect to time-varying nonlinear perturbations of the form

$$dx(t) = (Ax(t) + B\Delta(t, Cx(t)))dt + A_0x(t)dw_1(t) + B_0\Delta(t, Cx(t))dw_2(t) \tag{79}$$

is then defined by (76) with (74) replaced by (79), and $\mathcal{D}(\mathbb{K})$ replaced by $\mathcal{D}_{tn}(\mathbb{K})$. Here the nonlinear system (79) is said to be *stable* (recall Definition 2.1) if the solutions $x_\Delta(\cdot, x^0)$ of (79) satisfy

$$\int_0^\infty \mathcal{E}\|x_\Delta(t, x^0)\|^2 dt \le c\|x^0\|^2, \qquad x^0 \in \mathbb{K}^n,$$

for some suitable constant c. In the deterministic context ($A_0 = 0$, $B_0 = 0$), it is known [14] that the complex stability radius is not changed by such an extension of the perturbation class, whereas the real stability radius is. In the special case where the nominal model (73) is deterministic and the perturbations are purely stochastic, i.e., $A_0 = 0$, $B = 0$, it has been shown in [2] that the real and the complex stability radii coincide and are equal to the inverse of the norm of the perturbation operator ($\|\mathbb{L}\|^{-1}$) if *nonlinear* disturbances are considered. In this case, it is also possible to analyze the effect of *blockdiagonal perturbations*, where the single stochastic perturbation term $B_0\Delta Cx(t)dw_2(t)$ is replaced by a sum of the form $\sum_{i=1}^N B_0^i\Delta^i C^i x(t)dw_2^i(t)$. In the deterministic context, the analysis of blockdiagonal perturbations is the object of μ-analysis. In [16] it was shown that, in the case of purely stochastic perturbations of a deterministic system, the real and complex radii coincide and, although they are not equal to $\|\mathbb{L}\|^{-1}$, they are equal to the inverse of the norm of a suitably scaled perturbation operator. Analogous results are *not* available for deterministic blockdiagonal perturbations.

PROPOSITION 5.3. *Suppose that* (73) *is stable and* $\Delta \in \mathcal{D}_{tn}(\mathbb{K})$ *is a time-varying nonlinearity satisfying* $\|\Delta\| = \sup\{\|\Delta(t,y)\|/\|y\|;\ t \geq 0,\ y \in \mathbb{K}^q,\ y \neq 0\} < \|\mathbb{L}\|^{-1}$, *where* $\mathbb{L}$ *is the perturbation operator associated with data* $(A, A_0, B_0, B, C, 0)$ (*see Definition* 2.3). *Then the perturbed system* (79) *is stable. In particular,*

$$r_{\mathbb{K}}^w(A, A_0; B, B_0, C) \geq \|\mathbb{L}\|^{-1}. \tag{80}$$

Proof. Since Δ is Lipschitz bounded and linearly bounded, for every $x^0 \in \mathbb{K}^n$, $T > 0$, there exists a unique solution $x_\Delta(\cdot) = x_\Delta(\cdot, x^0) \in L_w^2([0,T]; L^2(\Omega, \mathbb{K}^n))$ of (79) satisfying $x_\Delta(0) = x^0$ with bounded second moments [19]. $x_\Delta(\cdot)$ is a continuous nonanticipative stochastic process on $\mathbb{R}_+$ satisfying the Ito integral equation

$$\begin{aligned} x_\Delta(t) = x^0 &+ \int_0^t (Ax_\Delta(s) + B\Delta(s, Cx_\Delta(s)))ds \\ &+ \int_0^t [A_0 x_\Delta(s) \quad B_0\Delta(s, Cx_\Delta(s))]d\begin{bmatrix} w_1(s) \\ w_2(s) \end{bmatrix}, \qquad t \geq 0. \end{aligned}$$

So $x_\Delta(\cdot)$ satisfies (7) with $v(\cdot) = v_\Delta(\cdot) = \Delta(\cdot, Cx_\Delta(\cdot)) \in L_w^2([0,T]; L^2(\Omega, \mathbb{K}^\ell))$ for every $T > 0$. Since $\|\Delta\| < \|\mathbb{L}\|^{-1}$, there exists $\gamma > \|\mathbb{L}\|$ such that $\gamma\|\Delta\| < 1$. Applying Theorem 2.8 to (74), there are $\delta > 0$ and $P = P^* \prec 0$ satisfying $M(P) \succeq \delta^2 I$. By Lemma 2.4 we obtain, for every $x^0 \in \mathbb{K}^n$ and $T > 0$,

$$\begin{aligned} J_T^{\gamma^2}(x^0, v_\Delta) = \langle x^0, Px^0\rangle &- \mathcal{E}\langle x_\Delta(T), Px_\Delta(T)\rangle \\ &+ \int_0^T \mathcal{E}\left(\left\langle \begin{bmatrix} x_\Delta(t) \\ v_\Delta(t) \end{bmatrix}, M(P)\begin{bmatrix} x_\Delta(t) \\ v_\Delta(t) \end{bmatrix}\right\rangle\right) dt. \end{aligned}$$

Substituting $\Delta(\cdot, Cs_\Delta(\cdot))$ for $v_\Delta(\cdot)$ and making use of definition (12) and inequality $M(P) \succeq \delta^2 I$, we obtain

$$\begin{aligned} \langle x^0, Px^0\rangle - \mathcal{E}\langle x_\Delta(T), Px_\Delta(T)\rangle \leq \int_0^T &[\gamma^2\mathcal{E}\|\Delta(t, Cx_\Delta(t))\|^2 \\ &- \mathcal{E}\|Cx_\Delta(t)\|^2 - \delta^2\mathcal{E}\|x_\Delta(t)\|^2]dt. \end{aligned}$$

Now, $\gamma^2\mathcal{E}\|\Delta(t,Cx_\Delta(t))\|^2 \le \gamma^2\|\Delta\|^2\mathcal{E}\|Cx_\Delta(t)\|^2 \le \mathcal{E}\|Cx_\Delta(t)\|^2$ and $-P \ge \eta I$ for some $\eta > 0$. Hence,

$$\eta\mathcal{E}\|x_\Delta(T)\|^2 \le -\mathcal{E}\langle x_\Delta(T), Px_\Delta(T)\rangle \le \|P\|\,\|x^0\|^2 - \int_0^T \delta^2\mathcal{E}\|x_\Delta(t)\|^2 dt, \qquad T > 0,$$

and it follows that

$$\int_0^\infty \mathcal{E}\|x_\Delta(t)\|^2 dt \le \|P\|\,\|x^0\|^2/\delta^2,$$

i.e., (79) is stable. □

We illustrate the above result by considering the same scalar stochastic system as in Example 2.15. In this simple example, we will see that the estimate (80) is tight.

Example 5.4. Consider

$$dx(t) = -x(t)dt + x(t)dw_1(t) + v(t)dw_2(t) + v(t)dt, \qquad z(t) = x(t), \tag{81}$$

and assume first that $q_{11} = 1$, $q_{12} = q_{22} = 0$ so that the stochastic perturbation term $v(t)dw_2(t)$ is absent from (81). We have shown in Example 2.15 that $\|\mathbb{L}\| = 2$ in this case. The corresponding perturbed equation (74) takes the form

$$dx(t) = -(1-\Delta)x(t)dt + x(t)dw_1(t).$$

By (8), this stochastic equation is stable if and only if there exists $p < 0$ such that

$$-(2-\Delta-\Delta^*)p + p > 0, \quad \text{i.e., } -(2-\Delta-\Delta^*) + 1 < 0.$$

Hence, $\Delta = 1/2$ is a destabilizing *real* disturbance, and by Proposition 5.3 there is no smaller disturbance $\Delta \in \mathbb{C}$ which destabilizes. So $r^w_{\mathbb{R}} = r^w_{\mathbb{C}} = \|\mathbb{L}\|^{-1} = 1/2$.

Now suppose $w_1 = w_2 = w$ and $q_{11} = q_{12} = q_{22} = 1$. We have shown in Example 2.15 that $\|\mathbb{L}\|^2 = 9 + \sqrt{80}$. The perturbed model takes the form

$$dx(t) = -(1-\Delta)x(t)dt + (1+\Delta)x(t)dw(t).$$

By (8), this stochastic equation is stable if and only if there exists $p < 0$ such that

$$-(2-\Delta-\Delta^*)p + |1+\Delta|^2 p > 0, \quad \text{i.e., } -(2-\Delta-\Delta^*) + |1+\Delta|^2 < 0.$$

A short calculation shows that $\Delta = \sqrt{5}-2$ is the smallest disturbance $\Delta \in \mathbb{C}$ violating this condition. Hence $r^w_{\mathbb{R}} = r^w_{\mathbb{C}} = \sqrt{5}-2$, but $\|\mathbb{L}\|^{-1} = 1/\sqrt{9+\sqrt{80}} = \sqrt{9-\sqrt{80}} = \sqrt{5}-2$. Therefore, in this case we again have $r^w_{\mathbb{R}} = r^w_{\mathbb{C}} = \|\mathbb{L}\|^{-1}$.

As was to be expected, the presence of the stochastic disturbance term $\Delta x(t)dw(t)$ effectively decreases the stability radius of the system. This is not necessarily so if there is more than one disturbance parameter, i.e., $\max\{\ell, q\} > 1$.

We now turn to the synthesis problem. The perturbed closed loop equation obtained by setting $v = \Delta z$ in (44) is

$$d\bar{x}(t) = (A_{cl} + B_{cl}\Delta C_{cl})\bar{x}(t)dt + (A^0_{cl}dw_1(t) + B^0_{cl}\Delta C_{cl}dw_2(t))\bar{x}(t). \tag{82}$$

As an immediate corollary of Proposition 5.3 and Theorem 3.3 we have the following.

COROLLARY 5.5. *Let γ_{opt} be defined by* (55)*. Then for any $\gamma > \gamma_{\text{opt}}$, there exists a stabilizing compensator* (43) *such that the corresponding closed loop system has a stability radius $r^w_{\mathbb{K}}(A_{cl}, A^0_{cl}; B_{cl}, B^0_{cl}, C_{cl}) > \gamma^{-1}$.*

In the following examples, we indicate how the synthesis problem may be solved under simplifying assumptions. We first consider two system classes (4) for which the estimate (80) is tight.

Example 5.6. $A_0 = 0, B_0 = 0$. For this deterministic case, we have shown [12] that $r^w_{\mathbb{C}} = \|\mathbb{L}_{cl}\|^{-1}$, but in general $r^w_{\mathbb{R}} > r^w_{\mathbb{C}}$ [14]. Equations (60) and (61) are equivalent to

$$\Pi_R = AR + RA^* - RC_1^*C_1R - B_1B_1^*/\gamma^2 + \alpha^2 B_2B_2^* \succ 0. \tag{83}$$

$$\Pi_S = SA + A^*S - C_1^*C_1 - SB_1B_1^*S/\gamma^2 + \beta^2 C_2^*C_2 \succ 0. \tag{84}$$

Let

$$r_{\text{opt}} = \sup\{r^w_{\mathbb{C}}(A_{cl}, 0; B_{cl}, 0, C_{cl}); \exists \text{ compensator } M_K \text{ s.t. } A_{cl} \text{ is stable}\}.$$

Then we have $r_{\text{opt}} = \gamma^{-1}_{\text{opt}}$. For any $\gamma > 0$, $\gamma^{-1} < r_{\text{opt}}$, using the same procedure as that in section 4, it is easy to verify that, provided there exist $R, S \in \mathcal{H}_n(\mathbb{K})$, $\alpha, \beta > 0$ satisfying $S \prec R^{-1} \prec 0$ and (83), (84), the following compensator of order n achieves $r^w_{\mathbb{C}} > \gamma^{-1}$:

$$\begin{aligned} B_K &= \beta^2(R^{-1} - S)^{-1}C_2^*, \quad C_K = \alpha^2\beta_2^* R^{-1}, \quad D_K = 0, \\ A_K &= A - B_KC_2 + B_2C_K - B_1B_1^*R^{-1}/\gamma^2 - (I - RS)^{-1}\Pi_R R^{-1}, \end{aligned}$$

where Π_R is defined by (83).

Example 5.7. $A_0 = 0, B_1 = 0$. For this case, we have shown that if nonlinear perturbations Δ are allowed, then $r^w_{\mathbb{R}} = r^w_{\mathbb{C}} = \|\mathbb{L}_{cl}\|^{-1}$ [16]. Equations, (53) and (54) are equivalent to

$$AR + RA^* - RC_1^*C_1R \succ 0 \quad \text{on ker } B_2^* \tag{85}$$

$$SA + A^*S - C_1^*C_1 \succ 0 \quad \text{on ker } C_2, \qquad \Pi_\gamma(S) = \gamma^2 I + q_{22}B_0^*SB_0 \succ 0. \tag{86}$$

Let

$$r_{\text{opt}} = \sup\{r^w_{\mathbb{C}}(A_{cl}, 0; 0, B^0_{cl}, C_{cl}), \exists \text{ a compensator } M_K \text{ s.t. } A_{cl} \text{ is stable}\}.$$

Then again we have $r_{\text{opt}} = \gamma^{-1}_{\text{opt}}$, and for any $\gamma > 0$, $\gamma^{-1} < r_{\text{opt}}$ we can use (85) and (86) together with (52) to obtain a compensator which achieves $r^w_{\mathbb{C}} > \gamma^{-1}$. In fact, in [16] we have shown that (by scaling) it is possible to explicitly construct such compensators for more general perturbation structures where $B_0\Delta C_1 x(t)dw_2(t)$ is replaced by a sum of the form $\sum_{i=1}^N B_0^i\Delta^i C_1^i x(t)dw_2^i(t)$, and w_2^i are independent Wiener processes.

Finally we consider an example where we do not know whether or not (80) is tight.

Example 5.8: $A_0 = 0, B_0 = [B_0^0 \quad 0], B_1 = [0 \quad B_1^1]$, where $B_0^0 \in \mathbb{K}^{n\times \ell_1}$. For $v = \left[\begin{smallmatrix} v_0 \\ v_1 \end{smallmatrix}\right]$, (42) has the form

$$\begin{aligned} dx(t) &= Ax(t)dt + B_0^0 v_0(t)dw_2(t) + B_1^1 v_1(t)dt + B_2u(t)dt, \\ z(t) &= C_1x(t), \\ y(t) &= C_2x(t). \end{aligned}$$

Moreover, if

$$\begin{bmatrix} v_0 \\ v_1 \end{bmatrix} = \Delta z = \begin{bmatrix} \Delta_0 \\ \Delta_1 \end{bmatrix} C_1x,$$

then the perturbed equation is

$$dx(t) = (A + B_1^1 \Delta_1 C_1)x(t)dt + B_0^0 \Delta_0 C_1 x(t) dw_2(t) \; + \; B_2 u(t) dt.$$

Because of the presence of both stochastic and deterministic perturbations, we do not know whether the stability radius is equal to $\|\mathbb{L}_{cl}\|^{-1}$ or whether this is just a lower bound.

Since $B_1 B_0^* = 0$, we have $B_1(\gamma^2 I + q_{22} B_0^* S B_0)^{-1} B_1^* = B_1 B_1^*/\gamma^2$, and hence (60) and (61) are equivalent to

$$\Pi_R = AR + RA^* - RC_1^* C_1 R - B_1 B_1^*/\gamma^2 + \alpha^2 B_2 B_2^* \succ 0, \tag{87}$$

$$\Pi_S = SA + A^* S - C_1^* C_1 - S B_1 B_1^* S/\gamma^2 + \beta^2 C_2^* C_2 \succ 0. \tag{88}$$

For $\gamma > 0$, provided there exist R, $S \in \mathcal{H}_n(\mathbb{K})$, $\alpha, \beta > 0$ satisfying $S \prec R^{-1} \prec 0$, $\gamma^2 I + q_{22} B_0^* S B_0 \succ 0$ and (87), (88), the same compensator as that given in Example 5.6 achieves $r_{\mathbb{C}}^w > \gamma^{-1}$.

6. Concluding remarks. We have posed and solved an H^∞-type problem where both stochastic and deterministic perturbations are present. In opening up this field, which appears to be fruitful, we think that it would be interesting to pursue research into the following problems.

- Since our theory includes both cases where w_1, w_2 are independent and $w_1 = w_2$, we think we have laid the foundation for considering the stochastic multiperturbation H^∞ problem

$$\begin{aligned} dx(t) &= Ax(t)dt + \sum_{i=1}^N A_0^i x(t) dw^i(t) + \sum_{i=1}^N B_0^i v^i(t) dw^i(t) \\ &\quad + \sum_{i=1}^N B_1^i v^i(t) dt + B_2 u(t) dt, \\ z(t) &= C_1 x(t) + \sum_{i=1}^N D_{11}^i v^i(t) + D_{12} u(t), \\ y(t) &= C_2 x(t) + \sum_{i=1}^N D_{21}^i v^i(t), \end{aligned} \tag{89}$$

  where w^i are independent Wiener processes.
- In Proposition 5.3, we obtained the estimate $r_{\mathbb{K}}^w(A, A_0; B, B_0, C) \geq \|\mathbb{L}\|^{-1}$ for the stochastic stability radii. It would be interesting to know under what conditions equality holds. To do this, it is necessary to construct a destabilizing perturbation with norm as close as we like to $\|\mathbb{L}\|^{-1}$.
- A stability radius can be associated with the above multiperturbation problem (89) in a number of different ways. For example, if all the D's are zero and $v_i = \Delta_i z = \Delta_i C_1 x$, we get the so-called *full block* case, whereas if $v_i = \Delta_i z_i = \Delta_i C_1^i x$ we get the *blockdiagonal* case. In both cases, $\|\mathbb{L}_{cl}\|^{-1}$ will be a lower bound for the radii and, for full bock perturbations, it may well be tight. But we cannot expect this for blockdiagonal perturbations since in a deterministic setting this is a μ problem. The estimate can be improved by scaling $B_0^i \mapsto \alpha_i B_0^i$, $B_1^i \mapsto \alpha_i B_1^i$, $C_1^i \mapsto \alpha_i^{-1} C_1^i$, $\alpha_i > 0$ which does not

change the radius but does change the corresponding $\|\mathbb{L}_{cl}^{\alpha}\|$, and we have shown in [16] that for suitable α_i this estimate is tight for purely stochastic multiperturbations of a deterministic system. It would be interesting to know under what conditions this is true when deterministic perturbations are also present.

REFERENCES

[1] L. Arnold, *Stochastic Differential Equations: Theory and Applications*, John Wiley, New York, 1974.

[2] A. El Bouhtouri and A. J. Pritchard, *Stability radii of linear systems with respect to stochastic perturbations*, Systems Control Lett., 19 (1992), pp. 29–33.

[3] A. El Bouhtouri and A. J. Pritchard, *A Riccati equation approach to maximizing the stability radius of a linear system by state feedback under structured stochastic Lipschitzian perturbations*, Systems Control Lett., 21 (1993), pp. 475–484.

[4] G. Da Prato and J. Zabczyk, *Stochastic Equations in Infinite Dimensions*, Encyclopedia of Mathematics and Its Applications, Cambridge University Press, Cambridge, MA, 1992.

[5] J. Doyle, K. Glover, P. P. Khargonekar, and B. Francis, *State space solutions to standard H_2 and H^∞ control problems*, IEEE Trans. Automat. Control, AC-34 (1989), pp. 831–847.

[6] L. El Ghaoui, *State-feedback control of systems with multiplicative noise via linear matrix inequalities*, Systems Control Lett., 24 (1995), pp. 223–228.

[7] A. Friedman, *Stochastic Differential Equations and Applications*, Probability and Mathematical Statistics 28, Academic Press, New York, 1975.

[8] P. Gahinet and P. Apkarian, *A linear matrix inequality approach to H^∞ control*, Internat. J. Robust Nonlinear Control, 4 (1994), pp. 421–448.

[9] M. Green and D. J. N. Limebeer, *Linear Robust Control*, Prentice–Hall, Englewood Cliffs, NJ, 1995.

[10] R. Z. Has'minskii, *Stochastic Stability of Differential Equations*, Sijthoff & Noordhoff, Alphen aan den Rijn, 1980. (Translation of Russian edition, Nauka, Moscow, 1969.)

[11] U. G. Haussmann, *Optimal stationary control with state and control dependent noise*, SIAM J. Control Optim., 9 (1971), pp. 184–198.

[12] D, Hinrichsen and A. J. Pritchard, *Stability radius for structured perturbations and the algebraic Riccati equation*, Systems Control Lett., 8 (1986), pp. 105–113.

[13] D. Hinrichsen and A. J. Pritchard, *Riccati equation approach to maximizing the complex stability radius by state feedback*, Internat. J. Control, 52 (1990), pp. 769–794.

[14] D. Hinrichsen and A. J. Pritchard, *Real and complex stability radii: A survey*, in Control of Uncertain Systems, Progress in System and Control Theory 6, D. Hinrichsen and B. Martensson, eds., Birkhäuser, Basel, 1990, pp. 119–162.

[15] D. Hinrichsen and A. J. Pritchard, *Stability margins for systems with deterministic and stochastic uncertainty* in Proc. 33rd IEEE Conf. Decision and Control, Florida, 1994, IEEE Computer Society Press, Los Alamitos, CA, pp. 3825–3836.

[16] D. Hinrichsen and A. J. Pritchard, *Stability radii of systems with stochastic uncertainty and their optimization by output feedback*, SIAM J. Control Optim., 34 (1996), pp. 1972–1998.

[17] T. Iwasaki and R. E. Skelton, *All controllers for the general H^∞ control problem: LMI existence conditions and state space formulas*, Automatica, 30 (1994), pp. 1307–1317.

[18] N. V. Krylov, *Introduction to the Theory of Diffusion Processes*, Translations of Mathematical Monographs 142, AMS, Providence, RI, 1995.

[19] N. V. Krylov, *Controlled Diffusion Processes*, Springer-Verlag, New York, 1980.

[20] P. J. McLane, *Optimal stochastic control of linear systems with state and control-dependent disturbances*, IEEE Trans. Automat. Control, AC-16 (1971), pp. 292–299.

[21] T. Morozan, *Stability radii for some stochastic differential equations*, Stochastics, Stochastics Rep., 54 (1995), pp. 281–291.

[22] T. Morozan, *Parametrized Riccati Equations Associated to Input-Output Operators for Time-Varying Stochastic Differential Equations with State-Dependent Noise*, Institutul de Matematica al Academiei Romane, preprint no. 37, Bucarest, 1995.

[23] A. Packard, P. Zhou, P. Pandey, and G. Becker, *A collection of robust control problems leading to LMIs*, in Proc. 30th IEEE Conf. Decision and Control, Brighton 1991, IEEE Comput. Society Press, Los Alamitos, CA, pp. 1245–1250.

[24] L. Qiu, B. Bernhardsson, A. Rantzer, E. J. Davison, P. M. Young, and J. C. Doyle, *On the real structured stability radius*, in Proc. 12th IFAC World Congress, IFAC, Sydney, Australia, 1993, pp. 71–78.

[25] C. Scherer, *H^∞ optimization without assumptions on finite or infinite zeros*, SIAM J. Control Optim., 30 (1992), pp. 123–142.
[26] A. A. Stoorvogel, *The H^∞ Control Problem*, Prentice–Hall, New York, 1992.
[27] J. L. Willems and J. C. Willems, *Feedback stabilizability for stochastic systems with state and control dependent noise*, Automatica, 12 (1976), pp. 277–283.
[28] J. Willems and J. C. Willems, *Robust stabilization of uncertain systems*, SIAM J. Control Optim., 21 (1983), pp. 352–374.
[29] W. M. Wonham, *Optimal stationary control of a linear system with state dependent noise*, SIAM J. Control Optim., 5 (1967), pp. 486–500.
[30] W. M. Wonham, *Random differential equations in control theory*, in Probabilistic Methods in Applied Mathematics, 2, A. T. Bharucha-Reid, ed., Academic Press, New York, 1970, pp. 131–212.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 5, pp. 1539–1575, September 1998

003

SOLVING SCHEDULING PROBLEMS BY SIMULATED ANNEALING*

OLIVIER CATONI†

Abstract. We define a general methodology to deal with a large family of scheduling problems. We consider the case where some of the constraints are expressed through the minimization of a loss function. We study in detail a benchmark example consisting of some jigsaw puzzle problem with additional constraints. We discuss some algorithmic issues typical of scheduling problems, such as the apparition of small unused gaps or the representation of proportionality constraints. We also carry on an experimental comparison between the Metropolis algorithm, simulated annealing, and the iterated energy transformation method to see whether asymptotical theoretical results are a good guide towards practically efficient algorithms.

Key words. scheduling problems, Metropolis algorithm, simulated annealing, IET algorithm

AMS subject classifications. 60J10, 90C42, 82C80, 65C05

PII. S0363012996307813

Introduction. The aim of this paper is to describe a general strategy to deal with scheduling problems and to illustrate its use on the resolution of jigsaw puzzles. We will assume that we can put our scheduling problem in the form of a task assignment problem, and we will turn it into the minimization of a cost function defined on a suitable search space. This cost function will be minimized by a Monte Carlo algorithm of the Metropolis kind: either simulated annealing or our recently introduced iterated energy transformation (IET) method. We have already studied some of the theoretical aspects of these two methods in previous papers (see [6], [7]).

We have chosen to experiment on a jigsaw puzzle problem with rectangular pieces because this is a typical instance of the kind of difficulties encountered when building time tables, and because it is in itself a difficult problem (it is NP-complete) which deserves special attention. In the course of this experimentation, we will compare four algorithms: a randomized descent algorithm (the Metropolis dynamic at temperature zero), the Metropolis algorithm, simulated annealing, and the iterated energy transformation algorithm.

1. An abstract task assignment framework. Let B be a finite set of tasks. Let E be a set of resources needed to perform these tasks. The set E may be any kind of set, a finite set, a domain in $\mathbb{R}^n$, etc. In applications it can represent various things, such as a set of people who are to perform the tasks, in which case it is natural to see it as a finite set, or it can also represent space and time needed for the tasks, in which case it is sometimes natural to see it as a domain in $\mathbb{R}^n$. More often it is a product space of both kinds. Anyhow, we will only consider a finite collection of subsets of E; therefore it will always be possible to consider that E is a finite set from the theoretical point of view. This is reasonable, because a computer can only handle a finite number of possible ways to allocate resources and also because, in many problems of the time-table type, continuous quantities, such as time, are discretized (for instance when one tries to schedule lectures, they are usually constrained to start at full hours). Anyhow, the reader should think of E as a large set and our methods,

*Received by the editors August 5, 1996; accepted for publication (in revised form) August 20, 1997; published electronically June 3, 1998.
http://www.siam.org/journals/sicon/36-5/30781.html
†D.I.A.M. - Laboratoire de Mathématiques de l'Ecole Normale Supérieure, U.A. 762 du C.N.R.S., 45 rue d'Ulm, 75005 Paris, France (catoni@dmi.ens.fr).

inherited from statistical mechanics, are precisely meant to cope with a large state space.

The abstract scheduling problem we will consider is to allocate to each task in B a set of resources in a way which satisfies a set of constraints.

At this level of generality, we will not represent the constraints by equations or logical relations; we will merely view them as a subset $\mathcal{S}$ of $\mathcal{P}(B \times E)$ (where $\mathcal{P}(A)$ is the set of subsets of the set A). We will call $\mathcal{S}$ the "solution space." A solution x in $\mathcal{S}$ is a subset of the product space $B \times E$. We will use the notations π_B and π_E for projections on B and E. The fact that $(b, e) \in x$ means that the task b uses the resource e. The set of resources used by b is $\pi_E(\pi_B^{-1}(b) \cap x)$, for which we will use the functional notation $x(b)$.

We will assume that each solution $x \in \mathcal{S}$ is a complete assignment, in the sense that all the tasks are scheduled:

$$\pi_B(x) = B \quad \text{for any } x \in \mathcal{S}.$$

Our scheduling problem is to construct a solution x belonging to the solution space $\mathcal{S}$.

The idea of considering scheduling problems as putting objects in boxes in a multi-dimensional space is not new and can be found, for instance, in Abramson [1], where a specialized simulated annealing hardware is described for handling some generic types of cost functions.

2. The jigsaw puzzle example. This example is meant to be a benchmark, where the main algorithmic issues of scheduling problems are present.

The set of resources E will be a discretized rectangular frame

$$E = \{0, \dots, M-1\} \times \{0, \dots, N-1\} \subset \mathbb{Z}^2.$$

The set of tasks B will be the set of pieces of the jigsaw puzzle. Each piece r has a rectangular shape defined by its width $w_r \in \mathbb{N}^*$ and by its height $h_r \in \mathbb{N}^*$. The constraint is that pieces should not overlap. Thus the solution space is

$$\mathcal{S} = \{x \subset B \times E \ : \ x(r) = [a_r, a_r + w_r[\times[b_r, b_r + h_r[, \quad (a_r, b_r) \in E, \quad r \in B, \\ \text{and } x(r) \cap x(r') = \emptyset, \quad r \neq r' \in B\}.$$

The problem is to build the jigsaw puzzle; that is, to construct $x \in \mathcal{S}$. Although the shape of pieces is very simple, this problem can be seen to be very complex. In fact, it is easy to see that it is NP complete, because it contains the partition problem among its instances (see [14]). Indeed, the partition of given integers $\{c_1, \dots, c_N\}$ into two sets I and J such that

$$\sum_{i \in I} c_i = \sum_{j \in J} c_j$$

can be viewed as a jigsaw puzzle with N pieces, respectively, of width c_i and height 1, and a frame of width $(1/2) \sum_{i=1}^N c_i$ and height 2 (see Fig. 2.1).

3. A method of resolution based on the Metropolis dynamic. In this section we will sketch a methodology to solve the abstract problem of section 1. The general idea is to perform a random search for a solution in a state space larger than the solution space. This search space should be easy to describe and easy to search

FIG. 2.1.

by a Markov chain performing a succession of elementary moves. Of course, we will not use a Markov chain which uniformly samples the search space because, usually, the search space we will be able to build will be very large when compared to the solution space, and drawing points at random in the search space would seldom lead to discovering a solution.

Instead we will use a Markov chain with rare transitions, whose invariant measure is concentrated in a neighborhood of the solution space. This optimization technique is well known, but its improvement is still a subject of active research. The prototype algorithm we will start from is the Metropolis dynamic at low temperature. The Metropolis dynamic has been designed to simulate statistical mechanics systems, and not for optimization purposes. In order to improve its performance as an optimization algorithm, some speed-up techniques have been proposed. The most famous one is simulated annealing [15], [18]. We have also proposed recently another technique, which we called *the iterated energy transformation method* (IET) [7]. We will describe and use both of these.

3.1. Choice of a search space. The first step of the method is to choose a search space $\tilde{\mathcal{S}}$ containing the solution space $\mathcal{S}$. The most popular way to construct $\tilde{\mathcal{S}}$ is to relax some constraints about the solution and to measure, instead, how much the constraints have been violated by a score function one has afterwards to minimize. For instance, in circuits placement applications (one of the earliest applications of simulated annealing) the constraint that circuits should not overlap is often relaxed, and the overlapping of circuits is instead merely discouraged by some score function of the surface of the overlap. Our strategy will be somewhat of the same kind, with the difference that we will not relax a constraint which is specific to the problem. Instead, we will allow partial solutions, where only some proportion of the tasks have been scheduled. Defining partial solutions is usually very easy and very natural. Most of the time, this is how the problem is posed from the beginning. Indeed the constraints come usually from incompatibilities between tasks, such as sharing the same resource or needing to be performed in a given order, and can be expressed without assuming that all the tasks are already scheduled.

From the technical point of view, we will assume that the search space (the space of partial solutions) satisfies the following properties:

- The empty solution is in the search space: $\emptyset \in \tilde{\mathcal{S}}$.
- There is a path from the empty solution leading to any partial solution $x \in \tilde{\mathcal{S}}$ along which tasks are scheduled one after the other. This can be expressed in the following way:

  For any $x \in \tilde{\mathcal{S}}, x \neq \emptyset$, there is $b \in \pi_B(x)$ such that $x \setminus \pi_B^{-1}(b) \in \tilde{\mathcal{S}}$.
- All complete solutions in the search space satisfy the constraints. In other words, the solution space is exactly made of the complete solutions of the

search space. This is expressed by the following equation:

$$\mathcal{S} = \{x \in \tilde{\mathcal{S}} \ : \ \pi_B(x) = B\}.$$

Let us notice that the "best" choice for $\tilde{S}$ would be $\{x \cap \pi_B^{-1}(C) \ : \ x \in \mathcal{S}, C \subset B\}$, the set of all partial solutions contained in global solutions. Anyhow, this set is, in practical situations, never defined by simple relations, because when you have scheduled some of the tasks, it is never possible (except for trivial problems) to foretell whether there will remain suitable resources to schedule the remaining ones. Therefore the search space $\tilde{S}$ is, most of the time, much broader than $\mathcal{S}$ and contains many dead ends.

3.2. Building the dynamic: Constructions and destructions. The next idea is to define on $\tilde{S}$ two kinds of dynamics, a *constructive dynamic* and a *destructive dynamic*. These two random dynamics are characterized by two Markov matrices q_C and q_D,

$$\begin{aligned} q_C &: \tilde{\mathcal{S}} \times \tilde{\mathcal{S}} \longrightarrow [0,1], \\ q_D &: \tilde{\mathcal{S}} \times \tilde{\mathcal{S}} \longrightarrow [0,1]. \end{aligned}$$

We will assume that the transitions allowed by q_C consist of either keeping the current partial solution or scheduling one more task. In a similar way the transitions allowed by q_D consist of unscheduling a given number of tasks. We allow unscheduling of more than one task at a time, because it is in some situations more sensible to do so. For instance, if many tasks have to share the same resource, it may sometimes speed up the allocation process to unschedule all of them at the same time (think of students sharing the same teacher).

This conception of constructions and destructions can be expressed by the following equations, where $|A|$ is the number of elements in the finite set A:

$$\left\{ \begin{array}{l} \bullet \ \{(x,y) \ : \ q_C(x,y) > 0, x \neq y\} \\ \qquad = \{(x,y) \ : \ y \cap \pi_B^{-1}(\pi_B(x)) = x, |\pi_B(y)| = |\pi_B(x)| + 1\}, \\ \bullet \ \{(x,y) \ : \ q_C(y,x) > 0, x \neq y\} \\ \qquad\qquad \subset \{(x,y) \ : \ q_D(x,y) > 0\} \\ \qquad\qquad\qquad \subset \bigcup_{n=1}^{+\infty} \{(x,y) \ : \ q_C^n(y,x) > 0\}. \end{array} \right. \tag{1}$$

Let us remark that, usually, constructions will decompose into two steps, one being to choose an unscheduled task $b \in B \setminus \pi_B(x)$ and the second one being to try to allocate a set of resources to it. This second step is sometimes unsuccessful (either because it is impossible or the proper allocation has not been discovered); therefore as a rule, we have $q_C(x,x) > 0$ for a substantial number of partial solutions. On the contrary, destructions are simple moves, where you have only to choose a scheduled task b in $\pi_B(x)$ and to remove it. Therefore as a rule, we will have $q_D(x,x) = 0$, except when $x = \emptyset$, for which $q_D(\emptyset, \emptyset) = 1$.

When the two above assumptions are satisfied, the whole search space $\tilde{\mathcal{S}}$ can be constructed by q_C starting from the empty solution $\emptyset$, and reversely, any solution can

be shrunk to the empty solution by successive applications of q_D. More precisely, the following proposition holds.

PROPOSITION 3.1.

$$\begin{aligned}\tilde{\mathcal{S}} &= \bigcup_{n=0}^{+\infty} \{y \ : \ q_C^n(\emptyset, y) > 0\} \\ &= \bigcup_{n=0}^{+\infty} \{x \ : \ q_D^n(x, \emptyset) > 0\}.\end{aligned}$$

3.3. Building the cost function. Now we will build a cost, or energy, function defined on the search space, which penalizes partial solutions: we will call it $U \ : \ \tilde{\mathcal{S}} \longrightarrow \mathbb{R}$. Namely, we will require the following properties to hold:

(2)
$$\begin{cases} \bullet \ \arg\min_{x \in \tilde{\mathcal{S}}} U(x) = \mathcal{S}. \\ \bullet \ \text{There is a positive constant } \gamma \text{ such that } U(y) \geq U(x) + \gamma \text{ when } x \neq y \text{ and } q_D(x, y) > 0. \end{cases}$$

A typical example for U is

$$U(x) = \mu(B \setminus \pi_B(x)),$$

where μ is a positive measure on B. In this case the assumptions on U are satisfied and the largest choice of γ is

$$\gamma = \min_{b \in B} \mu(b).$$

3.4. Building a Metropolis dynamic. From Proposition 3.1, we see that any Markov matrix of the form $\lambda q_C + (1 - \lambda) q_D$ with $\lambda \in]0, 1[$ is irreducible. Therefore a straightforward way to build a Metropolis dynamic would be to consider the Markov matrix

$$p_T(x, y) = \begin{cases} (\lambda \, q_C(x, y) + (1 - \lambda) \, q_D(x, y)) e^{-(U(y) - U(x))^+ / T}, & x \neq y \\ 1 - \sum_{z \neq x} p_T(x, z), & x = y. \end{cases}$$

In fact, we can do better because we know in advance that, during a construction, the energy will decrease, and that during a destruction, the energy will increase by a quantity at least equal to γ. This avoids applying uselessly the kernel q_D at low temperatures in situations where we know that it will, most of the time, generate a move to be rejected.

More precisely, we will use the following Markov matrix:

(3)
$$\begin{aligned} p_T(x, y) &= \lambda \ e^{-\gamma / T} q_D(x, y) e^{-(U(y) - U(x) - \gamma)^+ / T} \\ &\quad + q_C(x, y) \left(1 - \lambda \ \sum_{z \in \tilde{\mathcal{S}}} q_D(x, z) e^{-(U(z) - U(x) - \gamma)^+ / T - \gamma / T} \right), \end{aligned}$$

where λ is again a positive parameter in the interval $0 < \lambda \leq 1$. A choice of $\lambda < 1$ avoids that destructions should always be chosen at high temperatures. Usually we will take $\lambda = 1/2$ or $\lambda = 1$. The positive part in $(U(y) - U(x) - \gamma)^+$ is needed only to cover the case where $x = y$.

The computer implementation of this Metropolis dynamic is the following: starting from the state x,

- first flip a coin with odds $\lambda e^{-\gamma/T}$ and $1 - \lambda e^{-\gamma/T}$ to decide whether or not to try a destruction.
- in case a destruction is tried,
  - choose a transition (x, y), drawing y according to the probability distribution $q_D(x, y)$.
  - then flip a second coin with odds $\exp -((U(y) - U(x) - \gamma)^+/T)$ and $1 - \exp -((U(y) - U(x) - \gamma)^+/T)$ to decide whether or not to apply this move.
- if the answer to one of the two previous tosses was *no*, then choose a transition (x, y), where y is chosen according to the distribution $q_C(x, y)$, and apply it.

The hypotheses we made about q_C, q_D, and U are what are needed to prove the following proposition.

PROPOSITION 3.2. *For any temperature $T > 0$, the matrix p_T is an irreducible Markov matrix.*

Considering the rate function $V : \tilde{\mathbb{S}} \times \tilde{\mathbb{S}} \to \mathbb{R}_+ \cup \{+\infty\}$ defined by

$$V(x, y) = \begin{cases} (U(y) - U(x))^+ & \text{if } q_D(x, y) + q_C(x, y) > 0 \text{ and } x \neq y, \\ +\infty & \text{otherwise,} \end{cases}$$

we see that there is a positive constant κ such that, whenever $x, y \in \tilde{\mathbb{S}}$, $x \neq y$,

$$\kappa\, e^{-V(x,y)/T} \leq p_T(x, y) \leq \frac{1}{\kappa}\, e^{-V(x,y)/T}.$$

Moreover V satisfies the weak reversibility condition of Hajek–Trouvé with respect to U. More precisely, if $\Gamma_{x,y}$ is the set of paths from x to y, we put for any $\gamma = (\gamma_1 = x, \ldots, \gamma_r = y) \in \Gamma_{x,y}$

$$H(\gamma) = \max_{i=1,\ldots,r-1} U(\gamma_i) + V(\gamma_i, \gamma_{i+1})$$

and

$$H(x, y) = \min_{\gamma \in \Gamma_{x,y}} H(\gamma).$$

The weak reversibility condition of Hajek–Trouvé states that for any $x, y \in \tilde{\mathbb{S}}$

$$H(x, y) = H(y, x).$$

Due to this reversibility property, U is a quasi-potential for p_T. We mean by this statement that the (unique) invariant probability measure μ_T of p_T satisfies for some positive constant α (independent of T) and for any $x \in \tilde{\mathbb{S}}$

$$\alpha \leq \mu_T(x) e^{(U(x) - \min U)/T} \leq 1/\alpha.$$

COROLLARY 3.1. *We can build optimization algorithms based on p_T following the results of Catoni* [7] *and Trouvé* [22]. *More precisely, for any fixed value of the temperature T, the homogeneous Markov chain with transition matrix p_T is a generalized Metropolis algorithm with quasi-potential function U. In the same way, for any decreasing sequence of temperatures $(T_n)_{n\in\mathbb{N}}$, the nonhomogeneous Markov chain $(X_n)_{n\in\mathbb{N}}$ on $\tilde{\mathcal{S}}$ with transitions*

$$P(X_n = y \ : \ X_{n-1} = x) = p_{T_n}(x, y)$$

is a generalized simulated annealing algorithm. Its behavior has been studied in [22] *and* [24] *and is very similar to the behavior of classical simulated annealing as studied in* [6].

We can also apply the iterated energy transformation method to p_T, which will be described in a further section of this paper and is studied in [7].

Proof. The only nonstraightforward point to check is the Hajek–Trouvé weak reversibility condition. Let us consider $x, y \in \tilde{\mathcal{S}}$, and $\gamma \in \Gamma_{x,y}$. We build a path from y to x in the following way. Replace any edge $(z,t) \in \gamma$ by the edge (t,z) if $q_C(z,t) > 0$ or $p_T(z,t) = 0$. If neither of the above two conditions is true, this means that $q_D(z,t) > 0$; then there is a path $\varphi \in \Gamma_{t,z}$ such that $q_C(u,v) > 0$ for any edge $(u,v) \in \varphi$, and we replace (z,t) by φ. The path φ is such that $H(\varphi) = U(z) + V(z,t) = U(t)$ because for any $(u,v) \in \varphi$, $U(v) < U(u)$ and $V(u,v) = 0$. Therefore by concatenating all these reversed edges and paths in reverse order we get a path $\psi \in \Gamma_{y,x}$ such that $H(\psi) = H(\varphi)$. Therefore $H(y,x) = H(x,y)$. □

Remarks.

- In the search space we consider, there is a natural starting point for optimization algorithms, which is the empty schedule $\emptyset$.
- In many scheduling problems, it is not known in advance whether a complete solution exists or whether one can possibly be found within the available computer time. Our method has the advantage of finding at least a partial solution, where some proportion of the tasks are scheduled in a coherent way. This is not the case if other constraints are relaxed, as is usually done. For instance, if the aim is to schedule the lectures at a university, a solution where some lectures share the same room at the same time has no practical interest, whereas a solution where some proportion of the lectures are scheduled in a coherent way can be applied.
- A slight variant of the present setup is the case where the search space satisfies condition (1), but one does not know whether it is possible to schedule all the tasks, and wants instead to schedule as many tasks as possible. In this situation the energy can weigh (through a positive measure) the relative importance of tasks.

In the three following sections, we are going to recall briefly some theoretical results about the speed of convergence of three optimization algorithms.

3.5. Rate of convergence of the Metropolis algorithm. In this section we consider the canonical process $(X_n)_{n\in\mathbb{N}}$ on the canonical space $(\tilde{\mathcal{S}}^{\mathbb{N}}, \mathcal{B})$, where $\mathcal{B}$ is the sigma field generated by the events depending on a finite number of coordinates.

For any temperature $T \in \mathbb{R}_+$, P_T will be the probability distribution on $(\tilde{\mathcal{S}}^{\mathbb{N}}, \mathcal{B})$ of a Markov chain with transition matrix p_T (where p_T is as in Proposition 3.2). Under this distribution, $(X_n)_{n\in\mathbb{N}}$ is a Metropolis algorithm and has the following convergence speed.

PROPOSITION 3.3. *There exists a positive constant d, depending only on the choice of the search space $\tilde{\mathcal{S}}$, of the constructive and destructive dynamics q_C and q_D and of the parameter λ, $0 < \lambda < 1$, such that for any energy function U satisfying the hypothesis* (2) *of section* 3.3, *for any positive constant η,*

$$\max_{x\in\tilde{\mathcal{S}}} P_T(U(X_N) \geq U_{\min} + \eta \,|\, X_0 = x) \leq d\left(\exp - \left(\frac{N}{d} e^{-H_1/T}\right) + e^{-\eta/T}\right),$$

where $H_1(V)$ is the first critical depth of the rate function V defined in Proposition 3.2. *The exponent $H_1(V)/T$ is optimal when η is small and when T tends to* 0 *and N tends to $+\infty$. With the notations of Proposition* 3.2,

$$H_1(V) = \max_{x\notin\mathcal{S}} \min_{y\in\mathcal{S}} H(x,y) - U(x).$$

As a consequence, considering $1/T = (1/H_1)\log(N\, H_1/d\,\eta\, \log N)$, we see that there is a constant d (independent of U and η), such that

$$\inf_{T\in\mathbb{R}_+} \max_{x\in\tilde{\mathcal{S}}} P_T(U(X_N) \geq U_{\min} + \eta \,|\, X_0 = x) \leq d\left(\frac{d\,\eta}{H_1(V)}\,\frac{\log N}{N}\right)^{\eta/H_1(V)}.$$

Moreover the exponent $\eta/H_1(V)$ is optimal for small enough values of $\eta \in (U(E) - U_{\min})$.

For a proof see, for instance, Cot and Catoni [9].

This proposition does not give a quantitative upper bound for the probability of failure for N iterations, since it does not give an estimate for constant d; nevertheless, the fact that this constant is independent of U allows us to compare the probability of failure for different energy functions U for a large *finite* number of iterations N and not only when N tends to infinity. More precisely, it shows that the convergence speed of the Metropolis algorithm is slow when there are states with energies close to $U_{\min}$. Indeed if one wants to study the convergence to $\mathcal{S}$, one has to choose

$$\eta = \min\{U(x) - U_{\min}\,,\, x \in \tilde{\mathcal{S}} \setminus \mathcal{S}\}.$$

If η is small, then it will reflect on the exponent $\eta/H_1(V)$.

This is a theoretical justification for the introduction of simulated annealing, which will not suffer from this drawback, when proper robust cooling schedules are used.

3.6. Rate of convergence of simulated annealing. We consider now a non-increasing triangular sequence $T_1^N \geq T_2^N \geq \cdots \geq T_N^N$ of temperatures and the measure $P_{(T_1^N,\ldots,T_N^N)}$ on $\tilde{\mathcal{S}}^N$ of the nonhomogeneous Markov chain with transitions

$$P_{(T_1^N,\cdots,T_N^N)}(X_n = y \,|\, X_{n-1} = x) = p_{T_n^N}(x,y).$$

The rate of convergence of such an algorithm has been studied in [6], [8], and [22] (see also [24] in English, translated from [22]). We give here a simple result; for more precise estimates, we refer to the original papers.

PROPOSITION 3.4 ([6], [22], [24]). *There is a positive constant K such that*

$$K^{-1}N^{-D^{-1}} \leq \inf_{T_1^N \geq \cdots \geq T_N^N} \max_{x \in \tilde{\mathcal{S}}} P_{(T_1^N,\ldots,T_N^N)}(U(X_N) > U_{\min} \mid X_0 = x) \leq KN^{-D^{-1}},$$

where the constant $D = D(V)$ is the difficulty of the rate function V. With the notations of Proposition 3.2, *the definition of $D(V)$ is*

$$D(V) = \max_{x \in \tilde{\mathcal{S}} \setminus \mathcal{S}} \min_{y \in \mathcal{S}} \frac{H(x,y) - U(x)}{U(x) - \min U}.$$

For any $A > 0$, there is a positive constant K such that the triangular exponential schedule

$$T_n^N = \frac{1}{A}\left(\frac{A}{(\log N)^2}\right)^{n/N}$$

gives a convergence speed of

$$\max_{x \in \tilde{\mathcal{S}}} P_{T^N}(U(X_N) > U_{\min} \mid X_0 = x) \leq K\left(\frac{\log N \, \log\log N}{N}\right)^{D(V)^{-1}},$$

for N large enough.

In the case of simulated annealing, we have a probability of failure for N iterations of order $\epsilon \asymp_{\log} (1/N)^{1/D}$ (meaning that the logarithms on both sides of this equation are equivalent when N tends to infinity). The important feature of this theoretical result is that the exponent $1/D$ is independent of the precision with which we want to reach $U_{\min}$ but depends, on the contrary, only on the structure of the local minima of U.

One other interesting point is that the exponential triangular cooling schedule $T_n^N = A^{-1}(A/(\log N)^2)^{n/N}$ is robust: it gives a convergence rate with the optimal exponent $1/D(V)$ for any energy function U.

3.7. Rate of convergence of the energy transformation method. We introduced in [7] the iterated energy transformation method as another mean to discourage uphill moves from low energy states more than from high energy states. In simulated annealing this effect is produced by an exogenous control of the temperature parameter: in "typical" successful runs of simulated annealing, the energy of the current state is moving downwards on the average, and at the same time uphill moves are more and more discouraged. In the iterated energy transformation method, a temporary hypothesis is made about the value of $U_{\min}$, and a concave transformation is applied to U on the basis of this hypothesis. Then the algorithm is run at constant temperature using the transformed energy. This produces the desired effect of discouraging more uphill moves from low energy states. Of course, in the beginning, the hypothesis about $U_{\min}$ is necessarily grossly underestimated, so that the energy transform is not very efficient, but after some iterations, it can be improved (this will work with a probability close to one) depending on the values of the energies of the explored states.

The convergence of the lower bound estimate for $U_{\min}$ towards the true value of $U_{\min}$ is exponentially fast (with a probability close to one), and therefore the energy transformation is quickly tuned to an efficient value.

The iterated energy transformation method applied to our problem is described as follows. For any strictly concave, strictly increasing energy transformation $F : \mathbb{R} \longrightarrow \mathbb{R} \cup \{-\infty\}$, we consider the Markov matrix

$$p_F(x,y) = \mathbf{1}_{(F(x)>-\infty)} \Bigg\{ \lambda e^{(F(U(x))-F(U(x)+\gamma))} q_D(x,y) e^{(F(U(x)+\gamma)-F(U(y)))^-} + q_C(x,y) \left(1 - \lambda \sum_z q_D(x,z) e^{F(U(x))-F(U(x)+\gamma)+(F(U(x)+\gamma)-F(U(z)))^-} \right) \Bigg\} + \mathbf{1}_{(F(x)=-\infty)} \mathbf{1}_{(x=y)}.$$

Consider for any positive constant α, any real shift τ, and any positive temperature T, the transformation

$$F_{\alpha,T,\tau}(U) = \begin{cases} \alpha\, U + \dfrac{1}{T} \log(U+\tau) & \text{if } U+\tau > 0, \\ -\infty & \text{otherwise .} \end{cases}$$

Let us introduce the simplified notation $\overline{p}_{\alpha,T,\tau} = p_{F_{\alpha,T,\tau}}$.

Given parameters $M \in \mathbb{N}$ (number of iterations performed with each energy transform), two real numbers $\rho > 0$ and $\eta_0 \geq 0$ (two parameters for the update of the shift τ), and an initial lower bound $\delta < U_{\min}$, we consider the canonical process $(X_n)_{n\in\mathbb{N}}$ on $\tilde{\mathcal{S}}^{\mathbb{N}}$ with probability distribution $P_{\alpha,T,M,\rho,\eta_0}$ defined by the following conditional distributions:

$$\begin{aligned} P_{\alpha,T,M,\rho,\eta_0}(X_n = y \,|\, (X_0,\dots,X_{n-1}) = (x_0,\dots,x_{n-1})) \\ = \overline{p}_{\alpha,T,\tau_n(x_0,\dots,x_{n-1})}(x_{n-1}, y), \end{aligned}$$

with

$$\begin{cases} \tau_r &= \eta_0 - \delta, \quad 0 < r \leq M, \\ \tau_{kM+r} &= \tau_{k\,M} - \dfrac{1}{1+\rho} \left(\min\limits_{n, n\leq k\,M} U(X_n) + \tau_{k\,M} \right) + \eta_0, \quad 0 < k, 0 < r \leq M. \end{cases}$$

We have proved in [7] the following theorem.

THEOREM 3.1 (Catoni). *For any fixed $\alpha > 0$, the family of processes described above satisfies for some positive constants B and K, for any choice of $r \in \mathbb{N}$, $\eta_0 \geq 0$, $\rho > 0$,*

$$\max_{x\in\tilde{\mathcal{S}}} P_\theta(U(X_{r\,M}) \geq \eta \,|\, X_0 = x) \leq \epsilon$$

where $\theta = (\alpha, T, M, \rho, \eta_0)$,

$$\begin{aligned} T &= \frac{\log(1+\rho)}{\log(Kr/\epsilon)} \\ M &= B \left(\frac{\epsilon}{K\,r} \right)^{-\log(1+\tilde{D}_{\eta_0})/\log(1+\rho)} \log \frac{K\,r}{\epsilon}, \\ &= \frac{B}{T} \log(1+\rho) \left(1 + \tilde{D}_{\eta_0} \right)^{1/T}, \\ \eta &= U_{\min} + \rho \left(\frac{\rho}{1+\rho} \right)^{r-1} (U_{\min} - \delta + \eta_0) + \eta_0 \rho (1+\rho), \end{aligned}$$

and where the constant $\tilde{D}_{\eta_0}(V)$ *is given by*

$$\tilde{D}_{\eta_0}(V) = \max_{x \in \tilde{\mathcal{S}} \setminus \mathcal{S}} \min_{y \in \mathcal{S}} \frac{H(x,y) - U(x)}{U(x) - U_{\min} + \eta_0} < D(V).$$

COROLLARY 3.2.

$$\limsup_{N \to +\infty} (\log N)^{-2} \log \inf_{T,M,\eta_0,\rho} P(U(X_N) > U_{\min} \mid X_0 = x) \leq -\frac{1}{4\ \log(1+D)}.$$

The interest of this theorem lies mainly in its corollary, which shows that a proper tuning of the parameters leads to a faster scale of convergence speed than the one achieved by simulated annealing (see [7]). This remark of course deals with the comparison of two long runs of both algorithms. For repeated trials of bounded length, which we will consider in section 6.3, the question of knowing which algorithm is faster is open.

We will discuss practical means of choosing the parameters in connection with the jigsaw puzzle benchmark.

4. Solving jigsaw puzzles. We will illustrate on jigsaw puzzles the different steps of the general method of resolution.

First of all, we have to choose a search space. This will be the set of partial solutions where only some of the pieces are put in the frame.

$$\tilde{\mathcal{S}} = \{x \subset B \times E : x(r) = [a_r, a_r + w_r[\times[b_r, b_r + h_r[\,, \\ r \in \pi_B(x), x(r) \cap x(r') = \emptyset, r \neq r' \in \pi_B(x)\}.$$

Let us define now $q_C(x,.)$, the constructive dynamic starting from state x:

- First choose $r \in B \setminus \pi_B(x)$ according to the uniform distribution on this set.
- Then choose $(z,t) \in E \setminus \pi_E(x)$ according to the uniform distribution.
- Then try to expend this germ to a rectangle $[a_r, a_r + w_r[\times[b_r, b_r + h_r[$ of the desired size by adding alternatively a column to the left (or else to the right) and a line to the top (or else to the bottom). If it is not possible to grow the germ to its final size, just abandon the construction.
- Then draw a number k at random in the interval $[0, \text{max\_drift}[$ and move the location of $[a_r, b_r[$ k steps along the direction $(-1,-1)$ (that is, to the upper left corner, according to usual image indexing) if there is enough room to do so, or else move it as far as possible in this direction (until it bumps into other pieces).

The last two actions are better described by the following self-explanatory pseudo-C code, where $[a, c\,[\times[b, d\,[$ is the current germ:

```
int expend() {
  a=z; c=z+1; b=t; d=t+1;
  while((test1=(c-a<w))||(test2=(d-b<h))) {
  if (test1&&grow_left()&&grow_right()) return 1;
  if (test2&&grow_up()&&grow_down()) return 1;
  }
  for (k=rand(0,max_drift);k;k--) {
  if (move_left()&move_up()) break;
  } return 0;
}
```

where the functions `expend()`, `grow_left()`, `grow_right()`, `grow_up()`, `grow_down()`, `move_left()`, and `move_up()` return 0 on success and 1 on failure.

The destructive dynamic q_D is simpler:

- Draw $r \in \pi_B(x)$ at random,
- Form $y = x \cap \pi_B^{-1}(B \setminus \{r\})$, the partial solution where the piece labeled "r" has been removed from the frame.

The mechanism that was chosen for the constructions is meant to discourage the formation of small gaps between pieces. If nothing were done, when the discretization step of the grid is fine, small gaps would be left between the pieces with a large probability, and a complete solution to the puzzle, where pieces necessarily stick together, would never be discovered.

We have now to choose an energy function. Here again we will discourage the formation of gaps between pieces by introducing a term proportional to the contact length. By contact length we mean the sum of the contact lengths between pieces and between pieces and the edge of the frame.

Let μ be the counting measure on E. We take

$$U(x) = -\mu(\pi_E(x)) - \alpha \times \text{contact-length}.$$

For this choice of U, we can take the constant γ in equation (2) to be equal to the size of the smallest piece:

$$\gamma = \min_{r \in B} w_r \, h_r.$$

5. Minimizing a loss function.

5.1. Statement of the problem. We will discuss in the next two sections the case where some loss function $V : \mathcal{S} \longrightarrow \mathbb{R}$ has to be minimized on the state space $\mathcal{S}$ of global solutions of a task assignment problem. We consider the same framework as in the first section, with the difference that the problem is now to find a solution x belonging to $\arg\min_{y \in \mathcal{S}} V(y)$.

5.2. A general method of resolution. We will extend the method of section 3 to deal with a loss function.

The two first steps, building the search space and the constructive and destructive dynamics, will be the same as in section 3.

The change comes from the choice of the energy function. First we need to extend the loss function V to the search space $\tilde{\mathcal{S}}$ of partial solutions. Ideally, we would like to use the extension $\overline{V} : \tilde{\mathcal{S}} \longrightarrow \mathbb{R}$ defined by

$$\overline{V}(x) = \min\{V(y) \,:\, y \in \mathcal{S}, x \subset y\}.$$

Usually this is not an easily computable function, but in many situations there is a natural way to define a loss function for partial solutions. A simple way to do so, if there is nothing else at hand, is to set $V(y) = c$ for $y \in \tilde{\mathcal{S}} \setminus \mathcal{S}$, where c is a constant and $c \geq \max_{x \in \mathcal{S}} V(x)$. Then we build a compound energy function

$$W(x) = \alpha \, U(x) + V(x), \qquad x \in \tilde{\mathcal{S}},$$

where the real positive coefficient α is chosen such that, for some positive constant γ,

$$\begin{cases} \arg\min\limits_{x \in \tilde{\mathcal{S}}} W(x) \subset \mathcal{S}, \\ W(y) - W(x) \leq -\gamma < 0, \quad x \subset y, \ x \neq y \in \tilde{\mathcal{S}}. \end{cases}$$

These conditions are always satisfied for α large enough. However, the difficulty D of the energy landscape, related to the performance of simulated annealing, tends to $+\infty$ when α tends to $+\infty$. Therefore it is better to keep α as small as possible. In the next section, we will give an example for which we can take α arbitrarily small, and even $\alpha = 0$ if we are satisfied with $\gamma = 0$.

Equipped with this new energy function, we can proceed just as in the simpler case of section 3.

5.3. Some example of useful loss function. Often in task assignment problems, we would like some resources to be distributed according to some prescribed distribution. For instance, in a time-table problem we may want to schedule an equal number of hours in each week of the year.

This can be formalized in the following way. We consider first some function $\Phi : B \times E \longrightarrow F$, where F is a finite set (which may be the discretization of a domain in $\mathbb{R}^n$). Typically, Φ will be the projection on the time axis in a time-table problem. Then we consider a target distribution ρ defined on F. Let us consider some reference measure μ on $B \times E$ (such as the counting measure). To each partial solution $x \subset B \times E$, we may associate the restriction μ_x of μ to x, defined by

$$\mu_x(A) = \mu(x \cap A).$$

This induces a measure $\mu_x \circ \Phi^{-1}$ on F. The constraint we would like to represent by a loss function is that $\mu_x \circ \Phi^{-1}$ is approximately proportional to the reference measure ρ. This can be reflected in a loss function of the type

$$V(x) = \int h\left(\frac{\mu_x \circ \Phi^{-1}}{\rho}\right) d\rho,$$

where $h(x) = (1-x)^2$ or $h(x) = 1-x+x\log x$. The function h is in both cases strictly convex, satisfies $h(1) = h'(1) = 0$, and h' is strictly increasing; therefore $\mu_x \circ \Phi^{-1} = \rho$ if and only if $V(x) = 0$ and the minimum of $V(x)$ on the set $\mu_x(B \times E) = \textit{constant}$ is attained when $\mu_x \circ \Phi^{-1}$ is proportional to ρ, when this is feasible.

The following proposition holds.

PROPOSITION 5.1. *Assume that the total weight $\mu_x(B \times E)$ of any solution is a function of the tasks to be scheduled only. This means that there is a measure $\tilde{\mu}$ on B such that*

$$\mu_x(B \times E) = \mu(x) = \tilde{\mu}(\pi_B(x)), \quad x \in \tilde{\mathcal{S}}.$$

Then for all global solutions $x \in \mathcal{S}$, $\mu_x(B \times E) = \mu(x) = \tilde{\mu}(B)$ is a constant.

Assume moreover that the measure ρ defining the constraint is such that $\rho(F) \geq \tilde{\mu}(B)$, and assume also that

$$\left\{ x \in \mathcal{S} \,:\, \frac{\mu_x \circ \Phi^{-1}}{\rho} \equiv \textit{constant} \right\} \neq \emptyset.$$

Then

$$\arg\min_{s \in \tilde{\mathcal{S}}} V(x) = \left\{ s \in \mathcal{S} \,:\, \frac{\mu_x \circ \Phi^{-1}}{\rho} \equiv \textit{constant} \right\},$$

meaning that the partial solutions minimizing V are exactly the global solutions x for which $\mu_x \circ \Phi^{-1}$ is proportional to the constraint ρ.

The assumptions of the proposition will be satisfied when $\mu_x(B\times E)$ measures the amount of assigned resources, and the amount of resources to be allocated to a task depends only on the task and not on the way it is scheduled. Typically, for instance, the number of hours of a course of teaching will be prescribed in advance and will not depend on the choice of a schedule for the lectures.

Now let us make the supplementary assumption that $(\mu_x \circ \Phi^{-1}/\rho) \leq 1$ for any $x \in \tilde{\mathcal{S}}$. We can always make this assumption true by increasing ρ by a suitable multiplicative factor (at least when ρ is strictly positive on F). In some cases we may, on the contrary, want to restrict $\tilde{\mathcal{S}}$ by adding the new constraint $(\mu_x \circ \Phi^{-1}/\rho) \leq 1$. This will be done when the constraint has a practical meaning for the problem. For instance, if $\mu_x \circ \Phi^{-1}$ measures the number of lectures taking place in each hour of time in the week, we may want to fix ρ to a constant equal to the total number of available lecture rooms, add the constraint $(\mu_x \circ \Phi^{-1}/\rho) \leq 1$ to indicate that there is to be enough rooms to schedule all the lectures, and use the loss function $\int h(\mu_x \circ \Phi^{-1}/\rho)d\rho$ to indicate that we would like the rooms to be evenly occupied during the week (in a weekly time-table problem).

If the assumption $(\mu_x \circ \Phi^{-1}/\rho) \leq 1$, $x \in \tilde{\mathcal{S}}$ holds, then only the decreasing part of h is used, and the loss function V is always increasing during a destruction and decreasing during a construction. Therefore if γ is the constant corresponding to U in Eq. (2), we will have

$$W(y) \geq W(x) + \alpha\,\gamma, \quad x, y \in \tilde{\mathcal{S}}, \quad x \neq y, q_D(x,y) > 0.$$

6. The practical issue of the choice of parameters. In practical situations, the critical constants of the energy landscape are usually unknown. Therefore it is not possible to rely on the theoretical results we recalled in preceding sections to choose the parameters of algorithms. In the following subsections, we explain how we set the parameters in the experiments about jigsaw puzzles.

6.1. Simulated annealing. The cooling schedule can be written as

$$\frac{1}{T_n^N} = \beta_{\min}\left(\frac{\beta_{\max}}{\beta_{\min}}\right)^{n/N}.$$

We choose $\beta_{\min}$ and $\beta_{\max}$ by looking at the repartition function of the energies of the explored states in simulations at constant temperatures. We keep a value of $\beta_{\min}$ for which the slope of the repartition function stays large up to the largest values of the energies, meaning that states with high energies have a significant probability to be explored. For $\beta_{\max}$ we require, on the contrary, a repartition function concentrated on the lowest energy values.

The theory tells us that we can safely underestimate $\beta_{\min}$ and overestimate $\beta_{\max}$, which makes their choice possible from a qualitative inspection of repartition functions.

Figures 6.1 and 6.2 are two examples of repartition functions, corresponding to values of $\beta_{\min}$ and $\beta_{\max}$ which have been retained during the experiments.

6.2. The iterated energy transformation method. In this case, the choice of parameters is perhaps less straightforward. The analogy with simulated annealing can serve as a guideline: the high temperature regime corresponds to the case $\tau = \tau_1$ (i.e., to the first energy transform used). The low temperature regime corresponds to

$$\tau = (1+\rho)\eta_0 - U_{\min}.$$

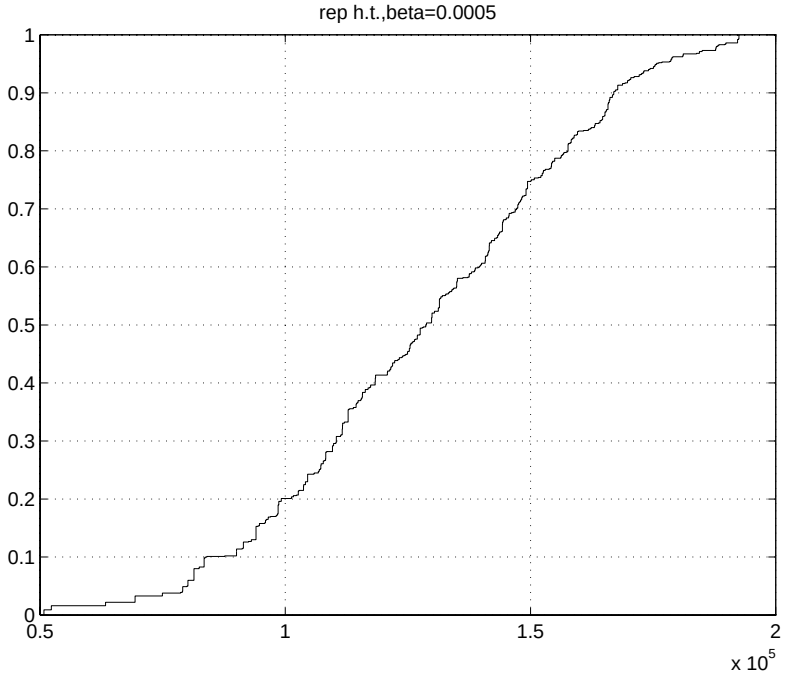

rep h.t.,beta=0.0005
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.5
1
1.5
2
x 10^5

FIG. 6.1.

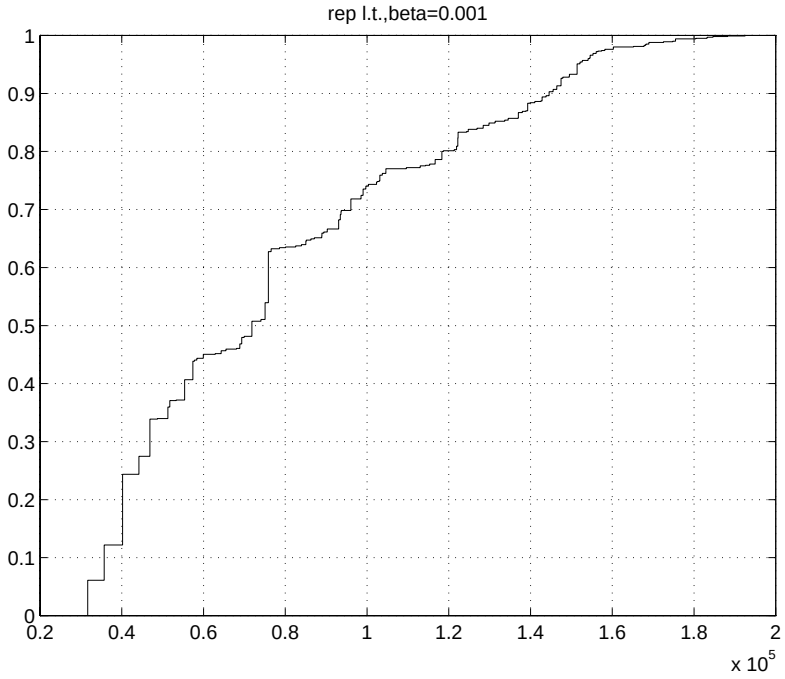

rep l.t.,beta=0.001
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
x 10^5

FIG. 6.2.

In order to test the behavior of the algorithm in these two configurations, we make a short test using a small value of ρ ($\rho << 1$). The law of evolution of τ_k shows that, for a small value of ρ, the algorithm will quickly switch from the high temperature regime during the first step to a low temperature regime during the following steps. In fact the value $\rho = 0$ may sometimes even be used. However, when this is done the algorithm sometimes encounters a state with a nondefined energy transform too quickly, and there are not enough iterations to compute a reliable repartition function for the low temperature regime. This problem, when encountered, can be circumvented by using a low but nonzero value of ρ.

We compute the repartition function of energies during the first step of the test run and during the last. The first function describes the equivalent of the "high temperature regime" and is tuned by the choice of the constant α and of the temperature parameter T; the second function corresponds to the "low temperature regime" and is tuned by a proper choice of $\tilde{\eta_0} = \eta_0(1+\rho)$.

Once these two choices are made, there remains a free parameter, namely, ρ.

The theory [7] indicates an optimal choice of ρ of order $\sqrt{N}$ and an optimal choice of $r = N/M$ of order $\sqrt{N}\log(N)$. On the other hand, as soon as $\log(1+\rho) > 1$, the convergence rate will be better than for simulated annealing. This indicates that a large value of ρ may safely be chosen and that r can then be set to make

$$(U_{\min} - \gamma + \eta_0)\rho\left(\frac{\rho}{1+\rho}\right)^{(r-1)}$$

small. This will ensure a small dependence of the final value of the shift τ_N with respect to its initial value $\tau_1 = \delta - \eta_0$.

6.3. Repeated optimizations. In this section, we will consider that N iterations are to be divided into N/M trials of length M, and that we will keep the best solution found out of these N/M trials. In this context, the probability of failure in the worst case with respect to the starting point of each trial is $\epsilon_1(M)^{N/M}$, where

$$\epsilon_1(M) = \max_{x\in\tilde{S}} P\left(U(X_N) > U_{\min} \mid X_0 = x\right).$$

The first remark to be made (see Azencott [2], [3]) is that for all the algorithms we have considered, $\lim_{M\to+\infty}(1/M)\log\epsilon_1(M) = 0$. Therefore when N is large enough, the optimal value for M is independent of N.

We will discuss here the choice of the length M of each run of the algorithm. For simulated annealing, we can, on the basis of the theoretical bound on the probability of failure, namely, $(A/M^\alpha)^{N/M}$ for N iterations divided into N/M runs of length M, conjecture that an overestimation of M will be relatively harmless, whereas an underestimation would be more penalizing. This can be seen on the derivative

$$\frac{\partial}{\partial M}\left(\frac{A}{M^\alpha}\right)^{N/M} = \left(\frac{A}{M^\alpha}\right)^{N/M}\frac{N\left(\alpha(\log M - 1) - \log A\right)}{M^2},$$

but is may be more vividly illustrated by a small numerical application. If we take, for example, $A = e^4$, $\alpha = 1$, and $N = 1000$, and if we put $\epsilon(M) = (A/M^\alpha)^{N/M}$, we see that

$$\min_M \epsilon(M) = \epsilon(148) \simeq 0.0012,$$

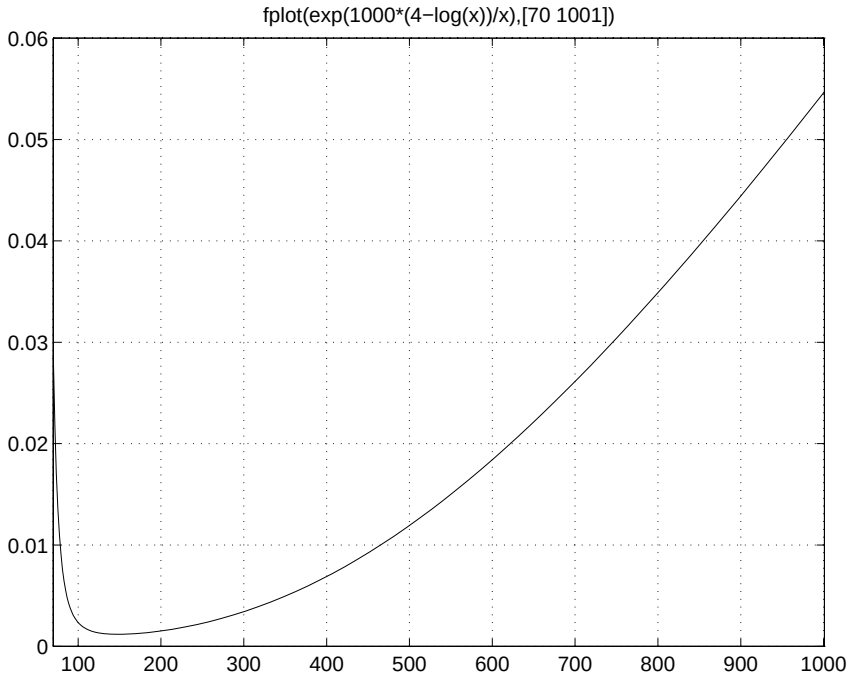


FIG. 6.3.

and that for this optimal value the probability of failure in each run is $\simeq 0.37$.

Here are some values taken by $\epsilon(M)$

M	74	100	148	300	500	1000
$\epsilon(M)$	0.016	0.0024	0.0012	0.0034	0.012	0.055

and a graphic of this function is shown in Fig. 6.3.

These figures show that, as far as this rough theoretical bound is a good guideline, there is a clear benefit in performing multiple runs instead of one long run, but that an overestimation of a factor two of the length of each run is relatively harmless. We remark also that a quite low confidence level for each run is favorable in this example where the difficulty is one.

The same kind of reasoning would also hold for the theoretical bound of order $\epsilon(M) = (A/M^{\alpha \log M})^{N/M}$ obtained for the iterated energy transformation method. In this case the derivative of the confidence level $\epsilon(M)$ is

$$\frac{\partial}{\partial M}\epsilon(M) = \epsilon(M)\frac{N\left(\alpha((\log M)^2 - 2\log M) - \log A\right)}{M^2}.$$

Figure 6.4 shows a plot of this function for some choice of the parameters α, N, and A: the tolerance with respect to an overestimation of M is even better than for simulated annealing.

For a comparison between repeated searches and interacting parallel searches, we refer to Graffigne [16] and Azencott and Graffigne [4].

6.4. A partial freezing method. In [7] we saw that the simulated annealing algorithm is not efficient to deal with a state space made of a large number of independent components. By "independent components," we mean the case when the energy

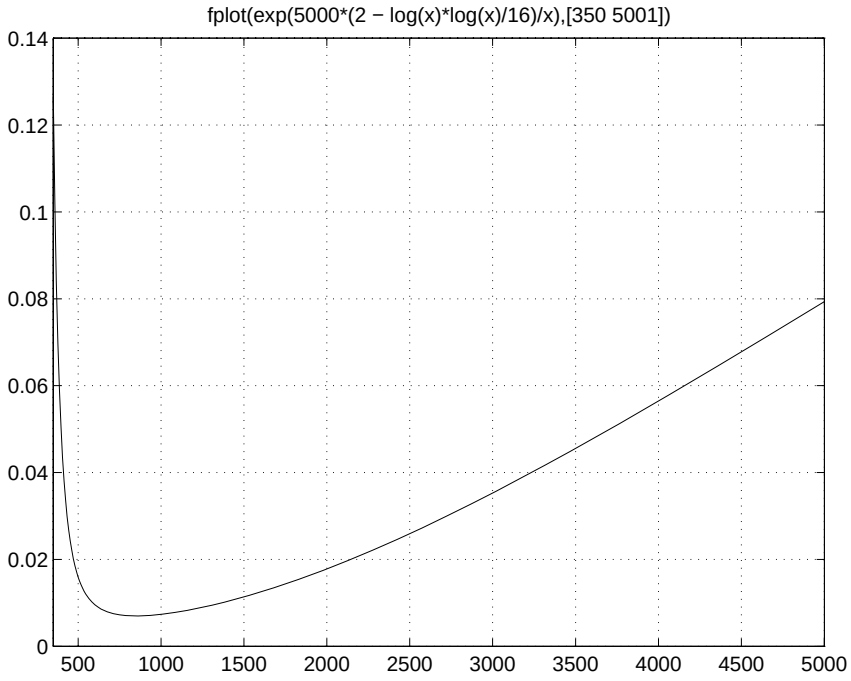


FIG. 6.4.

function is a sum of terms depending on distinct coordinates. Here the different tasks to be scheduled interact through the constraints and through the contact length term in the energy; therefore we cannot describe their assignment by independent coordinates. Nevertheless, in the end of the optimization process, we would like to be able to perform some last small local improvements on the current solution depending on distinct small subsets of tasks, with the assignment of the other tasks remaining untouched. We can write (formally) the energy function in a suitable neighborhood of the current solution as a sum of such possible local improvements. In the end of the optimization, we will be working at low temperature; therefore the current solution will, most of the time, be a local minimum and we will typically not try more than one local improvement at a time. Therefore (this is only a heuristic reasoning) we can expect the optimization process to behave approximately the same as in the independent component case when it draws to the end (that is, at low temperatures). It is easy to show (see [7]) that an efficient way to deal with independent components is to perform a series of local optimizations, resetting after each step the current solution to the best solution found. This will do much better than the global algorithms we described so far when the number of components is large. The reason is that a global algorithm cannot efficiently move one task around without disturbing the others. These considerations suggest adding a postprocessing to global optimization, made up of a series of local optimizations. When we put these things and the use of repeated optimizations together, we end up with a partial freezing method, which can be symbolically described by the following nested loops:

```
repeat
  reset the current solution to the empty assignment
  for (n taking increasing values from 0 to max)
    repeat
      choose at random a set of n ‘‘frozen’’ tasks among
      the scheduled tasks
      repeat
        run a stochastic optimization algorithm during
        which the frozen tasks stay untouched
      endrepeat
      reset the current solution to the best solution
      encountered in the previous loop
    endrepeat
  endfor
endrepeat
return the best solution encountered in the outer loop.
```

The stochastic algorithm used in the inner loop may be one of the three algorithms we studied here. It is applied to the subset of the state space defined by the current assignment of the frozen tasks. When the number of currently scheduled tasks is less than n, we freeze all the scheduled tasks. In the experimental section we will show results obtained with this partial freezing method applied to the iterated energy transformation algorithm. One advantage of the partial freezing method is that it is less demanding on the global optimization step and is therefore tolerant of a looser choice of the parameters of the algorithm.

7. Experimental results. We tried to solve two kinds of puzzles: a small "tight" puzzle with nine pieces and no loss function, and a big "loose" 60-piece puzzle with a loss function. By "tight" we mean that there is just enough room in the frame to put all the pieces, and by "loose" we mean, on the contrary, that there is some extra room left in the frame, the difficulty being then to minimize the loss function.

7.1. Small "tight" jigsaw puzzle. Our small jigsaw puzzle is a nine-piece problem. The algorithm we used to solve it corresponds to the description given in section 4. The frame is a 40×50 grid. The size of the pieces are $(14, 27)$, $(8, 36)$, $(8, 9)$, $(6, 14)$, $(34, 5)$, $(18, 9)$, $(22, 9)$, $(18, 21)$, $(18, 15)$. The problem has several solutions, due to symmetry properties. Figure 7.1 shows one solution: the parameters of the algorithms were set using the heuristics described in section 6.

We performed 40 runs of the simulated annealing algorithm and the same number of runs of the IET algorithm. For each algorithm, we computed the repartition function of the energy of the best solution encountered during each run and computed the repartition function of the energy of the final state of the algorithm. Of course, the former repartition function is always above the latter; therefore we can unambiguously plot them on the same diagram. In order to perform a "fair" comparison, we allowed the same number of iterations in both cases, namely, $N = 5000$ iterations per run.

The results (Figs. 7.2 and 7.3) are of the same order, with some advantage in favor of the IET method. This is especially true when the energy of the final state is considered. An interpretation of this fact is that the IET algorithm is more efficient in preventing the process from leaving the global minimum once it has reached it.

We were also able to check the influence of the drift towards the upper left corner.

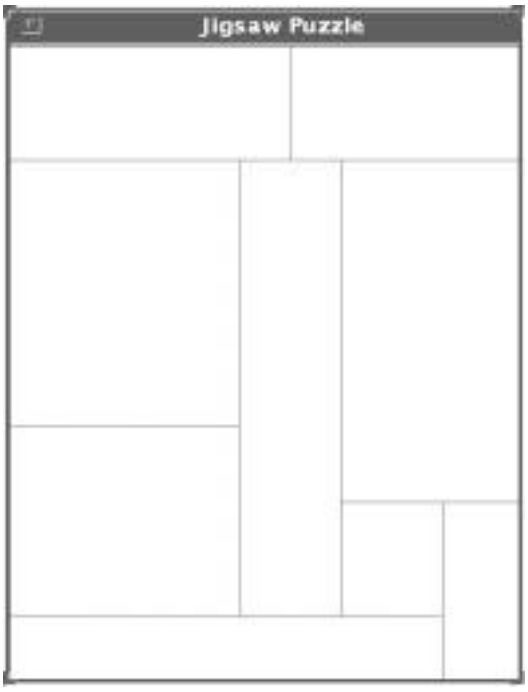


FIG. 7.1.

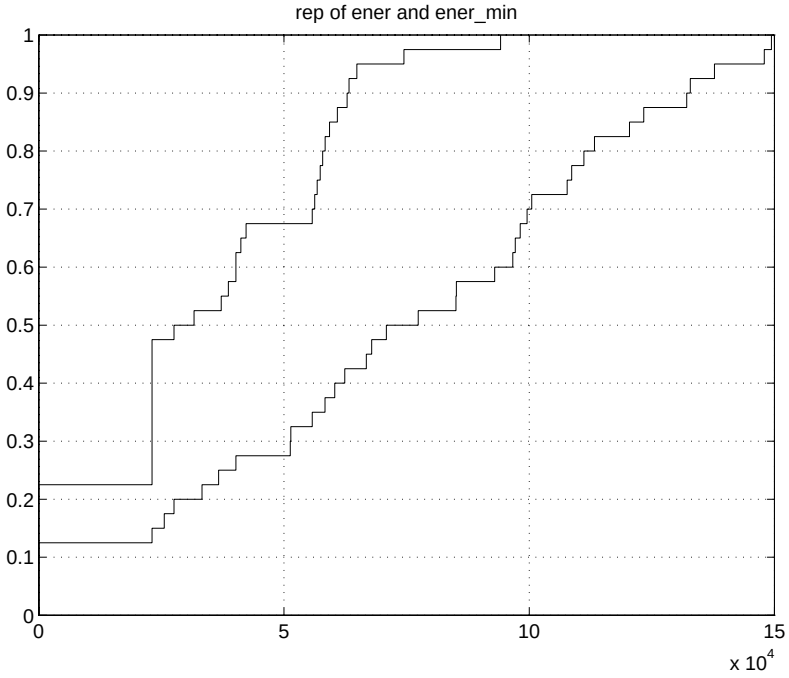


FIG. 7.2. *Performance of simulated annealing.*

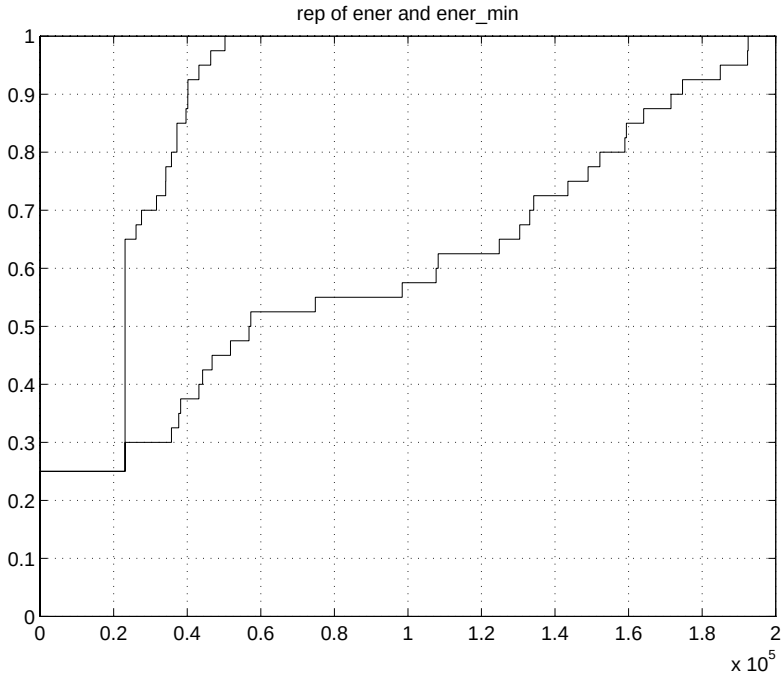


FIG. 7.3. *Performance of the iterated energy transformation method.*

In the two previous experiments, the maximum number of steps of the drift (the constant `max_drift` in the pseudocode of section 4) is 10. We have also tried a maximum number of steps of 50, for simulated annealing. We obtained on 40 runs the improvement in the performance shown in Fig. 7.4.

7.2. A big "loose" jigsaw puzzle. Our big jigsaw puzzle has 60 pieces, covering an area of 230 unit squares. The frame is a grid of size 30×10. The sizes of the pieces are the following:

number of pieces	width	height
15	3	2
15	2	1
5	5	2
5	2	4
20	1	1

The loss function is of the type described in the previous section. The function Φ here is the projection on the second axis, $\Phi((r, a, b)) = b$, $(r, a, b) \in B \times E$, so that the constraint indicates how much of each line the pieces should fill. On the following diagram, we have plotted the constraint function ρ (see Fig. 7.5).

With this choice of ρ, the constraint is tight, meaning that $\rho(F)$ is equal to the area of the pieces. When we use tight constraints, we build problems of the partition type, which are therefore NP complete. We chose the size of the pieces such that the set of global solutions is not empty. However, for a 60-piece problem, it is very difficult to find a (complete) solution.

We have chosen a coarse discretization step to keep the difficulty of the problem

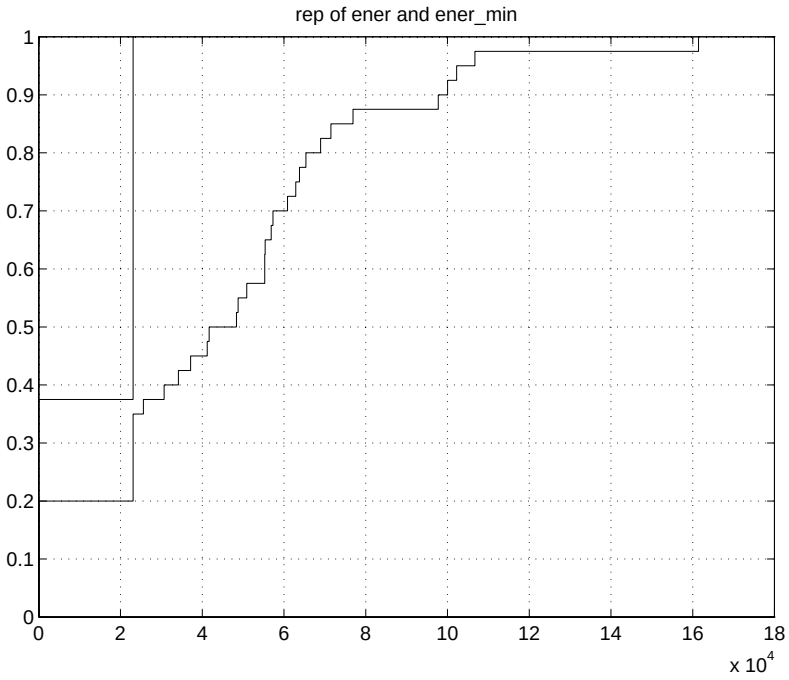

rep of ener and ener_min
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
2
4
6
8
10
12
14
16
18
x 10^4

FIG. 7.4.

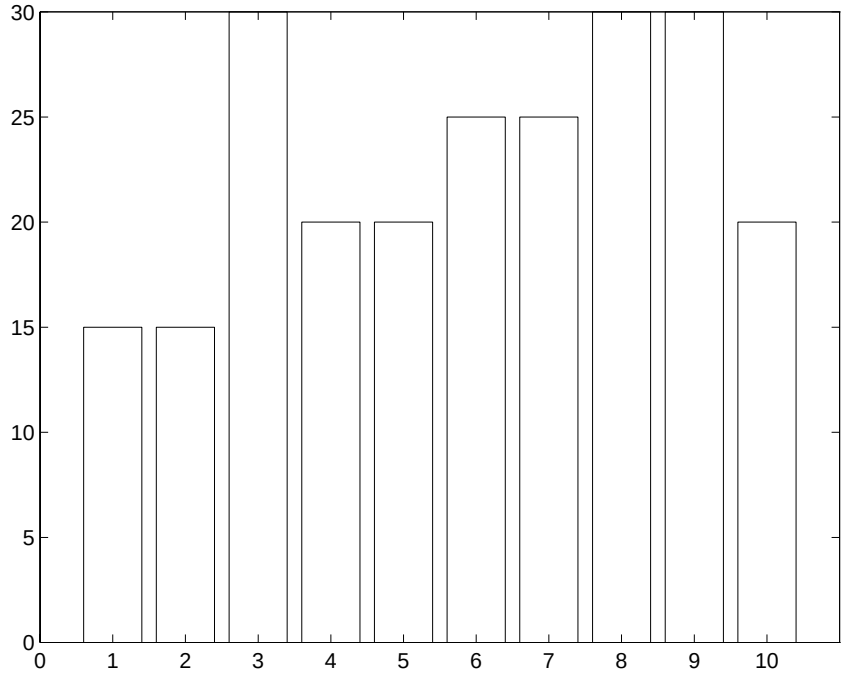

30
25
20
15
10
5
0
0
1
2
3
4
5
6
7
8
9
10

FIG. 7.5.

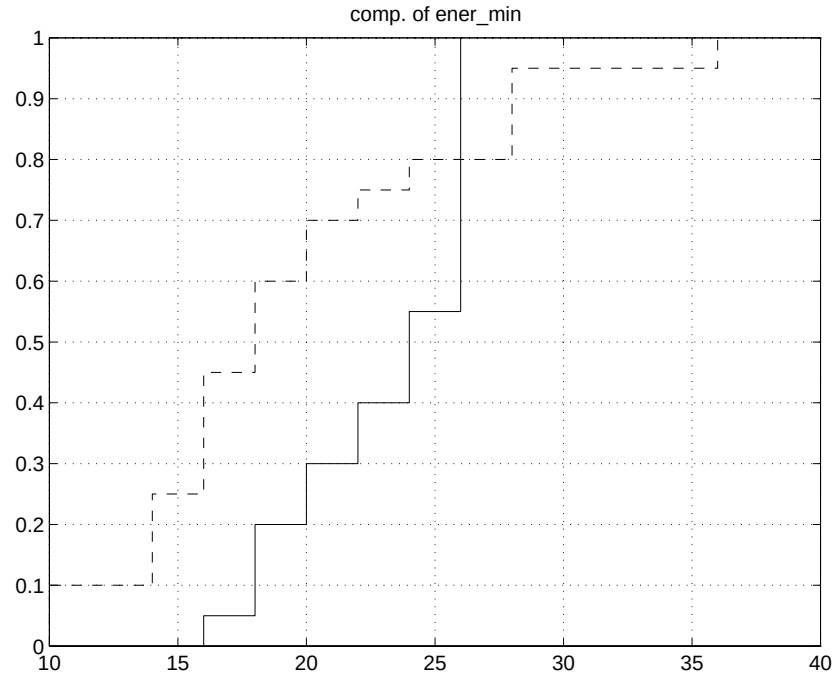


FIG. 7.6.

to a reasonable level, since we had to switch off the vertical drift. Indeed keeping a vertical drift would have decreased the stability of minimizing configurations in an unfavorable way.

We tried two kinds of energies. In order to have a point of comparison, we tried to use the simple energy $U(x) = -\mu(x) + \max_{y \in \tilde{\mathrm{S}}} \mu(y)$, where μ is the counting measure.

Then we tried a compound energy $W(x) = U(x) + \alpha V(x)$ for a large value of α and for $V(x) = \int (1 - (\mu_x \circ \Phi/\rho)^2 d\rho$.

Eventually, we tried to relax the constraint, changing ρ to $\tilde{\rho} = 6/5 \times \rho$.

7.2.1. Experiments with a simple energy function. In order to have a point of comparison, we recorded first the performance of repeated relaxations. The relaxation algorithm we used corresponds to a choice of $\lambda = 0$, or equivalently to a choice of $\beta = +\infty$ in the Metropolis algorithm.

Then we considered the Metropolis algorithm for different values of λ and of β. We tried $\lambda = 1$ and $\lambda = 0.5$, two "natural" choices for λ. The former let us inhibit destructions only according to the energy increment, whereas the latter let constructions and destructions have equal frequencies at infinite temperature.

The first conclusion we reached was that a significant improvement over the relaxation scheme could be obtained using the Metropolis algorithm with a moderate number of steps. We compared relaxation with 300 steps (for which convergence was always reached) with Metropolis with $\lambda = 1$, $\beta = 1$, and $N = 4000$. In order to compare methods using the same number of iterations, we repeated Metropolis 20 times and the relaxation algorithm $\lfloor (4000 \times 20)/300 \rfloor = 266$ times. On the following diagram (Fig. 7.6) we plotted the repartition functions of the best solution found for each of the 20 runs of Metropolis (dashed lines), along with the best 20 results out of the 266 runs of the relaxation algorithm (solid lines).

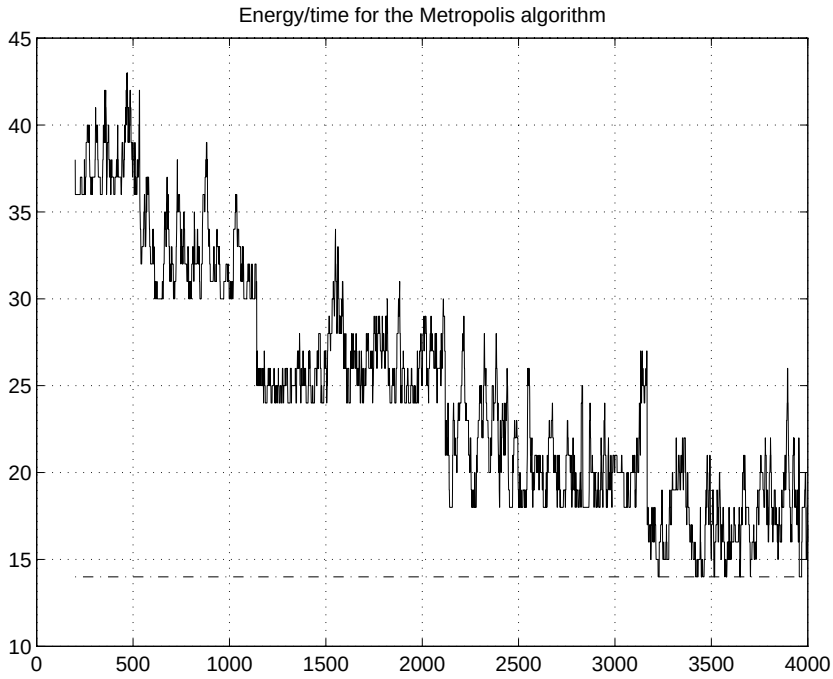


FIG. 7.7.

We obtained very suggestive evolutions for the Metropolis algorithm, such as the following (Fig. 7.7).

On this plot of $u_n = U(X_n)$ for $n = 300, \dots, 4000$, we see the "staircase" shape of the trajectories of the Metropolis algorithm. The algorithm "falls" into deeper and deeper maximal cycles [13] of the domain $\tilde{\mathcal{S}} \setminus \mathcal{S}$. We refer to [5] for a theoretical study of the exit path of the Metropolis algorithm from a domain at low temperature. For a study of the trajectories of simulated annealing algorithms, we refer to [6] and [22], which rely on more complex but also more general induction proofs which cover the time inhomogeneous case. For a semigroup approach of the same question in the continuous time case, we refer to [10], [11], [12], [17], [19], [20], and [21].

The energy evolution can be decomposed into a decreasing part $\underline{u}_n = \min_{k \le n} u_k$ and a "wandering" part $\overline{u}_n = u_n - \underline{u}_n$, as in the following diagram (Fig. 7.8).

The repartition function of the wandering part gives information about the depth of secondary attractors from which the algorithm is able to escape within the time of the simulation. It is a useful tool to choose the inverse temperature parameter β. Following is the repartition function corresponding to the preceding plot (Fig. 7.9).

The best results for the Metropolis algorithm of time length $N = 4000$ were obtained for $\beta = 1$ and $\lambda = 1$ or for $\beta = 0.8$ and $\lambda = 0.5$. This shows that in this case, the choice of λ is not crucial. In the following, we will use $\lambda = 1$, because we can hope to take better advantage of the discrimination made by the energy function between small and big pieces when we use this value of λ.

Then we used the Metropolis algorithm and simulated annealing on long time intervals. Namely, we took $N = 20000$, $\beta_{\min} = 0.7$, $\beta_{\max} = 1.1$ for simulated annealing and $\beta = 1$ for the Metropolis algorithm. On 10 runs of each algorithm, we could notice a clear gain in performance in favor of simulated annealing.

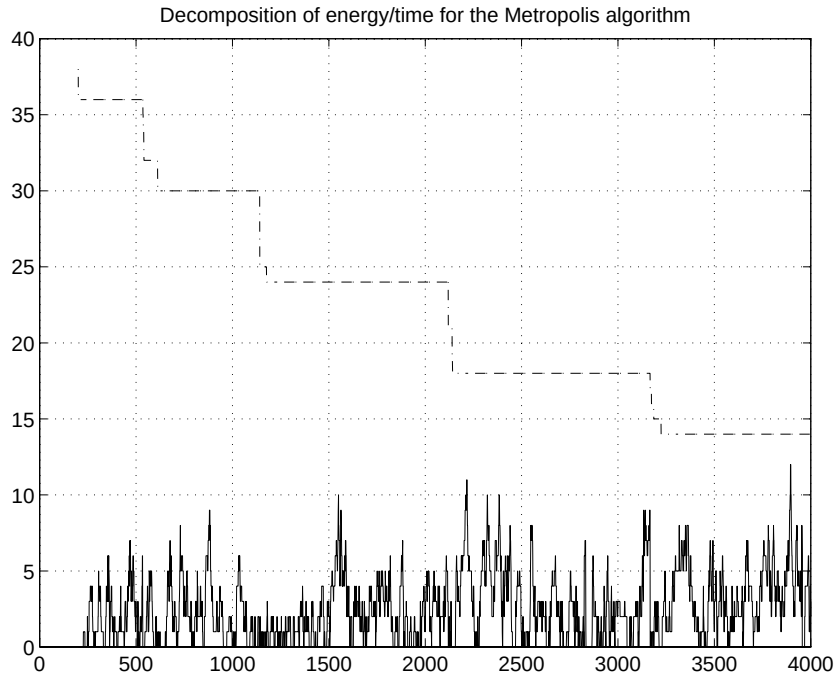

Decomposition of energy/time for the Metropolis algorithm
40
35
30
25
20
15
10
5
0
0
500
1000
1500
2000
2500
3000
3500
4000

FIG. 7.8.

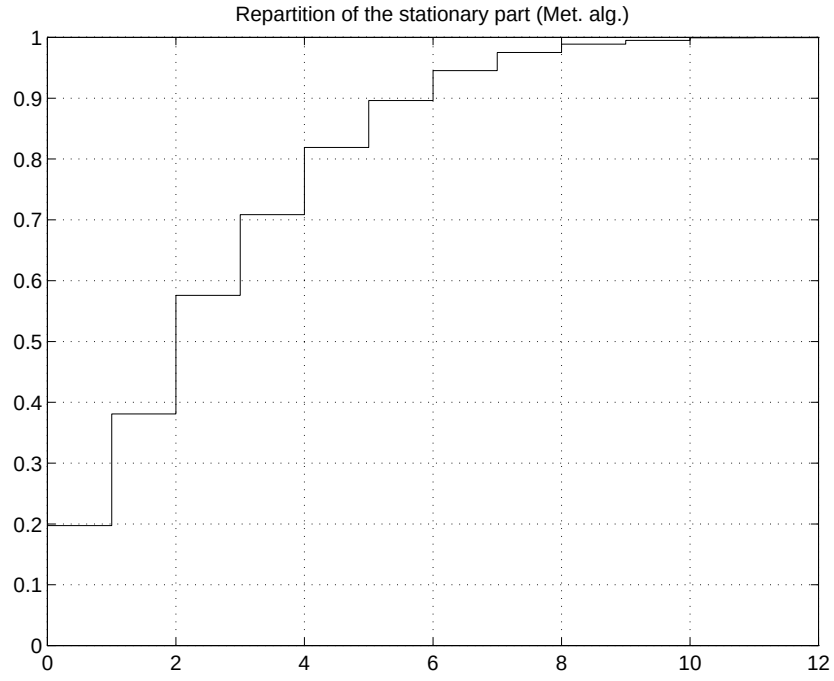

Repartition of the stationary part (Met. alg.)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
2
4
6
8
10
12

FIG. 7.9.

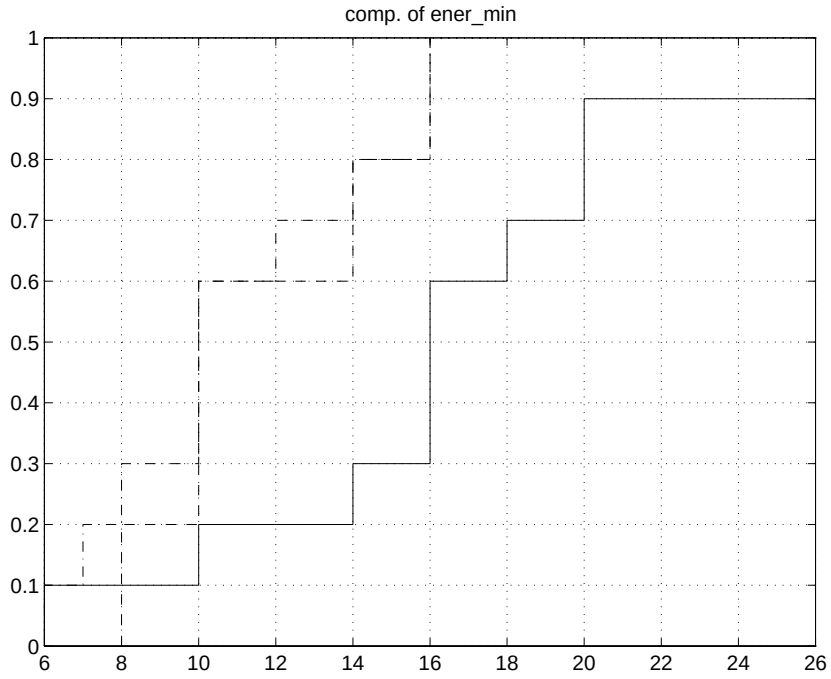


FIG. 7.10.

We tried eventually to get a better improvement using the IET algorithm. Since the state space is already rather large, we followed the idea introduced in [7] to use transformations $F_{\alpha,T,\tau}$ with a nonzero value of α.

We took $\alpha = 0.3$, $\beta = 1/T = 30$, $\frac{1}{1+\rho} = 0.5$, $r = 4$, and $\eta_0 = 15$. We obtained the following comparative results for the best energy found in each of 10 runs of each algorithm. The mean values are 16.2 for the Metropolis algorithm (solid lines), 11.6 for simulated annealing (dashed lines), and 10.9 for the IET algorithm (dash-dot lines). The repartition functions are plotted on the next diagram (Fig. 7.10).

7.2.2. Experiments with a compound energy function. We used the energy

$$W(x) = U(x) + \alpha V(x),$$

with a huge value of $\alpha = 10000$.

The range of this energy is very large, when compared with the previous one, since $W_{\max} = 2300230$, whereas $W_{\min} = 0$ and removing a piece of size 1×1 from a complete solution in a line of weight $\rho(y) = 30$ costs $\Delta W \simeq 334.33$. Therefore we may expect more spectacular improvements from the speed-up techniques.

We tried different temperatures for the Metropolis algorithm with $N = 20000$. The best results were obtained when $\beta = 8 \times 10^{-4}$. On 10 runs, the average best value was 15853.

Using simulated annealing with $\beta_{\min} = 10^{-4}$, $\beta_{\max} = 10^{-3}$, we improved the performance on the average, as shown in the next diagram. On 10 runs, the average best energy value was 8765.

We obtained some more improvement using the IET algorithm (with $\gamma = 5 \times 10^{-5}$, $\beta = 10$, and $\eta_0 = 2000$). On 10 runs the average best energy value was 6280.

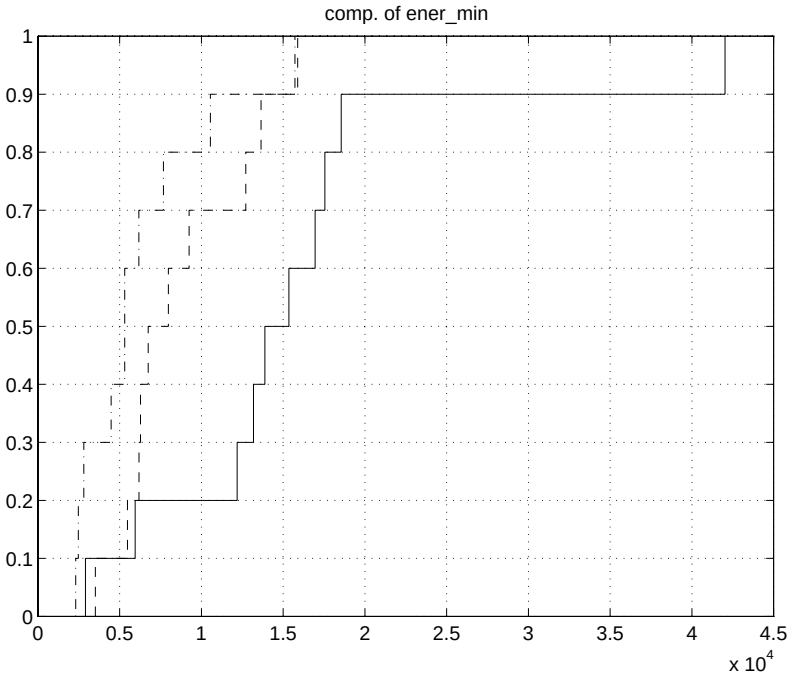


FIG. 7.11.

Figure 7.11 shows a diagram of the repartition functions of the best energy value for ten runs of the Metropolis algorithm (solid lines), simulated annealing (dashed lines), and the IET algorithm (dash-dot lines).

7.2.3. Experiments with a relaxed constraint. We explored also an alternative in the optimization design, which consists of replacing ρ by $\tilde{\rho} = \frac{6}{5}\rho$. We considered accordingly a larger search space $\tilde{\mathcal{S}}$ where the constraint $\frac{\mu_x \circ \Phi^{-1}}{\rho} \leq 1$ is relaxed to $\frac{\mu_x \circ \Phi^{-1}}{\tilde{\rho}} \leq 1$. We took again a compound energy of the type $W(x) = U(x) + \alpha\, V(x)$, with $\alpha = 10000$. The range of W is between $W_{\min} = 76666.66$ and $W_{\max} = 2760230$.

In this example, we can perform the same kind of comparison as in the case of tight constraints. We made 10 runs of length $N = 20000$ of each algorithm. The average of the best energy value found in each run is 8731 for the Metropolis algorithm, 8685 for simulated annealing, and 8567 for the IET algorithm. Figure 7.12 shows a diagram of the corresponding repartition functions (solid lines for the Metropolis algorithm, dashed lines for simulated annealing, and dash-dot lines for the IET algorithm).

The best solution was found by the IET algorithm. It has an energy of $W(x) = 78750$ and is shown in Fig. 7.13.

In this solution, all the pieces are set in the frame. We can judge the quality of the solution with respect to the proportionality constraint on the following diagram (Fig. 7.14), where we have plotted $\tilde{\rho}$ (dashed lines), the measure expressing the constraint, and $\mu_x \circ \Phi^{-1}$ (solid lines), giving the number of unit squares actually filled on each line by the solution. The optimum would be $\mu_x \circ \Phi^{-1} = \rho = 5/6 \times \tilde{\rho}$. We are not too far from that: the two entries $\mu_x \circ \Phi^{-1}(2)$ and $\mu_x \circ \Phi^{-1}(4)$ are one unit too large, and $\mu_x \circ \Phi^{-1}(9)$ is two units short from the optimum. This is the best approximation to an optimal solution we were able to compute on this example. This seems to show that

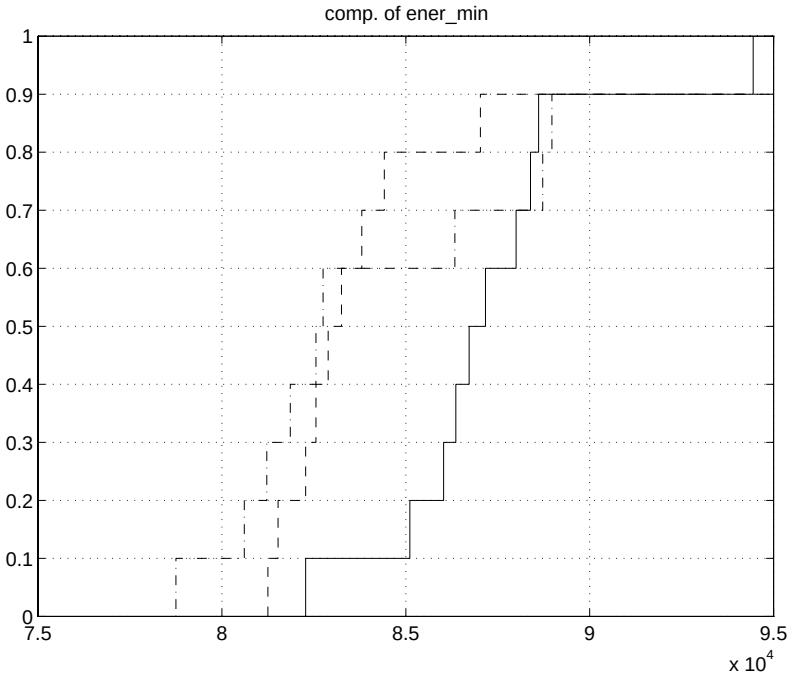


FIG. 7.12.

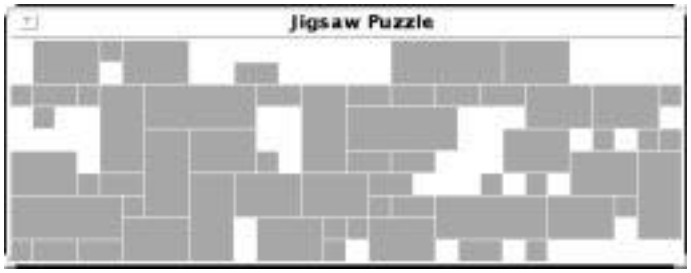


FIG. 7.13.

relaxing the constraint slightly and introducing the loss function $V(x)$ in the energy eases the optimization process.

This should be compared with the best solution found without relaxing the constraint (Fig. 7.15) and its constraint diagram (Fig. 7.16).

For this solution, $U(x) = 6$. Solutions of energy $U(x) = 6$ were also found using the simple energy U to guide the search. Therefore the advantage of introducing the V component in the energy function is not obvious when the constraint is really tight.

It is also interesting to consider typical energy evolutions of those three algorithms. On the following diagrams (Figs. 7.17–7.19), we have plotted the sequence

$$u_n = U(X_n).$$

As we have already mentioned, these sequences of energy values can be decomposed

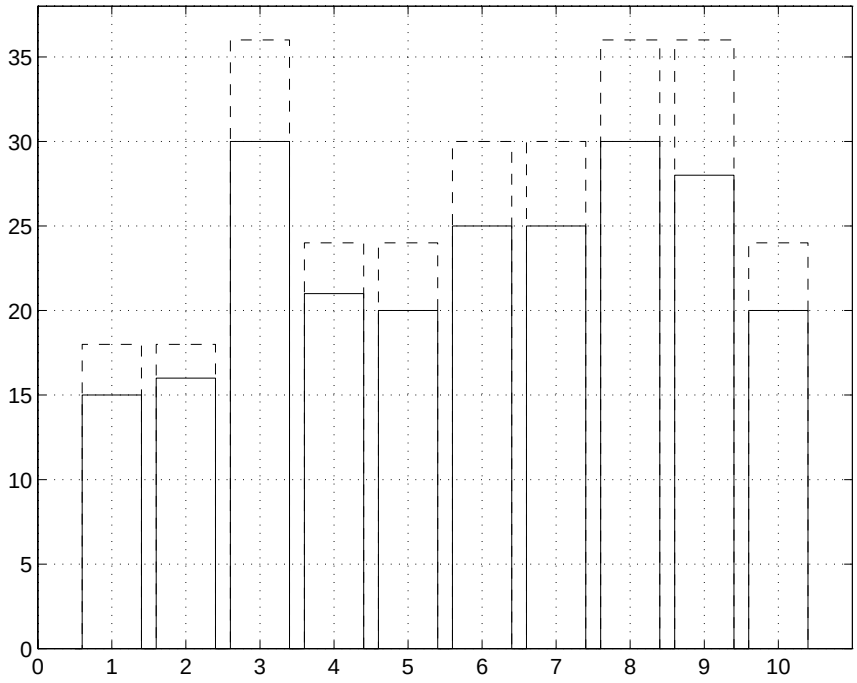


FIG. 7.14.

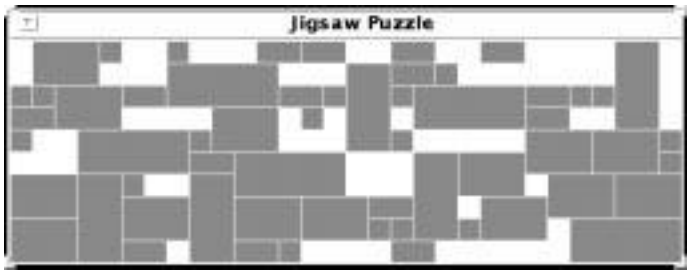


FIG. 7.15.

into a decreasing component

$$\underline{u}_n = \min\{u_k \,:\, k \leq n\},$$

and a wandering component

$$\overline{u}_n = u_n - \underline{u}_n.$$

The repartition functions of $(\overline{u}_n, n = 1, \ldots, N)$ can help to properly set the parameters. It indicates the depth of the attractors from which the algorithm is able to escape.

It is interesting to compare the energy evolutions of the three algorithms. The comparison between the Metropolis algorithm and simulated annealing shows clearly that the temperature used in Metropolis is too low during the first 4000 iterations and too high during the last 8000 iterations. As for the IET algorithm, we can see

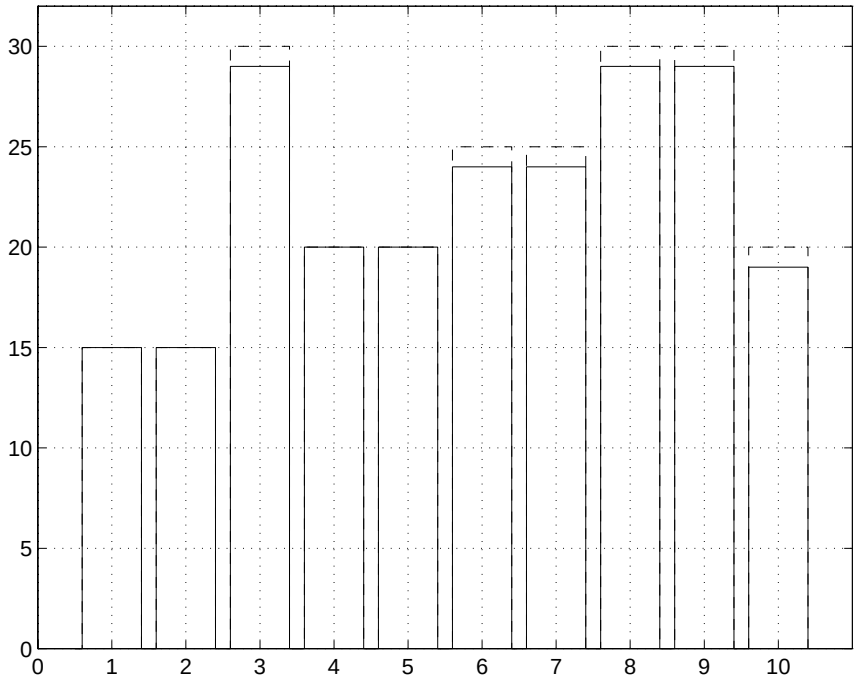


FIG. 7.16.

that the fluctuations of the wandering part are decreasing with time, as in the case of simulated annealing, but that the evolution of the energy is more unstable: it can go up and down faster (in other words, its peaks are sharper). This explains why it is able to sample more efficiently a state space containing many local minima.

7.2.4. Experiment with the partial freezing method. In this experiment, we took the IET algorithm, which had proved to be the best when used globally, and we added a postprocessing stage where we froze all but three of the tasks. At the same time we decreased the global optimization step from 20000 iterations to 3000 iterations and kept 50 times 300 iterations for postprocessing (we drew 50 different frozen configurations and made 300 iterations for each; in each frozen configuration, only three pieces were left unfrozen). Thus we decreased slightly the total number of iterations from 20000 to 18000 (in answer to a suggestion of one of our referees that a more complex algorithm should be allowed less iterations). At the same time, we got an improvement on 10 trials for both the mean value of the energy and its minimum value over the 10 trials (Fig. 7.20). We also found that it was much easier to tune the parameters of the IET algorithm.

Figure 7.21 shows an energy evolution typical of the partial freezing method, where the lower plot shows the evolution of $\eta_0 - \tau_n$: we see that the partial freezing allows us to go faster up and down the energy landscape (in other words, it allows us to work at a higher temperature).

7.2.5. Is the number of iterations a fair measure of complexity? In the case of the three algorithms in this paper, the inner loop is the same except for the computation of the rejection probabilities. When one goes from the Metropolis algorithm to simulated annealing, one has to add the cost of updating the temperature,

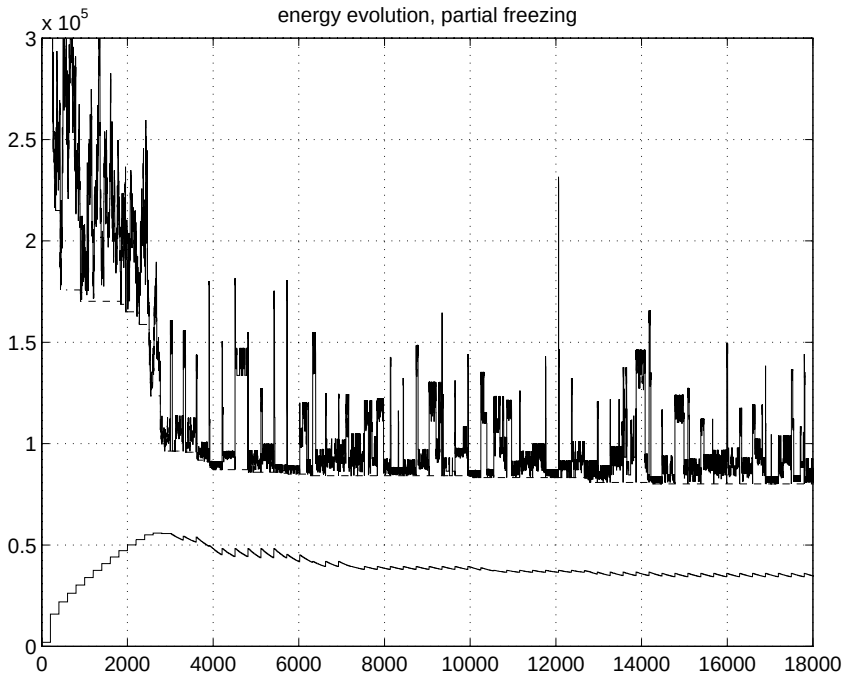


FIG. 7.17.

that is, the cost of one multiplication (in fact, it is even possible to use piecewise constant temperature schedules with the same theoretical properties; see [9], for which the update of the temperature is not in the inner loop and therefore has little influence on the computer time). When one goes from the Metropolis algorithm to the IET algorithm, one has to compute the energy transformation, that is, $\log \frac{U+\Delta U}{U}$, this means one addition, one multiplication, and a logarithm. The fact that one uses the value of U and not only the value of ΔU, the energy increment, does not really make a difference in practice since one will, anyhow, want to record the value of the energy U in order to keep the best solution encountered and not systematically keep the last current solution. Anyhow, computing U from the accumulated energy increments requires only one addition per iteration.

In fact, in our experiments, these differences in the computing time of the rejection probability does not seem to be the leading factor in the variations of the cpu time. The unix function "time" gave us the following figures: 66.3 seconds of user cpu time for 20000 iterations of the simulated annealing algorithm, 61.1 seconds for 20000 iterations of the IET algorithm, and 35.2 seconds for 18000 iterations of the IET algorithm combined with the partial freezing method. These figures are somewhat unexpected. Our interpretation is that all the moves do not have the same complexity. This is particularly true with the partial freezing method, where during most of the time (15000 iterations) the state space is restricted; therefore the choice of a task to schedule or to destroy and the choice of the resources to allocate are made from smaller sets and therefore are faster. To a minor degree the same happens with the IET algorithm and Simulated Annealing: the IET algorithm spends more time at low energy levels where more tasks are scheduled and where scheduling a new task is done from a smaller set of available tasks and resources.

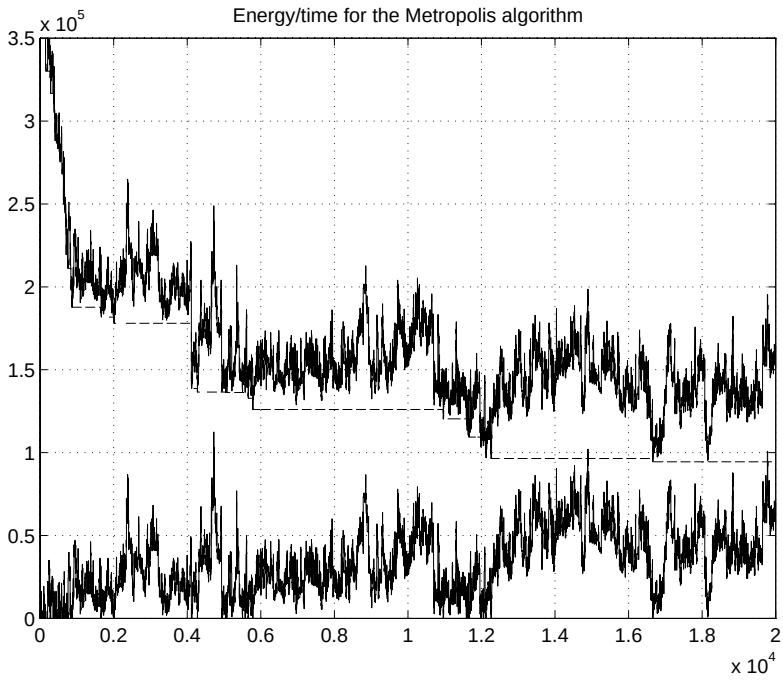


FIG. 7.18.

In conclusion, the number of iterations is not a perfect measure of complexity but has the advantage of being machine independent. To get a real cpu time complexity study, one would need to analyze dynamically the complexity of moves, which depends on the current configuration, and to gather statistics from profile files, which we have not done in the framework of this study.

8. Comparison with other energy landscapes. The most common way to enlarge the space of solutions to create a search space is to allow overlaps. In the task assignment formulation, this means that the same resource is allowed to be used by more than one task at the same time. In the jigsaw puzzle formulation, this means that we allow pieces to sit on top of each other. We will maintain the jigsaw puzzle terminology in the following discussion.

Let us discuss first the case where the aim is simply to find a complete admissible solution. We will discuss afterwards the case where a cost function has to be optimized on the set of complete solutions.

So for the moment, the energy in the overlap case will be made up of the total area of the overlaps and, in the partial solution case, will be made up of the total area of unused pieces.

We can expect the overlap approach to generate the same kind of energy barriers as ours. Indeed if the problem is "tight," meaning that there is just enough room to put all the pieces in the frame, it will be necessary, in order to move a task from one location to another distant location in the frame, to put it in an already occupied location, creating an overlap of the order of the area of the piece to be moved. In the partial solution approach, one has to remove two pieces from the frame. This means that the energy barrier will be from one to two times the energy barrier of the overlap

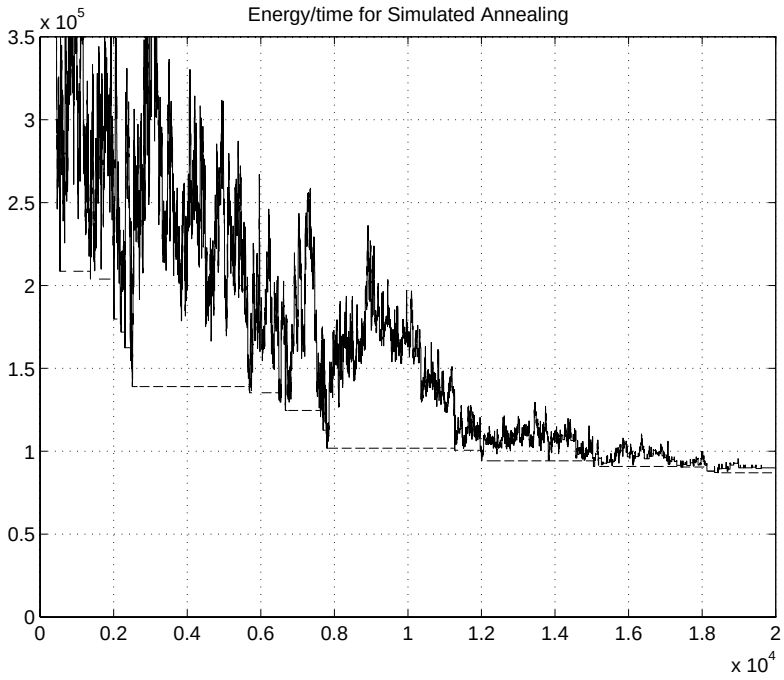


FIG. 7.19.

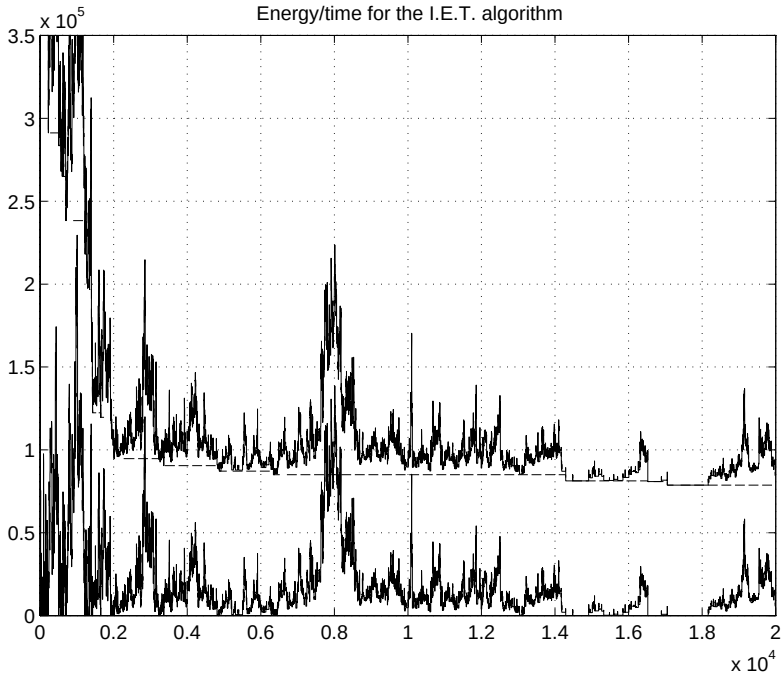


FIG. 7.20.

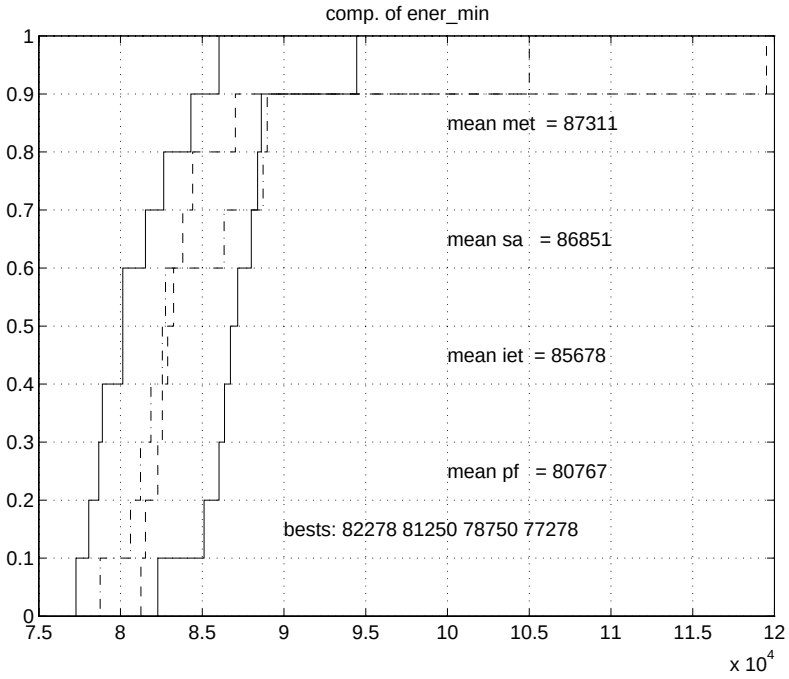


FIG. 7.21.

approach, depending on whether or not the pieces to be moved around are of equal bulk. If one has to exchange a large piece with a number of small pieces, the energy barrier will be the area of the large piece plus the largest area of the small pieces, because once the large piece is removed, the small ones can be transferred one at a time. This is what we can say about the "H" term in the difficulty. Now we have to compare it with the "U" term. For the U term, our approach is better since the energy is quantized: a local minimum solution has at least one piece out; therefore the "U" term is larger than the area of the smallest piece. On the contrary, if space is not discretized, or is finely discretized, the overlap in an imperfect solution of the overlap approach may be arbitrarily small, leading to an arbitrarily large difficulty. Thus the overlap energy cannot safely be used alone. Usually, what is done is to allow pieces to overlap not only between themselves, but also with the outside of the frame—this last kind of overlap being given a specific weight (increasing with time). This is very much like allowing partial solutions, leads to a more complicated algorithm where the balance between two energy terms has to evolve with time, and does not solve completely the problem posed by almost perfect solutions with small scattered overlaps which cannot be improved by local moves.

Let us discuss now the case where some other cost function has to be optimized on the state space of complete solutions. In the partial solution approach, it is usually quite easy to build a cost function which decreases with the number of scheduled tasks and therefore can be used alone. It will be the case, for instance, when the cost function is a sum of negative terms depending on (small) clusters of scheduled tasks. When a task is removed, some of the scheduled clusters are suppressed, the others being unchanged, and so the cost function is increased. It is not so easy and natural to build cost functions which are decreasing with the area of the overlap. Therefore

in the overlap approach an "artificial" overlap term with a big enough weight has to be added to the cost function in the definition of the energy function, creating new energy barriers and increasing the difficulty of the energy landscape.

For all these (qualitative) reasons, we think that the partial solution approach we propose here has some merits, at least in the case of "tight" problems, when it is compared with the traditional overlap approach. This would of course require confirmation by quantitative experimental comparisons.

Conclusion. We touched in this paper on three related topics with various degrees of generality. Our first aim was to bring experimental evidences comforting theoretical results about the behavior of algorithms. We wanted to show that the theory was not concerned with a "never-reached" asymptotic and led to the same qualitative ranking of performances to which an experimental benchmark would lead. Our second aim was to describe a general purpose methodology to deal with scheduling tasks. We insisted on two problems that are likely to be encountered in many situations: the creation of small gaps in the allocation of resources and the way to handle "proportionality constraints."

The third aspect of the paper was to account for experiments on a benchmark of the "jigsaw puzzle" type. Here we were confronted with the practical problem of the choice of parameters and of optimization design options (such as relaxing some of the constraints). Our conclusion on this third point is that we have acquired some know-how about the choice of parameters, which we tried to reflect in section 6, but that we have presently no systematic rule to choose them. We worked very much in a trial and error way, looking at the repartition functions we mentioned, to guide our intuition. A trial and error procedure is somehow justified by the theoretical result that many trials of moderate length are preferable to a long one. This gives us the opportunity to tune the parameters trial after trial.

Anyhow, we have to admit that the choice of parameters requires some skill, especially for the simulated annealing and IET algorithms, where there are more than one parameter to tune. What we did not find too hard to do was, starting from a given Metropolis algorithm at inverse temperature β_{Met}, to find $\beta_{\min} < \beta_{\mathrm{Met}} < \beta_{\max}$ for which simulated annealing performs better than Metropolis. Then we could get some more improvement using the IET algorithm, where again we chose the parameters in relation with those used for simulated annealing. We are not sure at all that this is the best way to tune simulated annealing or the IET algorithm, but it shows at least that the theoretical gains of one algorithm upon the previous one could be obtained in practice. Finally, in the partial freezing method the choice of parameters is easier because the algorithm runs, most of the time, on a restricted state space for which the tuning of parameters is less crucial.

Another positive result of these experiments is that it is possible to get good, if not optimal, solutions even in the case where very nonmonotonous evolutions of the energy are needed, as it is the case here, since the only way to move a piece of the puzzle is to remove it and put it somewhere else afterwards, a succession of two moves the first of which implies an energy increase.

Of course we have touched in this paper on only a limited number of questions. For instance, we leave open the practical question of the best choice of parameters for simulated annealing and for the IET algorithm, since we used only a robust "all purpose" set of parameters, namely, exponential temperature sequences in the case of simulated annealing and logarithmic energy transforms for the IET algorithm. Another question we left purposely in the dark is the choice of elementary moves.

Although it is clear that a benefit can be obtained from the use of more complex compound moves, we felt such an investigation would have been too dependent on the precise examples we chose to study. Rather we tried to lay the stress on general ideas and tools, with the hope that they could be useful in a variety of situations.

Acknowledgments. I wish to thank Professor Robert Azencott for many helpful discussions about this paper. I wish also to thank him for having involved me in the development of an algorithm for an industrial scheduling problem a few years ago. I am also pleased to acknowledge the useful comments of the referees which helped me to improve a first draft of this paper.

REFERENCES

[1] D. Abramson (1992), *A very high speed architecture for simulated annealing,* IEEE Comput., pp. 27–36.

[2] R. Azencott (1988), *Simulated annealing*, Séminaire Bourbaki 40ième année, 1987–1988, p. 697.

[3] R. Azencott (1992), *Sequential simulated annealing: Speed of convergence and acceleration techniques*, in Simulated Annealing: Parallelization Techniques, Wiley Interscience Ser. Discrete Math., R. Azencott, ed., Wiley Interscience, New York, pp. 1–10.

[4] R. Azencott and C. Graffigne (1992), *Parallel annealing by periodically interacting multiple searches: Acceleration rates*, in Simulated Annealing: Parallelization Techniques, Wiley Interscience Ser. Discrete Math., R. Azencott, ed., Wiley Interscience, New York, pp. 81–90.

[5] O. Catoni and R. Cerf (1997), *The Exit path of a Markov chain with rare transitions*, ESAIM Probab. Statist., 1, pp. 95–144; also available online from http://www.emath.fr/Maths/Ps/ps.html.

[6] O. Catoni (1992), *Rough large deviation estimates for simulated annealing: Application to exponential schedules*, Ann. Probab., 20, pp. 1109–1146.

[7] O. Catoni (1998), *The energy transformation method for the Metropolis algorithm Compared with simulated annealing*, Probab. Theory Related Fields, 110, pp. 69–89.

[8] O. Catoni (1991), *Exponential triangular cooling schedules for simulated annealing algorithms: A case study*, in Applied Stochastic Analysis, Proc. US–French Workshop, Rutgers University, New Brunswick, NJ, April 29–May 2, 1991, Lecture Notes in Control and Inform. Sci. 177, I. Karatzas and D. Ocone, eds., Springer–Verlag, Berlin, 1992, pp. 74–89.

[9] C. Cot and O. Catoni (1996), *Piecewise Constant Triangular Cooling Schedules for Generalized Simulated Annealing Algorithms*, preprint, LMENS 96-19; Ann. Appl. Probab., to appear; also available online from http://www.dmi.ens.fr/dmi/preprints.

[10] J. D. Deuschel and C. Mazza (1994), L^2 *convergence of time nonhomogeneous Markov processes:* I. *Spectral Estimates*, Ann. Appl. Probab., 4, pp. 1012–1056.

[11] P. Diaconis and D. Stroock (1991), *Geometric Bounds for Eigenvalues of Markov Chains*, Ann. Appl. Probab., 1, pp. 36–61.

[12] M. Duflo (1996), *Algorithmes Stochastiques*, Mathématiques & Applications (Paris), Springer-Verlag, New York.

[13] M. I. Freidlin and A. D. Wentzell (1984), *Random Perturbations of Dynamical Systems*, Springer-Verlag, New York.

[14] M. R. Garey and D. S. Johnson (1979), *Computers and Intractability: A guide to the theory of NP-completeness*, W. H. Freeman, New York.

[15] S. Geman and D. Geman (1984), *Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images*, IEEE Trans. Pattern Anal. Mach. Intelligence, 6, pp. 721–741.

[16] C. Graffigne (1992), *Parallel annealing by periodically interacting multiple searches: An experimental study*, in Simulated Annealing: Parallelization Techniques, Wiley Interscience Ser. Discrete Math., R. Azencott, ed., Wiley Interscience, New York, pp. 47–79.

[17] R. Holley and D. Stroock (1988), *Annealing via Sobolev inequalities*, Comm. Math. Phys., 115, pp. 553–559.

[18] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi (1983), *Optimization by simulated annealing*, Science, 220, pp. 621–680.

[19] L. Miclo (1991), *Evolution de l'énergie libre. Application à l'étude de la convergence des algorithmes du recuit simulé*, Doctoral Dissertation, Université d'Orsay, February 1991.
[20] L. Miclo (1996), *Sur les problèmes de sortie discrets inhomogènes*, Ann. Appl. Probab., 6, pp. 1112–1156.
[21] L. Miclo (1995), *Sur les temps d'occupations des processus de Markov finis inhomogènes à basse température*, Stochastics Stochastics Rep., submitted.
[22] A. Trouvé (1993), *Parallélisation massive du recuit simulé*, Doctoral Dissertation, Université Paris 11, January 5, 1993.
[23] A. Trouvé (1994), *Cycle decompositions and simulated annealing*, SIAM J. Control Optim., 34, 1996, pp. 966–986.
[24] A. Trouvé (1995), *Rough large deviation estimates for the optimal convergence speed exponent of generalized simulated annealing algorithms*, Ann. Inst. H. Poincaré Probab. Statist., 32, 1996, pp. 299–348.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 5, pp. 1576–1595, September 1998

004

BOUNDARY CONTROLLABILITY OF A HYBRID SYSTEM CONSISTING IN TWO FLEXIBLE BEAMS CONNECTED BY A POINT MASS*

CARLOS CASTRO† AND ENRIQUE ZUAZUA†

Abstract. We consider a hybrid system consisting of two flexible beams connected by a point mass. The constant of rotational inertia is assumed to be nonzero. In a previous paper we have proved that, in the presence of the point mass, the system is well posed in asymmetric spaces in which solutions have one more degree of regularity to one side of the mass.

We are interested in the problem of controllability when the control acts on the free extreme of one of the beams. We prove that when the control time is large enough the system is exactly controllable in an asymmetric space. This result is sharp. The proofs combine classical techniques from asymptotic analysis and the theory of nonharmonic Fourier series.

Key words. flexible beams, point mass, asymmetric spaces, Fourier series, controllability

AMS subject classifications. 35L30, 35P15, 42A55

PII. S0363012997316378

1. Introduction.

In this paper we study the boundary controllability of a linear system modelling the vibrations of two flexible beams connected by a point mass.

We assume that the beams occupy the intervals $(-1,0)$ and $(0,1)$ and that the point mass is located at $x=0$. By means of the scalar function $u=u(x,t)$ defined for $x\in(-1,1)$ and $t>0$, we describe the vertical displacements of the beams and the point mass. The linear equations describing the small vibrations of this system can be written as follows:

$$
\begin{cases}
\gamma\partial^2 u_{tt} - u_{tt} - \partial^4 u = 0, & \text{for } x\in(-1,0), t>0,\\
\gamma\partial^2 u_{tt} - u_{tt} - \partial^4 u = 0, & \text{for } x\in(0,1), t>0,\\
[u](0,t) = [\partial u](0,t) = 0, & \text{for } t>0,\\
M u_{tt}(0,t) + [\partial^3 u](0,t) = 0, & \text{for } t>0,\\
M\gamma_0 \partial u_{tt}(0,t) - [\partial^2 u](0,t) = 0 & \text{for } t>0,
\end{cases}
\tag{1}
$$

where ∂ denotes partial derivation with respect to x and the index t derivation with respect to time. $[u](0)=u(0^+)-u(0^-)$ denotes the jump of the function u at the point $x=0$ where the mass is located. Assuming that the beams are hinged at their extremes, system (1) has to be completed with the following boundary conditions:

$$
u(\pm 1,t) = \partial^2 u(\pm 1,t) = 0, \quad \text{for } t>0. \tag{2}
$$

In (1) the dynamic of the beams is described by the Rayleigh beam equation, where $\gamma\geq 0$ represents the constant of rotational inertia. The third equation guarantees

* Received by the editors February 10, 1997; accepted for publication (in revised form) September 22, 1997; published electronically June 3, 1998.
http://www.siam.org/journals/sicon/36-5/31637.html

† Departamento de Matemática Aplicada, Universidad Complutense de Madrid, 28040 Madrid, Spain (ccastro@sunma4.mat.ucm.es, zuazua@sunma4.mat.ucm.es). This research was partially supported by grant PB93-1203 from the DGICYT (Spain) and grant CHRX-CT94-0471 from the European Union. The research of the first author was also supported by a doctoral fellowship from the "Universidad Complutense de Madrid."

that u and ∂u are continuous across $x = 0$ while the last two equations describe the vibrations of the point mass at $x = 0$. M is the total mass concentrated in $x = 0$ and γ_0 is the rotational inertia at this point.

To simplify the exposition we assume $M = 1$ and $\gamma = \gamma_0$, although the analysis is valid for other values of the parameters.

It is worth noting that, in the particular case in which the constant γ of rotational inertia vanishes ($\gamma = 0$), $\partial^2 u$ is continuous across $x = 0$, also. This implies that the effect of the mass point is weaker on the behavior of the system when $\gamma = 0$ than when $\gamma > 0$. Thus the properties of system (1.1)–(1.2), when $\gamma = 0$, are much closer to the case in which the point mass is not present. We refer to [5] and [9] for precise statements and the details of the proofs.

Throughout this paper we assume that $\gamma > 0$.

System (1)–(2) has to be completed with suitable initial conditions for $u(x,t)$, $u(0,t)$, and $\partial u(0,t)$. The last two quantities will be denoted by y and z, resp., i.e.,

$$u(0,t) = y(t); \qquad \partial u(0,t) = z(t). \tag{3}$$

The initial conditions are then

$$\begin{cases} u(x,0) = u^0(x) & \text{in } (-1,0)\cup(0,1)\,; y(0) = y^0, \quad z(0) = z^0, \\ \\ u_t(x,0) = u^1(x) & \text{in } (-1,0)\cup(0,1);\ y_t(0) = y^1, \quad z_t(0) = z^1. \end{cases} \tag{4}$$

System (1)–(2) has been studied in [4] where it was proven that, with appropriate regularity and compatibility conditions, on the initial data it admits a unique solution in a suitable class. On the other hand, its energy

$$E(t) = \int_{-1}^{1} \left[\left|\partial^2 u(x,t)\right|^2 + \gamma \left|\partial u_t(x,t)\right|^2 + \left|u_t(x,t)\right|^2 \right] dx + \left|u_t(0,t)\right|^2 + \gamma \left|\partial u_t(0,t)\right|^2 \tag{5}$$

is constant along trajectories.

In this paper we assume that a control function $q = q(t)$ acts on the system through the extreme $x = 1$ on the quantity $\partial^2 u(1,t)$. Then the boundary conditions in (2) have to be replaced by

$$u(\pm 1,t) = \partial^2 u(-1,t) = 0; \quad \partial^2 u(1,t) = q(t) \quad \text{for } t > 0. \tag{6}$$

The problem of exact controllability can be formulated as follows: *Given $T > 0$, find the class H of initial conditions for which there exists a control q, say, in $L^2(0,T)$ such that the solution of* (1), (3) *with boundary conditions* (6) *is at rest at time $t = T$, i.e., it satisfies*

$$\begin{cases} u(x,T) = 0 & \textit{for } x \in (-1,0)\cup(0,1), \quad y(T) = 0, \quad z(T) = 0, \\ \\ u_t(x,T) = 0 & \textit{for } x \in (-1,0)\cup(0,1), \quad y_t(T) = 0, \quad z_t(T) = 0. \end{cases} \tag{7}$$

In this formulation of the control problem we have chosen the control to belong to $L^2(0,T)$. This is not, of course, the unique choice, but it is the one that comes more naturally when studying the problem of controllability by means of J.-L. Lions' HUM method (see [8]).

It turns out that the space H of controllable initial data cannot be found among the family of energy spaces in which system (1)–(2) is well posed. Indeed, all the energy

spaces have in common the fact that solutions in those classes have the same regularity on both sides of the point mass. For instance, the energy E in (5) corresponds to solutions u in H^2 to both sides of $x = 0$ and such that u_t belongs to $H^1(-1,1)$. However, the space of controllable data turns out to be asymmetric in the sense that its elements have one more degree of regularity to the left of $x = 0$.

The same phenomena was observed in [6] in the case of two flexible strings connected by a point mass. In [6] this was proved by using the explicit formula for solutions of the one-dimensional wave equation in terms of its initial data, and it was seen that this is a consequence of the fact that solutions gain one derivative when crossing the mass. In [6] it was also observed that the spectral gap of the wave equation vanishes in the presence of a point mass, and it was conjectured these two facts (i.e., the asymmetry of the controllable space and the lack of the spectral gap) to be closely related. Later on, in [3] it was proved that these two properties are equivalent (see also [1]).

For the fourth order system that we are considering here it has been observed that, in the presence of the point mass, the spectral gap vanishes, also (see [4]). Using Fourier developments of solutions, it is also proved in [4] that system (1)–(2) is well posed in asymmetric spaces in which the solutions have one more degree of regularity to one side of $x = 0$. This result applies only when $\gamma > 0$ since, as we said above, when $\gamma = 0$ the presence of the mass has a much weaker effect on the behavior of the system. In this case ($\gamma = 0$), system (1)–(2) is not well posed in asymmetric spaces of this kind.

Thanks to the existence of the asymmetric spaces in which system (1)–(2) is well posed and by means of the theory of nonharmonic Fourier series, and more precisely, of some results by D. Ulrich [10], we prove sharp observability results. These results establish the equivalence between a suitable asymmetric norm of the initial data and the quantity $\int_0^T |\partial u(1,t)|^2\, dt$ which measures the amount of energy concentrated at $x = 1$ during the time interval $t \in (0,T)$. This result requires the time T to be sufficiently large and, more precisely, $T \geq 4\sqrt{\gamma}$. This is due to the finite speed of propagation underlying in system (1) when $\gamma > 0$, and therefore it is a natural restriction to the observability to hold.

By means of HUM and as a direct consequence of this observability result, we prove the exact controllability of the system. We show, roughly, that when $\gamma > 0$ the space of controllable data coincides with the subspace of $H^2(-1,1) \times H^1(-1,1)$ of those elements that, restricted to $(-1,0)$, have one more degree of regularity, i.e., belong to $H^3(-1,0) \times H^2(-1,0)$.

It is worth mentioning that, in the absence of mass, the space of controllable data coincides with $H^2(-1,1) \times H^1(-1,1)$ (see [7]). Thus, as in the context of flexible strings with a point mass (see [6]), the presence of the point mass reduces the space of controllable data by one derivative on the opposite side of the mass with respect to the extreme in which the control is located.

The rest of the paper is organized as follows. In section 2 we recall without proofs some basic analytical results of the uncontrolled system (1)–(2) given in [4]. In particular we state the well posedness of the system in asymmetric spaces using Fourier series. In section 3 we prove suitable observability properties. In section 4 we solve system (1) with nonhomogeneous boundary conditions to give sense to the solutions of the controlled problem, where we have to consider the boundary condition (6). Finally, in section 5 we obtain the main controllability results.

2. Preliminary results. In this section, we set some analytical properties of the solutions of system (1) which will be used along this work. The proofs of the results can be seen in [4].

2.1. Spectral analysis. When decomposing solutions of (1)–(2) in Fourier series, one is led to consider solutions in separated variables $u = e^{i\lambda t}\varphi(x)$. In this class of solutions, system (1.1)–(1.2) becomes

$$\begin{cases} \partial^4\varphi = \lambda^2\varphi - \gamma\lambda^2\partial^2\varphi\,, & \text{for } x \in (-1,0),\\ \partial^4\varphi = \lambda^2\varphi - \gamma\lambda^2\partial^2\varphi\,, & \text{for } x \in (0,1),\\ [\varphi](0) = [\partial\varphi](0) = 0,\\ [\partial^2\varphi](0) = -\gamma\lambda^2\partial\varphi(0),\\ [\partial^3\varphi](0) = \lambda^2\varphi(0),\\ \varphi(\pm1) = \partial^2\varphi(\pm1) = 0. \end{cases} \tag{8}$$

We introduce the operator $(I-\gamma\partial^2)^{-1} : L^2(-1,1) \to H^2 \cap H_0^1(-1,1)$ such that $u = (I-\gamma\partial^2)^{-1}F$ if and only if $u \in H^2 \cap H_0^1(-1,1)$ and satisfies $u - \gamma\partial^2 u = F$.

If we define the vector valued eigenfunction $\phi = (\varphi, \varphi(0), \partial\varphi(0))$, system (8) can be written as

$$\begin{cases} K\phi = \lambda^2\phi,\\ \partial^2\varphi(\pm1) = 0, \end{cases} \tag{9}$$

where K is the linear operator given by

$$K = \begin{pmatrix} (I-\gamma\partial^2)^{-1}\partial^4_{(-1,1)/\{0\}} & 0 & 0\\ \partial^3_+ - \partial^3_- & 0 & 0\\ -\frac{1}{\gamma}\left(\partial^2_+ - \partial^2_-\right) & 0 & 0 \end{pmatrix}.$$

Here $\partial^4_{(-1,1)/\{0\}}$ represents the operator which assigns to each function, not necessarily continuous in $x = 0$, the fourth order derivative to both sides of $x = 0$, and $\partial^k_\pm$ represents the distribution which assigns to a function u the value $\partial^k u(0^\pm)$.

The following result is proved in [4].

PROPOSITION 2.1. *The eigenvalues $\{\lambda_k\}_{k\in\mathbb{N}}$ of system (9) are simple and constitute a sequence of positive real numbers:*

$$0 < \lambda_1 < \lambda_2 < \cdots < \lambda_k < \cdots \to \infty.$$

Moreover, the corresponding eigenfunctions $\{\phi_k\}_{k\in}\,\mathbb{N}$ may be normalized to form an orthonormal basis of the space

$$H = \left\{\phi = (\varphi, \varphi(0), \partial\varphi(0)) \in H^2 \cap H_0^1(-1,1) \times \mathbb{R} \times \mathbb{R}\right\}$$

with the norm

$$\|\phi\|_H = \left[\int_{-1}^{1} |\partial^2\varphi|^2 dx\right]^{1/2}.$$

REMARK 1. *All the results of the above proposition except the simplicity of the eigenvalues can be proved using classical theory on compact self-adjoint operators. The simplicity of the eigenvalues requires a detailed analysis of the system under consideration.*

From Proposition 2.1, the space H can also be written as follows:

$$H = \left\{ u : u = \sum_{k\in\mathbb{N}} a_k \phi_k, \| u \|_H^2 = \sum_{k\in\mathbb{N}} | a_k |^2 < \infty \right\}. \tag{10}$$

We can also define the following fractional Hilbert spaces $(H_\alpha, \| \cdot \|_\alpha)_{\alpha\in\mathbb{R}}$:

$$H_\alpha = \left\{ u = \sum_{k\in\mathbb{N}} a_k \phi_k : \| u \|_\alpha^2 = \sum_{k\in\mathbb{N}} | a_k |^2 \lambda_k^{4\alpha} < \infty \right\}. \tag{11}$$

We will denote by $< \cdot, \cdot >_\alpha$ the scalar product in H_α.

Clearly $H_0 = H$ and $\| \cdot \|_H = \| \cdot \|_0$.

Observe that, if $u = \sum_{k\in\mathbb{N}} a_k \phi_k$, then $Ku = \sum_{k\in\mathbb{N}} \frac{a_k}{\lambda_k^2} \phi_k$. Clearly K is an isomorphism from H_α into $H_{\alpha+1}$. We can also write explicitly K^{-1}:

$$K^{-1} u = \sum_{k\in\mathbb{N}} \lambda_k^2 a_k \phi_k$$

which is continuous from $H_{\alpha+1}$ into H_α.

We need to identify the spaces H_α for some values of the parameter $\alpha \in \mathbb{R}$. To do that, we denote by $H^s((-1,1)\setminus\{0\}) \cap H^2 \cap H_0^1(-1,1)$ the subspace of $H^2 \cap H_0^1(-1,1)$ constituted by the elements such that its restrictions to $(-1,0)$ and $(0,1)$ belong to H^s.

We have the following characterizations of the fractional spaces H_α.

PROPOSITION 2.2. (a) $H_{1/2}$ *coincides algebraically and topologically with the subspace of*

$$H^3((-1,1)\setminus\{0\}) \cap H^2 \cap H_0^1(-1,1) \times \mathbb{R} \times \mathbb{R}$$

constituted by the elements (u, y, z) *such that*

$$\partial^2 u(\pm 1) = 0, \quad u(0) = y, \quad \partial u(0) = z. \tag{12}$$

(b) $H_{-1/2}$ *coincides with the subspace of* $H_0^1(-1,1) \times \mathbb{R} \times \mathbb{R}$ *constituted by the elements* (u, y, z) *such that* $u(0) = y$.

Moreover,

$$\| (u, y, z) \|_{-1/2}^2 = \int_{-1}^{1} \left[\gamma | \partial u |^2 + | u |^2 \right] dx + | y |^2 + \gamma | z |^2. \tag{13}$$

(c) H_{-1} *coincides algebraically and topologically with the quotient space of* $L^2(-1,1) \times \mathbb{R} \times \mathbb{R}$ *constituted by the classes* (u, y, z) *characterized in the following way: Two elements* (u^1, y^1, z^1) *and* (u^2, y^2, z^2) *belong to the same class if and only if*

$$(u^1 - u^2, y^1 - y^2, z^1 - z^2) = \alpha(m, -1, 0) + \beta(n, 0, \gamma^{-1}),$$

where $\alpha, \beta \in \mathbb{R}$, *and* m *and* n *are the functions*

$$m(x) = \begin{cases} \frac{\sinh(\frac{1+x}{\sqrt{\gamma}})}{2\sqrt{\gamma}\cosh(\frac{1}{\sqrt{\gamma}})} & \text{if } x \in [-1,0], \\ \frac{\sinh(\frac{1-x}{\sqrt{\gamma}})}{2\sqrt{\gamma}\cosh(\frac{1}{\sqrt{\gamma}})} & \text{if } x \in [0,1], \end{cases} \qquad n(x) = \begin{cases} \frac{\sinh(\frac{1+x}{\sqrt{\gamma}})}{2\gamma\sinh(\frac{1}{\sqrt{\gamma}})} & \text{if } x \in [-1,0], \\ \frac{-\sinh(\frac{1-x}{\sqrt{\gamma}})}{2\gamma\sinh(\frac{1}{\sqrt{\gamma}})} & \text{if } x \in [0,1]. \end{cases}$$

(14)

(d) $H_{-3/2}$ *coincides with the quotient space of* $H^{-1}(-1,1) \times \mathbb{R} \times \mathbb{R}$ *constituted by the classes* (u,y,z) *characterized in the following way: Two elements* (u^1,y^1,z^1) *and* (u^2,y^2,z^2) *belong to the same class if and only if*

$$(u^1-u^2, y^1-y^2, z^1-z^2) = \alpha(m,-1,0) + \beta(n,0,\gamma^{-1}),$$

where $\alpha, \beta \in \mathbb{R}$, *and* m *and* n *are the functions given in* (14).

Let us recall now how solutions of (1)–(2) can be developed in Fourier series.

Consider the energy space $\mathcal{H} = H_0 \times H_{-1/2}$ and define $\bar{\phi}_k = (\phi_k, i\lambda_k\phi_k)$ where $\lambda_{-k} = -\lambda_k$ and $\phi_{-k} = \phi_k$. The set $(\bar{\phi}_k)_{k\in\mathbb{Z}}$ constitutes an orthonormal basis in $\mathcal{H}$. Then, for any initial data $((u^0,y^0,z^0),(u^1,y^1,z^1)) \in \mathcal{H}$ we can find a sequence of coefficients (a_k) such that

$$((u^0,y^0,z^0),(u^1,y^1,z^1)) = \sum_{k\in\mathbb{Z}} a_k\bar{\phi}_k,$$

and the vector valued solution $U = ((u,y,z),(u_t,y_t,z_t))$ of (1), (2), (3), and (4) is given by

$$U(t) = \sum_{k\in\mathbb{Z}} a_k e^{i\lambda_k t}\bar{\phi}_k. \tag{15}$$

The conservation of the energy E in (5) is equivalent to the fact that system (1)–(2) generates a group of isometries in $\mathcal{H}$. More precisely,

$$E(t) = \| U(t) \|^2_{\mathcal{H}} = \sum_{k\in\mathbb{Z}} \left| a_k e^{i\lambda_k t}\bar{\phi}_k \right|^2 = \sum_{k\in\mathbb{Z}} | a_k |^2 = \| U(0) \|^2_{\mathcal{H}} = E(0). \tag{16}$$

Obviously, one can also obtain developments in Fourier series of the form (15) for solutions of (1)–(2) in other classes $\mathcal{H}_\alpha = H_\alpha \times H_{\alpha-1/2}$.

2.2. Asymptotics of the spectrum. In this section we recall the main results concerning the asymptotic behavior of the eigenvalues and eigenfunctions of (9) that we will need later to prove the observability inequalities.

PROPOSITION 2.3. *We have*

$$\lambda_{2k-1} = \frac{k\pi - \pi/2}{\sqrt{\gamma}} - \frac{c_1(\gamma)}{\sqrt{\gamma}\,(k\pi-\pi/2)} + O(k^{-2}), \ \text{as } k \to \infty, \tag{17}$$

$$\lambda_{2k} = \frac{k\pi - \pi/2}{\sqrt{\gamma}} - \frac{c_2(\gamma)}{\sqrt{\gamma}\,(k\pi-\pi/2)} + O(k^{-2}), \ \text{as } k \to \infty, \tag{18}$$

where $c_1(\gamma) = \left(2\gamma + \sqrt{\gamma}\tanh\left(\gamma^{-1/2}\right)\right)^{-1} + (2\gamma)^{-1}$ *and* $c_2(\gamma) = \gamma^{-1/2}\coth\gamma^{-1/2} - 2 + (2\gamma)^{-1}$.

Moreover,

$$\lambda_{2k} - \lambda_{2k-1} = \frac{C(\gamma)}{(k\pi-\pi/2)\,\gamma^{1/2}} + O(k^{-2}), \ \text{as } k \to \infty, \tag{19}$$

where $C(\gamma) = c_1(\gamma) - c_2(\gamma) > 0$, *for all* $\gamma > 0$.

REMARK 2. *In the absence of mass the asymptotic behavior of eigenvalues is as follows:*

$$\lambda_k = \frac{k\pi}{2\sqrt{\gamma}} + O(k^{-2}), \text{ as } k \to \infty.$$

This shows that the first term of the asymptotic expansion of λ_{2k} *is affected by the presence of the mass but not the first term of the asymptotic expansion of* λ_{2k-1}.

In view of (19) *the asymptotic gap* $\lambda_{2k} - \lambda_{2k-1}$ *decays like* $1/k$ *as* $k \to \infty$ *in the presence of the mass. Note, however, that in the absence of the mass the eigenvalues are uniformly separated and the gap is of the order of* $\pi/(2\sqrt{\gamma})$ *for all* k.

Concerning the eigenfunctions we have the following.

PROPOSITION 2.4. *The eigenfunctions of* (9) *normalized in* $H^2 \cap H_0^1(-1,1)$ *are*

$$\varphi_{2k-1}(x) = \frac{\rho_{2k-1}}{(\mu^+_{2k-1})^2}\left[\sin\left(\mu^+_{2k-1}(1-|x|)\right) - \frac{\mu^+_{2k-1}\cos\mu^+_{2k-1}}{\mu^-_{2k-1}\cosh\mu^-_{2k-1}}\sinh\left(\mu^-_{2k-1}(1-|x|)\right)\right], \tag{20}$$

$$\varphi_{2k}(x) = -\frac{\rho_{2k}}{(\mu^+_{2k})^2}\left[\sin\left(\mu^+_{2k}\left(\frac{x}{|x|}-x\right)\right) - \frac{\sin\mu^+_{2k}}{\sinh\mu^-_{2k}}\sinh\left(\mu^-_{2k}\left(\frac{x}{|x|}-x\right)\right)\right], \tag{21}$$

where $\rho_k = 1 + \mathcal{O}(k^{-1})$ *and*

$$\mu^+_k = \lambda_k\sqrt{\gamma}\sqrt{\frac{1}{2}+\frac{1}{2}\sqrt{1+\frac{4}{\gamma^2\lambda_k^2}}}, \quad \mu^-_k = \frac{1}{\sqrt{\gamma}\sqrt{\frac{1}{2}+\frac{1}{2}\sqrt{1+\frac{4}{\gamma^2\lambda_k^2}}}}. \tag{22}$$

Moreover, $\partial\varphi_k(1) \neq 0$.

REMARK 3. *We observe that the eigenfunctions* φ_{2k-1} *are even while* φ_{2k} *are odd. This fact is due to the symmetry of the problem with respect to* $x = 0$. *On the other hand, as the mass only affects the first term in the asymptotic expansion of the eigenvalues corresponding to odd eigenfunctions, these are the only eigenfunctions which are affected in a significant way by the point mass.*

2.3. Asymmetric spaces. In this section we are going to introduce and characterize some asymmetric spaces. It is easy to see that these spaces are stable under the flow generated by system (1)–(2), and they are natural spaces to solve the boundary control problem.

With the notations of section 2.1 we set

$$Y_\alpha = \left\{ U = \sum_{k\in\mathbb{Z}\setminus\{0\}} a_k\bar{\phi}_k \in H \ : \right.$$

$$\left. \| U \|^2_{Y_\alpha} = \sum_{k\in\mathbb{Z}\setminus\{0\}} \left(\frac{|\, a_{2k-\sigma_k}\,|^2}{\delta_k^{4\alpha}} + \frac{|\, a_{2k} - a_{2k-\sigma_k}\,|^2}{\delta_k^{4\alpha+2}} \right) < \infty \right\}, \tag{23}$$

where $\delta_k = \lambda_{2k} - \lambda_{2k-\sigma_k}$, $\sigma_k = \operatorname{sgn} k$, i.e., $\sigma_k = 1$ if $k > 0$ and $\sigma_k = -1$ if $k < 0$.

Clearly Y_α endowed with the norm $\|\cdot\|_{Y_\alpha}$ is a Hilbert space. On the other hand, it is clear that if all the δ_k were uniformly positive and bounded above, then Y_α would coincide algebraically and topologically with one of the energy spaces $\mathcal{H}_\alpha$.

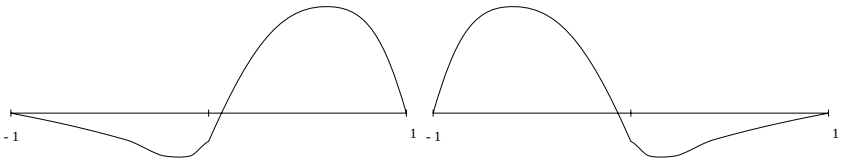


FIG. 1. $(\varphi_1 - \varphi_2)/2$ *and* $(\varphi_1 + \varphi_2)/2$.

Since, in view of Proposition 2.3, $\delta_k = \mathcal{O}(k^{-1}) = \mathcal{O}(\lambda_k^{-1}) \to 0$ as $k \to \infty$, we deduce that Y_α is a strict subspace of $\mathcal{H}_\alpha$.

We have the following result.

PROPOSITION 2.5. *Let* $U^0 = \big((u^0, y^0, z^0), (u^1, y^1, z^1)\big)$ *be an element of* Y_α. *Then, the solution* $U(t) = ((u(t), y(t), z(t)), (u_t(t), y_t(t), z_t(t)))$ *of* (1)–(2) *with initial data* U^0 *belongs to* Y_α *for every* $t > 0$ *and* $\alpha \in \mathbb{R}$. *Furthermore, for any* $T > 0$ *there exists a constant* $C(T) > 0$ *such that*

$$\| U(t) \|_{Y_\alpha} \leq C(T) \| U^0 \|_{Y_\alpha}, \quad \forall\ 0 \leq t \leq T, \quad \forall\ U^0 \in Y_\alpha. \tag{24}$$

The following theorem provides a precise characterization of the spaces Y_0 and Y_{-1}.

THEOREM 2.6. (a) Y_0 *is the subspace of elements* $U^0 = \big((u^0, y^0, z^0), (u^1, y^1, z^1)\big)$ *of* $\mathcal{H}$ *such that the restriction of* (u^0, u^1) *to* $(0,1)$ *belongs to* $H^3(0,1) \times H^2(0,1)$ *and, in addition to the compatibility conditions of* $\mathcal{H}$ $\big(u^0(0) = y^0, \partial u^0(0) = z^0, u^1(0) = y^1\big)$, *the following hold:*

$$\partial u^1(0^+) = z^1, \qquad \partial^2 u^0(1) = 0. \tag{25}$$

Furthermore, the norm $\| \cdot \|_{Y_0}$ *is equivalent to*

$$\Big[\| U \|_{\mathcal{H}}^2 + \big\| \big(u^0 \mid_{(0,1)}, u^1 \mid_{(0,1)}\big) \big\|_{H^3 \times H^2(0,1)}^2 \Big]^{1/2}.$$

(b) Y_{-1} *is the subspace of elements* $U^0 = \big((u^0, y^0, z^0), (u^1, y^1, z^1)\big)$ *of* $\mathcal{H}_{-1}$ *such that the restriction of* (u^0, u^1) *to* $(0,1)$ *belongs to* $H^1(0,1) \times L^2(0,1)$ *and verify*

$$u^0(0^+) = y^1, \qquad u^0(1) = 0. \tag{26}$$

REMARK 4. *The spaces* Y_α *are asymmetric in the sense that their elements have one more degree of regularity to the left of* $x = 0$.

The characterization of Y_α *as asymmetric spaces can be explained as follows: The vectors* $p_k = \delta_k^{2\alpha}(\bar{\phi}_{2k-1} + \bar{\phi}_{2k})/2$, $q_k = \delta_k^{2\alpha+1}(\bar{\phi}_{2k-1} - \bar{\phi}_{2k})/2$ *constitute a Riesz basis of* Y_α. *We observe that* p_k *and* q_k *are constituted by the functions* $(\varphi_{2k-1} + \varphi_{2k})/2$, $(\varphi_{2k-1} - \varphi_{2k})/2$ *weighted differently. As it can be seen in Figure* 1 *for* $k = 1$, *due to the presence of the mass, the profiles of* $(\varphi_{2k-1} + \varphi_{2k})/2$ *and* $(\varphi_{2k-1} - \varphi_{2k})/2$ *are essentially one reflection of the other with respect to* $x = 0$.

This explains the asymmetric structure of Y_α.

3. Observability. As we mentioned in the introduction, using HUM [8], the controllability of system (1) with controls of the form (6) can be reduced to the obtention of suitable observability estimates for the system in the absence of control:

$$
\left\{
\begin{array}{ll}
\gamma\partial^2 u_{tt} - u_{tt} - \partial^4 u = 0, & \text{for } x \in (-1,0),\ 0<t<T \\
\gamma\partial^2 u_{tt} - u_{tt} - \partial^4 u = 0, & \text{for } x \in (0,1),\ 0<t<T \\
[u](0,t) = [\partial u](0,t) = 0, & \text{for } 0<t<T \\
u_{tt}(0,t) + \left[\partial^3 u\right](0,t) = 0, & \text{for } 0<t<T \\
\gamma\partial u_{tt}(0,t) - \left[\partial^2 u\right](0,t) = 0, & \text{for } 0<t<T \\
u(\pm 1,t) = \partial^2 u(\pm 1,t) = 0, & \text{for } 0<t<T \\
(u(x,0), u(0,0), \partial u(0,0)) = (u^0, y^0, z^0), & \\
(u_t(x,0), u_t(0,0), \partial u_t(0,0)) = (u^1, y^1, z^1). &
\end{array}
\right. \tag{27}
$$

As in previous sections, we identify the solution u of (1) with the vector valued unknown U. The following holds.

LEMMA 3.1. *For any $T>0$, there exists $C(T)>0$ such that*

$$
\int_0^T |\partial u(1,t)|^2\, dt \le C \parallel U^0 \parallel^2_{Y_{-1}},\ \forall\ U^0 \in Y_{-1}. \tag{28}
$$

Moreover, if $T \ge 4\sqrt{\gamma}$, there exists $C(T)>0$ such that

$$
\parallel U^0 \parallel^2_{Y_{-1}} \le C \int_0^T |\partial u(1,t)|^2\, dt,\ \forall\ U^0 \in Y_{-1}. \tag{29}
$$

REMARK 5. *The first estimate* (28) *of this lemma establishes a hidden regularity result since, in view of Proposition* 2.5 *and Theorem* 2.6, *the fact that $U^0 \in Y_{-1}$ implies $u\big|_{(0,1)} \in C\big([0,T]; H^1(0,1)\big) \cap C^1\big([0,T]; L^2(0,1)\big)$, but this is not sufficient to guarantee that $\partial u(1,t) \in L^2(0,T)$.*

The second estimate (29) *of the lemma guarantees that the norm of the initial data in Y_{-1} can be observed continuously in terms of the $L^2(0,T)$-norm of $\partial u(1,t)$. In view of Proposition* 2.5 *this implies that the norm of the solution U in $C([0,T], Y_{-1})$ can be observed, also. Due to the finite speed of propagation of the system, the time required for the uniform observability* (29) *has to be large enough. The lower bound $4\sqrt{\gamma}$ is sharp.*

In order to prove Lemma 3.1, we need the following result on nonharmonic Fourier series due to Ulrich [10].

THEOREM 3.2. *Let $(\sigma_n)_{n\in\mathbb{Z}}$ and $(\tau_n)_{n\in\mathbb{Z}}$ be two sequences of distinct complex numbers such that $\sigma_n \ne \tau_n$ for all $n \in \mathbb{Z}$ and*

$$
\lim_{|n|\to\infty} |\sigma_n - n| = \lim_{|n|\to\infty} |\tau_n - n| = 0. \tag{30}
$$

Then $\left\{e^{i\sigma_n t}\right\}_{n\in\mathbb{Z}}$ forms a Riesz basis of $L^2(0,2\pi)$ and moreover,

$$
\left\{e^{i\sigma_n t}\right\}_{n\in\mathbb{Z}} \cup \left\{\frac{e^{i\sigma_n t} - e^{i\tau_n t}}{\sigma_n - \tau_n}\right\}_{n\in\mathbb{Z}} \tag{31}
$$

forms a Riesz basis of $L^2(0,4\pi)$.

REMARK 6. *We refer to* [3] *and* [1] *for a generalization of this result.*

As an immediate consequence of this result the following holds.

COROLLARY 3.3. *Let $\gamma > 0$. Then, if $\{\lambda_k^2\}_{k\in\mathbb{Z}}$ denote the eigenvalues of system* (27)*,*

$$\{e^{i\lambda_{2k}t}\}_{k\in\mathbb{Z}} \cup \left\{\frac{e^{i\lambda_{2k}t} - e^{i\lambda_{2k-\sigma_k}t}}{\lambda_{2k} - \lambda_{2k-\sigma_k}}\right\}_{k\in\mathbb{Z}} \tag{32}$$

form a Riesz basis of $L^2\left(0, 4\sqrt{\gamma}\right)$.

Proof. We introduce the change of variables $s = t\pi/\sqrt{\gamma}$ that transforms the functions in (32) into

$$e^{-is/2}\left\{e^{i\left(\lambda_{2k}\sqrt{\gamma}/\pi+1/2\right)s}\right\}_{k\in\mathbb{Z}} \cup e^{-is/2}\left\{\frac{e^{i\left(\lambda_{2k}\sqrt{\gamma}/\pi+1/2\right)s} - e^{i\left(\lambda_{2k-\sigma_k}\sqrt{\gamma}/\pi+1/2\right)s}}{\delta_k}\right\}_{k\in\mathbb{Z}},$$

and the interval $t \in \left(0, 4\sqrt{\gamma}\right)$ into $s \in (0, 4\pi)$. Obviously the common multiplicative factor $e^{-is/2}$ does not affect whether or not these functions constitute a Riesz basis of $L^2(0, 4\pi)$.

By setting

$$\tau_k = \lambda_{2k}\frac{\sqrt{\gamma}}{\pi} + \frac{1}{2}, \qquad \sigma_k = \lambda_{2k-\sigma_k}\frac{\sqrt{\gamma}}{\pi} + \frac{1}{2},$$

we are in the conditions of Theorem 3.2 in view of the asymptotic form of the eigenvalues proved in Propositions 2.3.

Undoing the change of variables, we deduce that Corollary 3.3 holds. □

The following characterization of Riesz basis (which can be seen in Young [11]) will also be used.

THEOREM 3.4. *Let H be a separable Hilbert space. The following two properties are equivalent:*

(a) *$\{e_n\}_{n\in\mathbb{Z}}$ forms a Riesz basis of H;*

(b) *$\{e_n\}_{n\in\mathbb{Z}}$ is a complete sequence in H and there exists two positive constants $A, B > 0$ such that*

$$A\sum_{i=1}^{n} |\, c_i \,|^2 \leq \left\|\sum_{i=1}^{n} c_i e_i\right\|_H^2 \leq B\sum_{i=1}^{n} |\, c_i \,|^2$$

for any $n \in \mathbb{N}$ and $c_i, i = 1, \cdots, n$.

We are now ready to prove Lemma 3.1.

Proof of Lemma 3.1. Recall that the solution u of (27) can be written as

$$(u, u(0), \partial u(0)) = \sum_{k\in\mathbb{Z}\setminus\{0\}} a_k e^{i\lambda_k t}\phi_k(x),$$

where $\phi_k = (\varphi_k, \varphi_k(0), \partial\varphi_k(0))$ are the eigenfunctions of the eigenvalue problem (8) with φ_k normalized in $H^2 \cap H_0^1(-1, 1)$ such that $(-1)^k\partial\varphi_k(1) > 0$.

The eigenfunctions are of the form

$$\varphi_{2k-\sigma_k}(x) = \frac{\rho_{2k-\sigma_k}}{\left(\mu^+_{2k-\sigma_k}\right)^2}\left[\sin\left(\mu^+_{2k-\sigma_k}(1-x)\right) - \frac{\mu^+_{2k-\sigma_k}\cos\mu^+_{2k-\sigma_k}}{\mu^-_{2k-\sigma_k}\cosh\mu^-_{2k-\sigma_k}}\sinh\left(\mu^-_{2k-\sigma_k}(1-x)\right)\right],$$

$$\varphi_{2k}(x) = -\frac{\rho_{2k}}{\left(\mu^+_{2k}\right)^2}\left[\sin\left(\mu^+_{2k}(1-x)\right) - \frac{\sin\left(\mu^+_{2k}\right)}{\sinh\left(\mu^-_{2k}\right)}\sinh\left(\mu^-_{2k}(1-x)\right)\right]$$

in the interval $(0, 1)$.

We introduce the following change in the Fourier coefficients of the solutions: $\widetilde{a}_k = a_k \lambda_k \, |\partial\varphi_k(1)|$. Then,

$$u(x,t) = \sum_{k\in\mathbb{Z}\setminus\{0\}} \frac{\widetilde{a}_k}{\lambda_k \mid \partial\varphi_k(1) \mid} e^{i\lambda_k t}\varphi_k(x).$$

Observe that $\{a_k\} \in \ell^2$ if and only if $\{\widetilde{a}_k\} \in \ell^2$. For that it is sufficient to see that there exist positive constants $A, B > 0$ such that $A \leq \mid \lambda_k \mid\mid \partial\varphi_k(1) \mid \leq B$ for all k. To do that, we observe that

$$|\lambda_{2k-\sigma_k}\partial\varphi_{2k-\sigma_k}(1)| = \frac{\lambda_{2k-\sigma_k}\rho_{2k-\sigma_k}}{\left(\mu^+_{2k-\sigma_k}\right)^2}\left(\mu^+_{2k-\sigma_k} - \mu^+_{2k-\sigma_k}\frac{\cos\mu^+_{2k-\sigma_k}}{\cosh\mu^-_{2k-\sigma_k}}\right)$$

$$= \frac{1}{\sqrt{\gamma}} + O(k^{-1}), \tag{33}$$

$$|\lambda_{2k}\partial\varphi_{2k}(1)| = \frac{\lambda_{2k}\rho_{2k}}{\left(\mu^+_{2k}\right)^2}\left(\mu^+_{2k} - \frac{\sin\left(\mu^+_{2k}\right)\mu^-_{2k}}{\sinh\left(\mu^-_{2k}\right)}\right) = \frac{1}{\sqrt{\gamma}} + O(k^{-1}) \tag{34}$$

in view of (22) and the asymptotic results of section 2.2. Thus $\mid \lambda_k\partial\varphi_k(1) \mid \to 1/\sqrt{\gamma}$ as $\mid k \mid \to \infty$ and, on the other hand, $\mid \lambda_k\partial\varphi_k(1) \mid \neq 0$ for all $k \neq 0$. Therefore constants $A, B > 0$ exist.

Consequently

$$A \sum_{k\in\mathbb{Z}\setminus\{0\}} \mid a_k \mid^2 \leq \sum_{k\in\mathbb{Z}\setminus\{0\}} \mid \widetilde{a}_k \mid^2 \leq B \sum_{k\in\mathbb{Z}\setminus\{0\}} \mid a_k \mid^2. \tag{35}$$

On the other hand, the norm in the asymmetric space Y_{-1} can also be written in an equivalent form in terms of the coefficients $\widetilde{a}_k$. Indeed,

$$\begin{aligned} &A^2\left(\sum_{k\in\mathbb{Z}\setminus\{0\}} \delta_k^4 \, |a_{2k-\sigma_k}|^2 + \sum_{k\in\mathbb{Z}\setminus\{0\}} \delta_k^2 \mid a_{2k} - a_{2k-\sigma_k} \mid^2\right) \\ &\quad\leq \sum_{k\in\mathbb{Z}\setminus\{0\}} \delta_k^4 \, |\widetilde{a}_{2k-\sigma_k}|^2 + \sum_{k\in\mathbb{Z}\setminus\{0\}} \delta_k^2 \mid \widetilde{a}_{2k} - \widetilde{a}_{2k-\sigma_k} \mid^2 \\ &\quad\leq B^2\left(\sum_{k\in\mathbb{Z}\setminus\{0\}} \delta_k^4 \mid a_{2k-\sigma_k} \mid^2 + \sum_{k\in\mathbb{Z}\setminus\{0\}} \delta_k^2 \mid a_{2k} - a_{2k-\sigma_k} \mid^2\right). \end{aligned} \tag{36}$$

We start with the second inequality:

$$\begin{aligned} &\sum_{k\in\mathbb{Z}\setminus\{0\}} \delta_k^2 \mid \widetilde{a}_{2k} - \widetilde{a}_{2k-\sigma_k} \mid^2 \\ &\quad= \sum_{k\in\mathbb{Z}\setminus\{0\}} \delta_k^2 \mid a_{2k}\lambda_{2k} \mid \partial\varphi_{2k}(1) \mid - a_{2k-\sigma_k}\lambda_{2k-\sigma_k} \mid \partial\varphi_{2k-\sigma_k}(1) \mid\mid^2 \\ &\quad\leq 4 \sum_{k\in\mathbb{Z}\setminus\{0\}} \left[\mid a_{2k} \mid^2 \delta_k^4 \left|\frac{\lambda_{2k} \mid \partial\varphi_{2k}(1) \mid -1/\sqrt{\gamma}}{\delta_k}\right|^2\right. \\ &\quad\left. + \mid a_{2k-\sigma_k} \mid^2 \delta_k^4 \left|\frac{\lambda_{2k-\sigma_k} \mid \partial\phi_{2k-\sigma_k}(1) \mid -1/\sqrt{\gamma}}{\delta_k}\right|^2\right] \end{aligned}$$

$$
\begin{aligned}
&+\frac{2}{\gamma}\sum_{k\in\mathbb{Z}\setminus\{0\}}\delta_k^2|\ a_{2k}-a_{2k-\sigma_k}\ |^2\\
&\le C\left\{\sum_{k\in\mathbb{Z}}\delta_k^4\,|a_{2k-\sigma_k}|^2+\sum_{k\in\mathbb{Z}\setminus\{0\}}\delta_k^4\mid a_{2k}\mid^2\right\}\\
&+\frac{2}{\gamma}\sum_{k\in\mathbb{Z}\setminus\{0\}}\delta_k^2|\ a_{2k}-a_{2k-\sigma_k}\ |^2 \qquad (37)
\end{aligned}
$$

since, in view of (33)–(34), $\mid \lambda_{2k}\mid \partial\varphi_{2k}(1)\mid -1/\sqrt{\gamma}\mid$, $\mid \lambda_{2k-\sigma_k}\mid \partial\varphi_{2k-\sigma_k}(1)\mid -1/\sqrt{\gamma}\mid$, and δ_k are of the order of $1/k$.

Clearly the last term in (37) can be bounded in terms of

$$\sum_{k\in\mathbb{Z}\setminus\{0\}}\delta_k^4\mid a_{2k-\sigma_k}\mid^2+\sum_{k\in\mathbb{Z}\setminus\{0\}}\delta_k^2\mid a_{2k}-a_{2k-\sigma_k}\mid^2.$$

The first inequality in (36) can be proved in a similar way.

Now, taking into account that $\partial u(1,t)=\sum_{k\in\mathbb{Z}\setminus\{0\}}a_k\partial\varphi_k(1)e^{i\lambda_k t}$ we deduce that

$$
\begin{aligned}
&\int_0^T|\partial u(1,t)|^2\,dt=\int_0^T\left|\sum_{k\in\mathbb{Z}\setminus\{0\}}a_k\partial\varphi_k(1)e^{i\lambda_k t}\right|^2dt=\int_0^T\left|\sum_{k\in\mathbb{Z}\setminus\{0\}}(-1)^k\frac{\widetilde{a}_k}{\lambda_k}e^{i\lambda_k t}\right|^2dt\\
&=\int_0^T\left|\sum_{k\in\mathbb{Z}\setminus\{0\}}\left[\left(\frac{\widetilde{a}_{2k}}{\lambda_{2k}}-\frac{\widetilde{a}_{2k-\sigma_k}}{\lambda_{2k-\sigma_k}}\right)e^{i\lambda_{2k}t}+\delta_k\left(\frac{\widetilde{a}_{2k-\sigma_k}}{\lambda_{2k-\sigma_k}}\right)\frac{e^{i\lambda_{2k}t}-e^{i\lambda_{2k-\sigma_k}t}}{\delta_k}\right]\right|^2dt\\
&\le C\sum_{k\in\mathbb{Z}\setminus\{0\}}\left(\left|\frac{\widetilde{a}_{2k}}{\lambda_{2k}}-\frac{\widetilde{a}_{2k-\sigma_k}}{\lambda_{2k-\sigma_k}}\right|^2+\delta_k^2\left|\frac{\widetilde{a}_{2k-\sigma_k}}{\lambda_{2k-\sigma_k}}\right|^2\right)\\
&\le C\sum_{k\in\mathbb{Z}\setminus\{0\}}\left(\frac{|\widetilde{a}_{2k}-\widetilde{a}_{2k-\sigma_k}|^2}{\lambda_{2k}^2}+|\widetilde{a}_{2k-\sigma_k}|^2\left|\frac{1}{\lambda_{2k}}-\frac{1}{\lambda_{2k-\sigma_k}}\right|^2+\frac{\delta_k^2}{\lambda_{2k-\sigma_k}}\,|\widetilde{a}_{2k-\sigma_k}|^2\right)\\
&\le C\sum_{k\in\mathbb{Z}\setminus\{0\}}\left(\delta_k^2\,|\widetilde{a}_{2k}-\widetilde{a}_{2k-\sigma_k}|^2+\delta_k^4\,|\widetilde{a}_{2k-\sigma_k}|^2\right)\le C\parallel U^0\parallel_Y^2
\end{aligned}
$$

in view of Corollary 3.3, inequalities (36), and the fact that $\lambda_{2k}\sim 1/\delta_k$ and $\lambda_{2k-\sigma_k}\sim 1/\delta_k$.

A similar computation shows that (29) holds, also.

Inequalities (36) imply the statement of Lemma 3.1 in view of Theorem 3.4. □

4. On the solvability of the system with nonhomogeneous data. In this section we analyze the existence, uniqueness, and regularity of nonhomogeneous boundary value problems that appear when addressing the control problem.

In order to state the main results of this section it is convenient to introduce the following asymmetric spaces:

$$V_0^+=\left\{(\varphi,\eta,\xi)\in H_{-3/2}\ :\ \varphi|_{(0,1)}\in L^2(0,1)\right\}; \qquad (38)$$

$$V_1^+=\left\{(\varphi,\eta,\xi)\in H_{-1}\ :\ \varphi|_{(0,1)}\in H^1(0,1),\ \varphi(0^+)=\eta,\ \varphi(1)=0\right\}. \qquad (39)$$

In these notations the superscripts + indicate that the elements of these spaces are more regular to the right of $x = 0$, while the subscripts 0 (resp., 1) indicates that the maximal regularity is L^2 (resp., H^1).

REMARK 7. *In view of the characterization of* Y_{-1} *given in Theorem* 2.6, *we observe that* $Y_{-1} = V_1^+ \times V_0^+$.

This section is divided in two parts. In the first one we analyze systems with nonzero right-hand side terms. In the second one we address nonhomogeneous boundary value problems by transposition.

4.1. Systems with nonzero right-hand side. Let us consider the nonhomogeneous system:

$$
\left\{
\begin{array}{l}
\gamma\partial^2 u_{tt} - u_{tt} - \partial^4 u = f,\ x \in (-1,0), 0 < t < T,\\
\gamma\partial^2 u_{tt} - u_{tt} - \partial^4 u = f,\ x \in (0,1), 0 < t < T,\\
[u](0,t) = [\partial u](0,t) = 0,\ 0 < t < T,\\
u_{tt}(0,t) + \left[\partial^3 u\right](0,t) = g,\ 0 < t < T,\\
\gamma\partial u_{tt}(0,t) - \left[\partial^2 u\right](0,t) = h,\ 0 < t < T,\\
u(\pm 1,t) = \partial^2 u(\pm 1,t) = 0,\ 0 < t < T,\\
(u(x,0), u(0,0), \partial u(0,0)) = \left(u^0, y^0, z^0\right),\\
(u_t(x,0), u_t(0,0), \partial u_t(0,0)) = (u^1, y^1, z^1).
\end{array}
\right. \tag{40}
$$

Observe that the boundary conditions at $x = \pm 1$ vanish.

We have the following result.

THEOREM 4.1. *Assume that* $U^0 = \left((u^0, y^0, z^0), (u^1, y^1, z^1)\right) \in H_{-1/2} \times H_{-1}$ *and*

$$
((1 - \gamma\partial^2)^{-1} f, g, h) \in L^2\left(0, T; H_{-1}\right). \tag{41}
$$

Then, there exists a unique solution $U \in C\left([0,T]; H_{-1/2} \times H_{-1}\right)$ *of* (40).

Moreover, there exists $C(T) > 0$ *such that*

$$
\begin{aligned}
&\int_0^T |\partial u(1,t)|^2\, dt\\
&\qquad \le C\left(\| ((1 - \gamma\partial^2)^{-1} f, g, h) \|^2_{L^1(0,T;H_0(-1,1))} + \| U^0 \|^2_{H_{-1/2}\times H_{-1}}\right)
\end{aligned} \tag{42}
$$

for all U^0 *and* (f, g, h) *as above.*

REMARK 8. *The first result of this theorem is classical and provides the existence of solutions with values in* $H_{-1/2} \times H_{-1}$, *which is a natural symmetric energy space for solving* (40).

Inequality (42) *extends the "hidden regularity" result of Lemma* 3.1 *to the solutions of the nonhomogeneous system. However,* (42) *is not sharp since it requires the same degree of regularity at both sides of* $x = 0$, *while in Lemma* 3.1 *this degree of regularity is only required on* $(0,1)$. *The next theorem provides a sharp result.*

THEOREM 4.2. *Assume that* $U^0 \in Y_{-1}$ *and* $((1 - \gamma\partial^2)^{-1} f, g, h) \in L^1(0,T; V_0^+)$. *Then, there exists a unique solution* $U \in C\left([0,T]; Y_{-1}\right)$ *of* (40).

Moreover, there exists $C(T) > 0$ *such that*

$$
\int_0^T |\partial u(1,t)|^2\, dt \le C\left[\, \| ((1 - \gamma\partial^2)^{-1} f, g, h) \|^2_{L^1(0,T;V_0^+)} + \| U^0 \|^2_{Y_{-1}} \,\right], \tag{43}
$$

for every U^0 *and* (f, g, h) *as above.*

The proof of both theorems is rather similar. For simplicity we only prove Theorem 4.2.

Proof of Theorem 4.2. Taking into account that the system is linear, and in view of Lemma 3.1, it is sufficient to prove it when $U^0 \equiv 0$.

To simplify the notation we identify $(I-\gamma\partial^2)^{-1}f$ and the vector valued function $\big((I-\gamma\partial^2)^{-1}f, g, h\big)$.

We observe that $(I-\gamma\partial^2)^{-1}f \in L^1(0,T;V_0^+)$ and then, in view of Remark 7, we have $\big(0,(I-\gamma\partial^2)^{-1}f\big) \in L^1(0,T;Y_{-1})$.

On the other hand, composing system (27) with the operator $(I-\gamma\partial^2)^{-1}$ and identifying the unknown vector $(u,u(0),\partial u(0))$ with u, it can be written as

$$u_{tt}+Au=(I-\gamma\partial^2)^{-1}f,$$

where $A=K^{-1}$ is the underlying elliptic operator.

Since $U^0\equiv 0$, by the variation of constants formula $u(t)=\int_0^t v(t-s;s)ds$, where $v(\cdot,\cdot;s)$ satisfies

$$\left\{\begin{array}{l} v_{tt}+Av=0,\\ v(0;s)=0, \quad v_t(0;s)=(I-\gamma\partial^2)^{-1}f(s).\end{array}\right. \tag{44}$$

In view of Proposition 2.5 it is easy to see that u or, more precisely, its corresponding vector valued solution U, belongs to $C([0,T];Y_{-1})$.

On the other hand, in view of Lemma 3.1, we have

$$\begin{aligned}\int_0^T |\partial v(1,t;s)|^2\,dt &\le C\left\|\left(0,\left[(I-\gamma\partial^2)^{-1}f\right](s)\right)\right\|^2_{Y_{-1}}\\ &= C\parallel\left[(I-\gamma\partial^2)^{-1}f\right](s)\parallel^2_{V_0^+},\quad \forall\, s\in[0,T].\end{aligned}$$

By Minkowski's inequality we deduce that

$$\begin{aligned}\parallel\partial u(1,t)\parallel_{L^2(0,T)} &= \left\|\int_0^t \partial v(1,t-s;s)ds\right\|_{L^2(0,T)}\\ &\le C\left\|(I-\gamma\partial^2)^{-1}f\right\|_{L^1(0,T;V_0^+)}.\end{aligned}$$

□

The following result is also needed.

THEOREM 4.3. *Assume that* $U^0\equiv 0$ *and* $(f,g,h)=\partial_t(F,G,H)$ *satisfying* $((I-\gamma\partial^2)^{-1}F,G,H)\in L^1\left(0,T;V_1^+\right)$. *Then the solution of* (40) *verifies* $U\in C([0,T];Y_{-1})$. *Moreover, there exists* $C(T)>0$ *such that*

$$\int_0^T |\partial u(1,t)|^2\,dt \le C\parallel((I-\gamma\partial^2)^{-1}F,G,H)\parallel^2_{L^1(0,T;V_1^+)} \tag{45}$$

for every (F,G,H) *as above.*

Proof of Theorem 4.3. As in Theorem 4.2 above, we identify $(I-\gamma\partial^2)^{-1}F$ (resp., $(I-\gamma\partial^2)^{-1}F_t$) with $((I-\gamma\partial^2)^1F,G,H)$ (resp., $(I-\gamma\partial^2)^{-1}F_t,G_t,H_t)$) to simplify the notation.

On the other hand, $u=v_t$ where v, which is also identified with the unknown vector $(v,v(0),\partial v(0))$, is the solution of

$$\left\{\begin{array}{l} v_{tt}+Av=(I-\gamma\partial^2)^{-1}F,\\ v(0)=v_t(0)=0.\end{array}\right. \tag{46}$$

Observe that we have taken null initial data for v. This is due to the fact that in this proof we may assume $((1-\gamma\partial^2)^{-1}F, G, H)$ to be of compact support in time. Indeed, if Theorem 4.3 is proved for those $((1-\gamma\partial^2)^{-1}F, G, H)$, it can then be extended by density to all $((1-\gamma\partial^2)^{-1}F, G, H) \in L^1(0,T;V_1^+)$.

With this in mind we see that the appropriate initial conditions for v are as follows:

$$v(0) = u_t(0) = 0;\ v_t(0) = u_{tt}(0) = (I-\gamma\partial^2)^{-1}F(0) - Au(0) = 0.$$

To complete the proof of Theorem 4.3 it is sufficient to prove that the following lemma holds.

LEMMA 4.4. *Assume that* $U^0 = 0$ *and* $((1-\gamma\partial^2)^{-1}f, g, h) \in L^1(0,T;V_1^+)$. *Then, there exists* $C > 0$ *such that the solution of* (40) *satisfies*

$$\int_0^T |\partial u_t(1,t)|^2\, dt \le C \parallel ((1-\gamma\partial^2)^{-1}f, g, h) \parallel^2_{L^1(0,T;V_1^+)} \tag{47}$$

for all (f, g, h) *as above.*

Proof of Lemma 4.4. First of all we observe that, in view of the characterization of the asymmetric space Y_{-1} given in Theorem 2.6, it is easy to see that $((I-\gamma\partial^2)^{-1}f, 0) \in L^1(0,T;Y_{-1})$. Note that we identify $(I-\gamma\partial^2)^{-1}f$ with the vector $((I-\gamma\partial^2)^{-1}f, g, h)$ to simplify the notation.

As in Theorem 4.2 above, $u = \int_0^t v(x, t-s; s)ds$ where v solves (44). Then $\omega = v_t$ verifies

$$\begin{cases} \omega_{tt} + A\omega = 0, \\ \omega(0;s) = (I-\gamma\partial^2)^{-1}f(s), \quad \omega_t(0;s) = 0. \end{cases} \tag{48}$$

In view of Lemma 3.2, we have

$$\int_0^T |\partial\omega(1,t)|^2\, dt \le C \parallel ((I-\gamma\partial^2)^{-1}f, 0) \parallel^2_{Y_{-1}} = C \parallel (I-\gamma\partial^2)^{-1}f \parallel^2_{V_1^+}. \tag{49}$$

On the other hand,

$$\partial u_t = \int_0^t \partial v_t(t-s;s)ds = \int_0^t \partial\omega(t-s;s)ds,$$

and therefore,

$$\int_0^T |\partial u_t(1,t)|^2\, dt = \left\| \int_0^t \partial\omega(1, t-s; s)ds \right\|^2_{L^2_t(0,T)}. \tag{50}$$

Now, by Minkowski's inequality and (49) we deduce that

$$\left\| \int_0^t \partial\omega(1, t-s; s)ds \right\|_{L^2_t(0,T)} \le C \parallel (I-\gamma\partial^2)^{-1}f \parallel_{L^1(0,T;V_1^+)}. \tag{51}$$

Combining (50)–(51) we deduce that (47) holds.

This concludes the proof of Lemma 4.4 and Theorem 4.3. ∎

4.2. Nonhomogeneous boundary conditions. Let us consider now the system

$$
\left\{
\begin{array}{l}
\gamma\partial^2 u_{tt} - u_{tt} - \partial^4 u = 0,\ x \in (-1,0),\ 0 < t < T,\\
\gamma\partial^2 u_{tt} - u_{tt} - \partial^4 u = 0,\ x \in (0,1),\ 0 < t < T,\\
[u](0,t) = [\partial u](0,t) = 0,\ 0 < t < T,\\
u_{tt}(0,t) + \left[\partial^3 u\right](0,t) = 0,\ 0 < t < T,\\
\partial u_{tt}(0,t) - \left[\partial^2 u\right](0,t) = 0,\ 0 < t < T,\\
u(\pm 1,t) = \partial^2 u(-1,t) = 0,\ \partial^2 u(1,t) = q(t),\ 0 < t < T,\\
(u(x,0), u(0,0), \partial u(0,0)) = \left(u^0, y^0, z^0\right),\\
(u_t(x,0), u_t(0,0), \partial u_t(0,0)) = \left(u^1, y^1, z^1\right).
\end{array}
\right. \tag{52}
$$

Observe that the boundary conditions in (52) are nonhomogeneous since the boundary condition $\partial^2 u(1,t)$ takes the value q.

We assume that

$$
q \in L^2(0,T) \tag{53}
$$

and

$$
\begin{aligned}
U^0 \in Y^- = \big\{ U^0 \in H_0 \times H_{-1/2}\ :\ & u^0|_{(-1,0)} \in H^3(-1,0),\ \partial^2 u^0(-1) = 0,\\
& u^1|_{(-1,0)} \in H^2(-1,0),\ \partial u^1(0^-) = z^1 \big\}.
\end{aligned} \tag{54}
$$

Observe that Y^- is the reflection of the space Y_0 (characterized in Theorem 2.6) with respect to $x = 0$. In other words, $U^0 \in Y^-$ if and only if $V^0(x) = U^0(-x) \in Y_0$.

In view of Proposition 2.5, and taking into account that system (52) is symmetric with respect to $x = 0$ when $q = 0$, the following holds.

LEMMA 4.5. *When $q \equiv 0$ and $U^0 \in Y^-$, system (6.16) admits a unique solution $U \in C([0,T]; Y^-)$. Moreover, there exists $C(T) > 0$ such that*

$$
\| U \|_{L^\infty(0,T;Y^-)} \leq C(T) \| U^0 \|_{Y^-},\ \forall\ U^0 \in Y^-. \tag{55}
$$

As a consequence of this lemma, and since system (52) is linear, it is sufficient to analyze solutions of (52) when $U^0 \equiv 0$ and q is as in (53). Therefore in the following we assume that $U^0 \equiv 0$.

Solutions of (52) can be understood in the sense of transposition. To make this notion precise we consider the adjoint system

$$
\left\{
\begin{array}{l}
\gamma\partial^2 \varphi_{tt} - \varphi_{tt} - \partial^4 \varphi = f,\ x \in (-1,0),\ 0 < t < T,\\
\gamma\partial^2 \varphi_{tt} - \varphi_{tt} - \partial^4 \varphi = f,\ x \in (0,1),\ 0 < t < T,\\
[\varphi](0,t) = [\partial\varphi](0,t) = 0,\ 0 < t < T,\\
\varphi_{tt}(0,t) + \left[\partial^3 \varphi\right](0,t) = g(t),\ 0 < t < T,\\
\gamma\partial\varphi_{tt}(0,t) - \left[\partial^2 \varphi\right](0,t) = h(t),\ 0 < t < T,\\
\varphi(\pm 1,t) = \partial^2 \varphi(\pm 1,t) = 0,\ 0 < t < T,\\
(\varphi(x,T), \varphi(0,T), \partial\varphi(0,T)) = (\varphi_t(x,T), \varphi_t(0,T), \partial\varphi_t(0,T)) \equiv 0.
\end{array}
\right. \tag{56}
$$

Given $((I - \gamma\partial^2)^{-1} f, g, h) \in L^1(0,T; V_0^+)$, and taking into account that system (56) is time reversible, in view of Theorem 4.2 we deduce that (56) admits a unique solution $\Phi \in C([0,T]; Y_{-1})$ (by Φ we denote the vector valued unknown associated with φ).

On the other hand, $\partial\varphi_t(1,t) \in L^2(0,T)$.

Multiplying in (52) by φ and integrating by parts we get, at least formally, the following identity:

$$\int_0^T < (I-\gamma\partial^2)u, (I-\gamma\partial^2)^{-1}f >_{H_0^1,H^{-1}} dt + \int_0^T y(t)g(t)dt + \int_0^T z(t)h(t)dt$$

$$= \int_0^T \partial\varphi(1,t)q(t)dt, \tag{57}$$

where $y(t) = u(0,t)$, $z(t) = \partial u(0,t)$, and $<,>_{H_0^1,H^{-1}}$ represents the duality product between $H_0^1(-1,1)$ and H^{-1}.

Note that we have used the self-adjointness of $(I-\gamma\partial^2)^{-1}$ in the identity

$$\int_{-1}^1 uf = \int_{-1}^1 (I-\gamma\partial^2)u(I-\gamma\partial^2)^{-1}f =< (I-\gamma\partial^2)u, (I-\gamma\partial^2)^{-1}f >_{H_0^1,H^{-1}}.$$

We adopt (57) as a definition of solution of (52) when $U^0 \equiv 0$.

DEFINITION 4.6. *We say that $U \in C(0,T;Y^-)$ is a solution of* (52) *with $U^0 \equiv 0$ in the sense of transposition if* (57) *holds for any $((I-\gamma\partial^2)^{-1}f, g, h) \in L^1(0,T;V_0^+)$.*

REMARK 9. *As we will see below, solutions in the sense of transposition are more regular on the left-hand side of $x = 0$. They satisfy*

$$u|_{(-1,0)} \in C\left([0,T];H^3(-1,0)\right) \cap C^1\left([0,T];H^2(-1,0)\right) \tag{58}$$

and the compatibility conditions

$$\partial u_t(0^-,t) = z_t(t), \quad \partial^2 u(-1) = 0. \tag{59}$$

Observe that the initial condition $U^0 \equiv 0$ is implicit in (57).

THEOREM 4.7. *When $U^0 \equiv 0$ and q is as in* (53)*, system* (52) *admits a unique solution in the sense of transposition.*

Moreover, there exists $C(T) > 0$ such that

$$\| U \|_{C([0,T];Y^-)} \leq C(T) \| q \|_{L^2(0,T)}, \tag{60}$$

for every q as above.

Proof of Theorem 4.7. In view of Theorem 4.2, the right-hand side of (57) defines a linear continuous operator in $L^1(0,T;V_0^+)$. Therefore, by duality we deduce that there exists a unique $(u,y,z) \in L^\infty\left(0,T;(V_0^+)'\right)$ solution of (57) where $\left(V_0^+\right)'$ denotes the dual of V_0^+. Moreover, there exists $C > 0$ such that

$$\| ((I-\gamma\partial^2)u, y, z) \|_{L^\infty\left(0,T;(V_0^+)'\right)} \leq C \| q \|_{L^2(0,T)}. \tag{61}$$

Furthermore, as a consequence of Theorem 4.2 we deduce that $(u_t, y_t, z_t) \in L^\infty(0,T;(V_1^+)')$, and the estimate

$$\left\| \left((I-\gamma\partial^2)u_t, y_t, z_t\right) \right\|_{L^\infty\left(0,T;\left(V_1^+\right)'\right)} \leq C \| q \|_{L^2(0,T)} \tag{62}$$

holds for every q as in (53).

Observe now that the duals of V_0^+, $\left(V_0^+\right)'$, and V_1^+ coincide, resp., with the spaces

$$V_1^- = \left\{(v,y,z) \in L^2(-1,1) \times \mathbb{R} \times \mathbb{R} \ : v|_{(-1,0)} \in H^1(-1,0),\ v(-1) = 0,\right.$$
$$\left.[(I-\gamma\partial^2)^{-1}v](0) = y,\ [\partial(I-\gamma\partial^2)^{-1}v](0) = z\right\},$$
$$V_0^- = \left\{(v,y,z) \in H^{-1}(-1,1) \times \mathbb{R} \times \mathbb{R} \ :\ v|_{(-1,0)} \in L^2(-1,0),\right.$$
$$\left.[(1-\gamma\partial^2)^{-1}v](0) = y,\ [\partial(1-\gamma\partial^2)^{-1}v](0^-) = z\right\}.$$

On the other hand, $V_1^- \times V_0^-$ is the image of the space Y^- by the operator $\mathcal{L}$ defined as

$$\mathcal{L}((u^0,y^0,z^0),(u^1,y^1,z^1)) = (((I-\gamma\partial^2)u^0,y^0,z^0),((I-\gamma\partial^2)u^1,y^1,z^1)). \tag{63}$$

We observe that $\mathcal{L}$ is in fact an isomorphism from Y^- to $V_1^- \times V_0^-$.

With the above considerations and (61)–(62), we deduce

$$\| U \|_{L^\infty(0,T;Y^-)} \leq C(T) \| q \|_{L^2(0,T)} .$$

Now by density it follows that (60) holds. To see this, it is sufficient to observe that when q is smooth enough and of compact support, solutions of (52) belong to $C\left([0,T];H_{1/2}\times H_0\right)$. □

5. Controllability. In this section we prove the main controllability result for the system

$$\begin{cases} \gamma\partial^2 u_{tt} - u_{tt} - \partial^4 u = 0,\ x\in(-1,0),\ 0<t<T, \\ \gamma\partial^2 u_{tt} - u_{tt} - \partial^4 u = 0,\ x\in(0,1),\ 0<t<T, \\ [u](0,t) = [\partial u](0,t) = 0,\ 0<t<T, \\ u_{tt}(0,t) + \left[\partial^3 u\right](0,t) = 0,\ 0<t<T, \\ \gamma\partial u_{tt}(0,t) - \left[\partial^2 u\right](0,t) = 0,\ 0<t<T, \\ u(\pm1,t) = \partial^2 u(-1,t) = 0,\ \partial^2 u(1,t) = q(t),\ 0<t<T, \\ (u(x,0),u(0,0),\partial u(0,0)) = \left(u^0,y^0,z^0\right), \\ (u_t(x,0),u_t(0,0),\partial u_t(0,0)) = \left(u^1,y^1,z^1\right). \end{cases} \tag{64}$$

The following holds.

THEOREM 5.1. *Assume that $T \geq 4\sqrt{\gamma}$. Then for every $((u^0,y^0,z^0),(u^1,y^1,z^1)) \in Y^-$ there exists a control $q \in L^2(0,T)$ such that the solution of* (64) *in the sense of transposition satisfies*

$$((u(x,T),u(0,T),\partial u(0,T)),(u_t(x,T),u_t(0,T),\partial u_t(0,T))) \equiv 0. \tag{65}$$

Moreover, there exists $C>0$ such that

$$\| q \|_{L^2(0,T)} \leq C \| U^0 \|_{Y^-},\ \forall\ U^0 \in Y^-. \tag{66}$$

REMARK 10. *Theorem* 5.1 *states the exact controllability of* (64) *in the space Y^- with controls in $L^2(0,T)$, provided $T \geq 4\sqrt{\gamma}$.*

The functional frame we have chosen for the control problem ($U^0 \in Y^-$ and $q \in L^2(0,T)$) is not unique. A similar result holds for U^0 in Y_{-1}^-, where

$$\begin{aligned} Y_{-1}^- = \big\{ U^0 \in H_{-1}\times H_{-3/2}\ :\ & u^0|_{(-1,0)} \in H^1(-1,0), \\ & u^1|_{(-1,0)} \in L^2(-1,0), u^0(0^-) = y^0,\ u^0(-1) = 0 \big\}, \end{aligned}$$

with controls $q \in H^{-2}(0,T)$. In this case exact controllability holds at time $T > 4\sqrt{\gamma}$, because it is convenient to take controls q of compact support in order to avoid further singularities in the solutions at $t=0$ and $t=T$.

Proof of Theorem 5.1. In view of the observability results of Lemma 3.1, it is a direct application of HUM (see [8]).

Given $T \geq 4\sqrt{\gamma}$, for any $\Phi^0 = \big((\varphi^0, \psi^0, \xi^0), (\varphi^1, \psi^0, \xi^1)\big) \in Y_{-1}$ we solve the adjoint system

$$
\left\{
\begin{array}{l}
\gamma\partial^2\varphi_{tt} - \varphi_{tt} - \partial^4\varphi = 0,\ -1 < x < 0,\ 0 < t < T,\\
\gamma\partial^2\varphi_{tt} - \varphi_{tt} - \partial^4\varphi = 0,\ 0 < x < 1,\ 0 < t < T,\\
[\varphi](0,t) = [\partial\varphi](0,t) = 0,\ 0 < t < T,\\
\varphi_{tt}(0,t) + \left[\partial^3\varphi\right](0,t) = 0,\ 0 < t < T,\\
\gamma\partial\varphi_{tt}(0,T) - \left[\partial^2\varphi\right](0,t) = 0,\ 0 < t < T,\\
\varphi(\pm 1,t) = \partial^2\varphi(\pm 1,t) = 0,\ 0 < t < T,\\
\Phi(0) \equiv \big((\varphi(x,0), \varphi(0,0), \partial\varphi(0,0)\big), \big(\varphi_t(x,0), \varphi_t(0,0), \partial\varphi_t(0,0))\big)\\
\quad = \big((\varphi^0, \psi^0, \xi^0), (\varphi^1, \psi^0, \xi^1)\big) = \Phi^0.
\end{array}
\right. \tag{67}
$$

In view of Proposition 2.5, system (67) admits a unique solution $\Phi \in C([0,T; Y_{-1})$. Moreover, thanks to Lemma 3.1, $\partial\varphi(1,t) \in L^2(0,T)$.

We then solve

$$
\left\{
\begin{array}{l}
\gamma\partial^2 u_{tt} - u_{tt} - \partial^4 u = 0,\ -1 < x < 0,\ 0 < t < T,\\
\gamma\partial^2 u_{tt} - u_{tt} - \partial^4 u = 0,\ 0 < x < 1,\ 0 < t < T,\\
[u](0,t) = [\partial u](0,t) = 0,\ 0 < t < T,\\
u_{tt}(0,t) + \left[\partial^3 u\right](0,t) = 0,\ 0 < t < T,\\
\gamma\partial u_{tt}(0,T) - \left[\partial^2 u\right](0,t) = 0,\ 0 < t < T,\\
u(\pm 1,t) = 0, \partial^2 u(-1,t) = 0, \partial^2 u(1,t) = \partial\varphi(1,t),\\
(u(x,T), u(0,T), \partial u(0,T)) \equiv (u_t(x,T), u_t(0,T), \partial u_t(0,T)) \equiv 0.
\end{array}
\right. \tag{68}
$$

In view of Theorem 4.7 and the time reversibility of system (68), we deduce that it has a unique solution defined by transposition.

We define the linear map

$$\Lambda\Phi^0 = \left((u(x,0), u(0,0), \partial u(0,0)\right), \left(u_t(x,0), u_t(0,0), \partial u_t(0,0)\right)).$$

Multiplying in (68) by φ and integrating by parts (this is a formal computation that may be done rigorously by the definition of the solution in the sense of transposition), it follows that

$$\langle L\Lambda\Phi^0, \Phi^0\rangle = \int_0^T |\partial\varphi(1,t)|^2\, dt,\ \forall\ \Phi^0 \in Y_{-1}, \tag{69}$$

where

$$L\Phi^0 = \big(\big((I - \gamma\partial^2)\varphi^1(x), \psi^1, \xi^1\big), -\big((I - \gamma\partial^2)\varphi^0(x), \psi^0, \xi^0)\big)\big).$$

In view of identity (69), it follows that $L\Lambda$ is an isomorphism from Y_{-1} into its dual Y'_{-1}, and therefore Λ is an isomorphism from Y_{-1} into $L^{-1}Y'_{-1}$.

Now we observe that, as was pointed out in Remark 7, $Y_{-1} = V_1^+ \times V_0^+$ and then $Y'_{-1} = V_0^- \times V_1^-$, where V_0^- and V_1^-, are the spaces introduced in the proof of Theorem 4.7.

It is easy to see that $L^{-1}(V_0^- \times V_1^-) \equiv Y^-$ algebraically and topologically.

This implies that $\Lambda : Y_{-1} \to Y^-$ is an isomorphism. Therefore, for any $U^0 \in Y^-$ there exists a unique $\Phi^0 \in Y_{-1}$ such that $\Lambda\Phi^0 = U^0$. This means that the solution U of (68) with control $q = \partial\varphi(1,t)$, where φ is the solution of (67) with initial data $\Phi^0 = \Lambda^{-1}U^0$, is such that $U(0) \equiv U^0$. Therefore, q is the control we were looking for.

We also have by construction, and in view of identity (69) and the observability inequalities of Lemma 3.1, that

$$\int_0^T |\partial\varphi(1,t)|^2\,dt = \left|\langle L\Lambda\Phi^0, \Phi^0\rangle\right| = \left|\langle LU^0, \Phi^0\rangle\right|$$
$$\leq C \parallel \Phi^0 \parallel_{Y_{-1}} \parallel U^0 \parallel_{Y^-} \leq C \left(\int_0^T |\partial\varphi(1,t)|^2\,dt\right)^{1/2} \parallel U^0 \parallel_{Y^-} .$$

Therefore,

$$\int_0^T |\partial\varphi(1,t)|^2\,dt \leq C \parallel U^0 \parallel_{Y^-}^2$$

and consequently (66) holds. □

REFERENCES

[1] C. Castro, *Asymptotic analysis and control of a hybrid system composed by two vibrating strings connected by a point mass*, European Ser. Appl. Indust. Math. Control Optim. Calc. Var. (ESAIM COCV), 2 (1997), pp. 231–280, also available online from http://www.emath.fr/cocv/.

[2] C. Castro and E. Zuazua, *Analyse spectrale et contrôle d'un système hybride composé de deux poutres connectées par une masse ponctuelle*, C. R. Acad. Sci. Paris, Sér. I, 322 (1996), pp. 351–356.

[3] C. Castro and E. Zuazua, *Une remarque sur les séries de Fourier non-harmoniques et son application à la contrôlabilité des cordes avec densité singulière*, C. R. Acad. Sci. Paris, Sér. I, 323 (1996), pp. 365–370.

[4] C. Castro and E. Zuazua, *A hybrid system consisting of two flexible beams connected by a point mass: Spectral analysis and well-posedness in asymmetric spaces*, in Elasticité, Viscoélasticité et Contrôle Optimal: ESAIM Proc., 2 (1997), pp. 17–53, also available online from http://www.emath.fr/proc/Vol.2/.

[5] C. Castro and E. Zuazua, *Exact boundary controllability of two Euler-Bernoulli beams connected by a point mass*, Math. Comput. Modelling, to appear.

[6] S. Hansen and E. Zuazua, *Exact controllability and stabilization of a vibrating string with an interior point mass*, SIAM J. Control Optim., 33 (1995), pp. 1357–1391.

[7] J. E. Lagnese and J.-L. Lions, *Modelling, Analysis and Control of Thin Plates*, Research in Applied Math. 6, Masson, Paris, 1988.

[8] J.-L. Lions, *Contrôlabilité exacte, stabilisation et perturbations de systèmes distribués. Tome 1. Contrôlabilité exacte*, Research in Applied Math. 8, Masson, Paris, 1988.

[9] S. W. Taylor, *Exact boundary controllability of a beam and mass system*, in Computation and Control IV, Prog. Systems Control Theory 20, K. L. Bowers and J. Land, eds., Birkhauser, Boston, 1995, pp. 305–321.

[10] D. Ulrich, *Divided differences and systems of nonharmonic Fourier series*, Proc. Amer. Math. Soc., 80 (1980), pp. 47–57.

[11] R. M. Young, *An Introduction to Nonharmonic Fourier Series*, Academic Press, New York, 1980.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 5, pp. 1596–1617, September 1998

005

THE MAXIMUM PRINCIPLE FOR PARTIALLY OBSERVED OPTIMAL CONTROL OF STOCHASTIC DIFFERENTIAL EQUATIONS*

SHANJIAN TANG†

Abstract. This paper concerns partially observed optimal control of possibly degenerate stochastic differential equations, with correlated noises between the system and the observation. The control is allowed to enter into all the coefficients. A general maximum principle is proved for the partially observed optimal control, and the relations among the adjoint processes are established. Adjoint vector fields, which are adapted to the past and present observations, are introduced as the solutions to some backward stochastic partial differential equations (BSPDEs), and their relations are established. Under suitable conditions, the adjoint processes are characterized in terms of the adjoint vector fields, their differentials and Hessians, along the optimal state process. Some other formulations of the partially observed stochastic maximum principle are then derived.

Key words. partially observed optimal control, adjoint processes, adjoint vector fields, maximum principle, backward stochastic partial differential equations

AMS subject classification. 93E20

PII. S0363012996313100

1. Formulation of the problem and some historical comments.

Throughout this paper, we use the Einstein convention for summation over repeated indices. For a matrix, we use superscripts to indicate (when necessary) the number of its columns or its rows or the position of its components, and the precise meaning can be specified from the context; the range of the superscripts will not be explicitly stated unless there is a danger of confusion. $\langle , \rangle$ denotes the product of two vectors in an Euclidean space, and $|\cdot|$ denotes the square root of the sum of all the squares of components of the underlying matrix. $*$ appearing in the superscripts denotes the transpose of a matrix, and $\mathbb{R}^n$ the n-dimensional Euclidean space. For a $\mathbb{R}^m$-valued vector function f on $\mathbb{R}^n$, we use the notation

$$f_x := \begin{pmatrix} \frac{\partial f^1}{\partial x^1} & \frac{\partial f^1}{\partial x^2} & \cdots & \frac{\partial f^1}{\partial x^n} \\ \frac{\partial f^2}{\partial x^1} & \frac{\partial f^2}{\partial x^2} & \cdots & \frac{\partial f^2}{\partial x^n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f^m}{\partial x^1} & \frac{\partial f^m}{\partial x^2} & \cdots & \frac{\partial f^m}{\partial x^n} \end{pmatrix}, \quad f^i_{xx} := \begin{pmatrix} \frac{\partial^2 f^i}{\partial (x^1)^2} & \frac{\partial^2 f^i}{\partial x^1 \partial x^2} & \cdots & \frac{\partial^2 f^i}{\partial x^1 \partial x^n} \\ \frac{\partial^2 f^i}{\partial x^2 \partial x^1} & \frac{\partial^2 f^i}{\partial (x^2)^2} & \cdots & \frac{\partial^2 f^i}{\partial x^2 \partial x^n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f^i}{\partial x^n \partial x^1} & \frac{\partial^2 f^i}{\partial x^n \partial x^2} & \cdots & \frac{\partial^2 f^i}{\partial (x^n)^2} \end{pmatrix}.$$

Let $(\Omega, \mathcal{F}, P)$ be a probability space on which are defined two independent standard Brownian motions $w(\cdot)$ and $Y(\cdot)$ valued in $\mathbb{R}^m$ and $\mathbb{R}^d$, respectively. Let x_0 be a random variable with the law P_0 and independent of $(w(\cdot), Y(\cdot))$. Let $\{\mathcal{F}^w_t\}$ and $\{\mathcal{F}^Y_t\}$ be the P-completed natural filtrations generated by $w(\cdot)$ and $Y(\cdot)$, respectively,

*Received by the editors December 6, 1996; accepted for publication (in revised form) September 12, 1997; published electronically June 9, 1998. This research was supported in part by the NSF of China, by the Laboratory of Mathematics for Nonlinear Sciences at Fudan University, by une bourse de Séjour Scientifique de Haut Niveau du gouvernement français, and by le Programme de Recherches Avancées franco-chinois, PRA M92-12.

http://www.siam.org/journals/sicon/36-5/31310.html

†Department of Mathematics, Fudan University, Shanghai 200433, People's Republic of China (sjtang@fudan.edu.cn).

and $\sigma(x_0)$ the σ-algebra generated by x_0. Set

$$\mathcal{F}_t := \sigma(x_0) \vee \mathcal{F}_t^w \vee \mathcal{F}_t^Y, \quad \mathcal{F} := \mathcal{F}_1.$$

Let U be a nonempty Borel subset of some Euclidean space. An admissible control is defined as a stochastic process $u : [0,1] \times \Omega \to U$ which is $\mathcal{F}_t^Y$-adapted and satisfies

$$\sup_{0 \le t \le 1} E|u(t,\cdot)|^i < \infty, \quad i = 1, 2, \ldots$$

For simplification of notation, we write $u(t)$ for $u(t,\cdot)$ in the following. A set U_{ad} of admissible controls is called an admissible class of controls if U_{ad} has the following property: $\forall\, v_1(\cdot), v_2(\cdot) \in U_{\mathrm{ad}}$, and a Borel subset K of $[0,1]$, define

$$v(t) = v_1(t)\chi_K(t) + v_2(t)\chi_{[0,1]\setminus K}(t), \quad t \in [0,1];$$

then $v(\cdot) \in U_{\mathrm{ad}}$.

In the literature, a partially observed optimal control problem has been studied at least for the following two admissible classes of controls: one is based on the present observation and defined as

$$\begin{aligned}\widetilde{U}_{\mathrm{ad}} := \{&v : v \text{ is a } U\text{-valued stochastic process such that } v(t) \text{ is } \sigma(Y(t))\text{-measurable}\\ &\text{for } \textit{almost every } t \in [0,1] \text{ and } \sup_{0\le t\le 1} E|v(t)|^i < \infty,\ i = 1, 2, \ldots\};\end{aligned}$$

the other one is based on the past and present observations and defined as

$$\begin{aligned}\overline{U}_{\mathrm{ad}} := \{&v : v \text{ is a } U\text{-valued } \mathcal{F}_t^Y\text{-adapted stochastic process}\\ &\text{and satisfies } \sup_{0\le t\le 1} E|v(t)|^i < \infty,\ i = 1, 2, \ldots\}.\end{aligned}$$

A control is said to be partially observed if the control is a nonanticipative functional of the observation $Y(\cdot)$. A set of controls is said to be partially observed if its every element is partially observed. Obviously, a set of admissible controls is partially observed.

We make the following hypothesis.

($\mathcal{A}1$) *Let U_{ad} be a given admissible class of controls. The functions $f : [0,1]\times\mathbb{R}^n \times U \to \mathbb{R}^n, g : [0,1]\times\mathbb{R}^n\times U \to \mathbb{R}^{n\times m}, \widetilde{g} : [0,1]\times\mathbb{R}^n\times U \to \mathbb{R}^{n\times d}, h : [0,1]\times\mathbb{R}^n\times U \to \mathbb{R}^d, l : [0,1]\times\mathbb{R}^n\times U \to \mathbb{R}$, and $m : \mathbb{R}^n \to \mathbb{R}$ are Borel measurable, continuous in v, and twice continuously differentiable in x, and for some constant C,*

$$\begin{aligned}&(1+|x|+|v|)^{-1}|f(t,x,v)| + |f_x(t,x,v)| + |f_{xx}^{i_1}(t,x,v)| \le C,\\ &(1+|x|+|v|)^{-1}|g^i(t,x,v)| + |g_x^i(t,x,v)| + |g_{xx}^{i_1 i}(t,x,v)| \le C,\\ &|\widetilde{g}^j(t,x,v)| + |\widetilde{g}_x^j(t,x,v)| + |\widetilde{g}_{xx}^{i_1 j}(t,x,v)| \le C,\\ &|h(t,x,v)| + |h_x(t,x,v)| + |h_{xx}^j(t,x,v)| \le C,\\ &(1+|x|^2+|v|^2)^{-1}|l(t,x,v)| + (1+|x|+|u|)^{-1}|l_x(t,x,v)| + |l_{xx}(t,x,v)| \le C,\\ &(1+|x|^2)^{-1}|m(x)| + (1+|x|)^{-1}|m_x(x)| + |m_{xx}(x)| \le C.\end{aligned}$$

x_0 has finite moments of arbitrary order.

Our general partially observed optimal control problem is stated as follows.

Consider the system

$$\begin{cases} dx(t) = f(t,x(t),v(t))\,dt + g^i(t,x(t),v(t))\,dw^i(t)\\ \qquad\qquad + \widetilde{g}^j(t,x(t),v(t))\,d\widetilde{w}^j(t), \quad t \in (0,1],\\ x(0) = x_0 \end{cases} \tag{1.1}$$

and the observation

$$\begin{cases} dY(t) = h(t, x(t), v(t))\,dt + d\widetilde{w}(t), \quad t \in (0,1], \\ Y(0) = 0. \end{cases} \tag{1.2}$$

Putting (1.2) into (1.1), we have

$$\begin{cases} dx(t) = (f - \widetilde{g}h)(t, x(t), v(t))\,dt + g^i(t, x(t), v(t))\,dw^i(t) \\ \qquad\quad + \widetilde{g}^j(t, x(t), v(t))\,dY^j(t), \qquad t \in (0,1], \\ x(0) = x_0. \end{cases} \tag{1.3}$$

For each $v(\cdot) \in U_{\mathrm{ad}}$, (1.3) has a unique strong solution, which will be denoted by $x^v(\cdot)$. From Girsanov's theorem, it follows that if

$$\rho^v(t) := \exp\biggl\{\int_0^t h^*(s, x^v(s), v(s))\,dY(s) - \frac{1}{2}\int_0^t |h(s, x^v(s), v(s))|^2\,ds\biggr\}, \tag{1.4}$$

$$\widetilde{w}(t) := Y(t) - \int_0^t h(s, x^v(s), v(s))\,ds, \tag{1.5}$$

and if $dP^v := \rho^v(1)\,dP$, then $(P^v, x^v, Y, w, \widetilde{w})$ is a weak solution on $(\Omega, \mathcal{F}, \mathcal{F}_t)$ of (1.1) and (1.2).

The cost functional is

$$J(v(\cdot)) = E^v\biggl[\int_0^1 l(t, x^v(t), v(t))\,dt + m(x^v(1))\biggr]. \tag{1.6}$$

Here, E^v denotes the expectation with respect to the probability space $(\Omega, \mathcal{F}, P^v)$. Our partially observed optimal control problem is to minimize the cost functional (1.6) over $v(\cdot) \in U_{\mathrm{ad}}$, i.e.,

$$\min_{v \in U_{\mathrm{ad}}} J(v). \tag{1.7}$$

Here, the words "partially observed" indicates that the admissible class U_{ad} in the underlying optimal control problem is partially observed. Our aim is to seek the necessary conditions for the partially observed optimal control $\widehat{u}(\cdot)$.

Such a subject has been discussed by many authors, such as Fleming [4]; Kwakernaak [7] (with an explorative style); Bensoussan [2]; Haussmann [5]; Baras, Elliott, and Kohlmann [1]; Zhou [13]; and Li and Tang [8]. Usually, they made at least one of the following five assumptions: 1) The diffusion term σ is nondegenerate (see Fleming [4], Kwakernaak [7], Bensoussan [2], and Zhou [13]). 2) The coefficients $f, g, h,$ and l are differentiable in control variable u, and the set U, in which the control takes values, is convex (see Fleming [4] and Bensoussan [2]). 3) The control does not appear in the diffusion term g and the observation h (see Bensoussan [2]; Haussmann [5]; and Baras, Elliott, and Kohlmann [1]). 4) $\widetilde{g} = 0$ (see Bensoussan [2]; Haussmann [5]; Baras, Elliott, and Kohlmann [1]; and Li and Tang [8]). 5) The initial state x_0 has a regular density function (see Bensoussan [2] and Zhou [13]).

In this paper, we consider the general case of the partially observed optimal control problem (1.1), (1.2), (1.6), (1.7), where the control is allowed to enter into all the coefficients, the diffusion term g is allowed to be degenerate, the correlation

coefficient $\widetilde{g}$ is present, the set U is not necessarily convex, and the initial state does not necessarily have a regular density function. A general maximum principle is proved for the partially observed optimal control, and the relations among the adjoint processes are established. Adjoint vector fields are introduced as the solutions to some BSPDEs, and their relations are established. Under suitable conditions, the adjoint processes are characterized in terms of the adjoint vector fields, their differentials and Hessians, along the optimal state process. Some other formulations of the partially observed stochastic maximum principle are then obtained, and our results are compared with those existing in the literature. Our approach does not involve the Zakai equation, and thus we can get around a lot of complicated stochastic calculus in infinite-dimensional spaces, in contrast with Bensoussan [2], Haussmann [5], and Zhou [13].

The rest of this paper is organized as follows. In section 2, we derive a general maximum principle for partially observed optimal controls from the general maximum principle for optimal controls with full information. In section 3, the relations are established among the adjoint processes, which are introduced in the general partially observed maximum principle and which are characterized as the unique $\mathcal{F}_t$-adapted square-integrable solutions of backward stochastic differential equations (BSDEs). In section 4, adjoint vector fields are introduced as the solutions to some BSPDEs, and their relations are established; under suitable conditions, the adjoint processes are characterized in terms of the adjoint vector fields, their differentials and Hessians, along the optimal state process. Finally in section 5, some other formulations of the partially observed stochastic maximum principle are derived, and our results are compared with the existing ones.

2. A general partially observed maximum principle. Let $\widehat{u}(\cdot)$ be an optimal control and $(\widehat{x}(\cdot), Y(\cdot), w(\cdot), \widehat{\widetilde{w}(\cdot)}, \widehat{P})$ be the corresponding weak solution of (1.1) – (1.2). We introduce the notation: $\widehat{E} = E^{\hat{u}}$, $\Delta f(t;v) := \Delta f(t, \widehat{x}(t), \widehat{u}(t); v) := f(t, \widehat{x}(t), v) - f(t, \widehat{x}(t), \widehat{u}(t))$, and similar notation will be made for other functions $g, \widetilde{g}, l, h,$ and $\mathcal{H}$ (see (2.10) below). For each $v(\cdot) \in U_{\text{ad}}$, the stochastic process $\rho^v(\cdot)$ can be characterized as the solution of the following stochastic differential equation (SDE):

$$\begin{cases} d\rho^v(t) = \rho^v(t) h^*(t, x^v(t), v(t))\ dY(t), \quad t \in (0,1], \\ \rho^v(0) = 1. \end{cases} \tag{2.1}$$

The cost functional (1.6) can be rewritten as

$$J(v(\cdot)) = E\bigg[\int_0^1 \rho^v(t) l(t, x^v(t), v(t))\ dt + \rho^v(1) m(x^v(1))\bigg]. \tag{2.2}$$

Set

$$\begin{aligned} &X := \begin{pmatrix} \rho \\ x \end{pmatrix}, \quad \widehat{X} := \begin{pmatrix} \widehat{\rho} \\ \widehat{x} \end{pmatrix} := \begin{pmatrix} \rho^{\hat{u}} \\ x^{\hat{u}} \end{pmatrix}, \quad X_0 := \begin{pmatrix} 1 \\ x_0 \end{pmatrix}, \\ &F(t, X, v) := \begin{pmatrix} 0 \\ f(t,x,v) - \widetilde{g}(t,x,v) h(t,x,v) \end{pmatrix}, \\ &G(t, X, v) := \begin{pmatrix} 0 \\ g(t,x,v) \end{pmatrix}, \quad \widetilde{G}(t, X, v) := \begin{pmatrix} \rho h^*(t,x,v) \\ \widetilde{g}(t,x,v) \end{pmatrix}, \\ &L(t, X, v) := \rho l(t,x,v), \qquad M(X) := \rho m(x). \end{aligned} \tag{2.3}$$

Equations (1.1), (1.2), and (2.1) can be compressed into the following form:

$$\begin{cases} dX(t) = F(t, X(t), v(t))\,dt + G^i(t, X(t), v(t))\,dw^i(t) \\ \qquad\quad + \widetilde{G}^j(t, X(t), v(t))\,dY^j(t), \quad t \in (0, 1], \\ X(0) = X_0. \end{cases} \tag{2.4}$$

The cost functional (2.2) is rewritten as

$$J(v(\cdot)) = E\bigg[\int_0^1 L(t, X(t), v(t))\,dt + M(X(1))\bigg]. \tag{2.5}$$

Our partially observed optimal control problem becomes the following minimization problem: to minimize $J(v(\cdot))$ over $v(\cdot) \in U_{\rm ad}$ subject to (2.4). The present formulation of the partially observed optimal control problem is quite similar to a completely observed optimal control problem; the only difference lies in the admissible class $U_{\rm ad}$ of controls. We can follow the same arguments to the case of full information to derive the following maximum principle. See Peng [11] and Tang and Li [12] for details.

Define the Hamiltonian $H : [0, 1] \times \mathbb{R}^{n+1} \times U \times \mathbb{R}^{n+1} \times \mathbb{R}^{(n+1)\times m} \times \mathbb{R}^{(n+1)\times d} \to \mathbb{R}$ as follows:

$$\begin{aligned} H(t, X, v, a, b, \widetilde{b}) := \langle a, F(t, X, v)\rangle + \langle b^i, G^i(t, X, v)\rangle \\ + \langle \widetilde{b}^j, \widetilde{G}^j(t, X, v)\rangle + L(t, X, v) \end{aligned} \tag{2.6}$$

$\forall\, t \in [0, 1], X \in \mathbb{R}^{n+1}, v \in U, a \in \mathbb{R}^{n+1}, b \in \mathbb{R}^{(n+1)\times m}, \widetilde{b} \in \mathbb{R}^{(n+1)\times d}$.

Let $(a(\cdot), b(\cdot), \widetilde{b}(\cdot))$ be the unique $\mathcal{F}_t$-adapted square integrable solution of the first-order adjoint equation

$$\begin{cases} da(t) = -\,H^*_X(t, \widehat{X}(t), \widehat{u}(t), a(t), b(t), \widetilde{b}(t))\,dt \\ \qquad\quad + b(t)\,dw(t) + \widetilde{b}(t)\,dY(t), \quad t \in [0, 1), \\ a(1) = M^*_X(\widehat{X}(1)) \end{cases} \tag{2.7}$$

and $(A(\cdot), B(\cdot), \widetilde{B}(\cdot))$ be the unique $\mathcal{F}_t$-adapted square integrable solution of the second-order adjoint equation

$$\begin{cases} dA(t) = -\,\{F^*_X(t, \widehat{X}(t), \widehat{u}(t))A(t) + A(t)F_X(t, \widehat{X}(t), \widehat{u}(t)) \\ \qquad\quad + G^{i*}_X(t, \widehat{X}(t), \widehat{u}(t))A(t)G^i_x(t, \widehat{X}(t), \widehat{u}(t)) \\ \qquad\quad + \widetilde{G}^{i*}_X(t, \widehat{X}(t), \widehat{u}(t))A(t)\widetilde{G}^i_X(t, \widehat{X}(t), \widehat{u}(t)) \\ \qquad\quad + G^{i*}_X(t, \widehat{X}(t), \widehat{u}(t))B^i(t) + B^i(t)G^i_X(t, \widehat{X}(t), \widehat{u}(t)) \\ \qquad\quad + \widetilde{G}^{j*}_X(t, \widehat{X}(t), \widehat{u}(t))\widetilde{B}^j(t) + \widetilde{B}^j(t)\widetilde{G}^j_X(t, \widehat{X}(t), \widehat{u}(t)) \\ \qquad\quad + H_{XX}(t, \widehat{X}(t), \widehat{u}(t), a(t), b(t), \widetilde{b}(t))\}\,dt \\ \qquad\quad + B^i(t)\ dw^i(t) + \widetilde{B}^j(t)\,dY^j(t), \quad t \in [0, 1), \\ A(1) = M_{XX}(\widehat{X}(1)). \end{cases} \tag{2.8}$$

Then the following maximum condition holds:

$$
\begin{aligned}
&E\int_0^1 \{H(t,\widehat{X}(t),v(t),a(t),b(t),\widetilde{b}(t)) - H(t,\widehat{X}(t),\widehat{u}(t),a(t),b(t),\widetilde{b}(t))\}\,dt \\
&+\frac{1}{2}E\int_0^1 \operatorname{tr}[A(t)(\Delta G(t;v(t))\Delta G^*(t;v(t)) + \Delta\widetilde{G}(t;v(t))\Delta\widetilde{G}^*(t;v(t)))]\,dt \\
&\geq 0 \qquad \forall\, v(\cdot)\in U_{\mathrm{ad}}.
\end{aligned} \tag{2.9}
$$

The reader will see later that since the state variable ρ appears in the optimal control problem in a linear way, some adjoint processes are superfluous in the above maximum principle. Now we begin to dispense with these adjoint processes and reformulate the above maximum principle.

We introduce a new Hamiltonian $\mathcal{H}: [0,1]\times\mathbb{R}^n\times U\times\mathbb{R}^n\times\mathbb{R}^{n\times m}\times\mathbb{R}^d\times\mathbb{R}^{n\times d}\to\mathbb{R}$ as follows:

$$
\begin{aligned}
\mathcal{H}(t,x,v,q,k,\widetilde{R},\widetilde{k}) :=\ & \langle q, f(t,x,v)\rangle + \langle k^i, g^i(t,x,v)\rangle \\
&+ \widetilde{R}^j h^j(t,x,v) + \langle \widetilde{k}^j, \widetilde{g}^j(t,x,v)\rangle + l(t,x,v)
\end{aligned} \tag{2.10}
$$

$\forall\, t\in[0,1], x\in\mathbb{R}^n, v\in U, q\in\mathbb{R}^n, k\in\mathbb{R}^{n\times m}, \widetilde{R}\in\mathbb{R}^d, \widetilde{k}\in\mathbb{R}^{n\times d}$.

Decompose the matrices $a(t); b(t); \widetilde{b}(t); A(t);\ B^i(t), i=1,\dots,m;$ and $\widetilde{B}^j(t), j=1,\dots,d,$ into blocks in the following manner:

$$
\begin{aligned}
a(t) &= \begin{pmatrix} a_1(t)\\ a_2(t)\end{pmatrix}\begin{matrix}\}1\\ \}n\end{matrix}, \quad b(t) = \begin{pmatrix} b_1(t)\\ b_2(t)\end{pmatrix}\begin{matrix}\}1\\ \}n\end{matrix}, \quad \widetilde{b}(t) = \begin{pmatrix} \widetilde{b}_1(t)\\ \widetilde{b}_2(t)\end{pmatrix}\begin{matrix}\}1\\ \}n\end{matrix}, \\
A(t) &= \begin{matrix} \\ 1\{\\ n\{\end{matrix}\begin{pmatrix} \overbrace{A_{11}(t)}^{1} & \overbrace{A_{12}(t)}^{n} \\ A_{21}(t) & A_{22}(t)\end{pmatrix}, \\
B^i(t) &= \begin{matrix} \\ 1\{\\ n\{\end{matrix}\begin{pmatrix} \overbrace{B^i_{11}(t)}^{1} & \overbrace{B^i_{12}(t)}^{n} \\ B^i_{21}(t) & B^i_{22}(t)\end{pmatrix}, \quad i=1,\dots,m, \\
\widetilde{B}^j(t) &= \begin{matrix} \\ 1\{\\ n\{\end{matrix}\begin{pmatrix} \overbrace{\widetilde{B}^j_{11}(t)}^{1} & \overbrace{\widetilde{B}^j_{12}(t)}^{n} \\ \widetilde{B}^j_{21}(t) & \widetilde{B}^j_{22}(t)\end{pmatrix}, \quad j=1,\dots,d.
\end{aligned} \tag{2.11}
$$

Then, we can check the following

$$
\begin{aligned}
F_X(t,X,v) &:= (F_\rho(t,X,v), F_x(t,X,v)) \\
&= \begin{pmatrix} 0 & 0 \\ 0 & f_x(t,x,v) - \widetilde{g}^j_x(t,x,v)h^j(t,x,v) - \widetilde{g}(t,x,v)h_x(t,x,v) \end{pmatrix}, \\
G^i_X(t,X,v) &:= (G^i_\rho(t,X,v), G^i_x(t,X,v)) = \begin{pmatrix} 0 & 0 \\ 0 & g^i_x(t,x,v) \end{pmatrix}, \\
\widetilde{G}^j_X(t,X,v) &:= (\widetilde{G}^j_\rho(t,X,v), \widetilde{G}^j_x(t,X,v)) = \begin{pmatrix} h^j(t,x,v) & \rho h^j_x(t,x,v) \\ 0 & \widetilde{g}^j_x(t,x,v) \end{pmatrix}, \\
L_X(t,X,v) &:= (L_\rho(t,X,v), L_x(t,X,v)) = (l(t,x,v), \rho l_x(t,x,v)), \\
M_X(X) &:= (M_\rho(X), M_x(X)) = (m(x), \rho m_x(x)),
\end{aligned} \tag{2.12}
$$

$$
\begin{aligned}
&H(t,X,v,a,b,\widetilde{b}) \\
&= \langle a_2, f(t,x,v) - \widetilde{g}h(t,x,v)\rangle + \langle b^i_2, g^i(t,x,v)\rangle \\
&\quad + \langle \widetilde{b}^j_2, \widetilde{g}^j(t,x,v)\rangle + \langle \widetilde{b}^*_1, \rho h(t,x,v)\rangle + \rho l(t,x,v), \\
&H_X(t,X,v,a,b,\widetilde{b}) := \Big(H_\rho(t,X,v,a,b,\widetilde{b}), H_x(t,X,v,a,b,\widetilde{b})\Big) \\
&= \Big(l(t,x,v) + \langle \widetilde{b}^*_1, h(t,x,v)\rangle, H_x(t,X,v,a,b,\widetilde{b})\Big), \\
&H_{XX}(t,X,v,a,b,\widetilde{b}) := \begin{pmatrix} H_{\rho\rho}(t,X,v,a,b,\widetilde{b}) & H_{x\rho}(t,X,v,a,b,\widetilde{b}) \\ H_{\rho x}(t,X,v,a,b,\widetilde{b}) & H_{xx}(t,X,v,a,b,\widetilde{b}) \end{pmatrix} \\
&= \begin{pmatrix} 0 & l_x(t,x,v) + \widetilde{b}_1 h_x(t,x,v) \\ l^*_x(t,x,v) + h^*_x(t,x,v)\widetilde{b}^*_1 & H_{xx}(t,X,v,a,b,\widetilde{b}) \end{pmatrix}, \\
&H_x(t,X,v,a,b,\widetilde{b}) \\
&= \rho \mathcal{H}_x(t,x,v,\rho^{-1}a_2, \rho^{-1}b_2, \widetilde{b}_1 - \rho^{-1}a^*_2\widetilde{g}(t,x,v), \rho^{-1}[\widetilde{b}_2 - a_2 h^*(t,x,v)]), \\
&H_{xx}(t,X,v,a,b,\widetilde{b}) \\
&= \rho \mathcal{H}_{xx}(t,x,v,\rho^{-1}a_2, \rho^{-1}b_2, \widetilde{b}_1 - \rho^{-1}a^*_2\widetilde{g}(t,x,v), \rho^{-1}[\widetilde{b}_2 - a_2 h^*(t,x,v)]) \\
&\quad - \widetilde{g}^{j*}_x(t,x,v)a_2 h^j_x(t,x,v) - h^{j*}_x(t,x,v)a^*_2\widetilde{g}^j_x(t,x,v).
\end{aligned} \tag{2.13}
$$

In view of the above calculations, equation (2.7) is decomposed into the following two equations:

$$
\begin{cases}
da_1(t) = -\left[l(t,\widehat{x}(t),\widehat{u}(t)) + \langle \widetilde{b}^*_1(t), h(t,\widehat{x}(t),\widehat{u}(t))\rangle\right] dt \\
\qquad\qquad + b_1(t)\, dw(t) + \widetilde{b}_1(t)\, dY(t), \quad t \in [0,1), \\
a_1(1) = m(\widehat{x}(1))
\end{cases} \tag{2.14}
$$

and

$$
\begin{cases}
da_2(t) = -H^*_x(t,\widehat{X}(t),\widehat{u}(t),a(t),b(t),\widetilde{b}(t))\, dt \\
\qquad\qquad + b_2(t)\, dw(t) + \widetilde{b}_2(t)\, dY(t), \quad t \in [0,1), \\
a_2(1) = \widehat{\rho}(1) m^*_x(\widehat{x}(1)),
\end{cases} \tag{2.15}
$$

while equation (2.8) is decomposed into the following four equations:

$$\begin{cases} dA_{11}(t) = -\{h^*h(t,\widehat{x}(t),\widehat{u}(t))A_{11}(t) + 2\widetilde{B}^j_{11}(t)h^j(t,\widehat{x}(t),\widehat{u}(t))\}\,dt \\ \qquad\qquad + B^i_{11}(t)\,dw^i(t) + \widetilde{B}^j_{11}(t)\,dY^j(t), \quad t\in[0,1), \\ A_{11}(1) = 0, \end{cases} \tag{2.16}$$

(2.17)

$$\begin{cases} dA_{21}(t) = -[\mathcal{H}^*_x(t,\widehat{x}(t),\widehat{u}(t);A_{21}(t),B_{21}(t),\widetilde{b}_1(t) - A^*_{21}(t)\widetilde{g}(t,\widehat{x}(t),\widehat{u}(t)),\widetilde{B}_{21}(t)) \\ \qquad\qquad + \widehat{\rho}(t)h^*_x(t,\widehat{x}(t),\widehat{u}(t))h(t,\widehat{x}(t),\widehat{u}(t))A_{11}(t) + \widehat{\rho}(t)h^{j*}_x(t,\widehat{x}(t),\widehat{u}(t))\widetilde{B}^j_{11}(t) \\ \qquad\qquad + \widetilde{B}^j_{21}(t)h^j(t,\widehat{x}(t),\widehat{u}(t))]\,dt \\ \qquad\qquad + B^i_{21}(t)\,dw^i(t) + \widetilde{B}^j_{21}(t)\,dY^j(t), \quad t\in[0,1), \\ A_{21}(1) = m^*_x(\widehat{x}(1)), \end{cases}$$

$$A_{12}(t) = A^*_{21}(t), \quad t\in[0,1], \tag{2.18}$$

and

$$\begin{cases} dA_{22}(t) = -\{f^*_x(t,\widehat{x}(t),\widehat{u}(t))A_{22}(t) + A_{22}(t)f_x(t,\widehat{x}(t),\widehat{u}(t)) \\ \qquad\qquad + g^{i*}_x(t,\widehat{x}(t),\widehat{u}(t))A_{22}(t)g^i_x(t,\widehat{x}(t),\widehat{u}(t)) \\ \qquad\qquad + \widehat{\rho}^2(t)h^{i*}_x(t,\widehat{x}(t),\widehat{u}(t))A_{11}(t)h^i_x(t,\widehat{x}(t),\widehat{u}(t)) \\ \qquad\qquad + \widehat{\rho}(t)\widetilde{g}^{j*}_x(t,\widehat{x}(t),\widehat{u}(t))A_{21}(t)h^j_x(t,\widehat{x}(t),\widehat{u}(t)) \\ \qquad\qquad + \widehat{\rho}(t)h^{j*}_x(t,\widehat{x}(t),\widehat{u}(t))A_{12}(t)\widetilde{g}^j_x(t,\widehat{x}(t),\widehat{u}(t)) \\ \qquad\qquad + \widetilde{g}^{j*}_x(t,\widehat{x}(t),\widehat{u}(t))A_{22}(t)\widetilde{g}^j_x(t,\widehat{x}(t),\widehat{u}(t)) \\ \qquad\qquad + B^i_{22}(t)g^i_x(t,\widehat{x}(t),\widehat{u}(t)) + g^{i*}_x(t,\widehat{x}(t),\widehat{u}(t))B^i_{22}(t) \\ \qquad\qquad + (\widehat{\rho}(t)\widetilde{B}_{21}(t) - A_{22}(t)\widetilde{g}(t,\widehat{x}(t),\widehat{u}(t)))h_x(t,\widehat{x}(t),\widehat{u}(t)) \\ \qquad\qquad + (\widetilde{B}^j_{22}(t) - A_{22}(t)h^j(t,\widehat{x}(t),\widehat{u}(t)))\widetilde{g}^j_x(t,\widehat{x}(t),\widehat{u}(t)) \\ \qquad\qquad + h^*_x(t,\widehat{x}(t),\widehat{u}(t))(\widehat{\rho}(t)\widetilde{B}_{12}(t) - \widetilde{g}^*(t,\widehat{x}(t),\widehat{u}(t))A_{22}(t)) \\ \qquad\qquad + \widetilde{g}^{j*}_x(t,\widehat{x}(t),\widehat{u}(t))(\widetilde{B}^j_{22}(t) - h^j(t,\widehat{x}(t),\widehat{u}(t))A_{22}(t)) \\ \qquad\qquad + H_{xx}(t,\widehat{X}(t),\widehat{u}(t),a(t),b(t),\widetilde{b}(t))\}\,dt \\ \qquad\qquad + B^i_{22}(t)\,dw^i(t) + \widetilde{B}^j_{22}(t)\,dY^j(t), \quad t\in[0,1), \\ A_{22}(1) = \widehat{\rho}(1)m_{xx}(\widehat{x}(1)). \end{cases} \tag{2.19}$$

We can obtain, from the uniqueness of the $\mathcal{F}_t$-adapted square integrable solution of the BSDE (2.16) (see Pardoux and Peng [9]),

$$A_{11} = 0, \quad B_{11} := (B^1_{11},\dots,B^m_{11}) = 0, \quad \widetilde{B}_{11} := (\widetilde{B}^1_{11},\dots,\widetilde{B}^d_{11}) = 0, \tag{2.20}$$

and then establish from (2.15), (2.17), via Itô's formula, the following relations:

$$\begin{aligned} &A_{21}(t) = \widehat{\rho}^{-1}(t)a_2(t), \quad B^i_{21}(t) = \widehat{\rho}^{-1}(t)b^i_2(t), \\ &\widetilde{B}^j_{21}(t) = \widehat{\rho}^{-1}(t)\widetilde{b}^j_2(t) - \widehat{\rho}^{-1}(t)a_2(t)h^j(t,\widehat{x}(t),\widehat{u}(t)). \end{aligned} \tag{2.21}$$

Moreover, if we set

$$
\begin{aligned}
&r(t) := a_1(t), \quad R(t) := b_1(t), \quad \widetilde{R}(t) := \widetilde{b}_1(t),\\
&q(t) := \widehat{\rho}^{-1}(t)a_2(t), \quad k(t) := \widehat{\rho}^{-1}(t)b_2(t),\\
&\widetilde{k}(t) := \widehat{\rho}^{-1}(t)\widetilde{b}_2(t) - \widehat{\rho}^{-1}(t)a_2(t)h^*(t,\widehat{x}(t),\widehat{u}(t)),\\
&Q(t) := \widehat{\rho}^{-1}(t)A_{22}(t), \quad K^i(t) := \widehat{\rho}^{-1}(t)B^i_{22}(t),\\
&\widetilde{K}^j(t) := \widehat{\rho}^{-1}(t)\widetilde{B}^j_{22}(t) - \widehat{\rho}^{-1}(t)A_{22}(t)h^j(t,\widehat{x}(t),\widehat{u}(t)),
\end{aligned}
\tag{2.22}
$$

then $(r, R^i,\, i = 1,\dots,m;\, \widetilde{R}^j,\, j = 1,\dots,d)$ is characterized as the unique $\mathcal{F}_t$-adapted solution of the following BSDE:

$$
\begin{cases}
dr(t) = -\{l(t,\widehat{x}(t),\widehat{u}(t)) + \widetilde{R}^j(t)h^j(t,\widehat{x}(t),\widehat{u}(t))\}\,dt\\
\qquad\quad + R^i(t)\,dw^i(t) + \widetilde{R}^j(t)\,dY^j(t), \quad t \in [0,1),\\
r(1) = m(\widehat{x}(1));
\end{cases}
\tag{2.23}
$$

$(q, k^i, i = 1,\dots,m; \widetilde{k}^j, j = 1,\dots,d)$ solves

(2.24)

$$
\begin{cases}
dq(t) = -\{\mathcal{H}^*_x(t,\widehat{x}(t),\widehat{u}(t);q(t),k(t),\widetilde{R}(t) - q^*(t)\widetilde{g}(t,\widehat{x}(t),\widehat{u}(t)),\widetilde{k}(t))\\
\qquad\quad + \widetilde{k}^j(t)h^j(t,\widehat{x}(t),\widehat{u}(t))\}\,dt + k^i(t)\,dw^i(t) + \widetilde{k}^j(t)\,dY^j(t), \quad t \in [0,1),\\
q(1) = m^*_x(\widehat{x}(1));
\end{cases}
$$

and $(Q, K^i, i = 1,\dots,m; \widetilde{K}^j, j = 1,\dots,d)$ solves

$$
\begin{cases}
\begin{aligned}
dQ(t) = -\{&f^*_x(t,\widehat{x}(t),\widehat{u}(t))Q(t) + Q(t)f_x(t,\widehat{x}(t),\widehat{u}(t))\\
&+ g^{i*}_x(t,\widehat{x}(t),\widehat{u}(t))Q(t)g^i_x(t,\widehat{x}(t),\widehat{u}(t))\\
&+ \widetilde{g}^{j*}_x(t,\widehat{x}(t),\widehat{u}(t))Q(t)\widetilde{g}^j_x(t,\widehat{x}(t),\widehat{u}(t))\\
&+ g^{i*}_x(t,\widehat{x}(t),\widehat{u}(t))K^i(t) + K^i(t)g^i_x(t,\widehat{x}(t),\widehat{u}(t))\\
&+ \widetilde{g}^{j*}_x(t,\widehat{x}(t),\widehat{u}(t))\widetilde{K}^j(t) + \widetilde{K}^j(t)\widetilde{g}^j_x(t,\widehat{x}(t),\widehat{u}(t))\\
&+ \mathcal{H}_{xx}(t,\widehat{x}(t),\widehat{u}(t),q(t),k(t),\widetilde{R}(t) - q^*(t)\widetilde{g}(t,\widehat{x}(t),\widehat{u}(t)),\widetilde{k}(t))\\
&+ \widetilde{K}^j(t)h^j(t,\widehat{x}(t),\widehat{u}(t))\\
&+ (\widetilde{k}(t) - Q(t)\widetilde{g}(t,\widehat{x}(t),\widehat{u}(t)))h_x(t,\widehat{x}(t),\widehat{u}(t))\\
&+ h^*_x(t,\widehat{x}(t),\widehat{u}(t))(\widetilde{k}^*(t) - \widetilde{g}^*(t,\widehat{x}(t),\widehat{u}(t))Q(t))\}\,dt\\
&+ K^i(t)\,dw^i(t) + \widetilde{K}^j(t)\,dY^j(t), \quad t \in [0,1),
\end{aligned}\\
Q(1) = m_{xx}(\widehat{x}(1)).
\end{cases}
\tag{2.25}
$$

The rest of this section is to rewrite the maximum condition (2.9). We can verify the following:

$$
\begin{aligned}
&\widehat{\rho}^{-1}(t)\{H(t,\widehat{X}(t),v(t),a(t),b(t),\widetilde{b}(t)) - H(t,\widehat{X}(t),\widehat{u}(t),a(t),b(t),\widetilde{b}(t))\}\\
&= \Delta\mathcal{H}(t,\widehat{x}(t),\widehat{u}(t),q(t),k(t),\widetilde{R}(t) - q^*(t)\widetilde{g}(t,\widehat{x}(t),\widehat{u}(t)),\widetilde{k}(t);v(t))\\
&\quad - \langle q(t), \Delta\widetilde{g}(t;v(t))\Delta h(t;v(t))\rangle,
\end{aligned}
\tag{2.26}
$$

(2.27)
$$
\begin{aligned}
A(t) &= \begin{pmatrix} 0 & q^*(t) \\ q(t) & \widehat{\rho}(t)Q(t) \end{pmatrix}, \\
\Delta G(t;v(t)) &= \begin{pmatrix} 0 \\ \Delta g(t;v(t)) \end{pmatrix}, \\
\Delta \widetilde{G}(t;v(t)) &= \begin{pmatrix} \widehat{\rho}(t)\Delta h^*(t;v(t)) \\ \Delta \widetilde{g}(t;v(t)) \end{pmatrix}, \\
\Delta G(t;v(t))\Delta G^*(t;v(t)) &= \begin{pmatrix} 0 & 0 \\ 0 & \Delta g(t;v(t))\Delta g^*(t;v(t)) \end{pmatrix}, \\
&\Delta \widetilde{G}(t;v(t))\Delta \widetilde{G}^*(t;v(t)) \\
&= \begin{pmatrix} \widehat{\rho}^{\,2}(t)\Delta h^*(t;v(t))\Delta h(t;v(t)) & \widehat{\rho}(t)\Delta h^*(t;v(t))\Delta \widetilde{g}^*(t;v(t)) \\ \widehat{\rho}(t)\Delta \widetilde{g}(t;v(t))\Delta h(t;v(t)) & \Delta \widetilde{g}(t;v(t))\Delta \widetilde{g}^*(t;v(t)) \end{pmatrix},
\end{aligned}
$$

$$
\begin{aligned}
&\operatorname{tr}[A(t)(\Delta G(t;v(t))\Delta G^*(t;v(t)) + \Delta \widetilde{G}(t;v(t))\Delta \widetilde{G}^*(t;v(t)))] \\
&= 2q^*(t)\Delta \widetilde{g}(t;v(t))\Delta h(t;v(t))\widehat{\rho}(t) \qquad (2.28) \\
&\quad + \widehat{\rho}(t)\operatorname{tr}[Q(t)(\Delta g(t;v(t))\Delta g^*(t;v(t)) + \Delta \widetilde{g}(t;v(t))\Delta \widetilde{g}^*(t;v(t)))].
\end{aligned}
$$

Then the maximum condition (2.9) can be rewritten as

$$
\begin{aligned}
&\widehat{E}\int_0^1 \Delta \mathcal{H}(t,\widehat{x}(t),\widehat{u}(t),q(t),k(t),\widetilde{R}(t) - q^*(t)\widetilde{g}(t,\widehat{x}(t),\widehat{u}(t)),\widetilde{k}(t);v(t))\,dt \\
&+ \frac{1}{2}\widehat{E}\int_0^1 \operatorname{tr}\left[Q(t)\left(\Delta g(t;v(t))\Delta g^*(t;v(t)) + \Delta \widetilde{g}(t;v(t))\Delta \widetilde{g}^*(t;v(t))\right)\right]\,dt \qquad (2.29) \\
&\geq 0 \qquad \forall\, v(\cdot) \in U_{\mathrm{ad}}.
\end{aligned}
$$

THEOREM 2.1. *Assume that the hypothesis* $(\mathcal{A}1)$ *holds. Let* $\widehat{u}(\cdot)$ *be an optimal control and* $(r, R^i, i = 1,\dots,m; \widetilde{R}^j, j = 1,\dots,d), (q, k^i, i = 1,\dots,m; \widetilde{k}^j, j = 1,\dots,d)$, *and* $(Q, K^i, i = 1,\dots,m; \widetilde{K}^j, j = 1,\dots,d)$ *be the corresponding* $\mathcal{F}_t$*-adapted square-integrable solutions of BSDEs* (2.23), (2.24), (2.25)*, respectively. Then the maximum condition* (2.29) *holds.*

Note that Theorem 2.1 applies to an arbitrary admissible class U_{ad} of controls. In particular, it contains the following two special cases.

REMARK 2.1. *When* $U_{\mathrm{ad}} = \overline{U}_{\mathrm{ad}}$*, the maximum condition* (2.29) *implies the following inequality of expectations conditioned on the past and present observations* $\{Y(s) : 0 \leq s \leq t\}$:

$$
\begin{aligned}
&\widehat{E}[\Delta \mathcal{H}(t,\widehat{x}(t),\widehat{u}(t),q(t),k(t),\widetilde{R}(t) - q^*(t)\widetilde{g}(t,\widehat{x}(t),\widehat{u}(t)),\tilde{k}(t);v)|\mathcal{F}_t^Y] \\
&+ \frac{1}{2}\widehat{E}[\operatorname{tr}[Q(t)(\Delta g(t;v)\Delta g^*(t;v) + \Delta \widetilde{g}(t;v)\Delta \widetilde{g}^*(t;v))]|\mathcal{F}_t^Y] \qquad (2.30) \\
&\geq 0 \qquad \forall\, v \in U,\ a.s.a.e.\ (\textit{almost surely, almost everywhere}).
\end{aligned}
$$

REMARK 2.2. *When* $U_{\mathrm{ad}} = \widetilde{U}_{\mathrm{ad}}$*, the maximum condition* (2.29) *implies the following inequality of expectations conditioned on the present observation* $Y(t)$:

$$
\begin{aligned}
&\widehat{E}[\Delta \mathcal{H}(t,\widehat{x}(t),\widehat{u}(t),q(t),k(t),\widetilde{R}(t) - q^*(t)\widetilde{g}(t,\widehat{x}(t),\widehat{u}(t)),\widetilde{k}(t);v)|\,Y(t)] \\
&+ \frac{1}{2}\widehat{E}[\operatorname{tr}[Q(t)(\Delta g(t;v)\Delta g^*(t;v) + \Delta \widetilde{g}(t;v)\Delta \widetilde{g}^*(t;v))]|\,Y(t)] \qquad (2.31) \\
&\geq 0 \qquad \forall\, v \in U,\ a.s.a.e.
\end{aligned}
$$

We have derived a partially observed maximum principle, without involving at all the well-known Zakai equation, which is a stochastic PDE driven by the observation $Y(\cdot)$. However, our formulation of the maximum principle seems to be unsatisfactory since the adjoint equations (2.23)–(2.25) seem not to tell us what kind of functionals the adjoint processes (as their solutions) are of the initial state x_0 and the system noise $\dot{w}$, which is not available in practice. In the next two sections, we shall show that the adjoint equations do imply how the adjoint processes depend on the initial state x_0 and the system noise $\dot{w}$ in a special way, which is crucial in the computation of the conditional expectation appearing in the maximum condition (see (2.30) and (2.31), for example).

3. Relations among the adjoint processes. Let $\phi^{t,x}$ be the solution of the SDE

$$\begin{cases} d\phi^{t,x}(s) = (f-\widetilde{g}h)(s,\phi^{t,x}(s),\widehat{u}(s))\,ds + g^i(s,\phi^{t,x}(s),\widehat{u}(s))\,dw^i(s) \\ \qquad\qquad + \widetilde{g}^j(s,\phi^{t,x}(s),\widehat{u}(s))\,dY^j(s), \quad s\in(t,1], \\ \phi^{t,x}(t) = x, \end{cases} \tag{3.1}$$

$(r^{t,x}, R^{i,t,x}, i=1,\dots,m; \widetilde{R}^{j,t,x}, j=1,\dots,d)$ be the solution of the BSDE

$$\begin{cases} dr^{t,x}(s) = -\left[l(s,\phi^{t,x}(s),\widehat{u}(s)) + \widetilde{R}^{j,t,x}(s)h^j(s,\phi^{t,x}(s),\widehat{u}(s))\right]ds \\ \qquad\qquad + R^{i,t,x}(s)\,dw^i(s) + \widetilde{R}^{j,t,x}(s)\,dY^j(s), \quad s\in[t,1), \\ r^{t,x}(1) = m(\phi^{t,x}(1)), \end{cases} \tag{3.2}$$

$(q^{t,x}, k^{i,t,x}, i=1,\dots,m; \widetilde{k}^{j,t,x}, j=1,\dots,d)$ be the solution of the vector-valued BSDE

$$\begin{cases} dq^{t,x}(s) \\ = -\left[\mathcal{H}_x^*(s,\phi^{t,x}(s),\widehat{u}(s);q^{t,x}(s),k^{t,x}(s),\widetilde{R}^{t,x}(s) - q^{t,x,*}(s)\widetilde{g}(s,\phi^{t,x}(s),\widehat{u}(s)),\widetilde{k}^{t,x}(s)) \right. \\ \quad \left. + \widetilde{k}^{j,t,x}(s)h^j(s,\phi^{t,x}(s),\widehat{u}(s))\right]ds + k^{i,t,x}(s)\,dw^i(s) + \widetilde{k}^{j,t,x}(s)\,dY^j(s), \quad s\in[t,1), \\ q^{t,x}(1) = m_x^*(\phi^{t,x}(1)), \end{cases} \tag{3.3}$$

and $(Q^{t,x}, K^{i,t,x}, i=1,\dots,m; \widetilde{K}^{j,t,x}, j=1,\dots,d)$ be the solution of the matrix-valued BSDE

(3.4)
$$\begin{cases} dQ^{t,x}(s) \\ = -\{f_x^*(s,\phi^{t,x}(s),\widehat{u}(s))Q^{t,x}(s) + Q^{t,x}(s)f_x(s,\phi^{t,x}(s),\widehat{u}(s)) \\ \quad + g_x^{i*}(s,\phi^{t,x}(s),\widehat{u}(s))Q^{t,x}(s)g_x^i(s,\phi^{t,x}(s),\widehat{u}(s)) \\ \quad + \widetilde{g}_x^{j*}(s,\phi^{t,x}(s),\widehat{u}(s))Q^{t,x}(s)\widetilde{g}_x^j(s,\phi^{t,x}(s),\widehat{u}(s)) \\ \quad + g_x^{i*}(s,\phi^{t,x}(s),\widehat{u}(s))K^{i,t,x}(s) + K^{i,t,x}(s)g_x^i(s,\phi^{t,x}(s),\widehat{u}(s)) \\ \quad + \widetilde{g}_x^{j*}(s,\phi^{t,x}(s),\widehat{u}(s))\widetilde{K}^{j,t,x}(s) + \widetilde{K}^{j,t,x}(s)\widetilde{g}_x^j(s,\phi^{t,x}(s),\widehat{u}(s)) \\ \quad + \mathcal{H}_{xx}(s,\phi^{t,x}(s),\widehat{u}(s),q^{t,x}(s),k^{t,x}(s),\widetilde{R}^{t,x}(s) - q^{t,x,*}(s)\widetilde{g}(s,\phi^{t,x}(s),\widehat{u}(s)),\widetilde{k}^{t,x}(s)) \\ \quad + \widetilde{K}^{j,t,x}(s)h^j(s,\phi^{t,x}(s),\widehat{u}(s)) \\ \quad + (\widetilde{k}^{t,x}(s) - Q^{t,x}(s)\widetilde{g}(s,\phi^{t,x}(s),\widehat{u}(s)))h_x(s,\phi^{t,x}(s),\widehat{u}(s)) \\ \quad + h_x^*(s,\phi^{t,x}(s),\widehat{u}(s))(\widetilde{k}^{t,x,*}(s) - \widetilde{g}^*(s,\phi^{t,x}(s),\widehat{u}(s))Q^{t,x}(s))\}\,ds \\ \quad + K^{i,t,x}(s)\,dw^i(s) + \widetilde{K}^{j,t,x}(s)\,dY^j(s), \quad s\in[t,1), \\ Q^{t,x}(1) = m_{xx}(\phi^{t,x}(1)). \end{cases}$$

Obviously,

$$\begin{aligned} \phi^{0,x_0} &= \widehat{x}, \\ (r^{0,x_0}, R^{0,x_0}, \widetilde{R}^{0,x_0}) &= (r, R, \widetilde{R}), \\ (q^{0,x_0}, k^{i,0,x_0}, \widetilde{k}^{j,0,x_0}) &= (q, k^i, \widetilde{k}^j), \\ (Q^{0,x_0}, K^{i,0,x_0}, \widetilde{K}^{j,0,x_0}) &= (Q, K^i, \widetilde{K}^j). \end{aligned} \tag{3.5}$$

Under the hypothesis $(\mathcal{A}1)$, the stochastic flows $\phi^{t,x}$, $(r^{t,x}, R^{t,x}, \widetilde{R}^{t,x})$, and $(q^{t,x}, k^{i,t,x}, i=1,\dots,m; \widetilde{k}^{j,t,x}, j=1,\dots,d)$ are continuously differentiable with respect to x in suitable spaces (see Pardoux and Peng [10]). Their differentials $\phi_x^{t,x}$, $(\nabla r^{t,x}, \nabla R^{i,t,x}, i=1,\dots,m; \nabla\widetilde{R}^{j,t,x}, j=1,\dots,d)$, and $(q_x^{t,x}, k_x^{i,t,x}, i=1,\dots,m; \widetilde{k}_x^{j,t,x}, j=1,\dots,d)$ satisfy, respectively, the following SDE and BSDEs (see Pardoux and Peng [10]):

$$\begin{cases} d\phi_x^{t,x}(s) = (f - \widetilde{g}h)_x(s,\phi^{t,x}(s),\widehat{u}(s))\phi_x^{t,x}(s)\,ds \\ \qquad\qquad + g_x^i(s,\phi^{t,x}(s),\widehat{u}(s))\phi_x^{t,x}(s)\,dw^i(s) \\ \qquad\qquad + \widetilde{g}_x^j(s,\phi^{t,x}(s),\widehat{u}(s))\phi_x^{t,x}(s)\,dY^j(s), \quad s\in(t,1], \\ \phi_x^{t,x}(t) = I_{n\times n} : \text{identity matrix of order } n\times n, \end{cases} \tag{3.6}$$

$$\begin{cases} d\nabla r^{t,x}(s) = -\{\phi_x^{t,x,*}(s)l_x^*(s,\phi^{t,x}(s),\widehat{u}(s)) \\ \qquad\qquad + \phi_x^{t,x,*}(s)h_x^{j*}(s,\phi^{t,x}(s),\widehat{u}(s))\widetilde{R}^{j,t,x}(s) \\ \qquad\qquad + \nabla\widetilde{R}^{j,t,x}(s)h^j(s,\phi^{t,x}(s),\widehat{u}(s))\}\,ds \\ \qquad\qquad + \nabla R^{i,t,x}(s)\,dw^i(s) + \nabla\widetilde{R}^{j,t,x}(s)\,dY^j(s), \quad s\in[t,1), \\ \nabla r^{t,x}(1) = \phi_x^{t,x,*}(1)m_x^*(\phi^{t,x}(1)), \end{cases} \tag{3.7}$$

and

(3.8)
$$\begin{cases} dq_x^{t,x}(s) \\ = -\{\mathcal{H}_{xx}(s,\phi^{t,x}(s),\widehat{u}(s);q^{t,x}(s),k^{t,x}(s),\widetilde{R}^{t,x}(s)-q^{t,x,*}(s)\widetilde{g}(s,\phi^{t,x}(s),\widehat{u}(s)),\widetilde{k}^{t,x}(s)) \\ \quad\times\phi_x^{t,x}(s)+f_x^*(s,\phi^{t,x}(s),\widehat{u}(s))q_x^{t,x}(s)+g_x^{i*}(s,\phi^{t,x}(s),\widehat{u}(s))k_x^{i,t,x}(s) \\ \quad+h_x^{j*}(s,\phi^{t,x}(s),\widehat{u}(s))[\widetilde{R}_x^{j,t,x}(s)-q^{t,x,*}(s)\widetilde{g}_x^j(s,\phi^{t,x}(s),\widehat{u}(s)) \\ \quad-\widetilde{g}^{j*}(s,\phi^{t,x}(s),\widehat{u}(s))q_x^{t,x}(s)] \\ \quad+\widetilde{g}_x^{j*}(s,\phi^{t,x}(s),\widehat{u}(s))\widetilde{k}_x^{j,t,x}(s)+\widetilde{k}^{j,t,x}(s)h_x^j(s,\phi^{t,x}(s),\widehat{u}(s))\phi_x^{t,x}(s) \\ \quad+\widetilde{k}_x^{j,t,x}(s)h^j(s,\phi^{t,x}(s),\widehat{u}(s))\}\,ds+k_x^{i,t,x}(s)\,dw^i(s)+\widetilde{k}_x^{j,t,x}(s)\,dY^j(s), \quad s\in[t,1), \\ q_x^{t,x}(1)=m_{xx}(\phi^{t,x}(1))\phi_x^{t,x}(1). \end{cases}$$

Here and in the following, $\nabla r^{t,x} := (r_x^{t,x})^*$ and similar notations are made for $R^{i,t,x}$, $\widetilde{R}^{j,t,x}$ and other functions W, Z, V.

Using Itô's formula, we obtain the equation for the stochastic process $\{[\phi_x^{t,x}(s)]^{-1}; t \le s \le 1\}$:

$$\begin{cases} d[\phi_x^{t,x}(s)]^{-1} = -[\phi_x^{t,x}(s)]^{-1}\{(f-\widetilde{g}h)_x(s,\phi^{t,x}(s),\widehat{u}(s)) \\ \qquad\qquad -g_x^i(s,\phi^{t,x}(s),\widehat{u}(s))^2-\widetilde{g}_x^j(s,\phi^{t,x}(s),\widehat{u}(s))^2\}\,ds \\ \qquad\qquad -[\phi_x^{t,x}(s)]^{-1}g_x^i(s,\phi^{t,x}(s),\widehat{u}(s))\,dw^i(s) \\ \qquad\qquad -[\phi_x^{t,x}(s)]^{-1}\widetilde{g}_x^j(s,\phi^{t,x}(s),\widehat{u}(s))\,dY^j(s), \quad s\in(t,1], \\ [\phi_x^{t,x}(t)]^{-1} = I_{n\times n}. \end{cases} \tag{3.9}$$

From the uniqueness of the solutions of (3.3) and (3.4), we can check, using Itô's formula, the following theorem.

THEOREM 3.1. *Let the hypothesis* $(\mathcal{A}1)$ *be satisfied. Then, for* $s\in[t,1]$,

$$\begin{aligned} q^{t,x}(s) &= [\phi_x^{t,x,*}(s)]^{-1}\nabla r^{t,x}(s), \\ k^{i,t,x}(s) &= [\phi_x^{t,x,*}(s)]^{-1}\nabla R^{i,t,x}(s)-g_x^{i*}(s,\phi^{t,x}(s),\widehat{u}(s))q^{t,x}(s), \\ \widetilde{k}^{j,t,x}(s) &= [\phi_x^{t,x,*}(s)]^{-1}\nabla \widetilde{R}^{j,t,x}(s)-\widetilde{g}_x^{j*}(s,\phi^{t,x}(s),\widehat{u}(s))q^{t,x}(s); \\ Q^{t,x}(s) &= q_x^{t,x}(s)[\phi_x^{t,x}(s)]^{-1}, \\ K^{i,t,x}(s) &= k_x^{i,t,x}(s)[\phi_x^{t,x}(s)]^{-1}-Q^{t,x}(s)g_x^i(s,\phi^{t,x}(s),\widehat{u}(s)), \\ \widetilde{K}^{j,t,x}(s) &= \widetilde{k}_x^{j,t,x}(s)[\phi_x^{t,x}(s)]^{-1}-Q^{t,x}(s)\widetilde{g}_x^j(s,\phi^{t,x}(s),\widehat{u}(s)). \end{aligned} \tag{3.10}$$

Note that the uncertainty of the solution to a BSDE is introduced by the uncertainty of the drift and the terminal value, rather than by the terms of stochastic integrals. There are two different sources of uncertainty in the drifts and the terminal conditions of BSDEs (3.2)–(3.4): one comes from the initial state x_0 and the system noise $\dot{w}(\cdot)$, and the other comes from the observation noise $\dot{Y}(\cdot)$. The former enter into the drifts and the terminal conditions via $\phi^{t,x}(\cdot)$. Thus, we have reason to expect that the corresponding solutions depend on x_0 and $\dot{w}(\cdot)$ in the same manner. In fact, it is true at least under some reasonable conditions. The following theorem reveals such an assertion. We use D_x to denote the differential operator along the direction x, namely, $D_x := x^*\nabla$.

THEOREM 3.2. *Let the hypothesis* $(\mathcal{A}1)$ *be satisfied. Then*

$$\begin{aligned} &r^{t,x}(s) = W(s,\phi^{t,x}(s)), \qquad R^{i,t,x}(s) = D_{g^i(s,\phi^{t,x}(s),\widehat{u}(s))}W(s,\phi^{t,x}(s));\\ &q^{t,x}(s) = \nabla W(s,\phi^{t,x}(s)), \quad k^{i,t,x}(s) = D_{g^i(s,\phi^{t,x}(s),\widehat{u}(s))}\nabla W(s,\phi^{t,x}(s));\\ &Q^{t,x}(s) = \nabla^2 W(s,\phi^{t,x}(s)) := W_{xx}(s,\phi^{t,x}(s)). \end{aligned} \tag{3.11}$$

Here, $W(t,x) := r^{t,x}(t)$ *is a stochastic flow which is adapted to the history (including the present) of the observation* $Y(\cdot)$*. Let* $\mu^{t,x}$ *be the solution of the SDE*

$$\begin{cases} d\mu^{t,x}(s) = \mu^{t,x}(s)h^*(s,\phi^{t,x}(s),\widehat{u}(s))\,dY(s), \quad s\in(t,1],\\ \mu^{t,x}(t) = 1. \end{cases} \tag{3.12}$$

Then, $W(\cdot,\cdot)$ *has the following probabilistic interpretation:*

$$W(t,x) = E\bigg[\int_t^1 \mu^{t,x}(s)l(s,\phi^{t,x}(s),\widehat{u}(s))\,ds + \mu^{t,x}(1)m(\phi^{t,x}(1)) \,\bigg|\, \mathcal{F}_t^Y\bigg]. \tag{3.13}$$

Proof of Theorem 3.2. From the uniqueness of the solutions of (3.1) and (3.2), we derive

$$\phi^{t,x}(\tau) = \phi^{s,\phi^{t,x}(s)}(\tau), \quad r^{t,x}(\tau) = r^{s,\phi^{t,x}(s)}(\tau), \quad t\le s\le\tau\le 1. \tag{3.14}$$

The first relation of (3.11) then follows.

It can be verified, via Malliavin's calculus for the BSDE (3.2), as in Pardoux and Peng [10], that

$$R^{i,t,x}(s) = g^{i*}(s,\phi^{t,x}(s),\widehat{u}(s))[\phi_x^{t,x,*}(s)]^{-1}\nabla r^{t,x}(s). \tag{3.15}$$

In view of the first relation of (3.11), we get the second relation of (3.11).

The third relation of (3.11) comes from the second relation of (3.14) and the first relation of (3.10), while the fourth relation of (3.11) comes from the second relation of (3.10), and the fifth relation of (3.11) comes from the fourth relation of (3.10).

The probabilistic interpretation (3.13) of W can be obtained from computing the quantity $r^{t,x}(s)\mu^{t,x}(s)$ with Itô's formula. □

It is worth noting that, if $\widehat{u}(t) = \beta(t,Y(t))\ \forall\, t\in[0,1]$, for some U-valued Borel function β on $[0,1]\times\mathbb{R}^d$, then $W(t,x) = \overline{W}(t,x,Y(t))$ with

$$\begin{aligned} \overline{W}(t,x,y) := {}& E\int_t^1 \zeta^{t,x,y}(s)l(s,\Phi^{t,x,y}(s),\beta(s,y+Y(s)-Y(t)))\,ds\\ &+ E[\zeta^{t,x,y}(1)m(\Phi^{t,x,y}(1))] \quad \forall\, t\in[0,1],\ x\in\mathbb{R}^n,\ y\in\mathbb{R}^d. \end{aligned} \tag{3.16}$$

Here, $\Phi^{t,x,y}$ and $\zeta^{t,x,y}$ are the solutions of the SDEs

$$\begin{cases} d\Phi^{t,x,y}(s) = (f-\widetilde{g}h)(s,\Phi^{t,x,y}(s),\beta(s,y+Y(s)-Y(t)))\,ds\\ \qquad\qquad + g^i(s,\Phi^{t,x,y}(s),\beta(s,y+Y(s)-Y(t)))\,dw^i(s)\\ \qquad\qquad + \widetilde{g}^j(s,\Phi^{t,x,y}(s),\beta(s,y+Y(s)-Y(t)))\,dY^j(s), \quad s\in(t,1],\\ \Phi^{t,x,y}(t) = x \end{cases} \tag{3.17}$$

and

$$\begin{cases} d\zeta^{t,x,y}(s) = \zeta^{t,x,y}(s)h^*(s,\Phi^{t,x,y}(s),\beta(s,y+Y(s)-Y(t)))\,dY(s), \quad s\in(t,1],\\ \zeta^{t,x,y}(t)=1, \end{cases} \tag{3.18}$$

respectively. It can be checked that $\overline{W}$ is the unique viscosity solution of the following Hamilton–Jacobi equation of second order:

$$\begin{cases} \dfrac{1}{2}\mathrm{tr}[(gg^*+\widetilde{g}\,\widetilde{g}^*)(t,x,\beta(t,y))\overline{W}_{xx}]\\ +\mathrm{tr}[\widetilde{g}(t,x,\beta(t,y))\overline{W}_{xy}]+\dfrac{1}{2}\mathrm{tr}(\overline{W}_{yy})\\ +\langle f(t,x,\beta(t,y)),\overline{W}_x^*\rangle+\langle h(t,x,\beta(t,y)),\overline{W}_y^*\rangle+l(t,x,\beta(t,y))\\ =0, \quad t\in[0,1),\ x\in\mathbb{R}^n,\ y\in\mathbb{R}^d,\\ \overline{W}(1,x,y)=m(x), \quad x\in\mathbb{R}^n,\ y\in\mathbb{R}^d, \end{cases} \tag{3.19}$$

at least when the function β is bounded and continuous and when the coefficients $f,g,\widetilde{g},h,$ and l are jointly continuous with respect to all their arguments.

4. BSPDEs of adjoint vector fields. In this section, adjoint vector fields are introduced as the solutions of BSPDEs, and their relations are established. The adjoint processes are then characterized in terms of the adjoint vector fields, their differentials and Hessians, along the optimal state process $\widehat{x}(\cdot)$.

For all $v\in U$, define the following operators:

$$\begin{aligned} \mathcal{L}^v(t,x)Z &:= \frac{1}{2}\mathrm{tr}[(gg^*+\widetilde{g}\,\widetilde{g}^*)(t,x,v)\nabla^2 Z]+\langle f(t,x,v),\nabla Z\rangle,\\ \mathcal{L}(t,x)Z &:= \mathcal{L}^{\hat{u}(t)}(t,x)Z \quad \forall\, Z\in C^2(\mathbb{R}^n,\mathbb{R});\\ \mathcal{M}^{j,v}(t,x)V &:= \langle\widetilde{g}^j(t,x,v),\nabla V\rangle+h^j(t,x,v)V,\\ \mathcal{M}^j(t,x)V &:= \mathcal{M}^{j,\hat{u}(t)}(t,x)V \quad \forall\, V\in C^1(\mathbb{R}^n,\mathbb{R}). \end{aligned} \tag{4.1}$$

For a matrix-valued smooth function of $x\in\mathbb{R}^n$, say $\mathcal{U}:=(\mathcal{U}^{ij})\in C^2(\mathbb{R}^n,\mathbb{R}^{n_1\times n_2})$, set $\mathcal{L}(t,x)\mathcal{U}:=(\mathcal{L}(t,x)\mathcal{U}^{ij})$ and $\mathcal{M}^{j_1}(t,x)\mathcal{U}:=(\mathcal{M}^{j_1}(t,x)\mathcal{U}^{ij})$.

Consider the following BSPDEs:

$$\begin{cases} dZ(t,x) = -\,[\mathcal{L}(t,x)Z(t,x)+l(t,x,\widehat{u}(t))+\mathcal{M}^j(t,x)V^j(t,x)]\,dt\\ \qquad\qquad +V^j(t,x)\,dY^j(t), \quad t\in(0,1],\ x\in\mathbb{R}^n,\\ Z(1,x)=m(x), \quad x\in\mathbb{R}^n, \end{cases} \tag{4.2}$$

$$\begin{cases} d\lambda(t,x)\\ =-\{\mathcal{L}(t,x)\lambda(t,x)+\mathcal{M}^j(t,x)\theta^j(t,x)\\ \quad +\mathcal{H}_x^*(t,x,\widehat{u}(t),\lambda(t,x),\lambda_x(t,x)g(t,x,\widehat{u}(t)),V(t,x),\theta(t,x)+\lambda_x(t,x)\widetilde{g}(t,x,\widehat{u}(t)))\}\,dt\\ \quad +\theta^j(t,x)\,dY^j(t), \quad t\in(0,1],\ x\in\mathbb{R}^n,\\ \lambda(1,x)=m_x^*(x), \quad x\in\mathbb{R}^n, \end{cases} \tag{4.3}$$

and

(4.4)
$$
\begin{cases}
d\Lambda(t,x) \\
\quad = -\,[\mathcal{L}(t,x)\Lambda(t,x) + \mathcal{M}^j(t,x)\Theta^j(t,x) \\
\qquad + \mathcal{H}_{xx}(t,x,\widehat{u}(t),\lambda(t,x),\lambda_x(t,x)g(t,x,\widehat{u}(t)),V(t,x),\theta(t,x)+\lambda_x(t,x)\widetilde{g}(t,x,\widehat{u}(t))) \\
\qquad + f_x^*(t,x,\widehat{u}(t))\Lambda(t,x) + \Lambda(t,x)f_x(t,x,\widehat{u}(t)) \\
\qquad + g_x^{i*}(t,x,\widehat{u}(t))\Lambda(t,x)g_x^i(t,x,\widehat{u}(t)) + \widetilde{g}_x^{j*}(t,x,\widehat{u}(t))\Lambda(t,x)\widetilde{g}_x^j(t,x,\widehat{u}(t)) \\
\qquad + g_x^{i*}(t,x,\widehat{u}(t))D_{g^i(t,x,\hat{u}(t))}\Lambda(t,x) + D_{g^i(t,x,\hat{u}(t))}\Lambda(t,x)g_x^i(t,x,\widehat{u}(t)) \\
\qquad + \widetilde{g}_x^{i*}(t,x,\widehat{u}(t))D_{\tilde{g}^i(t,x,\hat{u}(t))}\Lambda(t,x) + D_{\tilde{g}^i(t,x,\hat{u}(t))}\Lambda(t,x)\widetilde{g}_x^i(t,x,\widehat{u}(t)) \\
\qquad + \widetilde{g}_x^{j*}(t,x,\widehat{u}(t))(t,x,\widehat{u}(t))\Theta^j(t,x) + \Theta^j(t,x)\widetilde{g}_x^j(t,x,\widehat{u}(t)) \\
\qquad + \theta^j(t,x)h_x^j(t,x,\widehat{u}(t)) + h_x^{j*}(t,x,\widehat{u}(t))\theta^{j*}(t,x)]\,dt \\
\qquad + \Theta^j(t,x)\,dY^j(t), \quad t\in(0,1],\ x\in\mathbb{R}^n, \\
\quad \Lambda(1,x) = m_{xx}(x), \quad x\in\mathbb{R}^n.
\end{cases}
$$

There is a close relation between the solutions of (3.2)–(3.4) and the solutions of (4.2)–(4.4), which is stated in the following theorem:

THEOREM 4.1. *Assume that the SPDEs* (4.2)–(4.4) *have unique* $\mathcal{F}_t^Y$*-adapted smooth solutions* (Z,V), $(\lambda,\theta^j, j=1,\dots,d)$, *and* $(\Lambda,\Theta^j, j=1,\dots,d)$, *respectively, and their partial derivatives of arbitrary order with respect to* x *are uniformly bounded. Then we have the following relations:*

$$
\begin{aligned}
r^{t,x}(s) &= Z(s,\phi^{t,x}(s)), \\
R^{i,t,x}(s) &= D_{g^i(s,\phi^{t,x}(s),\hat{u}(s))}Z(s,\phi^{t,x}(s)), \quad i=1,\dots,m, \\
\widetilde{R}^{j,t,x}(s) &= V^j(s,\phi^{t,x}(s)) + D_{\tilde{g}^j(s,\phi^{t,x}(s),\hat{u}(s))}Z(s,\phi^{t,x}(s)), \quad j=1,\dots,d; \\
q^{t,x}(s) &= \lambda(s,\phi^{t,x}(s)), \\
k^{i,t,x}(s) &= D_{g^i(s,\phi^{t,x}(s),\hat{u}(s))}\lambda(s,\phi^{t,x}(s)), \quad i=1,\dots,m, \\
\widetilde{k}^{j,t,x}(s) &= \theta^j(s,\phi^{t,x}(s)) + D_{\tilde{g}^j(s,\phi^{t,x}(s),\hat{u}(s))}\lambda(s,\phi^{t,x}(s)), \quad j=1,\dots,d; \\
Q^{t,x}(s) &= \Lambda(s,\phi^{t,x}(s)), \\
K^{i,t,x}(s) &= D_{g^i(s,\phi^{t,x}(s),\hat{u}(s))}\Lambda(s,\phi^{t,x}(s)), \quad i=1,\dots,m, \\
\widetilde{K}^{j,t,x}(s) &= \Theta^j(s,\phi^{t,x}(s)) + D_{\tilde{g}^j(s,\phi^{t,x}(s),\hat{u}(s))}\Lambda(s,\phi^{t,x}(s)), \quad j=1,\dots,d.
\end{aligned}
\tag{4.5}
$$

Proof of Theorem 4.1. We use the generalized Itô–Kunita formula (see Kunita [6] for details) to compute the quantities $Z(s,\phi^{t,x}(s)), \lambda(s,\phi^{t,x}(s)), \Lambda(s,\phi^{t,x}(s))$, and we observe that the right hand sides of the equalities in (4.5) solve the BSDEs (3.2)–(3.4), respectively. According to the uniqueness of the solutions of (3.2)–(3.4) (see Pardoux and Peng [9]), we get the desired results. □

Note that if (Z,V) is a smooth solution of the BSPDE (4.2), then $Z=W$. Hence, W can be viewed as a probabilistic interpretation of Z. In a heuristic way, $V^j(t,x)$ should be interpreted as $\widetilde{R}^{j,t,x}(t) - D_{\tilde{g}^j(t,x,\hat{u}(t))}Z(t,x)$.

From (3.5) and Theorem 4.1, we see that the solutions of BSPDEs (4.2)–(4.4) are closely related with the adjoint processes: the former, if they exist, help to show how the latter depend on the initial state x_0 and the system noise $\dot{w}(\cdot)$ in a special way. For this reason, we call the former the adjoint vector fields. The relations among the adjoint vector fields are stated in the following theorem.

THEOREM 4.2. *Assume that* 1) *the coefficients* $f, g, \widetilde{g}, h, l$, *and* m *are smooth in the variable* x, *and they are bounded, together with their partial derivatives with respect to* x; 2) $gg^*(t,x,\widehat{u}(t)) \geq \delta I_{n\times n}$ *for some real number* $\delta > 0$. *Then the SPDEs* (4.2)–(4.4) *have unique smooth solutions* (Z,V), $(\lambda,\theta^j, j=1,\dots,d)$, *and* $(\Lambda,\Theta^j, j=1,\dots,d)$. *Moreover, the solutions have the following relations:*

$$\begin{aligned} &\lambda = \nabla Z, \quad \theta^j = \nabla V^j;\\ &\Lambda = \lambda_x = \nabla^2 Z, \quad \Theta^j = \theta^j_x = \nabla^2 V^j, \quad j=1,\dots,d. \end{aligned} \tag{4.6}$$

Proof of Theorem 4.2. The uniqueness is obtained from Theorem 4.1 and the uniqueness of the solutions of the BSDEs (3.2)–(3.4).

The existence of a smooth solution of the BSPDE (4.2), under the conditions of Theorem 4.2, is proved by Zhou [14, Remark 4.1, p. 290].

Let (Z,V) be a smooth solution of (4.2). Take differentials on both sides of (4.2), and we see that the differential $(\nabla Z, \nabla V^j,\ j=1,\dots,d)$ satisfies the BSPDE (4.3). Thus, $(\nabla Z, \nabla V^j,\ j=1,\dots,d)$ is a smooth solution of (4.3).

Since

(4.7)
$$\begin{aligned} &\{\mathcal{L}(t,x)\lambda(t,x)\}_x\\ &= \mathcal{L}(t,x)\lambda_x(t,x) + \lambda_x(t,x)f_x(t,x,\widehat{u}(t))\\ &\quad + D_{g^i(t,x,\hat{u}(t))}\lambda_x(t,x)g^i_x(t,x,\widehat{u}(t)) + D_{\tilde{g}^j(t,x,\hat{u}(t))}\lambda_x(t,x)\widetilde{g}^j_x(t,x,\widehat{u}(t)),\\ &\quad \{\mathcal{M}^j(t,x)\theta^j(t,x)\}_x\\ &= \mathcal{M}^j(t,x)\theta^j_x(t,x) + \theta^j_x(t,x)\widetilde{g}^j_x(t,x,\widehat{u}(t)) + \theta^j(t,x)h^j_x(t,x,\widehat{u}(t)),\\ &\quad \{\mathcal{H}^*_x(t,x,\widehat{u}(t),\lambda(t,x),\lambda_x(t,x)g(t,x,\widehat{u}(t)),V(t,x),\theta(t,x)+\lambda_x(t,x)\widetilde{g}(t,x,\widehat{u}(t)))\}_x\\ &= \mathcal{H}_{xx}(t,x,\widehat{u}(t),\lambda(t,x),\lambda_x(t,x)g(t,x,\widehat{u}(t)),V(t,x),\theta(t,x)+\lambda_x(t,x)\widetilde{g}(t,x,\widehat{u}(t)))\\ &\quad + f^*_x(t,x,\widehat{u}(t))\lambda_x(t,x) + g^{i*}_x(t,x,\widehat{u}(t))D_{g^i(t,x,\hat{u}(t))}\lambda_x(t,x)\\ &\quad + h^{j*}_x(t,x,\widehat{u}(t))V^j_x(t,x) + \widetilde{g}^{j*}_x(t,x,\widehat{u}(t))[D_{\tilde{g}^j(t,x,\hat{u}(t))}\lambda_x(t,x)+\theta^j_x(t,x)]\\ &\quad + g^*_x(t,x,\widehat{u}(t))\lambda_x(t,x)g_x(t,x,\widehat{u}(t)), \end{aligned}$$

we have that $(Q := \lambda_x, K^j := \theta^j_x, j=1,\dots,d)$ satisfies the BSPDE (4.4). Hence, it is a smooth solution of (4.4). The relation (4.6) is then obtained. □

Combining Theorems 4.1 and 4.2, we have the following theorem.

THEOREM 4.3. *Let the hypotheses of Theorem* 4.2 *be satisfied, and* $(Z, V^j, j=1,\dots,d)$ *be the unique solution of the BSPDE* (4.2). *Then*

$$\begin{aligned} r^{t,x}(s) &= Z(s,\phi^{t,x}(s)),\\ R^{i,t,x}(s) &= D_{g^i(s,\phi^{t,x}(s),\hat{u}(s))}Z(s,\phi^{t,x}(s)), \quad i=1,\dots,m,\\ \widetilde{R}^{j,t,x}(s) &= V^j(s,\phi^{t,x}(s)) + D_{\tilde{g}^j(s,\phi^{t,x}(s),\hat{u}(s))}Z(s,\phi^{t,x}(s)), \quad j=1,\dots,d;\\ q^{t,x}(s) &= \nabla Z(s,\phi^{t,x}(s)),\\ k^{i,t,x}(s) &= D_{g^i(s,\phi^{t,x}(s),\hat{u}(s))}\nabla Z(s,\phi^{t,x}(s)), \quad i=1,\dots,m,\\ \widetilde{k}^{j,t,x}(s) &= \nabla V^j(s,\phi^{t,x}(s)) + D_{\tilde{g}^j(s,\phi^{t,x}(s),\hat{u}(s))}\nabla Z(s,\phi^{t,x}(s)), \quad j=1,\dots,d;\\ Q^{t,x}(s) &= \nabla^2 Z(s,\phi^{t,x}(s)),\\ K^{i,t,x}(s) &= D_{g^i(s,\phi^{t,x}(s),\hat{u}(s))}\nabla^2 Z(s,\phi^{t,x}(s)), \quad i=1,\dots,m,\\ \widetilde{K}^{j,t,x}(s) &= \nabla^2 V^j(s,\phi^{t,x}(s)) + D_{\tilde{g}^j(s,\phi^{t,x}(s),\hat{u}(s))}\nabla^2 Z(s,\phi^{t,x}(s)), \quad j=1,\dots,d. \end{aligned} \tag{4.8}$$

Before closing this section, we remark that, if the coefficients $f, g, \widetilde{g}, h, l$, and m are smooth with respect to all their arguments, they are bounded together with their partial derivatives, and $\widehat{u}(t) = \beta(t, Y(t))\, \forall\, t \in [0,1]$, for some smooth function $\beta : [0,1] \times \mathbb{R}^d \to U$ with bounded derivatives of arbitrary order, then the Hamilton–Jacobi equation (3.19) has a unique bounded smooth solution $\overline{W}$ and the BSPDE (4.2) has the following unique bounded smooth solution: $Z(t,x) = \overline{W}(t,x,Y(t))$, $V^j(t,x) = \overline{W}_{y^j}(t,x,Y(t))$, $j = 1, \dots, d$.

5. Versions of Theorem 2.1 and comparison with the existing results. Combining Theorems 2.1 and 3.2, we have the following theorem.

THEOREM 5.1. *Let the hypothesis* $(\mathcal{A}1)$ *be satisfied, and* $h := h(t,x)$, $\widetilde{g} := \widetilde{g}(t,x)$. *Assume that* $\widehat{u}(\cdot)$ *is an optimal control. Let* $\mu^{t,x}(\cdot)$ *be the solution of* (3.12), *and set*

$$W(t,x) := E\bigg[\int_t^1 \mu^{t,x}(s) l(s, \phi^{t,x}(s), \widehat{u}(s))\, ds + \mu^{t,x}(1) m(\phi^{t,x}(1)) \,\bigg|\, \mathcal{F}_t^Y\bigg] \tag{5.1}$$

and

$$\overline{\mathcal{L}}^{\,v}(t,x) W(t,x) := \frac{1}{2}\mathrm{tr}[(gg^*)(t,x,v)\nabla^2 W(t,x)] + \langle f(t,x,v), \nabla W(t,x)\rangle. \tag{5.2}$$

Then, the following maximum condition holds:

$$\begin{aligned} \widehat{E}\int_0^1 \big\{ &\overline{\mathcal{L}}^{\,v(t)}(t, \widehat{x}(t)) W(t, \widehat{x}(t)) + l(t, \widehat{x}(t), v(t)) \\ &- \overline{\mathcal{L}}^{\,\widehat{u}(t)}(t, \widehat{x}(t)) W(t, \widehat{x}(t)) - l(t, \widehat{x}(t), \widehat{u}(t)) \big\}\, dt \\ &\geq 0 \qquad \forall v(\cdot) \in U_{\mathrm{ad}}. \end{aligned} \tag{5.3}$$

Proof of Theorem 5.1. Since the observation term h and the correlation term $\widetilde{g}$ do not depend on the control variable, we derive from Theorem 3.2 that

$$\begin{aligned} &\Delta\mathcal{H}(t, \widehat{x}(t), \widehat{u}(t), q(t), k(t), \widetilde{R}(t) - q^*(t)\widetilde{g}(t, \widehat{x}(t), \widehat{u}(t)), \widetilde{k}(t); v(t)) \\ &= \langle q(t), \Delta f(t; v(t))\rangle + \langle k^i(t), \Delta g^i(t; v(t))\rangle + \Delta l(t; v(t)) \\ &= \langle \nabla W(t, \widehat{x}(t)), \Delta f(t; v(t))\rangle + \Delta l(t; v(t)) \\ &\quad + \mathrm{tr}[\nabla^2 W(t, \widehat{x}(t))(g(t, \widehat{x}(t), \widehat{u}(t)) \Delta g^*(t; v(t)))], \\ &\quad \mathrm{tr}[Q(t)(\Delta g(t; v(t)) \Delta g^*(t; v(t)) + \Delta\widetilde{g}(t; v(t)) \Delta\widetilde{g}^*(t; v(t)))] \\ &= \mathrm{tr}[Q(t)(\Delta g(t; v(t)) \Delta g^*(t; v(t)))] \\ &= \mathrm{tr}[\nabla^2 W(t, \widehat{x}(t))(\Delta g(t; v(t)) \Delta g^*(t; v(t)))]. \end{aligned} \tag{5.4}$$

The maximum condition (5.3) then follows from Theorem 2.1. □

COROLLARY 5.1. *Assume that* 1) *the hypothesis* $(\mathcal{A}1)$ *holds;* 2) $\widetilde{g} := \widetilde{g}(t,x), h := h(t,x)$*;* 3) $U_{\mathrm{ad}} = \widetilde{U}_{\mathrm{ad}}$. *Let* $\widehat{u}(\cdot)$ *be an optimal control. Then the following maximum condition holds:*

$$\begin{aligned} \widehat{E}\Big[&\overline{\mathcal{L}}^{\,v}(t, \widehat{x}(t)) \overline{W}(t, \widehat{x}(t), Y(t)) + l(t, \widehat{x}(t), v) \\ &- \overline{\mathcal{L}}^{\,\widehat{u}(t)}(t, \widehat{x}(t)) \overline{W}(t, \widehat{x}(t), Y(t)) - l(t, \widehat{x}(t), \widehat{u}(t)) \,|\, Y(t) \Big] \\ &\geq 0 \qquad \forall\, v \in U, \text{ a.s.a.e.}, \end{aligned} \tag{5.5}$$

with $\overline{W}$ *being defined by* (3.16)–(3.18).

The partially observed optimal control with the admissible class $\widetilde{U}_{\rm ad}$ of controls has been studied by Fleming [4]. His observation model is of the following form:

$$\begin{cases} dY(t) = h(t,x(t),Y(t),v(t))\,dt + \sigma^i(t,x(t),Y(t),v(t))\,dw^i(t) \\ \qquad\qquad + \widetilde{\sigma}^j(t,x(t),v(t))\,d\widetilde{w}^j(t), \quad t\in(0,1], \\ Y(0) = y_0, \end{cases} \tag{5.6}$$

and it is more general than (1.2). His terminal time T is the least time of the system state going beyond a bounded domain, while ours is the fixed time $T=1$. He made the following nondegenerate hypothesis:

$$\begin{pmatrix} g & \widetilde{g} \\ \sigma & \widetilde{\sigma} \end{pmatrix} \begin{pmatrix} g^* & \sigma^* \\ \widetilde{g}^* & \widetilde{\sigma}^* \end{pmatrix} \geq \delta I_{(n+d)\times(n+d)} \quad \text{for some } \delta>0, \tag{5.7}$$

which implies, in our situation (i.e., $\sigma=0, \widetilde{\sigma}=I_{d\times d}$), the following condition: $gg^*\geq \delta I_{n\times n}$ for some $\delta>0$. Corollary 5.1 allows gg^* to be degenerate and is new.

COROLLARY 5.2. *Assume that* 1) *the hypothesis* $(\mathcal{A}1)$ *holds;* 2) $\widetilde{g}:=\widetilde{g}(t,x), h:=h(t,x)$*;* 3) $U_{\rm ad}=\overline{U}_{\rm ad}$*. Let* $\widehat{u}(\cdot)$ *be an optimal control. Then the following maximum condition holds:*

$$\begin{aligned} &\widehat{E}\Big[\overline{\mathcal{L}}^{\,v}(t,\widehat{x}(t))W(t,\widehat{x}(t)) + l(t,\widehat{x}(t),v) \\ &\quad - \overline{\mathcal{L}}^{\,\hat{u}(t)}(t,\widehat{x}(t))W(t,\widehat{x}(t)) - l(t,\widehat{x}(t),\hat{u}(t))\,|\,\mathcal{F}_t^Y\Big] \\ &\quad \geq 0 \qquad \forall\, v\in U,\ a.s.a.e. \end{aligned} \tag{5.8}$$

Combining Theorems 2.1 and 4.3, we have the following theorem.

THEOREM 5.2. *Let the hypotheses of Theorem* 4.2 *be satisfied. Let* $\widehat{u}(\cdot)$ *be an optimal control, and* (Z,V) *be the unique solution of the BSPDE*

$$\begin{cases} dZ(t,x) = -\left[\mathcal{L}(t,x)Z(t,x) + l(t,x,\widehat{u}(t)) + \mathcal{M}^j(t,x)V^j(t,x)\right]dt \\ \qquad\qquad + V^j(t,x)\,dY^j(t), \quad t\in(0,1],\ x\in\mathbb{R}^n, \\ Z(1,x) = m(x), \quad x\in\mathbb{R}^n. \end{cases} \tag{5.9}$$

Then the following maximum condition holds:

$$\begin{aligned} \widehat{E}\int_0^1 &\{\mathcal{L}^{v(t)}(t,\widehat{x}(t))Z(t,\widehat{x}(t)) + l(t,\widehat{x}(t),v(t)) + \mathcal{M}^{j,v(t)}(t,\widehat{x}(t))V^j(t,\widehat{x}(t)) \\ &- \mathcal{L}^{\hat{u}(t)}(t,\widehat{x}(t))Z(t,\widehat{x}(t)) - l(t,\widehat{x}(t),\widehat{u}(t)) - \mathcal{M}^{j,\hat{u}(t)}(t,\widehat{x}(t))V^j(t,\widehat{x}(t))\}\,dt \\ &\geq 0 \qquad \forall\, v(\cdot)\in U_{\rm ad}. \end{aligned} \tag{5.10}$$

Proof of Theorem 5.2. From Theorem 4.3, we derive

$$\begin{aligned} &\Delta\mathcal{H}(t,\widehat{x}(t),\widehat{u}(t),q(t),k(t),\widetilde{R}(t) - q^*(t)\widetilde{g}(t,\widehat{x}(t),\widehat{u}(t)),\widetilde{k}(t);v(t)) \\ =&\langle\nabla Z(t,\widehat{x}(t)),\Delta f(t;v(t))\rangle + \langle\nabla V^j(t,\widehat{x}(t)),\Delta g^j(t;v(t))\rangle \\ &+ V^j(t,\widehat{x}(t))\Delta h^j(t;v(t)) + \Delta l(t;v(t)) \\ &+ \mathrm{tr}[\nabla^2 Z(t,\widehat{x}(t))(g(t,\widehat{x}(t),\widehat{u}(t))\Delta g^*(t;v(t)) + \widetilde{g}(t,\widehat{x}(t),\widehat{u}(t))\Delta\widetilde{g}^*(t;v(t)))], \\ &\mathrm{tr}[Q(t)(\Delta g(t;v(t))\Delta g^*(t;v(t)) + \Delta\widetilde{g}(t;v(t))\Delta\widetilde{g}^*(t;v(t)))] \\ =&\mathrm{tr}[\nabla^2 Z(t,\widehat{x}(t))(\Delta g(t;v(t))\Delta g^*(t;v(t)) + \Delta\widetilde{g}(t;v(t))\Delta\widetilde{g}^*(t;v(t)))]. \end{aligned} \tag{5.11}$$

The maximum condition (5.10) then follows from Theorem 2.1. □

COROLLARY 5.3. *Let the hypotheses of Theorem* 4.2 *be satisfied and* $U_{\rm ad} = \overline{U}_{\rm ad}$. *Let* $\widehat{u}(\cdot)$ *be an optimal control and* (Z, V) *be the unique solution of the BSPDE* (5.9). *Then the following maximum condition holds:*

(5.12)
$$\begin{aligned}
&\widehat{E}\Big[\mathcal{L}^v(t,\widehat{x}(t))Z(t,\widehat{x}(t)) + l(t,\widehat{x}(t),v) + \mathcal{M}^{j,v}(t,\widehat{x}(t))V^j(t,\widehat{x}(t))\\
&\quad - \mathcal{L}^{\hat{u}(t)}(t,\widehat{x}(t))Z(t,\widehat{x}(t)) - l(t,\widehat{x}(t),\widehat{u}(t)) - \mathcal{M}^{j,\hat{u}(t)}(t,\widehat{x}(t))V^j(t,\widehat{x}(t))|\,\mathcal{F}_t^Y\Big]\\
&\quad \geq 0 \quad \forall\, v \in U,\ a.s.a.e.
\end{aligned}$$

The partially observed optimal control with the admissible control class $\overline{U}_{\rm ad}$, has been studied by Kwakernaak [7]; Bensoussan [2]; Haussmann [5]; Baras, Elliott, and Kohlmann [1]; Zhou [13]; and Li and Tang [8]. Corollary 5.2 essentially covers the partially observed maximum principles of Bensoussan [2]; Haussmann [5]; and Baras, Elliott, and Kohlmann [1]; and it generalizes them at least in two of the following respects: 1) gg^* may be degenerate; 2) the control may appear in the diffusion coefficient g; 3) correlated noises may be present between the system and the observation (i.e., the correlation coefficient $\widetilde{g}$ is not necessarily zero); 4) the initial state x_0 does not necessarily have a regular density function. Note that Bensoussan [2] considered the case of $\widetilde{g} = 0$, and characterized his adjoint processes via the BSPDE (5.9) but in the sense of strong solution; in a heuristic way, the formula (5.1) should be the probabilistic interpretation of his adjoint processes.

Zhou [13], like Kwakernaak [7] and Bensoussan [2], treated a partially observed optimal control problem as an optimal control problem with full information, but for the Zakai equation, which is a stochastic PDE driven by the observation. His result excludes both the case when the initial state has no regular density function and the case when gg^* is degenerate while $\widetilde{g} \neq 0$. Corollaries 5.2 and 5.3 consider both cases and therefore are new. They partially answer Fleming's question (see Fleming [4, p. 209]).

It is worth pointing out that our derivation of Corollary 5.2 does not involve the Zakai equation at all and avoids the complicated stochastic analysis in infinite-dimensional spaces of Bensoussan [2], Haussmann [5], and Zhou [13].

Combining Theorem 2.1 and Remark 2.1 with Theorem 3.1, we obtain the following theorem.

THEOREM 5.3. *Let the hypothesis* $(\mathcal{A}1)$ *be satisfied,* $U_{\rm ad} = \overline{U}_{\rm ad}$, *and* $\widehat{u}(\cdot)$ *be an optimal control. Let* $\phi^{t,x}(\cdot)$ *solve the SDE* (3.1) *and* $(r^{t,x}(\cdot), R^{t,x}(\cdot), \widetilde{R}^{t,x}(\cdot))$ *solve the BSDE* (3.2). *Then the maximum condition* (2.30) *holds with*

(5.13)
$$\begin{aligned}
q(s) &= [\phi_x^{0,x_0,*}(s)]^{-1}\nabla r^{0,x_0}(s),\\
k^i(s) &= [\phi_x^{0,x_0,*}(s)]^{-1}\nabla R^{i,0,x_0}(s) - g_x^{i*}(s,\phi^{0,x_0}(s),\widehat{u}(s))q(s),\\
\widetilde{k}^j(s) &= [\phi_x^{0,x_0,*}(s)]^{-1}\nabla \widetilde{R}^{j,0,x_0}(s) - \widetilde{g}_x^{j*}(s,\phi^{0,x_0}(s),\widehat{u}(s))q(s),\\
Q(s) &= q_x^{0,x_0}(s)[\phi_x^{0,x_0}(s)]^{-1}\\
&= [\phi_x^{0,x_0,*}(s)]^{-1}\nabla^2 r^{0,x_0}(s)[\phi_x^{0,x_0}(s)]^{-1}\\
&\quad - [\phi_x^{0,x_0,*}(s)]^{-1}\phi_{xx}^{i,0,x_0}(s)[\phi_x^{0,x_0,*}(s)]^{-1,i}\nabla r^{0,x_0}(s)[\phi_x^{0,x_0}(s)]^{-1}.
\end{aligned}$$

Note that $\phi^{i,t,x}(s)$ *is the* i*th component of the column vector* $\phi^{t,x}(s)$ *and* $[\phi_x^{t,x,*}(s)]^{-1,i}$ *is the* i*th row vector of the matrix* $[\phi_x^{t,x,*}(s)]^{-1}$.

Elliott and Kohlmann [3] considered an optimal stochastic control problem with full information, which is the special case of our partially observable optimal stochastic control problem with

$$g \equiv 0, \quad h \equiv 0, \quad l \equiv 0, \quad U_{\rm ad} = \overline{U}_{\rm ad}. \tag{5.14}$$

In this case, the BSDEs (3.2) and (3.7) reduce to

$$\begin{cases} dr^{t,x}(s) = R^{i,t,x}(s)\,dw^i(s) + \widetilde{R}^{j,t,x}(s)\,dY^j(s), \quad s \in [t,1), \\ r^{t,x}(1) = m(\phi^{t,x}(1)) \end{cases} \tag{5.15}$$

and

$$\begin{cases} d\nabla r^{t,x}(s) = \nabla R^{i,t,x}(s)\,dw^i(s) + \nabla \widetilde{R}^{j,t,x}(s)\,dY^j(s), \quad s \in [t,1), \\ \nabla r^{t,x}(1) = \phi_x^{t,x,*}(1) m_x^*(\phi^{t,x}(1)) \end{cases} \tag{5.16}$$

respectively. Hence, for $s \in [t,1]$,

$$\begin{aligned} r^{t,x}(s) &= E[m(\phi^{t,x}(1)|\,\mathcal{F}_s^Y], \\ \nabla r^{t,x}(s) &= E[\phi_x^{t,x,*}(1) m_x^*(\phi^{t,x}(1))|\,\mathcal{F}_s^Y] \\ &= E[\phi_x^{t,x,*}(1) m_x^*(\phi^{t,x}(1))] + \int_t^s \nabla \widetilde{R}^{j,t,x}(\tau)\,dY^j(\tau), \\ R^{i,t,x}(s) &\equiv 0, \qquad \nabla R^{i,t,x}(s) \equiv 0, \\ \nabla^2 r^{t,x}(s) &= E[\phi_x^{t,x,*}(1) \nabla^2 m(\phi^{t,x}(1))|\,\mathcal{F}_s^Y] \\ &\quad + E[\phi_{xx}^{i,t,x}(1) m_{x^i}(\phi^{t,x}(1))|\,\mathcal{F}_s^Y]. \end{aligned} \tag{5.17}$$

Putting the relations (5.17) and (5.13) into the maximum condition (2.30), we arrive at Theorem 4.2 of Elliott and Kohlmann [3, p. 36]. Thus, our Theorem 5.3 contains Theorem 4.2 of Elliott and Kohlmann [3] as a special case.

Finally, we remark that a version of Theorem 2.1 for the case of $\widetilde{g} = 0$ has been obtained by Li and Tang [8], but the relations among the adjoint processes in Theorem 2.1 were not discussed there at all.

Acknowledgments. The author would like to thank the referees for their helpful comments on the original version of this paper. The author is also grateful to Prof. Jiongmin Yong for his help in improving the original version. In addition, the author would like to thank Professor Etienne Pardoux for his kind invitation.

REFERENCES

[1] J. S. Baras, R. J. Elliott, and M. Kohlmann, *The partially observed stochastic minimum principle*, SIAM J. Control Optim., 27 (1989), pp. 1279–1292.

[2] A. Bensoussan, *Maximum principle and dynamic programming approaches of the optimal control of partially observed diffusions*, Stochastics, 9 (1983), pp. 169–222.

[3] R. J. Elliott and M. Kohlmann, *The second order minimum principle and adjoint processes*, Stochastics Stochastics Rep., 46 (1994), pp. 25–39.

[4] W. H. Fleming, *Optimal control of partially observable diffusions*, SIAM J. Control, 6 (1968), pp. 194–214.

[5] U. G. Haussmann, *The maximum principle for optimal control of diffusions with partial information*, SIAM J. Control Optim., 25 (1987), pp. 341–361.

[6] H. Kunita, *Stochastic Flows and Stochastic Differential Equations*, Cambridge Stud. Adv. Math. 24, Cambridge University Press, Cambridge, UK, 1990.

[7] H. Kwakernaak, *A minimum principle for stochastic control problems with output feedback*, Systems Control Lett., 1 (1981), pp. 74–77.
[8] X. Li and S. Tang, *General necessary conditions for partially observed optimal stochastic controls*, J. Appl. Probab., 32 (1995), pp. 1118–1137.
[9] E. Pardoux and S. Peng, *Adapted solution of backward stochastic equation*, Systems Control Lett., 14 (1990), pp. 55–61.
[10] E. Pardoux and S. Peng, *Backward stochastic differential equations and quasilinear parabolic partial differential equations*, in Stochastic Partial Differential Equations and Their Applications, Lecture Notes in Control and Inform. Sci. 176, B. L. Rozovskii and R. B. Sowers, eds., Springer-Verlag, Berlin, Heidelberg, New York, 1992, pp. 200–217.
[11] S. Peng, *A general stochastic maximum principle for optimal control problems*, SIAM J. Control Optim., 28 (1990), pp. 966–979.
[12] S. Tang and X. Li, *Necessary conditions for optimal control of stochastic systems with random jumps*, SIAM J. Control Optim., 32 (1994), pp. 1447–1475.
[13] X. Zhou, *On the necessary conditions of optimal controls for stochastic partial differential equations*, SIAM J. Control Optim., 31 (1993), pp. 1462–1478.
[14] X. Zhou, *A duality analysis on stochastic partial differential equations*, J. Funct. Anal., 103 (1992), pp. 275–293.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 5, pp. 1618–1637, September 1998

006

ROBUSTNESS OF NONLINEAR FILTERS OVER THE INFINITE TIME INTERVAL*

AMARJIT BUDHIRAJA† AND HAROLD J. KUSHNER†

Abstract. Nonlinear filtering is one of the classical areas of stochastic control. From the point of view of practical usefulness, it is important that the filter not be too sensitive to the assumptions made on the initial distribution, the transition function of the underlying signal process and the model for the observation. This is particularly acute if the filter is of interest over a very long or potentially infinite time interval. Then the effects of small errors in the model which is used to construct the filter might accumulate to make the output useless for large time. The problem of asymptotic sensitivity to the initial condition has been treated in several papers. We are concerned with this as well as with the sensitivity to the signal model, uniformly over the infinite time interval. It is conceivable that the effects of even small errors in the model will accumulate so that the filter will eventually be useless. The robustness is shown for three classes of problems. For the first two cases, the signal model is Markov and the observations are taken in discrete time, and the observation is the usual function of the signal plus noise. The last class treated is a continuous time Markov process, with a point process observation.

Key words. nonlinear filtering, model robustness, asymptotic stability, Hilbert metric, Birkhoff's contraction coefficient

AMS subject classifications. 93E11, 93E15, 60H10

PII. S0363012997318481

1. Introduction.

Nonlinear filtering is one of the classical areas of stochastic control, and a great deal of work has been done on it. Typically, in either discrete or continuous time, it is assumed that the signal process is Markov and that the observations are corrupted by white noise, assumptions that we retain. A fundamental question from the point of view of practical usefulness is the sensitivity of the filter to the assumptions made on the initial distribution, the transition function of the underlying signal process and the model for the observation. This is particularly acute if the filter is of interest over a very long or potentially infinite time interval. Then the effects of small errors in the model used to construct the filter might accumulate to make the output useless for large time. Suppose that the assumed transition function for the signal process is not correct. Direct methods of comparing the difference between the true optimal filter and the one actually constructed generally use crude bounds which might be useful over a bounded time interval but get at best exponentially growing error estimates as time goes to infinity. Clearly a more subtle analysis is called for. With the classical Kalman–Bucy filter, under observability and controllability, the effects of the initial condition disappear as time goes to infinity. But, regrettably, there is no workable analog of global observability for the nonlinear problem.

The earliest work on the subject of robustness over a long tome interval was that of Kushner and Huang [8]. They worked in continuous time and assumed only wide

*Received by the editors March 17, 1997; accepted for publication (in revised form) September 23, 1997; published electronically June 9, 1998. The research of the first author was supported by contracts N00014-96-1-0276 and N0014-96-1-0279 from the Office of Naval Research. The research of the second author was supported by contract DAAH04-96-1-0075 from the Army Research Office and contract N000140-96-1-0276 from the Office of Naval Research.

http://www.siam.org/journals/sicon/36-5/31848.html

†Division of Applied Mathematics, Brown University, Providence, RI 02912 (amarjit@cfm.brown.edu, hjk@dam.brown.edu).

bandwidth observation and system driving noise. The model for the filter was the natural one based on the weak convergence limit as the bandwidth went to infinity, and they were concerned with the average (mean square or other) errors per unit time for large time. Long term errors for numerical and other approximations to the signal process were also of interest. They reduced the problem to one concerning uniqueness of the invariant measure of the joint (signal, filter) process. If the signal is a Markov process in some locally compact space and one has the standard additive white noise model for the observations then the work of Kunita [7] and Stettner [12] showed that the ergodicity of the signal leads to a unique invariant measure for the filter, but nothing was said about the joint (signal, filter) process. See Stettner [13] for the existence and uniqueness of the invariant measure for the case where the signal is a finite state Markov chain.

We will use the term asymptotic stability to mean that the output of the filter is asymptotically insensitive to the initial condition, assuming that the signal model is fixed. Ocone and Pardoux [11] used the results and ergodicity assumptions of [12] to obtain the convergence, in an appropriate sense, of the output of the incorrectly initialized filter to that of the exact filter as time approaches infinity. In two fundamental papers, Delyon and Zeitouni [6] and Atar and Zeitouni [2, 1], studied a variety of signal–observation pairs with ergodicity hypothesis on the signal where they prove exponentially fast convergence of the output of an incorrectly initialized filter to that of the correct one. Another recent work on exponential asymptotic stability is Le Gland and Mevel [9] who study finite state Markov chains and under appropriate conditions prove geometric ergodicity of an extended chain, which includes as its states the filter and its gradient. The approach of [2, 1] is based on Hilbert's projective metric and Birkhoff's contraction inequality and provides some remarkable results on pathwise convergence. It requires a rather strong ergodicity condition on the signal, and in most situations it restricts the analysis to signals taking values in a compact state space. One can obtain asymptotic stability of the filter in the absence of the ergodicity of the signal. The classical example is the Kalman filter and some related problems, cf.[11]. Budhiraja and Ocone [5] derive exponential asymptotic stability of the filter for signals which are given as solutions to one dimensional stochastic difference equations. The observation noise is taken to be bounded; however, no assumption is made on the boundedness of the signal. However, in general the question of asymptotic stability in the absence of ergodicity of the signal process is a challenging problem and remains open.

All of the works cited (excluding [9]) in the last paragraph assumed that the correct transition function for the signal process was used in the construction of the filter. The only variable was the initial condition. In practice, one would rarely know the correct signal transition function, and it is important to know that small errors in the signal model do not have serious effects on the filter output, over an arbitrarily large time interval. Simultaneously, one still would like asymptotic insensitivity to the initial condition of the filter. This paper is devoted to this double robustness problem.

We consider three models, and compare the output of the optimal filter to that for a filter built with an incorrect signal model and initial condition (but with the same observation sequence). The first class, which we treat in section 3.1, is that of a discrete time Markov process observed via a nonlinear functional with additive white noise. The transition function of this Markov chain is assumed to satisfy the one step mixing condition of [2] (see (7)). The exponential stability for this class had been

derived in [2]. In the present work we show that for the general robustness problem the total variation distance between the filter for the misspecified model and the exact filter converges to zero, uniformly in time, as the misspecifications converge to zero. The uniformity in time is the key outcome.

The second result, derived in section 3.2, is for the class of nonbounded signals studied in [5]. Exponential asymptotic stability for this class is known from [5]. For the general robustness problem we show that the infinite time limit of the expected total variation distance between the exact filter and the filter for the misspecified model converges to zero as the misspecifications in the transition kernel and in the distribution of the observation noise go to zero. The result is not pathwise. The main difficulty is that we do not have a contraction in the distance between the filters at every observation update, and when there is a contraction it is random. It turns out (cf. [5]) that this is sufficient to yield a pathwise asymptotic stability result; however, for the problem of robustness, with respect to the signal model, we need to do an analysis of the contractions in the mean.

The final section of the paper is devoted to a continuous time Markov signal model, but with point process observations. The results are new even if the only misspecification is in the initial condition. The signal is assumed to satisfy a mixing type condition analogous to that used for the first case (see 30). Theorem 4.1 proves the asymptotic stability, and the general robustness problem is treated in Theorem 4.2. The main additional difficulties are due to the facts that the observations can occur at any time and the liklihood ratio is discontinuous at the times of the observations.

The central tools in all the arguments in this work are that of Hilbert's projective metric and Birkhoff's contraction inequality; cf. [3]. These were introduced to the study of asymptotic stability of filters in [1]. For the convenience of the reader we have included, in section 2, a brief overview of the central ideas concerning Hilbert metric which are important in filter analysis.

2. Hilbert's projective metric. In this section we present some preliminary definitions and results concerning Hilbert's projective metric which will be used in later sections. Let S be a Polish space and let $\mathcal{M}(S)$ (respectively, $\mathcal{M}^+(S)$) denote the space of finite signed measures (finite nonnegative measures, respectively) on S. For μ, $\nu \in \mathcal{M}^+(S)$ the Hilbert projective distance between them is defined as

$$h(\mu,\nu) := \ln\left[\sup_{A,A'\in\mathcal{S}} \frac{\mu(A)}{\nu(A)}\frac{\nu(A')}{\mu(A')}\right], \tag{1}$$

where $\mathcal{S}$ is the Borel $\sigma-$field on S and we employ the convention that $\alpha/0 = \infty$ for $\alpha \neq 0$ and $0/0 = 1$.

Observe that a necessary condition for $h(\mu,\nu)$ to be finite is that μ, ν are mutually absolutely continuous. In fact it can be shown (cf. [10]) that a necessary and sufficient condition for $h(\mu,\nu)$ to be finite for when μ and ν are positive measures is that there exist positive $c_i, i = 1, 2$, such that

$$c_1\nu \leq \mu \leq c_2\nu. \tag{2}$$

Then $h(\mu,\nu) = \inf \ln(c_2/c_1)$, where the infimum is taken over all pairs c_1, c_2 for which the above inequalities hold.

One of the important properties of the Hilbert metric from the point of view of nonlinear filtering problems is that of scale invariance; i.e., for μ, $\nu \in \mathcal{M}^+(S)$ and α and β positive numbers, $h(\mu,\nu) = h(\alpha\mu, \beta\nu)$. Thus, if μ and ν are conditional

distributions arising in a filtering problem, in computing the distance in the Hilbert projective metric it makes no difference whether they are normalized or unnormalized, and we will use this fact where convenient without further comment. The following inequality connects the Hilbert metric with the total variation norm. For μ, ν probability measures on $(S, \mathcal{S})$

$$||\mu - \nu||_{TV} \leq \frac{2}{\ln 3} h(\mu, \nu), \tag{3}$$

where $||.||_{TV}$ denotes the total variation norm on $\mathcal{M}(S)$. We refer the reader to [2] for a proof.

Another important property of Hilbert metric which makes it a very useful tool in stability analysis is the following contraction relation due to Birkhoff [3]. Let S_1, S_2 be Polish spaces and let $\mathcal{S}_1, \mathcal{S}_2$ be the respective Borel $\sigma-$ fields. Denote the Hilbert metric on $\mathcal{M}^+(S_1)$ and $\mathcal{M}^+(S_2)$ by the same symbol: namely, h. Let, $K : \mathcal{M}(S_1) \to \mathcal{M}(S_2)$ be a linear nonnegative operator. Then for μ, $\nu \in \mathcal{M}^+(S_1)$

$$h(K\mu, K\nu) \leq \tanh(C(K)/4) h(\mu, \nu), \tag{4}$$

where

$$\begin{aligned} C(K) &:= \sup\{h(K\mu, K\nu) : \mu, \nu \in \mathcal{M}^+(S_1)\} \\ &= \sup_{\mu,\nu \in \mathcal{M}^+(S_1)} \ln \left[\sup_{A,A' \in \mathcal{S}_2} \frac{K\mu(A)}{K\nu(A)} \frac{K\nu(A')}{K\mu(A')} \right]. \end{aligned} \tag{5}$$

We record one final observation for future use. Suppose that K is defined by

$$(K\mu)(A) = \int_A \int_{S_1} \mathcal{K}(x, y) \mu(dy) \lambda(dx), \ A \in \mathcal{S}_2,$$

where λ is a positive $\sigma-$ finite measure on $(S_2, \mathcal{S}_2)$ and $\mathcal{K} : S_2 \times S_1 \to [0, \infty)$ is a measurable map. Then

$$C(K) = \ln \left[\sup_{y,y' \in S_1} \operatorname{ess\,sup}_{x,x' \in S_2} \left(\frac{\mathcal{K}(x, y) \mathcal{K}(x', y')}{\mathcal{K}(x, y') \mathcal{K}(x', y)} \right) \right], \tag{6}$$

where same convention as before is employed for $\alpha/0; \alpha \neq 0$ and $0/0$. The essential supremum in (6) is with respect to the measure λ.

3. Discrete time signals. In this section we will consider the asymptotic errors when the observations are taken in discrete time. We will study the asymptotic sensitivity with respect to the incorrect initial condition, incorrect transition function, and incorrect observation noise distribution function. It will be seen that, under appropriate mixing-type conditions on the signal process, the effects of errors in the initial condition eventually disappear, and small errors in the transition or distribution functions cause only small errors in the filter output, uniformly over all time. We will work with two specific signal–observation pairs for which the Hilbert projective metric techniques can be applied. In our first result we consider signals whose transition kernels satisfy the boundedness condition (7), which was used in [2], which also proved the stability with respect to the initial condition. The second theorem is for a family of real valued signals observed in bounded noise, but which is not necessarily bounded. The stability of the filter for this class with respect to the misspecification of the initial condition alone had been studied in [5]. The second class of examples, although quite special, is interesting in that it shows that the boundedness conditions in [2] are not necessary.

3.1. The signal satisfies a one step mixing condition. Let (Ω, F, P) be a probability space. Let X_n be a Markov chain with a stationary transition probability distribution and a Polish state space S. Let $\mathcal{S}$ denote the Borel $\sigma-$ field on S. Assume that the signal admits a transition probability density $G(\cdot,\cdot)$, with respect to some $\sigma-$ finite measure λ on $(S,\mathcal{S})$.

Following the approach taken in [2], we assume that there exists a probability measure ρ on $(S,\mathcal{S})$ and finite positive constants, c_1, c_2, such that for all $A \in \mathcal{S}$:

$$c_1\rho(A) \leq \int_A G(x,y)\lambda(dy) \leq c_2\rho(A). \tag{7}$$

Since the left- and right-hand sides do not depend on the initial state x, the above key condition implies that the signal process has a stong one step mixing property. It would hold, for example, if the signal were a sampled nondegenerate diffusion on a compact state space. Let

$$Y_n = H(X_n) + \nu_n \tag{8}$$

be the observation sequence, where $H : S \to R^m$ is a measurable function and ν_n are R^m valued mutually independent and identically distributed random variables with a bounded density which we denote by g. (The condition that the distributions be independent of n can be weakened, but the assumption simplifies the notation.)

For fixed $y \in R^m$, define the following nonnegative operator $K(G,y)$ on $\mathcal{M}(S)$:

$$(K(G,y)\mu)(A) := \int_A \int_S g(y - H(z))G(x,z)\mu(dx)\lambda(dz). \tag{9}$$

Here, $\mu \in \mathcal{M}(S)$ and $A \in \mathcal{S}$.

If $K(G,y)\mu \neq 0$ (i.e., it is not the zero measure), then define the normalized measure $\tilde{K}(G,y) := K(G,y)\mu/[K(G,y)\mu](S)$. Otherwise set it equal to 0. Denote by Π_n the conditional distribution of X_n given $Y_1, \ldots, Y_n$. Then it is a simple verification (for a proof see Lemma 3.1 of [5]) that Π_n equals

$$\Pi_n = \tilde{K}(G,Y_n)o\tilde{K}(G,Y_{n-1})o\cdots o\tilde{K}(G,Y_1)p_0,$$

where p_0 is the distribution of X_0.

Let $\{G_k\}$ be a sequence of transition probability densities and $\{p_k\}$ a sequence of probability measures on S. Let us write

$$\Pi_n^{(k)} = \tilde{K}(G_k,Y_n)o\tilde{K}(G_k,Y_{n-1})o\cdots o\tilde{K}(G_k,Y_1)p_k.$$

The following theorem contains the main result of this subsection. It says essentially that if the filter is designed with an incorrect initial condition and incorrect signal transition function, then the pathwise difference between the true optimal filter and the incorrect one over an arbitrarily large or infinite time interval is bounded uniformly in the difference (in a suitable scale) between the correct and erroneous transition function, and initial condition.

THEOREM 3.1. *Suppose that,* $\forall x \in S$ *and* $\forall k \geq 1$, $G_k(x,\cdot)$ *and* $G(x,\cdot)$ *are positive and zero on the same sets. Let* $\ln G_k$ *converge to* $\ln G$ *uniformly on the* $(x,y)-$*set, where* $G(x,y) > 0$. *Then*

(a) $\lim_{k\to\infty} \limsup_{n\to\infty} h(\Pi_n^{(k)}, \Pi_n) = 0.$

(b) *If in addition, p_k converges to p_o in total variation norm as $k \to \infty$, then*

$$\lim_{k\to\infty} \sup_n ||\Pi_n^{(k)} - \Pi_n||_{TV} = 0.$$

Proof. For $q \in \mathcal{M}^+(S)$, define

$$\Pi_n^{[q]} := \tilde{K}(G, Y_n) o \tilde{K}(G, Y_{n-1}) o \cdots o \tilde{K}(G, Y_1) q.$$

We begin by noting that

$$h(\Pi_n^{(k)}, \Pi_n) \le h(\Pi_n^{(k)}, \Pi_n^{[p_k]}) + h(\Pi_n^{[p_k]}, \Pi_n). \tag{10}$$

Let us initially consider the second term on the right side of (10). Using the scale invariance property of Hilbert's projective metric, we have

$$\begin{aligned} h(\Pi_n^{[p_k]}, \Pi_n) &= h(K(G, Y_n)\Pi_{n-1}^{[p_k]}, K(G, Y_n)\Pi_{n-1}) \\ &\le \tanh\left(\frac{C(K(G, Y_n))}{4}\right) h(\Pi_{n-1}^{[p_k]}, \Pi_{n-1}). \end{aligned}$$

It is clear from (5) and (7) that $C(K(G, Y_n)) \le 2\ln(c_2/c_1)$. Using this observation in the above equality, we get

$$h(\Pi_n^{[p_k]}, \Pi_n) \le \delta h(\Pi_{n-1}^{[p_k]}, \Pi_{n-1}),$$

where $\delta := \tanh(\ln(c_2/c_1)/2)$.

Iterating the above inequality and observing from (7) and (1) that $h(\Pi_1^{[p_k]}, \Pi_1) \le 2\ln(c_2/c_1)$, we have

$$h(\Pi_n^{[p_k]}, \Pi_n) \le 2\delta^{n-1} \ln(c_2/c_1). \tag{11}$$

Consider now the first term on the right side of (10), namely,

$$\begin{aligned} h(\Pi_n^{(k)}, \Pi_n^{[p_k]}) &= h(K(G_k, Y_n)\Pi_{n-1}^{(k)}, K(G, Y_n)\Pi_{n-1}^{[p_k]}) \\ &\le h(K(G_k, Y_n)\Pi_{n-1}^{(k)}, K(G, Y_n)\Pi_{n-1}^{(k)}) \\ &\quad + h(K(G, Y_n)\Pi_{n-1}^{(k)}, K(G, Y_n)\Pi_{n-1}^{[p_k]}) \\ &\le h(K(G_k, Y_n)\Pi_{n-1}^{(k)}, K(G, Y_n)\Pi_{n-1}^{(k)}) + \delta h(\Pi_{n-1}^{(k)}, \Pi_{n-1}^{[p_k]}). \end{aligned} \tag{12}$$

The first term on the right side of (12) can be bounded as follows. Define ϵ_k by

$$\epsilon_k = \sup_{(x,y)\in B} |\ln G_k(x, y) - \ln G(x, y)|,$$

where $B := \{(x, y) : G(x, y) \ne 0\}$. By hypothesis, $\epsilon_k \to 0$ as $k \to \infty$. Also the following inequality holds $\forall (x, y) \in B$.

$$e^{-\epsilon_k} G(x, y) \le G_k(x, y) \le e^{\epsilon_k} G(x, y).$$

Using the above inequality and (1) it is easy to see that

$$h(K(G_k, Y_n)\Pi_{n-1}^{(k)}, K(G, Y_n)\Pi_{n-1}^{(k)}) \le 2\epsilon_k. \tag{13}$$

Combining (12) and (13) yields

$$h(\Pi_n^{(k)}, \Pi_n^{[p_k]}) \leq 2\epsilon_k + \delta h(\Pi_{n-1}^{(k)}, \Pi_{n-1}^{[p_k]}).$$

Iterating the above inequality and using the observation that $h(\Pi_1^{(k)}, \Pi_1^{[p_k]})$ is bounded by $2\epsilon_k$, we have that

$$h(\Pi_n^{(k)}, \Pi_n^{[p_k]}) \leq 2\epsilon_k/(1-\delta).$$

Combining the above inequality with (10) yields that

$$h(\Pi_n^{(k)}, \Pi_n) \leq \frac{2\epsilon_k}{1-\delta} + \delta^{n-1} h(\Pi_1^{(k)}, \Pi_1^{[p_k]}). \tag{14}$$

This proves (a).

To prove (b) it suffices to show that $||p_k - p||_{TV} \to 0$ implies that $h(\Pi_1^{(k)}, \Pi_1^{[p_k]})$ converges to zero. A straightforward inequality shows that

$$\frac{1}{1 + \frac{c_2}{c_1}||p_k - p_0||_{TV}} \leq \frac{\int G(x,y) p_0(dx)}{\int G(x,y) p_k(dx)} \leq 1 + \frac{c_2}{c_1}||p_k - p_0||_{TV},$$

a.e. y. This implies that

$$h(\Pi_1^{(k)}, \Pi_1^{[p_k]}) \leq 2\ln(1 + \frac{c_2}{c_1}||p_k - p_0||_{TV}).$$

The result now follows on combining this observation with (14). □

Remark 1. Theorem 3.1(a) can be shown to hold if G satisfies, instead of (7), the weaker condition

$$c_1 \rho(A) \leq \int_A G^k(x,y) \lambda(dy) \leq c_2 \rho(A), \tag{15}$$

for some $k \geq 1$, where G^k is the k-step transition kernel and c_1, c_2, A, ρ are as before. The proof is more involved mainly because the contraction coefficient(analogus to δ in the proof of Theorem 3.1) is now a random quantity. However (15) implies that the signal is ergodic and has a unique invariant measure. Also the observation noise sequence is ergodic. From this one can show that the contraction coefficients obtained for successive k-step updates of the filter form an ergodic sequence. One can then apply Birkhoff's ergodic theorem to complete the proof.

3.2. A difference equation model. The result in section 3.1 relies heavily on the mixing and boundedness properties of the signal and the observations played no role in the analysis. However, observations are a critical ingredient in the problem of nonlinear filtering, and should play a central role in any asymptotic analysis. Each time the filter is updated the observations recenter the distribution in an interval of the observed value and hence, intuitively, if the observations are "good" they should help the convergence of the output of an incorrectly initialized filter to that of the correct filter, and aid in stabilizing the effects of model error as well. One such model was studied in [5], and it was shown that the filter is asymptotically independent of its initial condition. In this subsection we revisit that example from the perspective of a more general robustness in the infinite time limit. Although the model is one dimensional, it is nonlinear and the observations play a crucial role.

Let $(\Omega, \mathcal{F}, P)$ be a probability space on which are defined two sequences, $\{\xi_n\}_{n=1}^{\infty}$ and $\{\nu_n\}_{n=1}^{\infty}$, which are mutually independent, and each has independent and identically distributed components. We assume that both ξ_1 and ν_1 have bounded densities with respect to Lebesgue measure, denoted by f and g, respectively. Furthermore, following the approach in [5] we assume that there is an $M < \infty$ such that

(A.1) $$\text{supp } g \subseteq [-M, M].$$

Let X_0 be another real-valued random variable on the above probability space, independent of both $\{\xi_i\}$ and $\{\nu_i\}$, with law p_0. The signal $\{X_n\}_{n=0}^{\infty}$ is defined as

$$X_{n+1} = m(X_n) + \sigma(X_n)\xi_{n+1}, \qquad n \geq 0. \tag{16}$$

We use the assumptions of [5], where m and σ are real-valued and Borel measurable and are assumed to satisfy

(A.2) $$0 < \underline{\sigma} := \inf_{x \in R} \sigma(x) \leq \sup_{x \in R} \sigma(x) =: \overline{\sigma} < \infty.$$

(A.3) $$C := \sup_{|z-z'| \leq 2M} |m(z) - m(z')| < \infty.$$

The observations on the signal are given by

$$Y_n = X_n + \nu_n, \qquad n \geq 1. \tag{17}$$

Next, in order to get the filter update formula we introduce the following linear operator. For $u, v \in R$, define the operator K by

$$K \equiv K(u, v, f, g, m) : \mathcal{M}[u - M, u + M] \to \mathcal{M}[v - M, v + M]$$

by

$$K\mu(A) = \int_A \int_{[u-M,u+M]} g(v - x) f\left(\frac{x - m(z)}{\sigma(z)}\right) \sigma^{-1}(z)\mu(dz)dx, \tag{18}$$

$A \in \mathcal{B}[v - M, v + M]$. As before, define the nonlinear operator $\tilde{K}$ to be the normalization of K. Define $\tilde{K}\mu = 0$ if $K\mu = 0$. For $n \geq 2$, let K_n denote the operator $K(Y_{n-1}, Y_n, f, g, m)$, and let $\tilde{K}_n$ denote its normalized form. Finally, for $P(R)$ denoting the family of probability measures on R, define

$$K_1 : P(R) \to \mathcal{M}[Y_1 - M, Y_1 + M]$$

by

$$K_1\mu(A) = \int_A \int_R g(v - x) f(\frac{x - m(z)}{\sigma(z)})\sigma^{-1}(z)\mu(dz)dx,$$

and let $\tilde{K}_1$ denote the normalized form. Let Π_n denote the conditional distribution of X_n given $Y_1, \ldots, Y_n$. Then it can be shown that, with probability one,

$$\Pi_n = \tilde{K}_n o \tilde{K}_{n-1} o \cdots o \tilde{K}_1 p_0.$$

Now let $\{g_k\}, \{f_k\}, \{m_k\}$ be sequences of maps from $R \to R$, where for every k, f_k is a probability density, and g_k is a probability density with support $[-M, M]$. Let $K_n^{(k)}$ denote the operator defined by $K(Y_{n-1}, Y_n, f_k, g_k, m)$ and let $K_n^{(k)\prime}$ denote the operator $K(Y_{n-1}, Y_n, f_k, g_k, m_k)$. Let $\tilde{K}_n^{(k)}, \tilde{K}_n^{(k)\prime}$ denote the respective normalizations. Let $\{p_k\}$ be a sequence of probability measures on R and define

$$\Pi_n^{(k)} = \tilde{K}_n^{(k)} o \tilde{K}_{n-1}^{(k)} o \cdots o \tilde{K}_1^{(k)} p_k$$

and

$$\Pi_n^{(k)\prime} = \tilde{K}_n^{(k)\prime} o \tilde{K}_{n-1}^{(k)\prime} o \cdots o \tilde{K}_1^{(k)\prime} p_k.$$

Define

$$\Pi_n^{[p_k]} := \tilde{K}_n o \tilde{K}_{n-1} o \cdots o \tilde{K}_1 p_k,$$

and set $S := \inf\{n : \Pi_n^{[p_k]} = 0 \text{ for some } k\}$.

Define $a(x) := [2M + C]/\underline{\sigma} + [\overline{\sigma}/\underline{\sigma}]|x| \equiv c_0 + c_1|x|$.

Then we have the following result. It will be seen that the conditions are not too restrictive.

THEOREM 3.2. *Assume that $S = \infty$ with probability one. Suppose that there is $\epsilon_0 > 0$ such that $\underline{f} := \inf\{f(u) : |u| \le c_0 + \epsilon_0\} > 0$, and*

$$\lim_{k\to\infty} \sup_{x\in[-M,M]} |\ln g_k(x) - \ln g(x)| = 0,$$

where $|\ln(0) - \ln(0)| = 0$. For all $k \in [0,\infty)$, define $\rho_k(\cdot)$ by

$$\sup_{x\in[-l,l]} |\ln f_k(x) - \ln f(x)| =: \rho_k(l), \tag{19}$$

and suppose that $E\rho_k(a + b|\xi_1|) \to 0$ as $k \to \infty$ for all positive a, b. Then

(i) $\lim_{k\to\infty} \limsup_{n\to\infty} E||\Pi_n - \Pi_n^{(k)}||_{TV} = 0$.

(ii) *Define $\rho^*(\cdot)$ by*

$$\sup_{x,y\in[-l,l];x\neq y} \frac{|\ln f(x) - \ln f(y)|}{|x-y|} =: \rho^*(l).$$

Suppose that for all positive a, b, $E\rho^*(a+b|\xi_1|) < \infty$. Finally assume that m_k converges to m uniformly on R. Then

$$\lim_{k\to\infty} \limsup_{n\to\infty} E||\Pi_n - \Pi_n^{(k)\prime}||_{TV} = 0.$$

Remark 2. We note that the condition $S = \infty$ with probability one is satisfied, for example, if p_k is mutually absolutely continuous with respect to p for every k. The above stated conditions on f, f_k are satisfied, if for example, $f \sim N(\mu,\sigma)$, $f_k \sim N(\mu_k,\sigma_k)$ and μ_k, σ_k converge to μ, σ, respectively, as $k \to \infty$. More generally, if we have the common forms $g_k(x) = e^{-\phi_k(x)}$ and $g(x) = e^{-\phi(x)}$ (and analogously for $f(\cdot)$ and $f_k(\cdot)$), then the conditions on the convergence of the logs of the densities become conditions on the convergence of the $\phi_k(\cdot)$, and we can see that they are not too stringent.

Proof of Theorem 3.2. By (3) and the triangle inequality for the total variation norm, we have

$$||\Pi_n^{(k)} - \Pi_n||_{TV} \le [2/\ln(3)]h(\Pi_n^{(k)}, \Pi_n^{[p_k]}) + ||\Pi_n^{[p_k]} - \Pi_n||_{TV}.$$

From Corollary 3.3 of [5] we know that for all k, $h(\Pi_n^{[p_k]}, \Pi_n)$ converges to zero with probability one as $n \to \infty$. This immediately yields L^1 convergence, of the second term in the above inequality, for each fixed k. Therefore it suffices to consider the first term in the above inequality.

Observe now that (as in Theorem 3.1)

$$h(\Pi_n^{(k)}, \Pi_n^{[p_k]}) \le h(K_n^{(k)}\Pi_{n-1}^{(k)}, K_n\Pi_{n-1}^{(k)}) + h(K_n\Pi_{n-1}^{(k)}, K_n\Pi_{n-1}^{[p_k]}). \tag{20}$$

We will now consider the second term above. From Birkhoff's contraction inequality, we know that it can be at most,

$$\tanh\left(\frac{C(K_n)}{4}\right) h(\Pi_{n-1}^{(k)}, \Pi_{n-1}^{[p_k]}).$$

Also, using the definition of $C(\cdot)$ and (18),

$$C(K_n) = \ln\left[\sup_{y,y'} \operatorname{ess\,sup}_{x,x'} \left(\frac{f(\frac{x-m(y)}{\sigma(y)})f(\frac{x'-m(y')}{\sigma(y')})}{f(\frac{x-m(y')}{\sigma(y')})f(\frac{x'-m(y)}{\sigma(y)})}\right)\right], \tag{21}$$

where the first supremum is taken over $x, x' \in [Y_{n-1} - M, Y_{n-1} + M]$ and the essential supremum is with respect to the Lebesgue measure on $[Y_n - M, Y_n + M]^2$. A straightforward computation using (A.2) and (A.3) shows that for $x \in [Y_n - M, Y_n + m]$ and $y \in [Y_{n-1} - M, Y_{n-1} + M]$, we have $\frac{x-m(y)}{\sigma(y)} < c_0 + c_1|\xi_n|$. Hence for $|\xi_n| < \epsilon_0/c_1$ and x, y as above, $f(\frac{x-m(y)}{\sigma(y)}) \ge \underline{f}$. In view of (21) it then follows that

$$C(K_n) I_{|\xi_n|<\epsilon_0/c_1} \le \ln\left(\frac{\overline{f}^2}{\underline{f}^2}\right) I_{|\xi_n|<\epsilon_0/c_1}.$$

This implies that

$$\tanh\left(\frac{C(K_n)}{4}\right) \le 1 - \left(1 - \tanh\left(\frac{1}{4}\ln\left(\frac{\overline{f}^2}{\underline{f}^2}\right)\right)\right) I_{|\xi_n|<\epsilon_0/c_1} := \delta(\xi_n).$$

Using the above inequality for the second term on the right side of (20) we have that

$$h(K_n\Pi_{n-1}^{(k)}, K_n\Pi_{n-1}^{[p_k]}) \le \delta(\xi_n) h(\Pi_{n-1}^{(k)}, \Pi_{n-1}^{[p_k]}) \tag{22}$$

Now we consider the first term on the right side of (20). For an arbitrary positive finite measure ν on $[Y_{n-1} - M, Y_{n-1} + M]$, and a Borel set A in $B[Y_n - M, Y_n + M]$, the term $K_n^{(k)}\nu(A)$ is equal to

$$\int_A \int g_k(Y_n - x) f_k\left(\frac{x - m(y)}{\sigma(z)}\right) \sigma^{-1}(y)\nu(dy)dx. \tag{23}$$

Now define

$$\overline{\rho}_k := \sup_{|x|<M} |\ln g_k(x) - \ln g(x)|.$$

Then clearly, for $x \in [Y_n - M, Y_n + M]$,

$$e^{-\overline{\rho}_k} g(Y_n - x) \le g_k(Y_n - x) \le e^{\overline{\rho}_k} g(Y_n - x). \tag{24}$$

Next note that in view of the convergence assumption on $\ln f_k$ (see (19)) and recalling that for $x \in [Y_n - M, Y_n + M]$and $y \in [Y_{n-1} - M, Y_{n-1} + M]$, we have $\frac{x-m(y)}{\sigma(y)} \le a(\xi_n)$, it follows that for such x, y,

$$e^{-\rho_k(a(\xi_n))} f\left(\frac{x - m(y)}{\sigma(y)}\right) \le f_k\left(\frac{x - m(y)}{\sigma(y)}\right) \le e^{\rho_k(a(\xi_n))} f\left(\frac{x - m(y)}{\sigma(y)}\right). \tag{25}$$

Using the inequalities, (24) and (25) in the representation for $K_n^{(k)}\nu(A)$ (see (23)), we have that

$$e^{-\rho_k(a(\xi_n))-\overline{\rho}_k}K_n\nu(A) \leq K_n^{(k)}\nu(A) \leq e^{\rho_k(a(\xi_n))+\overline{\rho}_k}K_n\nu(A).$$

The above inequality yields that

$$h(K_n^{(k)}\Pi_{n-1}^{(k)}, K_n\Pi_{n-1}^{(k)}) \leq 2\rho_k(a(\xi_n)) + 2\overline{\rho}_k. \tag{26}$$

This observation in conjunction with (22) when used on the right side of (20) gives

$$h(\Pi_n^{(k)}, \Pi_n^{[p_k]}) \leq 2\rho_k(a(\xi_n)) + 2\overline{\rho}_k + \delta(\xi_n)h(\Pi_{n-1}^{(k)}, \Pi_{n-1}^{[p_k]}).$$

Iterating the above inequality, we obtain

$$h(\Pi_n^{(k)}, \Pi_n^{[p_k]}) \leq \sum_{j=1}^{n}(2\rho_k(a(\xi_j)) + 2\overline{\rho}_k)\delta(\xi_{j+1})\cdots\delta(\xi_n).$$

Taking expectations we get

$$E[h(\Pi_n^{(k)}, \Pi_n^{[p_k]})] \leq 2(E\rho_k(a(\xi_1)) + \overline{\rho}_k)/(1-\delta),$$

where $\delta := E\delta(\xi_1)$ is strictly less than one, since in view of the assumption on the support of f, $P(|\xi_1| < \epsilon_0/c_1) > \underline{f}\epsilon_0/c_1 > 0$. (Note $c_1 \geq 1$.) This proves (i).

We now prove (ii). As in the proof of (i), we have the inequality:

$$||\Pi_n^{(k)\prime} - \Pi_n||_{TV} \leq [2/\ln(3)]h(\Pi_n^{(k)\prime}, \Pi_n^{[p_k]}) + ||\Pi_n^{[p_k]} - \Pi_n||_{TV}.$$

The second term converges in L^1 as $n \to \infty$ for each fixed k. Hence, it suffices to consider the first term. Again, an application of the triangle inequality as in (i) yields

$$h(\Pi_n^{(k)\prime}, \Pi_n^{[p_k]}) \leq h(K_n^{(k)\prime}\Pi_{n-1}^{(k)\prime}, K_n\Pi_{n-1}^{(k)\prime}) + \delta(\xi_n)h(\Pi_{n-1}^{(k)\prime}, \Pi_{n-1}^{[p_k]}). \tag{27}$$

Finally, consider the first term in the above inequality. By definition, $K_n^{(k)\prime}\Pi_{n-1}^{(k)\prime}$ equals

$$\int_{\cdot}\int g_k(Y_n - x)f_k\left(\frac{x - m_k(y)}{\sigma(z)}\right)\sigma^{-1}(y)\nu(dy)dx. \tag{28}$$

Let $\delta_k := \sup_{x\in R}|m_k(x) - m(x)|$. By hypothesis $\delta_k \to 0$ as $k \to \infty$. Now let $M_o > 0$ be such that, $\forall k \geq M, \delta_k < \underline{\sigma}$. Then $\forall k > M_o$, $x \in [Y_n - M, Y_n + M]$, $z \in [Y_{n-1} - M, Y_{n-1} + M]$,

$$|\frac{x - m_k(z)}{\sigma(z)}| \leq |\frac{x - m(z)}{\sigma(z)}| + |\frac{m(z) - m_k(z)}{\sigma(z)}| \leq a(\xi_n) + 1. \tag{29}$$

We have by a straightforward application of a triangle inequality that for all $k > M_o$, $|\ln f_k(\frac{x-m_k(z)}{\sigma(z)}) - \ln f(\frac{x-m(z)}{\sigma(z)})|$ is bounded above by the sum of

$$\left|\ln f_k\left(\frac{x - m_k(z)}{\sigma(z)}\right) - \ln f\left(\frac{x - m_k(z)}{\sigma(z)}\right)\right|$$

and

$$\left|\ln f\left(\frac{x-m_k(z)}{\sigma(z)}\right)-\ln f\left(\frac{x-m(z)}{\sigma(z)}\right)\right|.$$

In view of the definition (19), the first of these terms is bounded by $\rho_k(a(\xi_n)+1)$. In view of the Lipschitz condition on f, the second term is bounded by $\rho^*(a(\xi_n)+1)\delta_k/\underline{\sigma}$.

Using this observation in the representation of $K_n^{(k)\prime}\Pi_{n-1}^{(k)\prime}$(i.e., the expression (28) along with (24), we have that $\forall k > M_o$

$$e^{-\rho_k(a(\xi_n)+1)-\rho^*(a(\xi_n)+1)\delta_k/\underline{\sigma}-\overline{\rho}_k}K_n \le K_n^{(k)\prime} \le e^{\rho_k(a(\xi_n)+1)+\rho^*(a(\xi_n)+1)\delta_k/\underline{\sigma}+\overline{\rho}_k}K_n.$$

Therefore

$$h(K_n^{(k)\prime}\Pi_{n-1}^{(k)\prime}, K_n\Pi_{n-1}^{(k)\prime}) \le 2\left(\rho_k(a(\xi_n)+1)+\rho^*(a(\xi_n)+1)\delta_k/\underline{\sigma}+\overline{\rho}_k\right).$$

Using the above inequality in (27) we have that $\forall k > M_o$

$$h(\Pi_n^{(k)\prime}, \Pi_n^{[p_k]}) \le 2\left(\rho_k(a(\xi_n)+1)+\rho^*(a(\xi_n)+1)\delta_k/\underline{\sigma}+\overline{\rho}_k\right)+\delta(\xi_n)h(\Pi_{n-1}^{(k)\prime}, \Pi_{n-1}^{[p_k]}).$$

Iterating the above inequality and taking expectations, we get

$$E[h(\Pi_n^{(k)\prime}, \Pi_n^{[p_k]})] \le 2\left(E\left[\rho_k(a(\xi_1)+1)+\rho^*(a(\xi_1)+1)\delta_k/\underline{\sigma}+\overline{\rho}_k\right]\right)/(1-\delta),$$

The proof now follows on observing that

$$E\left[\rho_k(a(\xi_1)+1)+\rho^*(a(\xi_1)+1)\delta_k/\underline{\sigma}+\overline{\rho}_k\right] \to 0$$

as $k \to \infty$. □

4. Continuous time signals with point process observations. In this section we will examine the filter robustness properties for a continuous time signal where the observations are a point process. The signal satisfies one step mixing type properties similar to those used in Theorem 3.1. The first result on asymptotic stability (Theorem 4.1) shows that asymptotically the filter output does not depend on the initial condition. It is the point process analog of the results in [1], and is new. The proof uses the Hilbert projective metric and Birkhoff's contraction coefficient. The basic convergence result is pathwise. The main difficulty is that the contraction in the total variation distance between the filters is given from the time of one observation to the next, but the contraction is not uniform since the observations times are not equally spaced; the observations can occur at any time. Because of this the initial analysis is in the mean. Nevertheless, one can recover the almost sure convergence on noting that the distance in the Hilbert projective metric is nonincreasing. In the second theorem of this section we consider the general robustness problem. The analysis is more involved since one needs to keep track of the errors caused by the incorrect transition function, continuously in time, since the observations can occur at any time.

The precise model is the following. Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $\{X_t, t \ge 0\}$ be a cadlag Markov process on this space taking values in a Polish space S. Let $\mathcal{S}$ be the Borel σ-field on S. We assume that $\{X_t, t \ge 0\}$ has a stationary transition density, denoted by $G_t(x,y), t \ge 0$, with respect to some σ-finite measure m on $(S, \mathcal{S})$. The distribution of X_0 is denoted by p_0. We assume that there exist maps f_1, f_2 from $(0,\infty)$ to $[0,\infty)$ and $a, b \in (0,\infty), a < b$ such that

(H) $$\sup_{t\in[a,b]} \frac{f_2(t)}{f_1(t)} < \infty,$$

and $\forall t \in (0,\infty)$ and for $x, y \in S$

$$f_1(t) \le G_t(x,y) \le f_2(t), \tag{30}$$

where $0/0 = 0$.

Condition (H) is not very restrictive since a, b are arbitrary. The observation process $\{Y_t, t \ge 0\}$ is assumed to be a real-valued (right continuous) Poisson process with intensity $\lambda(X_t, Y_t)$, where $\lambda(\cdot)$ is assumed to be bounded from both above and below, i.e., for all $x \in S$:

$$0 < \lambda_1 \le \lambda(x,y) \le \lambda_2 < \infty.$$

We note that the vector-valued observation case is treated in the same way, and with the same result, but we wish to keep the notation simple. To obtain a representation for the filter, we use the usual measure transformation method and introduce another probability space $(\Omega_1, \mathcal{F}_1, P_1)$ on which we define a copy of the process X_t, denoted by $X_t^{(0)}$. Let $(\overline{\Omega}, \overline{\mathcal{F}}, \overline{Q})$ be the product space: $(\Omega, \mathcal{F}, P) \otimes (\Omega_1, \mathcal{F}_1, P_1)$. We define processes $\{X_t, Y_t, t \ge 0\}$ and $\{X_t^{(0)}, t \ge 0\}$ on the extended space in the usual manner. By construction, the process $\{X_t^{(0)}, t \ge 0\}$ is independent of the processes $\{X_t, Y_t, t \ge 0\}$.

Let $\Pi_t^{(p_0)}$ denote the conditional distribution of X_t given $\sigma\{Y_s, 0 \le s \le t\}$. Let q be a real-valued, measurable, and bounded function on S. It is well known (cf. [4]) that $\Pi_t^{(p_0)}$ can be written as

$$\Pi_t^{(p_0)} q := \frac{\tilde{\Pi}_t^{(p_0)} q}{\tilde{\Pi}_t^{(p_0)} 1},$$

where

$$\tilde{\Pi}_t^{(p_0)} q := E_{\overline{Q}} \left[L_t q(X_t^{(0)}) | Y_s : s \le t \right],$$

where

$$L_t := \left(\prod_{0 \le s \le t} \lambda(X_s^{(0)}, Y_{s^-}) \Delta Y_s \right) \exp\left(- \int_0^t \lambda(X_s^{(0)}, Y_s) ds \right),$$

where $\Delta Y_s = Y_s - Y_{s-}$. The empty product is defined to be unity.

Now let p_1 be an arbitrary probability measure on $(S, \mathcal{S})$ and introduce a Markov process on $(\Omega_1, \mathcal{F}_1, P_1)$, denoted by $\{X_t^{(1)}, t \ge 0\}$, with initial distribution p_1 and the same transition function as $\{X_t^{(0)}, t \ge 0\}$. By the construction of the product space, under $\bar{Q}$ the process $\{X_t^{(1)}, t \ge 0\}$ is independent of $\{X_t, Y_t, t \ge 0\}$. Define

$$\Pi_t^{(p_1)} q := \frac{\tilde{\Pi}_t^{(p_1)} q}{\tilde{\Pi}_t^{(p_1)} 1},$$

where

$$\tilde{\Pi}_t^{(p_1)} q := E_{\overline{Q}} \left[\tilde{L}_t q(X_t^{(1)}) | Y_s : s \le t \right],$$

and

$$\tilde{L}_t := \left(\prod_{0 \le s \le t} \lambda(X_s^{(1)}, Y_{s^-}) \Delta Y_s \right) \exp\left(- \int_0^t \lambda(X_s^{(1)}, Y_s) ds \right).$$

The following result states that the effects of the initial condition disappear as time goes to infinity.

THEOREM 4.1. (a) $h(\Pi_t^{(p_1)}, \Pi_t^{(p_0)})$ *converges to* 0 *with probability one as* $t \to \infty$.
(b) *If* $h(p_0, p_1) < \infty$ *then the above convergence is in* L^p *for every* $0 < p < \infty$.

Proof. We begin by noting that $\Pi_t^{(p_i)}, i = 0, 1$, can be recursively obtained as follows: Let $T_1, T_2, \ldots$ be the jump times of the Poisson process $\{Y_t, t \ge 0\}$, and define $T_0 = 0$. For $j \ge 1$, define $\tau_j = T_j - T_{j-1}$. Let q be a bounded real-valued measurable function. For $s, t \in [T_j, T_{j+1}), s < t, j \ge 0, i = 0, 1$, we have the following equality:

$$\Pi_t^{(p_i)} q = c \int_S E_{\overline{Q}} \left[q(X_{t-s}^{(i)}) \exp\left\{ - \int_0^{t-s} \lambda(X_u^{(i)}, Y_{T_j}) du \right\} \Big| X_0^{(i)} = x, Y_{T_j}, T_j \right] \Pi_s^{(p_i)}(dx),$$

where c is some normalizing constant. For $s \in [T_j, T_{j+1}), t = T_{j+1}$ we have

$$\Pi_{T_{j+1}}^{(p_i)} q = c \int_S E_{\overline{Q}} \Bigg[q(X_{T_{j+1}-s}^{(i)}) \lambda(X_{T_{j+1}-s}^{(i)}, Y_{T_j}) \times \exp\left\{ - \int_0^{T_{j+1}-s} \lambda(X_u^{(i)}, Y_{T_j}) du \right\} \Big| X_0^{(i)} = x, Y_{T_j}, T_j, T_{j+1} \Bigg] \Pi_s^{(p_i)}(dx).$$

The first equation yields that for any positive-valued and measurable function q and $s, t \in [T_j, T_{j+1}), s < t$ and $j \ge 0$,

$$\begin{aligned} &\exp(-\lambda_1(t-s)) \int_S G_{t-s}(x, y) q(y) m(dy) \\ &\ge E_{\overline{Q}} \left[q(X_{t-s}^{(i)}) \exp\left\{ - \int_0^{t-s} \lambda(X_u^{(i)}, Y_{T_j}) du \right\} \Big| X_0^{(i)} = x, Y_{T_j}, T_j \right] \\ &\ge \exp(-\lambda_2(t-s)) \int_S G_{t-s}(x, y) q(y) m(dy). \end{aligned} \tag{31}$$

By (4), (6), and (31) we have the inequality

$$h(\Pi_t^{(p_1)}, \Pi_t^{(p_0)}) \le \tanh \left[\frac{\ln \frac{f_2(t-s)}{f_1(t-s)} + (\lambda_2 - \lambda_1)(t-s)}{2} \right] h(\Pi_s^{(p_1)}, \Pi_s^{(p_0)}).$$

Similar considerations show that for $s \in [T_j, T_{j+1})$ and $t = T_{j+1}$,

$$h(\Pi_{T_{j+1}}^{(p_1)}, \Pi_{T_{j+1}}^{(p_0)}) \le \tanh \left[\frac{\ln \frac{f_2(T_{j+1}-s)}{f_1(T_{j+1}-s)} + (\lambda_2 - \lambda_1)(T_{j+1}-s) + \ln \frac{\lambda_2}{\lambda_1}}{2} \right] h(\Pi_s^{(p_1)}, \Pi_s^{(p_0)}). \tag{32}$$

The above inequalities show in particular that $h(\Pi_t^{(p_1)}, \Pi_t^{(p_0)})$ is nonincreasing in t. This observation and assertion (a) yields assertion (b) via the monotone convergence theorem.

Now letting δ_i denote

$$\tanh\left[\frac{\ln\frac{f_2(\tau_i)}{f_1(\tau_i)}+(\lambda_2-\lambda_1)\tau_i+\ln\frac{\lambda_2}{\lambda_1}}{2}\right],$$

we obtain

$$h(\Pi_{T_n}^{(p_1)},\Pi_{T_n}^{(p_0)})\le\delta_n h(\Pi_{T_{n-1}}^{(p_1)},\Pi_{T_{n-1}}^{(p_0)}). \tag{33}$$

Denoting $\sigma\{X_t,t\ge 0\}$ by $\mathcal{F}_X$, we have for every $m>1$,

$$P\{\exists n:\tau_n>a,n\le m\}=1-E\left\{E\left[\prod_{i=1}^{m}I_{(\tau_j\le a)}\big|\mathcal{F}_X\right]\right\}.$$

Since $E[I_{(\tau_j\le a)}|\mathcal{F}_X,T_1,\ldots,T_{j-1}]$ is at most $1-e^{-a\lambda_1}$, we must have that

$$P\{\exists n:\tau_n>a\}=1.$$

Next observe that $\tau_n(\omega)>a$ implies that $h(\Pi_{T_n}^{(p_1)},\Pi_{T_n}^{(p_0)})$ is finite, since from (30) and the proved monotonicity of the Hilbert distance it can be at most $2[\ln f_2(a)/f_1(a)+(\lambda_2-\lambda_1)a]$, which is finite by hypothesis. Hence, it follows that

$$P\left\{\exists n:h(\Pi_{T_n}^{(p_1)},\Pi_{T_n}^{(p_0)})<\infty\right\}=1.$$

Now let $m=\inf\{n:h(\Pi_{T_n}^{(p_1)},\Pi_{T_n}^{(p_0)})<\infty\}$, then $P\{m<\infty\}=1$. We show now that $E[h(\Pi_{T_n}^{(p_1)},\Pi_{T_n}^{(p_0)})|\mathcal{F}_0]$ converges to zero with probability one as $n\to\infty$, where $\mathcal{F}_0:=\mathcal{F}_X\vee\sigma\{m,T_1,\ldots,T_m\}$. This will imply that $h(\Pi_{T_n}^{(p_1)},\Pi_{T_n}^{(p_0)})$ converges in probability to 0. Then the proved monotonicity property of $h(\Pi_{T_n}^{(p_1)},\Pi_{T_n}^{(p_0)})$ implies that $h(\Pi_t^{(p_1)},\Pi_t^{(p_0)})$ converges with probability one as $t\to\infty$.

We begin by observing that for $n>m(\omega)$,

$$E\left[h(\Pi_{T_n}^{(p_1)},\Pi_{T_n}^{(p_0)})|\mathcal{F}_0\right]\le E\left(\left(\prod_{i=m+1}^{n}\delta_i\right)|\mathcal{F}_0\right)h(\Pi_{T_m}^{(p_1)},\Pi_{T_m}^{(p_0)}). \tag{34}$$

Observe that

$$\delta_i=1-2\left[\frac{\frac{f_1(\tau_i)\lambda_1}{f_2(\tau_i)\lambda_2}e^{-(\lambda_2-\lambda_1)\tau_i}}{1+\frac{f_1(\tau_i)\lambda_1}{f_2(\tau_i)\lambda_2}e^{-(\lambda_2-\lambda_1)\tau_i}}\right].$$

Therefore, for $\tau_i\in[a,b]$,

$$\delta_i\le 1-2\frac{e^{-(\lambda_2-\lambda_1)b}/C}{1+e^{-(\lambda_2-\lambda_1)b}/C},$$

where

$$C:=\frac{\lambda_2}{\lambda_1}\sup_{t\in[a,b]}f_2(t)/f_1(t),$$

which is finite by hypothesis.

Defining $\gamma = 1 - 2\frac{e^{-(\lambda_2-\lambda_1)b}/C}{1+e^{-(\lambda_2-\lambda_1)b}/C}$, we can write

$$P\left\{\delta_i \le \gamma \middle| \mathcal{F}_0, T_{m+1}, \ldots, T_{i-1}\right\} \ge P\left\{\tau_i \in [a,b] \middle| \mathcal{F}_0, T_{m+1}, \ldots, T_{i-1}\right\}.$$

Finally, observe that for every $i > m(\omega)$

$$P\left\{\tau_i \in [a,b] \middle| \mathcal{F}_0, T_{m+1}, \ldots, T_{i-1}\right\} \ge e^{-a\lambda_2}\left(1 - e^{-\lambda_1(b-a)}\right) > 0.$$

This implies that there exists a constant $\overline{\gamma} < 1$ such that

$$E[\delta_{m+j} | \mathcal{F}_0, T_{m+1}, \ldots, T_{m+j-1}] \le \overline{\gamma}$$

for all $j \ge 0$. This and (34) imply that $E[h(\Pi_{T_n}^{(p_1)}, \Pi_{T_n}^{(p_0)}) | \mathcal{F}_0] \to 0$ a.s., which completes the proof. □

Remark 3. The hypothesis (H) is satisfied if $\{X_t, t \ge 0\}$ is a diffusion on a compact Riemannian manifold with smooth drift and diffusion coefficients and a strictly elliptic generator, as shown in [2].

In the final part of this section we consider the above filtering problem, where in addition to an incorrect initial condition, there is a misspecification in the transition kernel. Let G_t be as before and replace (H) with the following stronger condition (H1).

(H1) There exist maps f_1, f_2 from $(0,\infty)$ to $[0,\infty)$ such that
$\forall \quad a, b \in (0,\infty)\,, \quad \sup_{t\in[a,b]} \frac{f_2(t)}{f_1(t)} < \infty,$
and $\forall t \in (0,\infty)$ and for $x, y \in S$

$$f_1(t) \le G_t(x,y) \le f_2(t), \tag{35}$$

where $0/0 = 0$.

Let $G_t^{(k)}, k \ge 1$, be a sequence of transition probability kernels, and let $p_k, k \ge 1$, be a sequence of probability measures on $(S, \mathcal{S})$. Let $(\Omega_1, \mathcal{F}_1, P_1)$ and $(\overline{\Omega}, \overline{\mathcal{F}}, \overline{Q})$ be as before.

For $k \ge 1$, define the Markov processes $\{X_t^{(k)}, t \ge 0\}$, on $(\Omega_1, \mathcal{F}_1, P_1)$ with initial distribution p_k and transition probability kernel $G_t^{(k)}$. Under $\overline{Q}$, they are independent of $\{X_t, Y_t, t \ge 0\}$. Define

$$L_t^{(k)} := \left(\prod_{0\le s\le t} \lambda(X_s^{(k)}, Y_{s-})\Delta Y_s\right) \exp\left\{-\int_0^t \lambda(X_s^{(k)}, Y_s) ds\right\}, \tag{36}$$

and define the probability measure $\Pi_t^{(k)}$ as follows: For $A \in \mathcal{S}$,

$$\Pi_t^{(k)}(A) := \frac{E_{\overline{Q}}[L_t^{(k)} I_A(X_t^{(k)}) | Y_s : s \le t]}{E_{\overline{Q}}[L_t^{(k)} | Y_s : s \le t]}. \tag{37}$$

Finally, for later notational convenience, on $(\Omega_1, \mathcal{F}_1, P_1)$ we define Markov processes $\{X_t^{[k]}, t \ge 0\}$ with initial distribution p_k and transition probability kernel G_t. They are also independent of $\{X_t, Y_t, t \ge 0\}$, under $\overline{Q}$. Define $L_t^{[k]}$ and $\Pi_t^{[k]}$ as in (36) and (37) by replacing (k) with $[k]$. We will assume the following condition on the kernels $G_t^{(k)}$:

(H2) $\limsup_{k\to\infty} \sup_{0\le t\le\delta_0} \sup_{(x,y)\in S\times S} |\ln G_t^{(k)}(x,y) - \ln G_t(x,y)| = 0$,
for some $\delta_0 > 0$, where $|\ln(0) - \ln(0)| = 0$.

Condition (H2) says that the distance (in a log scale) between $G_t(\cdot)$ and $G_t^{(k)}(\cdot)$ is small for large k uniformly for t in some interval containing the origin. The theorem says that the outputs of filters (with the same observation process) but built under different assumptions on the transition kernel for the signal process, will eventually be close with a high probability if the two kernels are close in the given metric, irrespective of the (different) initial conditions.

THEOREM 4.2. *Assume that* (H1) *and* (H2) *hold, then* $\forall\delta$, $0 < \delta < \delta_0$, *there exists a sequence of stopping times,* t_n, *with respect to the filtration* $\sigma\{Y_s : s \le t\}$, *increasing to* ∞ *and satisfying* $|t_{n+1} - t_n| < \delta$, *with probability one and such that*

$$\lim_{k\to\infty} \limsup_{n\to\infty} P\left\{h(\Pi_{t_n}, \Pi_{t_n}^{(k)}) > \epsilon\right\} = 0$$

for every $\epsilon > 0$. *In fact, given* t_n, t_{n+1} *equals the time of the next observation if this occurs no later than* δ *units of time later; otherwise it is* $t_n + \delta$.

Proof. For notational simplicity we present the proof for the case where $\lambda(x,y) \equiv \lambda(x)$. Let $\{T_n\}$ continue to denote the jump times of Y. Define t_n as in the theorem statement. More formally, set $t_1 := T_1 \wedge \delta$ and for $n > 1$, define $t_n := t_{n-1} + (T_n^* - t_{n-1}) \wedge \delta$, where $T_n^* := \inf\{T_j : T_j > t_{n-1}\}$. Observe that $t_n - t_{n-1} \le \delta$ and t_n increases to infinity with probability one. We begin by observing that the triangle inequality implies that

$$h(\Pi_{t_n}, \Pi_{t_n}^{(k)}) \le h(\Pi_{t_n}, \Pi_{t_n}^{[k]}) + h(\Pi_{t_n}^{[k]}, \Pi_{t_n}^{(k)}). \tag{38}$$

By Theorem 4.1, for each fixed k the first term converges to zero with probability one as $n \to \infty$. Therefore it suffices to consider the second term. For the second term we will show that

$$\lim_{k\to\infty} \sup_n E\left[h(\Pi_{t_n}^{(k)}, \Pi_{t_n}^{[k]})\right] = 0. \tag{39}$$

Clearly, this will give the desired result. Let $\tau_n^* := t_n - t_{n-1}$. Define a sequence of nonnegative operators, $\{K_n\}_{n\ge1}$, on $\mathcal{M}(S)$ as follows. For a measurable function $q : S \to [0,\infty)$ and $\mu \in \mathcal{M}(S)$, $t_n \in \{T_j; j \ge 1\}$, set

$$\begin{aligned} K_n(\mu)q := \int_S E_{\overline{Q}}\Big[& q(X_{\tau_n^*}^{(0)})\lambda(X_{\tau_n^*}^{(0)}) \\ & \exp\left\{-\int_0^{\tau_n^*} \lambda(X_u^{(0)})du\right\} \big| X_0^{(0)} = x, T_j, j \ge 1\Big]\mu(dx). \end{aligned} \tag{40}$$

Otherwise, for $t_n \notin \{T_j; j \ge 1\}$, set

$$K_n(\mu)q := \int_S E_{\overline{Q}}\left[q(X_{\tau_n^*}^{(0)}) \exp\left\{-\int_0^{\tau_n^*} \lambda(X_u^{(0)})du\right\} \big| X_0^{(0)} = x, T_j, j \ge 1\right] \mu(dx). \tag{41}$$

Define the operators $K_n^{(k)}$ in a similar fashion by replacing $X^{(0)}$ by $X^{(k)}$. Then clearly, $\Pi_{t_n}^{(k)} = K_n^{(k)}\Pi_{t_{n-1}}^{(k)}$ and $\Pi_{t_n}^{[k]} = K_n\Pi_{t_{n-1}}^{[k]}$. By applying the triangle inequality to the first term on the right side of (38), we have

$$h(\Pi_{t_n}^{(k)}, \Pi_{t_n}^{[k]}) \le h(K_n^{(k)}\Pi_{t_{n-1}}^{(k)}, K_n\Pi_{t_{n-1}}^{(k)}) + h(K_n\Pi_{t_{n-1}}^{(k)}, K_n\Pi_{t_{n-1}}^{[k]}). \tag{42}$$

Now we will obtain an upper bound for the first term on the right side of the above inequality. Assume initially that $t_n \in \{T_j; j \geq 1\}$. Then for $x \in S$ and q as before, and letting δ_x denote the Dirac measure at x, $K_n^{(k)}(\delta_x)q$ can be written as

$$e^{-\lambda_2 \tau_n^*} E_{\overline{Q}} \left[q(X_{\tau_n^*}^{(k)}) \lambda(X_{\tau_n^*}^{(k)}) \exp\left\{ \int_0^{\tau_n^*} (\lambda_2 - \lambda(X_u^{(k)}))du \right\} \big| X_0^{(k)} = x, T_j, j \geq 1 \right].$$

For $y \in S$, define $\lambda^*(y) = \lambda_2 - \lambda(y)$. Then we can write

$$\exp\left\{ \int_0^{\tau_n^*} (\lambda_2 - \lambda(X_u^{(k)}))du \right\} = 1 + \sum_{j=1}^{\infty} \int_0^{\tau_n^*} \int_0^{s_{j-1}} \cdots \int_0^{s_1} \lambda^*(X_{s_{j-1}}^{(k)}) \cdots \lambda^*(X_{s_0}^{(k)})\, ds_0 \cdots ds_{j-1}.$$

The above representation yields that $K_n^{(k)}(\delta_x)q$ equals

$$e^{-\lambda_2 \tau_n^*} \Bigg[E_{\overline{Q}}[q(X_{\tau_n^*}^{(k)}) \lambda(X_{\tau_n^*}^{(k)}] + \sum_{j=1}^{\infty} \int_0^{\tau_n^*} \int_0^{s_{j-1}} \int_0^{s_1} E_{\overline{Q}}[q(X_{\tau_n^*}^{(k)}) \lambda(X_{\tau_n^*}^{(k)}) \lambda^*(X_{s_{j-1}}^{(k)}) \cdots \lambda^*(X_{s_0}^{(k)})] ds_0 \cdots ds_{j-1} \Bigg], \tag{43}$$

where the expectation is over everything but τ_n^*.

Next observe that the expectation inside the multiple integral above can be written as

$$\int_{S^{j+1}} \lambda(x_{j+1}) q(x_{j+1}) \lambda^*(x_j) \cdots \lambda^*(x_1) G_{\tau_n^* - s_j}^{(k)}(x_j, x_{j+1}) \cdots G_{s_0}^{(k)}(x, x_1) dm(x_1) \cdots dm(x_{j+1}) \tag{44}$$

Define

$$\rho_k := \sup_{0 \leq t \leq \delta_0} \sup_{(x,y) \in S \times S} |\ln G_t^{(k)}(x,y) - \ln G_t(x,y)|.$$

By hypothesis $\rho_k \to 0$ as $k \to \infty$. Moreover, $\forall u \in [0, \delta_0]$ and $(x,y) \in S \times S$,

$$e^{-\rho_k} G_u(x,y) \leq G_u^{(k)}(x,y) \leq e^{\rho_k} G_u(x,y).$$

Using this observation in (44) we obtain that $K_n^{(k)}(\delta_x)q$ is bounded above by

$$\exp\{\rho_k + \lambda_2 \tau_n^* (e^{\rho_k} - 1)\} E_{\overline{Q}} \Bigg[q(X_{\tau_n^*}^{(0)}) \lambda(X_{\tau_n^*}^{(0)}) \times \exp\left\{ -e^{\rho_k} \int_0^{\tau_n^*} \lambda(X_u^{(0)}) du \right\} \big| X_0^{(0)} = x, T_j, j \geq 1 \Bigg].$$

and below by

$$\exp\{-\rho_k + \lambda_2 \tau_n^* (e^{-\rho_k} - 1)\} E_{\overline{Q}} \left[q(X_{\tau_n^*}^{(0)}) \lambda(X_{\tau_n^*}^{(0)}) \times \exp\left\{ -e^{-\rho_k} \int_0^{\tau_n^*} \lambda(X_u^{(0)}) du \right\} \big| X_0^{(0)} = x, T_j, j \geq 1 \right].$$

Furthermore, the upper bound is bounded above by

$$\exp\left\{\rho_k + \lambda_2 \tau_n^* (e^{\rho_k} - 1)\right\} K_n(\delta_x) q,$$

and the lower bound is bounded below by

$$\exp\left\{-\rho_k + \lambda_2 \tau_n^* (e^{-\rho_k} - 1)\right\} K_n(\delta_x) q.$$

On recalling the definition of the Hilbert projective distance, using (2), and noting that $\tau_n^* \leq \delta$, these bounds yield

$$h(K_n^{(k)} \Pi_{t_{n-1}}^{(k)}, K_n \Pi_{t_{n-1}}^{(k)}) \leq \lambda_2 \delta [e^{\rho_k} - e^{-\rho_k}] + 2\rho_k.$$

Recall that we have proved the above inequality for $t_n \in \{T_j : j \geq 1\}$. However, it is easy to see that the inequality continues to hold for $t_n \notin \{T_j : j \geq 1\}$. (Compare (40) and (41).) Define $\epsilon_k = \lambda_2 \delta [e^{\rho_k} - e^{-\rho_k}] + 2\rho_k$. Then $\epsilon_k \to 0$ as $k \to \infty$ and

$$h(K_n^{(k)} \Pi_{t_{n-1}}^{(k)}, K_n \Pi_{t_{n-1}}^{(k)}) \leq \epsilon_k \, \forall n \geq 1. \tag{45}$$

Now consider the second term on the right side of (42). Using Birkhoff's contraction inequality, we have

$$h(K_n \Pi_{t_{n-1}}^{(k)}, K_n \Pi_{t_{n-1}}^{[k]}) \leq \tanh(C(K_n)/4) h(\Pi_{t_{n-1}}^{(k)}, \Pi_{t_{n-1}}^{[k]}). \tag{46}$$

Moreover, the arguments in the proof of Theorem 4.1 show that

$$C(K_n) \leq 2 \left(\ln \frac{f_2(\tau_n^*)}{f_1(\tau_n^*)} + \lambda_2 \tau_n^* + \ln \lambda_2/\lambda_1 \right). \tag{47}$$

It follows now from (42), (45), (46), (47) that

$$h(\Pi_{t_n}^{(k)}, \Pi_{t_n}^{[k]}) \leq \epsilon_k \sum_{j=1}^{n} \delta_j \cdots \delta_n, \tag{48}$$

where

$$\delta_n := \tanh \left[\frac{\ln \frac{f_2(\tau_n^*)}{f_1(\tau_n^*)} + \lambda_2 \tau_n^* + \ln \lambda_2/\lambda_1}{2} \right].$$

Next note that

$$E_{\overline{Q}} [\delta_j \cdots \delta_n] \leq E_{\overline{Q}} \left\{ \delta_j \cdots \delta_{n-1} E_{\overline{Q}} \left[\delta_n \middle| \mathcal{F}_X, t_1, \cdots, t_{n-1} \right] \right\}.$$

Also, as in Theorem 4.1 (see the arguments following inequality (34)),

$$E_{\overline{Q}} \left[\delta_n \middle| \mathcal{F}_X, t_1, \ldots, t_{n-1} \right] \leq 1 - \gamma^* P \left\{ \tau_n^* \in [\delta/2, \delta] \middle| \mathcal{F}_X, t_1, \ldots, t_{n-1} \right\},$$

where

$$\gamma^* := 2 \left[\frac{e^{-\lambda_2 \delta / C}}{1 + e^{-\lambda_2 \delta / C}} \right],$$

and

$$C := [\lambda_1/\lambda_2] \sup_{t\in[\delta/2,\delta]} f_2(t)/f_1(t).$$

Also

$$P\left\{\tau_n^* \in [\delta/2,\delta]\middle|\mathcal{F}_X, t_1,\dots,t_{n-1}\right\} > P\left\{Y_{t_{n-1}+\delta} - Y_{t_{n-1}} = 0\middle|\mathcal{F}_X, t_1,\dots,t_{n-1}\right\}.$$

The last term on the right is obviously greater than $e^{-\lambda_2\delta}$. Therefore,

$$E_{\overline{Q}}\left[\delta_n\middle|\ \mathcal{F}_X, t_1,\dots,t_{n-1}\right] \le 1-\gamma^* e^{-\lambda_2\delta} =: \kappa.$$

Using this observation in (48), we get

$$E[h(\Pi_{t_n}^{(k)}, \Pi_{t_n}^{[k]})] \le \epsilon_k/(1-\kappa).$$

This proves (39) and hence the theorem. □

Acknowledgments. We would like to thank a careful referee for pointing out an error in the proof of Theorem 3.1 and suggesting an alternate proof.

REFERENCES

[1] R. Atar and O. Zeitouni, *Lyapunov exponents for finite state nonlinear filtering*, SIAM J. Control Optim., (1996).
[2] R. Atar and O. Zeitouni, *Exponential stability for nonlinear filtering*, Ann. Inst. H. Poincaré Probab. Statist., 36 (1997), pp. 691–725.
[3] G. Birkhoff, *Lattice Theory*, 3rd ed., Am. Math. Soc. Colloq. Publ. 25, AMS, Providence, RI, 1967.
[4] P. Brémaud, *Point Processes and Queues*, Springer-Verlag, New York, 1981.
[5] A. Budhiraja and D. Ocone, *Exponential stability of discrete time filters without signal ergodicity*, Systems Control Lett., 30 (1997), pp. 185–193.
[6] B. Delyon and O. Zeitouni, *Lyapunov exponents for filtering problems*, in Applied Stochastic Analysis, M.H.A. Davis and R.J. Elliot, eds., Gordon and Breach, New York, 1991, pp. 511–521.
[7] H. Kunita, *Asymptotic behavior of the nonlinear filtering errors of Markov processes*, J. Multivariate Anal., 1 (1971), pp. 365–393.
[8] H. Kushner and H. Huang, *Approximate and limit results for nonlinear filters with wide bandwidth observation noise*, Stochastics, 16 (1986), pp. 65–96.
[9] F. Le Gland and L. Mevel, *Geometric Ergodicity in Hidden Markov Models*, preprint.
[10] C. Liverani, *Decay of correlations*, Ann. Math., 142 (1995), pp. 239–301.
[11] D. Ocone and E. Pardoux, *Asymptotic stability of the optimal filter with respect to its initial condition*, SIAM J. Control Optim., 34 (1996), pp. 226–243.
[12] L. Stettner, *On invariant measures of filtering processes*, in Stochastic Differential Systems, Proc. 4th Bad Honnef Conf., 1988, Lecture Notes in Control and Inform Sci., K. Helmes, N. Christopeit, and M. Kohlmann, eds., 1989, pp. 279–292.
[13] L. Stettner, *Invariant measures of pair: State, approximate filtering process*, Colloq. Math. LXII, (1991), pp. 347–352.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 5, pp. 1638–1668, September 1998

007

NONLINEAR FILTERING AND CONTROL OF A SWITCHING DIFFUSION WITH SMALL OBSERVATION NOISE*

Q. ZHANG†

Abstract. This paper is concerned with nonlinear filtering and control of a switching diffusion coupled by an unknown Markov chain. Two statistical estimation methods are used to track the unknown Markov chain. Computable approximate filters are obtained based on these methods. The filters are then used to construct controls for the partially observed system. These controls are shown to be asymptotically optimal as the observation noise tends to zero. Finally an example is considered and numerical experiments are reported.

Key words. nonlinear filtering, hybrid system, interacting multiple model, small observation noise, nearly optimal control

AMS subject classifications. 93E20, 93E11

PII. S0363012997315440

1. Introduction.

Let $(\alpha(t), x(t))$, $t \geq 0$, denote a pair of signal (or state) processes which is not directly observable. Consider the case that a function of $(\alpha(s), x(s))$, $s \leq t$, (linear in $x(\cdot)$) with an additive noise is observable given by $y(t)$. Let $u(t)$, $t \geq 0$, be a control process depending on the observation $y(\cdot)$ up to time t. Assume both $x(\cdot)$ and $y(\cdot)$ are $\mathbb{R}^p$-valued stochastic processes satisfying the equations

$$\begin{cases} dx(t) = b(t, \alpha(t), x(t), u(t))dt + \sigma(t, \alpha(t))dw(t), & x(0) = x_0, \\ dy(t) = h(t, \alpha(t), x(t))dt + \varepsilon dv(t), & y(0) = 0, \end{cases} \tag{1.1}$$

for $0 \leq t \leq T$, where T is a finite number, x_0 is a given random variable, $\alpha(\cdot)$ is an unknown Markov process, $(w(\cdot), v(\cdot))$ is a standard Brownian motion, and $\varepsilon > 0$ is a small parameter. Here we only consider those $u(\cdot)$ under which the equations in (1.1) have a (strong) solution.

Let $\mathcal{Y}_t$ denote the σ-algebra generated by the observation process $y(\cdot)$ up to time t, i.e., $\mathcal{Y}_t = \sigma\{y(s) : 0 \leq s \leq t\}$. The objective of the problem is to choose a $\mathcal{Y}_t$ progressively measurable control $u(\cdot)$ to minimize the cost functional

$$J(u(\cdot)) = E \int_0^T L(t, \alpha(t), x(t), u(t))dt. \tag{1.2}$$

First of all, let us consider the case when $\alpha(t)$ is a constant over time, say $\alpha(t) = \alpha_0$. If α_0 is known and $\varepsilon = 0$ in (1.1), then the system reduces to a completely observable system provided that h is one-to-one in x. So for ε small but different from zero, the problem under consideration is a singular perturbation of a "trivial" situation. In Haussmann and Zhang [11], such optimal control problems were studied with the aid of the extended Kalman filter (EKF) and the Picard filter (PF). Controls

*Received by the editors January 27, 1997; accepted for publication (in revised form) November 11, 1997; published electronically June 9, 1998. This research was supported in part by Office of Naval Research grant N00014-96-1-0263 and in part by the University of Georgia Faculty Research Grant.

http://www.siam.org/journals/sicon/36-5/31544.html

†Department of Mathematics, University of Georgia, Athens, GA 30602 (qingz@math.uga.edu).

based on the filtering outcomes were obtained, which are shown to be asymptotically optimal as $\varepsilon \to 0$.

When α_0 is not available to the controller of the system, the situation becomes more complicated. One of the major difficulties lies in the nonlinearility of system (1.1). If (1.1) is a linear system, the optimal control problem was solved by Hijab [14] and Caines and Chen [3] assuming an a priori distribution of the unknown parameter. However, in some practical situations, such a priori knowledge of α_0 is not available. To deal with the problem in this case, Haussmann and Zhang [12] used two statistical hypothesis tests, the quadratic variation test (QVT) and the likelihood ratio (least-squares) test (LRT), to estimate the value of α_0 and to choose among competing filters on successive time intervals. Then a control policy is obtained by using the filtering outcomes which is shown to be asymptotically optimal. The QVT and the LRT schemes were introduced by Fleming and Pardoux [7] to identify the sign of the state variable $x(t)$ in a partially observed system; see also Fleming and Zhang [9, 10] and Fleming et al. [6] for the corresponding discrete-time models and related numerical results along this line.

In this paper we consider the case when $\alpha(\cdot)$ is an unobservable Markov chain. Typically, to solve the underlying control problem, one needs to solve the associated filtering problem first, i.e., to find a conditional expectation $(\hat{\alpha}(t), \hat{x}(t)) = E[(\alpha(t), x(t))|\mathcal{Y}_t]$. However, owing to the nonlinearity of the system, especially the presence of $\alpha(\cdot)$, obtaining $(\hat{\alpha}(t), \hat{x}(t))$ requires solving the associated Zakai equation (or the nonlinear filtering equation), which is inherently infinite dimensional. Much effort in the literature was devoted to finite dimensional approximations. In Blom and Bar-Shalom [2], a discrete-time version of the corresponding filtering problem was considered. They proposed a numerical algorithm to compute $(\hat{\alpha}(\cdot), \hat{x}(\cdot))$. The algorithm seems to perform well numerically. However, there is no theoretical justification for the optimality (or the near optimality) of these filters; see Li [18] for further discussions.

In this paper, in order to design an approximate filter, we use the QVT (or the LRT) to estimate the value of $(\alpha(t), x(t))$ over time. We show that the resulting filters are asymptotically optimal as the observation noise goes to 0. The random jumps of $\alpha(\cdot)$ create one of the major difficulties when verifying the near optimality of these filters.

When dealing with a switching diffusion coupled by an unknown Markov chain, most of the nonlinear filters in the literature require the generator of $\alpha(\cdot)$ to be given. In practice, it usually takes a certain period of time to estimate the generator matrix. This creates a major problem if the generator is time dependent. The advantage of the filtering methods used in this paper is that they do not require knowing the generator of $\alpha(\cdot)$. In fact, the methods used in this paper can be easily extended to deal with much more general models in which even the Markovian assumption of $\alpha(\cdot)$ is unnecessary! In this connection, we refer to Remark 3.4 for discussions.

This paper extends the results on filtering and control in [12] to incorporate the case when $\alpha(\cdot)$ is an unknown Markov chain. The Markovian property is only required when dealing with the feedback controls. We design approximate filters and feedback controls for the problem under consideration. The main contribution of the paper is the verification of the asymptotic optimality of these filters and controls.

This paper is concerned with nonlinear filtering and control of a partially observed system. There is substantial literature on many related models and problems. For classical results on nonlinear filtering, we refer the books by Kallianpur [15] and

Liptser and Shiryayev [19]. For recent developments and review of the literature on partially observed systems, we refer the reader to the books by Bensoussan [1], Elliott, Aggoun, and Moore [5], Kushner [17], and references therein.

The paper is organized as follows. In the next section we formulate the problem under consideration and make assumptions. In section 3, we study the nonlinear filtering problem by using the QVT and the LRT methods and prove the asymptotic optimality of these filters. Then in section 4, we consider control policies based on the filtering results together with the dynamic programming approach. We show that the constructed control policies are nearly optimal as $\varepsilon \to 0$. In these sections we use the idea of an EKF to design nonlinear filters. For small ε, an EKF can be further approximated by a PF. In section 5, we extend these results to a hybrid linear quadratic system. In section 6, we consider the PF and present further extensions of these results to the case when the EKF is replaced by the corresponding PF. In section 7, we give a simple example and a set of numerical simulations to demonstrate the performance of the schemes and to compare with an existing algorithm in filtering. Finally, we conclude the paper by making some remarks. Technical results used in the paper are given in the Appendix.

Before moving on to the next section, let us give a list of notation used in the paper:

A'	the transpose of a matrix A;
B^c	the complement of a set B;
I	the identity matrix;
I_D	the indicator function of a set D;
$O(x)$	a function of x such that $\sup_{x\neq 0} \lvert O(x)\rvert/\lvert x\rvert < \infty$;
$[a]$	the integer part of a number a;
$\mathrm{tr}(A)$	the trace of a matrix A;
$\nabla_x f$	the partial derivative $\partial f/\partial x$;
$\lvert\xi(\cdot)\rvert_T$	$:= E\int_0^T \lvert\xi(t)\rvert dt$ for a stochastic process $\xi(\cdot)$.

Also, K is used as a generic positive constant throughout. The values of the K may be different for each appearance, but it should be clear from the context.

2. Problem formulation. Let $(\Omega, \mathcal{F}, P)$ denote a probability space and let $(w(\cdot), v(\cdot))$ be a standard Brownian motion. Given a positive integer m, let $\mathcal{M} = \{1, 2, \ldots, m\}$ denote the state space of $\alpha(\cdot)$, i.e., $\alpha(t) \in \mathcal{M}$, $t \geq 0$. We assume that $\alpha(\cdot)$ is a finite state Markov chain generated by a Borel measurable and bounded matrix $Q(t) = (q_{ij}(t))$, $t \geq 0$, with $q_{ij}(t) \geq 0$ for $i \neq j$ and $q_{ii}(t) = -\sum_{j\neq i} q_{ij}(t)$. The construction of a Markov chain generated by $Q(t)$ can be given as in Davis [4]. A particular case of the generator is when $Q(t) = Q$, which is independent of t, and the resulting Markov chain $\alpha(\cdot)$ is stationary.

We make the following assumptions in this paper.

(A1) For each $i \in \mathcal{M}$, $b(t, i, x, u)$ is a Borel measurable function of (t, x, u). The gradients $\nabla_x b(t, i, x, u)$ and $\nabla_u b(t, i, x, u)$ exist and are bounded.

(A2) There exist a matrix $H(t, i)$ and a constant $c > 0$ such that $h(t, i, x) = H(t, i)x$ with $H'(t, i)H(t, i) \geq cI > 0$ for all $t \geq 0$ and $i \in \mathcal{M}$. Furthermore, for each $i \in \mathcal{M}$, $H(t, i)$ is bounded and continuously differentiable in t.

(A3) $\sigma(t, i) = F(t, i)H'(t, i)$ for symmetric matrices $F(t, i) \geq cI > 0$. Moreover, $F(t, i)$ is bounded and continuously differentiable in t.

(A4) The initial value x_0 is a Gaussian random variable and $E|x_0 - Ex_0|^4 = O(\varepsilon^2)$. Moreover, x_0, $\alpha(\cdot)$, $w(\cdot)$, and $v(\cdot)$ are independent.

Remark 2.1. For notational simplicity, we use the Gaussian initial conditions

in (A4). This requirement can be relaxed as in Haussmann and Zhang [11] to non-Gaussian initial cases. The form assumed for $\sigma(t,i)$ in (A3) is not as restrictive as it appears. In fact, in the one-dimensional case, it is equivalent to the condition that $\sigma(t,i) \neq 0$. For higher dimensional cases, we refer the papers Haussmann and Zhang [11, 12] for related discussions.

3. Nonlinear filtering. In this section we consider the nonlinear filtering of the problem under consideration. For simplicity in notation, we suppress the variable u in (1.1) and consider the system given as

$$\begin{cases} dx(t) = b(t,\alpha(t),x(t))dt + \sigma(t,\alpha(t))dw(t), & x(0) = x_0, \\ dy(t) = h(t,\alpha(t),x(t))dt + \varepsilon dv(t), & y(0) = 0. \end{cases} \tag{3.1}$$

Remark 3.1. In the context of target tracking and filtering, the model in (3.1) is rich enough to capture many practical scenarios. To illustrate, let us consider $x(t) = (x^1(t), x^2(t))$ with $x^1(t) \in \mathbb{R}^3$ representing the position of the target and $x^2(t) \in \mathbb{R}^3$ its velocity. If we take $\alpha(t)$ to be the driving force of the target, then $\alpha(t)$ is proportional to the acceleration rate of the target given by the derivative of $x^2(t)$. Viewing the problem in this way, it is reasonable to consider the observation function $h(t,\alpha,x)$ is dependent on α.

In practice the observation noise level mainly depends on the sensor measurement characteristics. The development of new technology (such as the use of infrared technology) makes it possible for having fairly small disturbances in observation. So it is not only reasonable but also practical to consider the models with small observation noise.

Let $D[0,T]$ denote the space of functions defined on $[0,T]$ that are right-continuous and have a left-hand limit. Let

$$\Theta = \{\theta(\cdot) \in D[0,T] : \text{ such that } \theta(t) \in \mathcal{M},\ 0 \le t \le T\}.$$

For notational convenience, we write $\theta = \theta(\cdot)$, for each $\theta(\cdot) \in \Theta$. Let $(\tilde{x}^\theta(t), R^\theta(t))$, $t \ge 0$, denote the output of the EKF under the condition that $\alpha(\cdot) = \theta$. Then

$$\begin{aligned} d\tilde{x}^\theta(t) =\ & b(t,\theta(t),\tilde{x}^\theta(t))dt \\ & + \frac{1}{\varepsilon^2} R^\theta(t) H'(t,\theta(t)) \left(dy(t) - h(t,\theta(t),\tilde{x}^\theta(t))dt\right), \\ \frac{dR^\theta(t)}{dt} =\ & \nabla_x b(t,\theta(t),\tilde{x}^\theta(t)) R^\theta(t) + R^\theta(t)(\nabla_x b(t,\theta(t),\tilde{x}^\theta(t)))' \\ & + F(t,\theta(t)) H'(t,\theta(t)) H(t,\theta(t)) F(t,\theta(t)) \\ & - \frac{1}{\varepsilon^2} R^\theta(t) H'(t,\theta(t)) H(t,\theta(t)) R^\theta(t), \end{aligned} \tag{3.2}$$

with $\tilde{x}^\theta(0) = Ex_0 \in \mathbb{R}^p$ and $R^\theta(0) = \mathrm{Cov}(x_0) \in \mathbb{R}^{p\times p}$.

Given a fixed number $0 < \sigma < 1$, let

$$\Theta_\sigma^\varepsilon = \left\{\theta(\cdot) \in \Theta : \text{ the number of jumps of } \theta(\cdot) \le [1/\varepsilon^\sigma]\right\}.$$

Then in view of Lemma A.6 in the Appendix, there exists a constant K such that

$$P(\alpha(\cdot) \notin \Theta_\sigma^\varepsilon) = P(\alpha(\cdot) \text{ jumps more than } [1/\varepsilon^\sigma] \text{ times}) \le K\varepsilon^2. \tag{3.3}$$

Thus, the set $\Theta_\sigma^\varepsilon$ can be regarded as an approximation to Ω because $P(\alpha(\cdot) \in \Theta_\sigma^\varepsilon)$ is close to 1 due to the fact that

$$P(\alpha(\cdot) \in \Theta_\sigma^\varepsilon) = 1 - P(\alpha(\cdot) \notin \Theta_\sigma^\varepsilon) \ge 1 - K\varepsilon^2.$$

Let $\gamma_0 > 0$ be a constant and define

$$\Theta^{\varepsilon}_{\sigma,\gamma_0} = \Big\{\theta = \theta(\cdot) \in \Theta^{\varepsilon}_{\sigma} : \text{ the duration between any two jumps of } \theta(\cdot) \geq \gamma_0\varepsilon\Big\}.$$

If we let $0 = t_0 < t_1 < \cdots < t_n < T$ denote the jump times of $\theta(\cdot)$, then $\theta(\cdot) \in \Theta^{\varepsilon}_{\sigma,\gamma_0}$ implies that $n \leq [1/\varepsilon^{\sigma}]$ and $t_{j+1} - t_j \geq \gamma_0\varepsilon$.

Let $0 = \tau_0 < \tau_1 < \cdots$ denote the sequence of the random jump times of $\alpha(\cdot)$. Then, for $j = 0, 1, \ldots$, the distribution of $\tau_{j+1} - \tau_j$ is exponential. Thus, for some constant K, $P(\tau_{j+1} - \tau_j < t) \leq Kt$ for $t \geq 0$. It follows that

$$\begin{aligned} P(\alpha(\cdot) \notin \Theta^{\varepsilon}_{\sigma,\gamma_0}) &\leq P\Big(\bigcup_{j=0}^{[1/\varepsilon^{\sigma}]} \{\tau_{j+1} - \tau_j < \gamma_0\varepsilon\}\Big) + P(\alpha(\cdot) \notin \Theta^{\varepsilon}_{\sigma}) \\ &\leq \sum_{j=0}^{[1/\varepsilon^{\sigma}]} P(\tau_{j+1} - \tau_j < \gamma_0\varepsilon) + O(\varepsilon^2) \leq \sum_{j=0}^{[1/\varepsilon^{\sigma}]} K\varepsilon + O(\varepsilon^2) = O(\varepsilon^{1-\sigma}). \end{aligned} \tag{3.4}$$

Therefore, $\Theta^{\varepsilon}_{\sigma,\gamma_0}$ can also be considered as an approximation to Ω.

The sets $\Theta^{\varepsilon}_{\sigma}$ and $\Theta^{\varepsilon}_{\sigma,\gamma_0}$ are defined so that their elements meet certain requirements on the number of jumps and the duration between consecutive jumps. These requirements are useful in the subsequent analysis.

Then as can be shown in Lemma A.3 that there exist $\gamma_0 > 0$ and K such that for each $0 < \sigma < 1$, $0 < \varepsilon < \varepsilon_0$, and for all $\theta = \theta(\cdot) \in \Theta^{\varepsilon}_{\sigma,\gamma_0}$,

$$E^{\theta} \int_0^T \left|\hat{x}^{\theta}(t) - \tilde{x}^{\theta}(t)\right|^2 dt \leq K\varepsilon^4,$$

where $\hat{x}^{\theta}(t) = E^{\theta}[x(t)|\mathcal{Y}_t]$ and E^{θ} is the conditional expectation given $\alpha(\cdot) = \theta$.

It is important to estimate the value of $\alpha(\cdot)$ in the filtering problem under consideration because if an estimate of $\alpha(\cdot)$ is given, one can use such estimate to choose the corresponding EKF as an estimate for $\hat{x}(\cdot)$ as in Lemma A.3. In this section we consider two statistical tests, the QVT and the LRT, to identify the value of the unknown parameter process $\alpha(\cdot)$ at a given time t.

In Haussmann and Zhang [12], they considered the case when $\alpha(t) = \alpha_0$, a constant parameter. So there are only a finite number of parameter values to examine. In this paper we need to carry out the parameter identification at each time instant t to incorporate random fluctuations of $\alpha(\cdot)$.

QVT. For $k = 0, 1, \ldots$, let

$$\zeta(k) = \frac{1}{\varepsilon}\left(y(\varepsilon(k+2)) - 2y(\varepsilon(k+1)) + y(\varepsilon k)\right).$$

We define a test statistic

$$\Lambda^{n_0,n} = \frac{1}{\varepsilon(n-n_0)} \sum_{k=n_0}^{n-1} |\zeta(k)|^2.$$

Let

$$\mu_i^{n_0,n} = \frac{1}{\varepsilon(n-n_0)} \int_{\varepsilon n_0}^{\varepsilon(n+1)} \rho^{n_0,n}(s)\mathrm{tr}(H(s,i)Q(s,i)H'(s,i))^2 ds + 2p,$$

where, for $n_0, n = 0, 1, \ldots,$

$$\rho^{n_0,n}(s) = \begin{cases} \phi^2(s)I_{[\varepsilon n_0, \varepsilon n]} + \phi^2(s-\varepsilon)I_{[\varepsilon(n_0+1),\varepsilon n]} & \text{if } (n-n_0) \text{ even} \\ \phi^2(s)I_{[\varepsilon n_0, \varepsilon(n+1)]} + \phi^2(s-\varepsilon)I_{[\varepsilon(n_0+1),\varepsilon(n+1)]} & \text{if } (n-n_0) \text{ odd,} \end{cases}$$

and $\phi(s)$ is a "sawtooth" function on $[0, T]$ such that for any $j = 0, 2, 4, \ldots$ even,

$$\phi(s) = \begin{cases} \dfrac{s - j\varepsilon}{\varepsilon} & \text{if } j\varepsilon \le s < (j+1)\varepsilon \\ \dfrac{(j+2)\varepsilon - s}{\varepsilon} & \text{if } (j+1)\varepsilon \le s < (j+2)\varepsilon. \end{cases}$$

(See Haussmann and Zhang [12] for interpretation of these functions.)

Given $\alpha(t) = i$, $\varepsilon n_0 \le t < \varepsilon n$, it can be shown as in [12] that for large $(n - n_0)$, $\Lambda^{n_0,n}/\mu_i^{n_0,n}$ is close to 1 by the law of large numbers. In order to distinguish the $\mu_i^{n_0,n}$'s, we impose a detectability condition as in Fleming and Pardoux [7].

(A5) There exists a constant $c > 0$ such that

$$\left| \operatorname{tr}\left(H(t,i)F(t,i)H'(t,i)\right)^2 - \operatorname{tr}\left(H(t,j)F(t,j)H'(t,j)\right)^2 \right| \ge c > 0$$

for $i \ne j$ and all $t \ge 0$.

In one-dimensional case, if $F(t,i) = 1$ and $H(t,i) = H(i)$, which is independent of t, then (A5) is equivalent to the condition $|H(i)| \ne |H(j)|$.

The QVT is given as follows: Let $\alpha^{n_0,n}$ denote a random variable such that $\alpha^{n_0,n} = i_0$ if

$$\left| \frac{\Lambda^{n_0,n}}{\mu_{i_0}^{n_0,n}} - 1 \right| = \min\left\{ \left| \frac{\Lambda^{n_0,n}}{\mu_j^{n_0,n}} - 1 \right|,\ j \in \mathcal{M} \right\}. \tag{3.5}$$

The next lemma gives the error probability of the QVT, which can be proved similarly as in [12, Lemma 3.1].

LEMMA 3.1. *Assume* (A1)–(A5). *Then for each* $j = 1, 2, \ldots,$ *there exist* $k_0 > 0,\ K > 0, \varepsilon_0 > 0$ *such that for* $0 < \varepsilon < \varepsilon_0,\ n - n_0 \ge l_0 := [k_0(\log \varepsilon)^2] + 1,$

$$P(\alpha^{n_0,n} \ne i_0 | \alpha(t) = i_0,\ t \in [\varepsilon n_0, \varepsilon n]) \le K\varepsilon^2. \tag{3.6}$$

It follows that

$$P(\{\alpha^{n_0,n} \ne i_0\} \cap \{\alpha(t) = i_0,\ t \in [\varepsilon n_0, \varepsilon n]\}) \le K\varepsilon^2. \tag{3.7}$$

Remark 3.2. As in Fleming and Pardoux [7], the error bounds in (3.6) and (3.7) can be improved to an order of ε^k for any given k by choosing ε_0 small enough and K sufficiently large. In this paper, we need only the estimate up to an order of ε^2.

In general, for $\varepsilon n \le t < \varepsilon(n+1)$, $n = 0, 1, \ldots$, we define

$$\tilde{\alpha}(t) = \begin{cases} 1 & \text{if } t \in [0, \varepsilon), \\ \alpha^{0,n} & \text{if } n \le l_0, \\ \alpha^{n-l_0,n} & \text{if } n \ge l_0. \end{cases} \tag{3.8}$$

We will show in Theorem 3.3 that $\tilde{\alpha}(\cdot)$ is indeed a good approximation to $\alpha(\cdot)$.

Next let us give another method for estimating the unknown process $\alpha(\cdot)$ using the outputs of several EKFs based on the principle of the well known least squares algorithm; see [12] for related discussions.

LRT. In the LRT, the detectability condition (A5) used for the QVT can be relaxed to the following condition:

(A6) $\text{tr}\Big(H(t,i)F(t,i)H'(t,i) - H(t,j)F(t,j)H'(t,j)\Big)^2 \geq c > 0$ for $i \neq j$.

Note that the condition

$$\text{tr}\Big(H(t,i)F(t,i)H'(t,i) - H(t,j)F(t,j)H'(t,j)\Big)^2 = 0$$

implies $H(t,i)F(t,i)H'(t,i) = H(t,j)F(t,j)H'(t,j)$. Clearly, (A5) implies (A6) because $F(t,i)$ is a symmetric matrix.

Given $0 < \sigma < 1$, let $l_1 = [1/\varepsilon^\sigma]$. For each $0 \leq t \leq T$, we consider the interval of a moving window $\mathcal{I}(t)$ defined as follows:

$$\mathcal{I}(t) = [\gamma_1(t), \gamma_2(t)) = \begin{cases} \left[0, \left[\frac{t}{\varepsilon}\right]\varepsilon\right) & \text{if } t \leq l_1, \\ \left[\left[\frac{t}{\varepsilon}\right]\varepsilon - l_1\varepsilon, \left[\frac{t}{\varepsilon}\right]\varepsilon\right) & \text{if } t > l_1. \end{cases}$$

On the interval $\mathcal{I}(t)$, we consider the output of the EKF under the condition that $\alpha(\cdot) = i$ on $\mathcal{I}(t)$, i.e., $\alpha(s) = i$ for $s \in \mathcal{I}(t)$. Then

$$\begin{aligned} d\tilde{x}^{(i)}(t) =& \ b(t,i,\tilde{x}^{(i)}(t))dt + \frac{1}{\varepsilon^2}R^{(i)}(t)H'(t,i)\left(dy(t) - h(t,i,\tilde{x}^{(i)}(t))dt\right), \\ \frac{dR^{(i)}(t)}{dt} =& \ \nabla_x b(t,i,\tilde{x}^{(i)}(t))R^{(i)}(t) + R^{(i)}(t)(\nabla_x b(t,i,\tilde{x}^{(i)}(t)))' \\ & + F(t,i)H'(t,i)H(t,i)F(t,i) - \frac{1}{\varepsilon^2}R^{(i)}(t)H'(t,i)H(t,i)R^{(i)}(t), \end{aligned}$$

with $\tilde{x}^i(\gamma_1(t)) = \tilde{x}(\gamma_1(t))$, $R^{(i)}(\gamma_1(t)) = \widetilde{R}(\gamma_1(t))$, and $\tilde{x}(\gamma_1(t))$ and $\widetilde{R}(\gamma_1(t))$ will be defined in what follows.

For each $i \in \mathcal{M}$, let $L^{(i)}(\mathcal{I}(t))$ denote a test statistics on $\mathcal{I}(t)$,

$$L^{(i)}(\mathcal{I}(t)) = \frac{1}{\varepsilon}\left(\int_{\gamma_1(t)}^{\gamma_2(t)} (H(t,i)\tilde{x}^{(i)}(t))'dy(t) - \frac{1}{2}\int_{\gamma_1(t)}^{\gamma_2(t)} |H(t,i)\tilde{x}^{(i)}(t)|^2 dt\right). \tag{3.9}$$

For each t, define

$$\tilde{\alpha}(t) = \begin{cases} 1 & \text{if } t \in [0,\varepsilon), \\ i & \text{if } t \geq \varepsilon \text{ and } L^{(i)}(\mathcal{I}(t)) = \max\{L^{(j)}(\mathcal{I}(t)) : \ j \in \mathcal{M}\}. \end{cases} \tag{3.10}$$

Using $\tilde{\alpha}(\cdot)$, we define the EKF based on the LRT as follows:

$$\begin{aligned} d\tilde{x}(t) =& \ b(t,\tilde{\alpha}(t),\tilde{x}(t))dt \\ & + \frac{1}{\varepsilon^2}\widetilde{R}(t)H'(t,\tilde{\alpha}(t))\left(dy(t) - h(t,\tilde{\alpha}(t),\tilde{x}(t))dt\right), \\ \frac{d\widetilde{R}(t)}{dt} =& \ \nabla_x b(t,\tilde{\alpha}(t),\tilde{x}(t))\widetilde{R}(t) + \widetilde{R}(t)(\nabla_x b(t,\tilde{\alpha}(t),\tilde{x}(t)))' \\ & + F(t,\tilde{\alpha}(t))H'(t,\tilde{\alpha}(t))H(t,\tilde{\alpha}(t))F(t,\tilde{\alpha}(t)) \\ & - \frac{1}{\varepsilon^2}\widetilde{R}(t)H'(t,\tilde{\alpha}(t))H(t,\tilde{\alpha}(t))\widetilde{R}(t), \end{aligned} \tag{3.11}$$

where $\tilde{x}(0) = Ex_0$ and $\widetilde{R}(0) = \text{Cov}(x_0)$.

Note that the processes $(\tilde{x}(\cdot), \widetilde{R}(\cdot))$, $(\tilde{x}^{(i)}(\cdot), \widetilde{R}^{(i)}(\cdot))$, and $\tilde{\alpha}(\cdot)$ are well defined. In fact, it is easy to see that the processes $\tilde{\alpha}(\cdot)$ and $(\tilde{x}^{(i)}(\cdot), \widetilde{R}^{(i)}(\cdot))$ are defined on $[0, (l_1+1)\varepsilon)$, so is $(\tilde{x}(\cdot), \widetilde{R}(\cdot))$ because the lower limit of $\mathcal{I}(t)$ equals $\gamma_1(t) = 0$ and $(\tilde{x}(0), \widetilde{R}(0)) = (Ex_0, \mathrm{Cov}(x_0))$, which is given. The EKF gives the value of $(\tilde{x}(\varepsilon), \widetilde{R}(\varepsilon))$. Then $(\tilde{x}^{(i)}(\varepsilon), \widetilde{R}^{(i)}(\varepsilon))$ is used for computing $(\tilde{x}^{(i)}(\cdot), \widetilde{R}^{(i)}(\cdot))$ on the next interval $\mathcal{I}(t)$ for $t \in [(l_1+1)\varepsilon, (l_1+2)\varepsilon)$ and that leads to $\tilde{\alpha}(\cdot)$ on this interval. Then using $\tilde{\alpha}(\cdot)$ on $[(l_1+1)\varepsilon, (l_1+2)\varepsilon)$, we can compute $(\tilde{x}(\cdot), \widetilde{R}(\cdot))$ on this interval. This procedure can be repeated on intervals $[j\varepsilon, (j+1)\varepsilon)$ for all j.

LEMMA 3.2. *Assume* (A1)–(A4) *and* (A6). *Then there exist positive constants* ε_0, K *such that for* $0 < \varepsilon < \varepsilon_0$

$$P\Big(\tilde{\alpha}(t) \neq i_0 | \alpha(\cdot) = i_0 \text{ on } \mathcal{I}(t)\Big) \leq K\varepsilon^2.$$

Proof. The proof can be given similarly as in [12, Lemma 3.3]. □

Remark 3.3. In this paper we use a moving window (fixed sample size) to estimate the value of $\alpha(\cdot)$. The length of the window is εl_0 for the QVT and εl_1 for the LRT. Typically the QVT requires less time than the LRT with a given error probability; see [6] and [13]. Moreover, a sequential test can be used to estimate the value of $\alpha(\cdot)$, which usually requires less time when compared with the fixed sample size test.

Approximate filters. Next we study asymptotic filters based on the QVT and the LRT and estimate the corresponding error bounds.

For any given stochastic process $\xi(t)$, $t \geq 0$, we define the norm of $\xi(\cdot)$ as follows:

$$|\xi(\cdot)|_T = E\int_0^T |\xi(t)|dt.$$

The next theorem is concerned with the asymptotic property of $\tilde{\alpha}(\cdot)$ and the associate error bound in terms of the $|\cdot|_T$ norm.

THEOREM 3.3. *Let* $\tilde{\alpha}(\cdot)$ *be a filter based on the QVT, defined in* (3.8), *and assume* (A1)–(A5) (*or based on the LRT, defined in* (3.10), *and assume* (A1)–(A4) *and* (A6)). *Then for each* $0 < \delta < 1$, *there exist positive constants* ε_0 *and* K *such that for* $0 < \varepsilon < \varepsilon_0$,

$$|\tilde{\alpha}(\cdot) - \alpha(\cdot)|_T \leq K\varepsilon^{1-\delta}.$$

Proof. Let $0 < \tau_1 < \tau_2 < \cdots$ denote the random jump times of $\alpha(\cdot)$. Then, for $j = 1, 2, \ldots$,

$$P\Big(\tau_{j+1} - \tau_j \leq k_0\varepsilon(\log\varepsilon)^2\Big) = O\left(\varepsilon(\log\varepsilon)^2\right). \tag{3.12}$$

For notational convenience, let $\xi(t) = |\tilde{\alpha}(t) - \alpha(t)|$. Then, for $0 < \sigma < \delta$,

$$E\int_0^T |\tilde{\alpha}(t) - \alpha(t)|dt = E\int_0^T \xi(t)dt = E\int_0^T \xi(t)dtI_{\{\alpha(\cdot)\in\Theta_\sigma^\varepsilon\}} + E\int_0^T \xi(t)dtI_{\{\alpha(\cdot)\notin\Theta_\sigma^\varepsilon\}}.$$

Recall the inequalities $P(\alpha(\cdot) \notin \Theta_\sigma^\varepsilon) \leq K\varepsilon^2$ given in (2.3) and $|\xi(t)| \leq m$. It follows that

$$E\int_0^T \xi(t)dtI_{\{\alpha(\cdot)\notin\Theta_\sigma^\varepsilon\}} \leq mTEI_{\{\alpha(\cdot)\notin\Theta_\sigma^\varepsilon\}} = mTP(\alpha(\cdot) \notin \Theta_\sigma^\varepsilon) \leq K\varepsilon^2.$$

Moreover, note that the process $\alpha(\cdot)$ jumps at most $[1/\varepsilon^\sigma]$ times if $\alpha(\cdot) \in \Theta_\sigma^\varepsilon$. Thus,

$$E\int_0^T \xi(t)dtI_{\{\alpha(\cdot)\in\Theta_\sigma^\varepsilon\}} \le \sum_{j=0}^{[1/\varepsilon^\sigma]} E\int_{\tau_j\wedge T}^{\tau_{j+1}\wedge T} \xi(t)dt.$$

If we show that, for each $j = 0, 1, \ldots, [1/\varepsilon^\sigma]$,

$$E\int_{\tau_j\wedge T}^{\tau_{j+1}\wedge T} \xi(t)dt = O(\varepsilon(\log\varepsilon)^2), \tag{3.13}$$

then

$$E\int_0^T \xi(t)dtI_{\{\alpha(\cdot)\in\Theta_\sigma^\varepsilon\}} \le \sum_{j=0}^{[1/\varepsilon^\sigma]} O(\varepsilon(\log\varepsilon)^2) = O(\varepsilon^{1-\sigma}|\log\varepsilon|^2) = O(\varepsilon^{1-\delta}),$$

because for $0 < \sigma < \delta$, $\varepsilon^{\delta-\sigma}(\log\varepsilon)^2 \to 0$ as $\varepsilon \to 0$. Hence, it suffices to show (3.13).

We start from $j = 0$. Note that, for $l_0 = [k_0(\log\varepsilon)^2] + 1$ as in Lemma 3.1,

$$E\int_0^{\tau_1\wedge T} \xi(t)dt = E\int_0^T \xi(t)I_{[0,\tau_1)}dt = E\left(\int_0^{l_0\varepsilon} + \int_{l_0\varepsilon}^T\right)\xi(t)I_{[0,\tau_1)}dt.$$

Moreover, the boundedness of $\xi(t)$ implies that

$$E\int_0^{l_0\varepsilon} \xi(t)I_{[0,\tau_1)}dt = O(l_0\varepsilon) = O(\varepsilon(\log\varepsilon)^2).$$

Write

$$E\int_{l_0\varepsilon}^T \xi(t)I_{[0,\tau_1)}dt = E\sum_{j=l_0}^{T_\varepsilon}\int_{j\varepsilon}^{(j+1)\varepsilon} \xi(t)I_{[0,\tau_1)}dt,$$

where $T_\varepsilon = [T/\varepsilon]$. For all $j \ge l_0$, we have

$$\begin{aligned} E\int_{j\varepsilon}^{(j+1)\varepsilon} \xi(t)I_{[0,\tau_1)}dt = {} & E\int_{j\varepsilon}^{(j+1)\varepsilon} \xi(t)I_{[0,\tau_1)}dtI_{\{\tau_1<j\varepsilon\}} \\ & + E\int_{j\varepsilon}^{(j+1)\varepsilon} \xi(t)I_{[0,\tau_1)}dtI_{\{j\varepsilon\le\tau_1<(j+1)\varepsilon\}} \\ & + E\int_{j\varepsilon}^{(j+1)\varepsilon} \xi(t)I_{[0,\tau_1)}dtI_{\{\tau_1\ge(j+1)\varepsilon\}}. \end{aligned} \tag{3.14}$$

The first term on the right side equals 0 because $[j\varepsilon, (j+1)\varepsilon] \cap [0, \tau_1) = \emptyset$. The second term in (3.14) is of order ε^2 due to the fact that

$$\begin{aligned} E\int_{j\varepsilon}^{(j+1)\varepsilon} & \xi(t)I_{[0,\tau_1)}dtI_{\{j\varepsilon\le\tau_1<(j+1)\varepsilon\}} \\ & \le m\int_{j\varepsilon}^{(j+1)\varepsilon} EI_{\{j\varepsilon\le\tau_1<(j+1)\varepsilon\}} \\ & = m\int_{j\varepsilon}^{(j+1)\varepsilon} P(j\varepsilon \le \tau_1 < (j+1)\varepsilon) = O(\varepsilon^2). \end{aligned}$$

To estimate the third term on the right side of (3.14), note that

(1) $[j\varepsilon,(j+1)\varepsilon]\subset[0,\tau_1]$, which implies $\alpha(t)=i_0$ (no jump) on $[0,\tau_1]$;
(2) For $j\le t<(j+1)\varepsilon$, $\tilde\alpha(t)$ is defined based on the information given on the interval $[(j-l_0+1)\varepsilon,j\varepsilon]$.

These imply the following inequality:

$$E\int_{j\varepsilon}^{(j+1)\varepsilon}\xi(t)I_{[0,\tau_1)}dtI_{\{\tau_1\ge(j+1)\varepsilon\}}\le m\sum_{i_0=1}^{m}\int_{j\varepsilon}^{(j+1)\varepsilon}P(\{\alpha(t)=i_0,\ t\in[(j-l_0+1)\varepsilon,j\varepsilon]\}\cap\tilde\alpha(t)\ne i_0)dt.$$

It follows that

$$E\int_0^{\tau_1\wedge T}\xi(t)dt=O(\varepsilon(\log\varepsilon)^2).$$

Next we estimate $E\int_{\tau_1\wedge T}^{\tau_2\wedge T}\xi(t)dt$, which can be written as $E\int_0^T\xi(t)I_{[\tau_1,\tau_2)}dt$. Note that $P(\tau_2-\tau_1\le l_0\varepsilon)=O(\varepsilon(\log\varepsilon)^2)$ and

$$E\int_0^T I_{[\tau_1,l_0\varepsilon+\tau_1]}dt=O(l_0\varepsilon)=O(\varepsilon(\log\varepsilon)^2).$$

It follows that

$$E\int_0^T\xi(t)I_{[\tau_1,\tau_2)}dt=E\int_0^T\xi(t)I_{[\tau_1,\tau_2)}I_{\{\tau_2-\tau_1\ge l_0\varepsilon\}}dt+O(\varepsilon(\log\varepsilon)^2).$$

Note also that if $\tau_1>\varepsilon T_\varepsilon$, then $E\int_0^T\xi(t)I_{[\tau_1,\tau_2)}dt=O(\varepsilon)$. Write

$$\begin{aligned}&E\int_0^T\xi(t)I_{[\tau_1+l_0\varepsilon,\tau_2)}I_{\{\tau_2-\tau_1\ge l_0\varepsilon\}}dt\\&=\sum_{i=0}^{T_\varepsilon}E\int_0^T\xi(t)I_{[\tau_1+l_0\varepsilon,\tau_2)}I_{\{\tau_2-\tau_1\ge l_0\varepsilon\}}dtI_{\{i\varepsilon\le\tau_1<(i+1)\varepsilon\}}\\&=\sum_{i=0}^{T_\varepsilon}E\int_{(i+l_0)\varepsilon}^T\xi(t)I_{[\tau_1+l_0\varepsilon,\tau_2)}I_{\{\tau_2-\tau_1\ge l_0\varepsilon\}}dtI_{\{i\varepsilon\le\tau_1<(i+1)\varepsilon\}},\end{aligned}$$

because $I_{[\tau_1+l_0\varepsilon,\tau_2)}=0$ for $t\le(i+l_0)\varepsilon$ given $\{i\varepsilon\le\tau_1<(i+1)\varepsilon\}$. Moreover,

$$\begin{aligned}&E\int_{(i+l_0)\varepsilon}^T\xi(t)I_{[\tau_1+l_0\varepsilon,\tau_2)}I_{\{\tau_2-\tau_1\ge l_0\varepsilon\}}dtI_{\{i\varepsilon\le\tau_1<(i+1)\varepsilon\}}\\&=\sum_{j=i+l_0}^{T_\varepsilon}E\int_{j\varepsilon}^{(j+1)\varepsilon}\xi(t)I_{[\tau_1+l_0\varepsilon,\tau_2)}I_{\{\tau_2-\tau_1\ge l_0\varepsilon\}}dtI_{\{i\varepsilon\le\tau_1<(i+1)\varepsilon\}}\\&=O(\varepsilon^2)+\sum_{j=i+l_0+1}^{T_\varepsilon}E\int_{j\varepsilon}^{(j+1)\varepsilon}\xi(t)I_{[\tau_1+l_0\varepsilon,\tau_2)}I_{\{\tau_2-\tau_1\ge l_0\varepsilon\}}dtI_{\{i\varepsilon\le\tau_1<(i+1)\varepsilon\}}.\end{aligned}$$

The last equality is due to the fact that $P(i\varepsilon\le\tau_1<(i+1)\varepsilon)=O(\varepsilon)$. Let

$$\zeta=E\int_{j\varepsilon}^{(j+1)\varepsilon}\xi(t)I_{[\tau_1+l_0\varepsilon,\tau_2)}I_{\{\tau_2-\tau_1\ge l_0\varepsilon\}}dtI_{\{i\varepsilon\le\tau_1<(i+1)\varepsilon\}}.$$

Then for $j\ge i+l_0+1$, we have

(1) If $\tau_2 < j\varepsilon$, then $\zeta = 0$;
(2) If $j\varepsilon \le \tau_2 < (j+1)\varepsilon$, then $\zeta = O(\varepsilon^2)$ because $P(j\varepsilon \le \tau_2 < (j+1)\varepsilon) = O(\varepsilon)$;
(3) If $\tau_2 \ge (j+1)\varepsilon$, then

$$\zeta \le m \sum_{i_0=1}^{m} \int_{j\varepsilon}^{(j+1)\varepsilon} P(\{\alpha(t) = i_0,\ t \in [(j-l_0+1)\varepsilon, j\varepsilon]\} \cap \{\tilde{\alpha}(t) \ne i_0\})dt = O(\varepsilon^3).$$

Continue this way, we can show (3.13), for all $j = 1, 2, \ldots, [1/\varepsilon^\sigma]$, and thus complete the proof. □

Similar to the proof of Theorem 3.3, one can show the following result.

COROLLARY 3.4. *For each* $j = 0, 1, \ldots,$

$$P(\tilde{\alpha}(t) \ne \alpha(t) : \quad \textit{for some } t \in [\tau_j + l_0\varepsilon, \tau_{j+1})) = O(\varepsilon).$$

Using $\tilde{\alpha}(\cdot)$, we define an approximate filter $\tilde{x}(\cdot)$ satisfying the following equations:

$$\begin{aligned} d\tilde{x}(t) =\ & b(t, \tilde{\alpha}(t), \tilde{x}(t))dt \\ & + \frac{1}{\varepsilon^2}\widetilde{R}(t)H'(t, \tilde{\alpha}(t))\left(dy(t) - h(t, \tilde{\alpha}(t), \tilde{x}(t))dt\right), \\ \frac{d\widetilde{R}(t)}{dt} =\ & \nabla_x b(t, \tilde{\alpha}(t), \tilde{x}(t))\widetilde{R}(t) + \widetilde{R}(t)(\nabla_x b(t, \tilde{\alpha}(t), \tilde{x}(t)))' \\ & + F(t, \tilde{\alpha}(t))H'(t, \tilde{\alpha}(t))H(t, \tilde{\alpha}(t))F(t, \tilde{\alpha}(t)) \\ & - \frac{1}{\varepsilon^2}\widetilde{R}(t)H'(t, \tilde{\alpha}(t))H(t, \tilde{\alpha}(t))\widetilde{R}(t), \end{aligned} \tag{3.15}$$

with $\tilde{x}(0) = Ex_0$ and $\widetilde{R}(0) = \mathrm{Cov}(x_0)$.

In order to verify the asymptotic optimality of $(\tilde{\alpha}(\cdot), \tilde{x}(\cdot)))$ and estimate the corresponding error bound, we need to consider an intermediate "filter" $(\check{x}(t), \check{R}(t))$, $t \ge 0$, assuming $\alpha(\cdot)$ is given, defined as follows:

$$\begin{aligned} d\check{x}(t) =\ & b(t, \alpha(t), \check{x}(t))dt \\ & + \frac{1}{\varepsilon^2}\check{R}(t)H'(t, \alpha(t))\left(dy(t) - h(t, \alpha(t), \check{x}(t))dt\right), \\ \frac{d\check{R}(t)}{dt} =\ & \nabla_x b(t, \alpha(t), \check{x}(t))\check{R}(t) + \check{R}(t)(\nabla_x b(t, \alpha(t), \check{x}(t)))' \\ & + F(t, \alpha(t))H'(t, \alpha(t))H(t, \alpha(t))F(t, \alpha(t)) \\ & - \frac{1}{\varepsilon^2}\check{R}(t)H'(t, \alpha(t))H(t, \alpha(t))\check{R}(t), \end{aligned} \tag{3.16}$$

with $\check{x}(0) = Ex_0$ and $\check{R}(0) = \mathrm{Cov}(x_0)$. Clearly, $(\check{x}(\cdot), \check{R}(\cdot))$ can be obtained by replacing $\alpha(\cdot)$ with $\tilde{\alpha}(\cdot)$ in (3.15).

THEOREM 3.5. *Assume the conditions in Theorem* 3.3. *Then for a given* $0 < \delta < 1$, *there exists* $\varepsilon_0 > 0$ *such that for* $0 < \varepsilon < \varepsilon_0$,

$$|\hat{\alpha}(\cdot) - \tilde{\alpha}(\cdot)|_T + |\hat{x}(\cdot) - \tilde{x}(\cdot)|_T = O(\varepsilon^{1-\delta}),$$

where $(\hat{\alpha}(t), \hat{x}(t))$ *denotes the conditional mean of* $(\alpha(t), x(t))$ *given* $\mathcal{Y}_t$.

Proof. First of all, note that $\tilde{\alpha}(t)$ is $\mathcal{Y}_t$ measurable. It follows from the Jensen's inequality that

$$|\hat{\alpha}(t) - \tilde{\alpha}(t)| = |E[\alpha(t) - \tilde{\alpha}(t)|\mathcal{Y}_t]| \le E[\ |\alpha(t) - \tilde{\alpha}(t)|\ \ |\mathcal{Y}_t].$$

Taking expectation on both sides of the above inequality yields

$$E|\hat{\alpha}(t) - \tilde{\alpha}(t)| \le E|\alpha(t) - \tilde{\alpha}(t)|,$$

which implies, in view of Theorem 3.3,

$$|\hat{\alpha}(\cdot) - \tilde{\alpha}(\cdot)|_T \le |\alpha(\cdot) - \tilde{\alpha}(\cdot)|_T = O(\varepsilon^{1-\delta}).$$

We next show that $|\hat{x}(\cdot) - \tilde{x}(\cdot)|_T = O(\varepsilon^{1-\delta})$. The basic idea of the proof is to show that $\hat{x}(\cdot)$ is close to $\check{x}(\cdot)$, which can be further approximated by $\tilde{x}(\cdot)$. Let $0 < \sigma < \delta$ and γ_0 be as given in Lemmas A.3 and A.4 in the Appendix. We divide the rest of the proof into several steps. We only consider the case when $\tilde{\alpha}(\cdot)$ is obtained via the QVT since the proof for the LRT case is similar.

Step 1. We show that

$$\left|(\hat{x}(\cdot) - \tilde{x}(\cdot))I_{\{\alpha(\cdot)\notin\Theta^{\varepsilon}_{\sigma,\gamma_0}\}}\right|_T = O(\varepsilon^{1-\sigma}).$$

In fact, note that for each $t \ge 0$,

$$E\left|(\hat{x}(t) - \tilde{x}(t))I_{\{\alpha(\cdot)\notin\Theta^{\varepsilon}_{\sigma,\gamma_0}\}}\right| = \int E(|\hat{x}(t) - \check{x}(t)\,|\,|\,\alpha(\cdot) = \theta)I_{\{\theta\notin\Theta^{\varepsilon}_{\sigma,\gamma_0}\}}P(\alpha(\cdot) \in d\theta).$$

It is easy to see that the conditional expectation $E[\,|\hat{x}(t) - \tilde{x}(t)|\,|\alpha(\cdot) = \theta]$ is uniformly bounded with respect to $\theta \in \Theta$ and $t \ge 0$. Thus, in view of (3.4),

$$E\left|(\hat{x}(t) - \tilde{x}(t))I_{\{\alpha(\cdot)\notin\Theta^{\varepsilon}_{\sigma,\gamma_0}\}}\right| \le KP(\alpha(\cdot) \notin \Theta^{\varepsilon}_{\sigma,\gamma_0}) = O(\varepsilon^{1-\sigma}).$$

Step 2. We next show that

$$\left|(\hat{x}(\cdot) - \check{x}(\cdot))I_{\{\alpha(\cdot)\in\Theta^{\varepsilon}_{\sigma,\gamma_0}\}}\right|_T = O(\varepsilon^2). \tag{3.17}$$

Note that

$$\left|(\hat{x}(\cdot) - \check{x}(\cdot))I_{\{\alpha(\cdot)\in\Theta^{\varepsilon}_{\sigma,\gamma_0}\}}\right|_T = \int E(|\hat{x}(t) - \check{x}(t)||\alpha(\cdot) = \theta)I_{\{\theta\in\Theta^{\varepsilon}_{\sigma,\gamma_0}\}}P(\alpha(\cdot) \in d\theta).$$

Moreover, under the condition $\alpha(\cdot) = \theta$, we have

$$\hat{x}(\cdot) = \hat{x}^{\theta}(\cdot) \text{ and } \check{x}(\cdot) = \tilde{x}^{\theta}(\cdot).$$

Thus using Lemma A.3, we obtain (3.17).

Step 3. We show that $dy(t) = H(t,\alpha(t))\check{x}(t)dt + \varepsilon d\hat{v}(t) + \eta(t)dt$, where $\hat{v}(\cdot)$ is an innovation process and $\eta(t) = E(H(t,\alpha(t))x(t)|\mathcal{Y}_t) - H(t,\alpha(t))\check{x}(t)$. We also show that

$$\int_0^T E\,|\eta(t)|\,I_{\{\alpha(\cdot)\in\Theta^{\varepsilon}_{\sigma,\gamma_0}\}}dt = O(\varepsilon^2). \tag{3.18}$$

To prove these, it suffices to verify (3.18). In fact, we have

$$\begin{aligned}
&\textstyle\int_0^T E\,|E(H(t,\alpha(t))x(t)|\mathcal{Y}_t) - H(t,\alpha(t)\check{x}(t)|\,I_{\{\alpha(\cdot)\in\Theta^{\varepsilon}_{\sigma,\gamma_0}\}}dt\\
&\quad= \int\!\!\int_0^T E[|E(H(t,\alpha(t))x(t)|\mathcal{Y}_t) - H(t,\alpha(t))\check{x}(t)|\,|\alpha(\cdot) = \theta]I_{\{\theta\in\Theta^{\varepsilon}_{\sigma,\gamma_0}\}}dt\\
&\qquad\qquad \times P(\alpha(\cdot) \in d\theta)\\
&\quad= \int\!\!\int_0^T E[\left|H(t,\theta(t))\hat{x}^{\theta}(t) - H(t,\theta(t))\tilde{x}^{\theta}(t)\right|\,|\alpha(\cdot) = \theta]I_{\{\theta\in\Theta^{\varepsilon}_{\sigma,\gamma_0}\}}dtP(\alpha(\cdot) \in d\theta)\\
&\quad\le K\int\!\!\int_0^T E^{\theta}\left|\hat{x}^{\theta}(t) - \tilde{x}^{\theta}(t)\right|dtI_{\{\theta\in\Theta^{\varepsilon}_{\sigma,\gamma_0}\}}P(\alpha(\cdot) \in d\theta)\\
&\quad= O(\varepsilon^2).
\end{aligned}$$

Step 4. Given $\theta = \theta(\cdot) \in \Theta$ and a, b being the multiples of ε such that $0 \le a < b$, let

$$F^\theta_{a,b} = \{\alpha(\cdot) = \theta(\cdot) \text{ on } [a,b)\},\ (:= \{\alpha(t) = \theta(t) \text{ for all } t \in [a,b)\}),$$

and let $E^\theta_{a,b}$ denote the conditional expectation given $F^\theta_{a,b}$. We show that

$$E^\theta_{a,b}|\tilde{x}(t) - \check{x}(t)| = O\left(\varepsilon^{1-\sigma} + \exp\left(-\frac{\kappa(t-a-l_0\varepsilon)}{\varepsilon}\right)\right)$$

uniformly for $t \in [a + l_0\varepsilon, b]$; here $l_0 = [k_0(\log\varepsilon)^2] + 1$ as in Lemma 3.1. (l_0 needs to be replaced by l_1 as in Lemma 3.2 when the LRT is used to define $\tilde{\alpha}(\cdot)$).

First of all, by considering the differentials of $H(t,i_0)(\tilde{x}(t)-\check{x}(t))$ and $|H(t,i_0)(\tilde{x}(t)-\check{x}(t))|^2$, we obtain

$$\begin{aligned}
d(H(t,i_0)(\tilde{x}(t)-\check{x}(t))) &= \tfrac{\partial H(t,i_0)}{\partial t}(\tilde{x}(t)-\check{x}(t))dt \\
&\quad + H(t,i_0)\Big\{(b(t,\tilde{\alpha}(t),\tilde{x}(t)) - b(t,i_0,\check{x}(t)))dt \\
&\qquad -\frac{1}{\varepsilon^2}\widetilde{R}(t)H'(t,\tilde{\alpha}(t))(H(t,\tilde{\alpha}(t))\tilde{x}(t) - H(t,i_0)\check{x}(t))dt \\
&\qquad +\frac{1}{\varepsilon^2}\left(\widetilde{R}(t)H'(t,\tilde{\alpha}(t)) - \check{R}(t)H'(t,i_0)\right)\eta(t)dt \\
&\qquad +\frac{1}{\varepsilon}\left(\widetilde{R}(t)H'(t,\tilde{\alpha}(t)) - \check{R}(t)H'(t,i_0)\right)d\hat{v}(t)\Big\}
\end{aligned}$$

and

$$\begin{aligned}
d|H(t,i_0)(\tilde{x}(t)-\check{x}(t))|^2 &= 2(\tilde{x}(t)-\check{x}(t))'H'(t,i_0)d(H(t,i_0)(\tilde{x}(t)-\check{x}(t))) \\
&\quad +\frac{1}{\varepsilon^2}\mathrm{tr}\Big(\Big(\widetilde{R}(t)H'(t,\tilde{\alpha}(t)) - \check{R}(t)H'(t,i_0)\Big)\Big(\widetilde{R}(t)H'(t,\tilde{\alpha}(t)) - \check{R}(t)H'(t,i_0)\Big)'\Big) \\
&=: A(t)dt + B(t)d\hat{v}(t),
\end{aligned}$$

where

$$\begin{aligned}
A(t) &= 2(\tilde{x}(t)-\check{x}(t))'H'(t,i_0)\frac{\partial H(t,i_0)}{\partial t}(\tilde{x}(t)-\check{x}(t)) \\
&\quad +2(\tilde{x}(t)-\check{x}(t))'H'(t,i_0)H(t,i_0)\Big\{(b(t,\tilde{\alpha}(t),\tilde{x}(t)) - b(t,i_0,\check{x}(t))) \\
&\quad -\frac{1}{\varepsilon^2}\widetilde{R}(t)H'(t,\tilde{\alpha}(t))(H(t,\tilde{\alpha}(t))\tilde{x}(t) - H(t,i_0)\check{x}(t)) \\
&\quad +\frac{1}{\varepsilon^2}\left(\widetilde{R}(t)H'(t,\tilde{\alpha}(t)) - \check{R}(t)H'(t,i_0)\right)\eta(t)\Big\} \\
&\quad +\frac{1}{\varepsilon^2}\mathrm{tr}\Big(\Big(\widetilde{R}(t)H'(t,\tilde{\alpha}(t)) - \check{R}(t)H'(t,i_0)\Big)\Big(\widetilde{R}(t)H'(t,\tilde{\alpha}(t)) - \check{R}(t)H'(t,i_0)\Big)'\Big)
\end{aligned}$$

and

$$B(t) = \frac{2}{\varepsilon}(\tilde{x}(t)-\check{x}(t))'H'(t,i_0)\left(\widetilde{R}(t)H'(t,\tilde{\alpha}(t)) - \check{R}(t)H'(t,i_0)\right).$$

Let $\phi(t) = |H(t,i_0)(\tilde{x}(t)-\check{x}(t))|^2$. Then

$$\phi(t) = \phi(a+l_0\varepsilon) + \int_{a+l_0\varepsilon}^t A(s)ds + \int_{a+l_0\varepsilon}^t B(s)d\hat{v}(s).$$

Write

$$A(t) = A(t)I_{\{\tilde{\alpha}(\cdot)=i_0 \text{ on } [a+l_0\varepsilon,b]\}} + A(t)I_{\{\tilde{\alpha}(\cdot)=i_0 \text{ on } [a+l_0\varepsilon,b]\}^c}.$$

As in Lemma A.5, we can show that $(E^\theta_{a,b}|A(t)|^n)^{1/n} = O(1/\varepsilon)$, for $n = 1, 2, \ldots$. In view of Lemma 3.1, it follows that

$$\begin{aligned} &E^\theta_{a,b}\left|A(t)I_{\{\tilde{\alpha}(\cdot)=i_0 \text{ on } [a+l_0\varepsilon,b]\}^c}\right| \\ &\quad\le (E^\theta_{a,b}|A(t)|^n)^{\frac{1}{n}}(E^\theta_{a,b}I_{\{\tilde{\alpha}(\cdot)=i_0 \text{ on } [a+l_0\varepsilon,b]\}^c})^{\frac{n-1}{n}} \\ &\quad\le \frac{K}{\varepsilon}(P(\tilde{\alpha}(t)\ne i_0 \text{ for some } t\in[a+l_0\varepsilon,b]|F^\theta_{a,b}))^{\frac{n-1}{n}} \\ &\quad\le K\varepsilon^{(2(n-1)/n)-1} \le K\varepsilon^{1-\sigma} \text{ for } n \text{ large enough.} \end{aligned} \tag{3.19}$$

Given $\{\tilde{\alpha}(\cdot) = i_0 \text{ on } [a + l_0\varepsilon, b]\}$, we have

$$\begin{aligned} A(t) &= 2(\tilde{x}(t)-\check{x}(t))'H'(t,i_0)\frac{\partial H(t,i_0)}{\partial t}(\tilde{x}(t)-\check{x}(t)) \\ &\quad +2(\tilde{x}(t)-\check{x}(t))'H'(t,i_0)H(t,i_0)\Big\{(b(t,i_0,\tilde{x}(t))-b(t,i_0,\check{x}(t))) \\ &\quad -\frac{1}{\varepsilon^2}\widetilde{R}(t)H'(t,i_0)(H(t,i_0)(\tilde{x}(t)-\check{x}(t)) \\ &\quad +\frac{1}{\varepsilon^2}\left(\widetilde{R}(t)-\check{R}(t))H'(t,i_0)\right)\eta(t)\Big\} \\ &\quad +\frac{1}{\varepsilon^2}\mathrm{tr}\left((\widetilde{R}(t)-\check{R}(t))H'(t,i_0)H(t,i_0)(\widetilde{R}(t)-\check{R}(t))\right). \end{aligned}$$

Recall that $H(t, i_0)$ is invertible and both $\widetilde{R}(t)/\varepsilon$ and $\check{R}(t)/\varepsilon$ are uniformly bounded. Moreover, in view of Lemma A.4, we have

$$\frac{\widetilde{R}(t)-\check{R}(t)}{\varepsilon} = O\left(\varepsilon + \exp\left(-\frac{\kappa(t-a-l_0\varepsilon)}{\varepsilon}\right)\right).$$

Given $\{\tilde{\alpha}(\cdot) = i_0 \text{ on } [a + l_0\varepsilon, b]\}$, it follows that

$$A(t) \le K_0\phi(t) - \frac{\kappa_0}{\varepsilon}\phi(t) + K\left(\varepsilon + \frac{|\eta(t)|^2}{\varepsilon^2} + \exp\left(-\frac{\kappa(t-a-l_0\varepsilon)}{\varepsilon}\right)\right).$$

Hence,

$$\begin{aligned} E^\theta_{a,b}A(t) &\le \left(K_0 - \frac{\kappa}{\varepsilon}\right)E^\theta_{a,b}\phi(t) \\ &\quad +K\left(\varepsilon^{1-\sigma} + \frac{E^\theta_{a,b}|\eta(t)|^2}{\varepsilon^2} + \exp\left(-\frac{\kappa(t-a-l_0\varepsilon)}{\varepsilon}\right)\right). \end{aligned} \tag{3.20}$$

Next, we claim that

$$E^\theta_{a,b}\int_{a+l_0\varepsilon}^t B(s)d\hat{v}(s) = 0.$$

In fact, if we let $\mathcal{F}^\alpha_{a,b} = \sigma\{\alpha(r) : a \le r < b\}$ and $\zeta(t) = \int_{a+l_0\varepsilon}^t B(s)d\hat{v}(s)$, then for $0 \le t < b$, $E[\zeta(t)|\mathcal{F}^\alpha_{0,b}] = 0$ because $\hat{v}(\cdot)$ is a Brownian motion given $\alpha(\cdot)$. Note that $E[\zeta(t)|\mathcal{F}^\alpha_{a,b}] = E[E[\zeta(t)|\mathcal{F}^\alpha_{0,b}]|\mathcal{F}^\alpha_{a,b}]$. It follows that $E[\zeta(t)|\mathcal{F}^\alpha_{a,b}] = 0$, a.s. Thus,

$$E^\theta_{a,b}\int_{a+l_0\varepsilon}^t B(s)d\hat{v}(s) = 0, \text{ almost everywhere with respect to } \widehat{P}(d\theta) = P(\alpha(\cdot)\in d\theta).$$

By virtue of this claim, we have

$$E^\theta_{a,b}\phi(t) = E^\theta_{a,b}\phi(a+l_0\varepsilon) + E^\theta_{a,b}\int_{a+l_0\varepsilon}^t A(s)ds.$$

Thus, in view of (3.20),

$$\begin{aligned}\frac{dE^\theta_{a,b}\phi(t)}{dt} = E^\theta_{a,b}A(t) &\le \left(K_0 - \frac{\kappa}{\varepsilon}\right)E^\theta_{a,b}\phi(t)\\ &+K\left(\varepsilon^{1-\sigma} + \frac{E^\theta_{a,b}|\eta(t)|^2}{\varepsilon^2} + \exp\left(-\frac{\kappa(t-a-l_0\varepsilon)}{\varepsilon}\right)\right).\end{aligned}$$

Using Gronwall's inequality and the uniform boundedness of $E^\theta_{a,b}\phi(a+l_0\varepsilon)$, we obtain, by integration by parts,

$$E^\theta_{a,b}\phi(t) \le K\left(\varepsilon^{2-\sigma} + \exp\left(-\frac{\kappa(t-a-l_0\varepsilon)}{\varepsilon}\right)\right)$$

for some $\kappa > 0$. Therefore,

$$(E^\theta_{a,b}\phi(t))^{\frac{1}{2}} \le K\left(\varepsilon^{1-\sigma/2} + \exp\left(-\frac{\kappa(t-a-l_0\varepsilon)}{\varepsilon}\right)\right).$$

Step 5. We show that

$$\int_0^T E|\tilde{x}(t) - \check{x}(t)|I_{\{\alpha(\cdot)\in\Theta^\varepsilon_{\sigma,\gamma_0}\}}dt = O(\varepsilon^{1-\delta}). \tag{3.21}$$

Let $\xi(t) = |\tilde{x}(t) - \check{x}(t)|I_{\{\alpha(\cdot)\in\Theta^\varepsilon_{\sigma,\gamma_0}\}}$. Note that

$$E\int_0^T \xi(t)dt \le \sum_{j=0}^{[1/\varepsilon^\sigma]} E\int_0^T \xi(t)I_{[\tau_j,\tau_{j+1})}.$$

It suffices to show

$$E\int_0^T \xi(t)I_{[\tau_j,\tau_{j+1})} \le O(\varepsilon^{1-\sigma})$$

because

$$E\int_0^T \xi(t)dt \le \sum_{j=0}^{[1/\varepsilon^\sigma]} K\varepsilon^{1-\sigma} \le O(\varepsilon^{1-2\sigma}) = O(\varepsilon^{1-\delta})$$

when $2\sigma < \delta$.

For $j = 0$,

$$E\int_0^T \xi(t)I_{[0,\tau_1)}dt = \int_0^\infty E\left[\int_0^{s\wedge T}\xi(t)dt\Big|\tau_1 = s\right]p_1(s)ds,$$

where $p_1(s)$ is the density function of τ_1.

It is easy to see that

$$E\left[\int_0^{l_0\varepsilon}\xi(t)dt\Big|\tau_1=s\right]=O(l_0\varepsilon)=O(\varepsilon(\log\varepsilon)^2).$$

Moreover, using the result in Step 4, we have

$$E\left[\int_{l_0\varepsilon}^{s\wedge T}\xi(t)dt\Big|\tau_1=s\right]=O\left(\varepsilon^{1-\sigma}+\exp\left(-\frac{\kappa(s-l_0\varepsilon)}{\varepsilon}\right)\right).$$

Thus,

$$E\int_0^T\xi(t)I_{[0,\tau_1)}dt=O(\varepsilon^{1-\sigma}).$$

For $j\geq 1$, we have

$$\begin{aligned}E\int_0^T\xi(t)I_{[\tau_j,\tau_{j+1})}dt&=\int_0^\infty E\left[\int_r^T\xi(t)dt\Big|\tau_j=r\right]p_j(r)dr\\&=\int_0^\infty\int_r^T E\left[\xi(t)dtI_{[r,\tau_{j+1})}\Big|\tau_j=r\right]p_j(r)dr.\end{aligned}$$

Similarly, as in the case for $j=0$, we can show

$$\int_r^T E\left[\xi(t)dtI_{[r,\tau_{j+1})}\Big|\tau_j=r\right]=O(\varepsilon^{1-\sigma}).$$

It follows that

$$E\int_0^T\xi(t)I_{[\tau_j,\tau_{j+1})}dt=O(\varepsilon^{1-\sigma}).$$

Step 6. Combining Steps 3–5, we have

$$\begin{aligned}\left|(\hat{x}(\cdot)-\tilde{x}(\cdot))I_{\{\alpha(\cdot)\in\Theta^\varepsilon_{\sigma,\gamma_0}\}}\right|_T\leq{}&\left|(\hat{x}(\cdot)-\check{x}(\cdot))I_{\{\alpha(\cdot)\in\Theta^\varepsilon_{\sigma,\gamma_0}\}}\right|_T\\&+\left|(\check{x}(\cdot)-\tilde{x}(\cdot))I_{\{\alpha(\cdot)\in\Theta^\varepsilon_{\sigma,\gamma_0}\}}\right|_T\\={}&O(\varepsilon^2)+O(\varepsilon^{1-\delta})=O(\varepsilon^{1-\delta}).\end{aligned}$$

It follows from Step 1 that

$$|\hat{x}(\cdot)-\tilde{x}(\cdot)|_T=\left|(\hat{x}(\cdot)-\tilde{x}(\cdot))I_{\{\alpha(\cdot)\in\Theta^\varepsilon_{\sigma,\gamma_0}\}}\right|_T+\left|(\hat{x}(\cdot)-\tilde{x}(\cdot))I_{\{\alpha(\cdot)\notin\Theta^\varepsilon_{\sigma,\gamma_0}\}}\right|_T=O(\varepsilon^{1-\delta}).$$

This completes the proof. □

Remark 3.4. Note that the proofs of the filtering results in this section do not require the Markovian property of $\alpha(\cdot)$. In fact, the results hold for any general stochastic process $\alpha(\cdot)$ provided it satisfies the following conditions:

$$P(\alpha(\cdot)\notin\Theta^\varepsilon_\sigma)=O(\varepsilon^2)\text{ and }P(\tau_{j+1}-\tau_j\leq t)\leq Kt,\ t\geq 0,$$

where $\{\tau_j\}$ denotes the sequence of random jump times of $\alpha(\cdot)$.

4. Nearly optimal control. Now we turn to the optimal control problems. We assume the following additional conditions.

(A7) For each $i \in \mathcal{M}$, $L(t,i,x,u) \in C^{1,2,2}([0,T]\times\mathbb{R}^p\times\Gamma)$. There exists a constant K such that

$$|L(t,i,x,u)| + |\nabla_x L(t,i,x,u)| + |\nabla_x^2 L(t,i,x,u)| \le K.$$

For some constant c,

$$\frac{\partial^2 L(t,i,x,u)}{\partial u^2} \ge cI > 0.$$

Moreover, there exist $b_1(t,i,x)$, $b_2(t,i,x) \in C^{1,2}([0,T]\times R^p)$ such that

$$b(t,i,x,u) = b_1(t,i,x) + b_2(t,i,x)u.$$

Furthermore, the generator $Q(\cdot)$ is continuous on $[0,T]$ and the set of control points $\Gamma \in \mathbb{R}^{p_1}$ is compact and convex.

Remark 4.1. In order to prove the main results without undue technical complexities, we impose conditions in (A7). These conditions are somewhat conservative. The conditions on the drift term b are used to obtain the Lipschitz property of the optimal control policies as in Lemma 4.1. These conditions can be relaxed if we know a priori the optimal control is Lipschitz. The results to follow can also be extended to the case when L is independent of u as in Haussmann and Zhang [12]. Moreover, the terminal cost in the control problem was suppressed because essentially the terminal cost could be written as an integration of a running cost (e.g., [12]).

Let us temporarily consider the case when the state $(\alpha(t), x(t))$, $t \ge 0$, is completely observable. Let $\mathcal{U}_{\alpha,x}$ denote a set of controls $u(\cdot)$ which is progressively measurable with respect to $\sigma\{(\alpha(r), x(r)) : r \le t\}$ and $u(t) \in \Gamma$, $t \ge 0$.

Let $0 \le s \le T$, $\alpha(s) = i$, and $x(s) = x$, and define the corresponding value function

$$v(s,i,x) = \inf_{u(\cdot)\in\mathcal{U}_{\alpha,x}} E_{s,\alpha,x}\int_s^T L(t,\alpha(t),x(t),u(t))dt,$$

where $E_{s,\alpha,x}$ is the conditional expectation given $\alpha(s) = i$ and $x(s) = x$. Then the optimal control is determined by the value function $v(t,i,x)$, which satisfies the following Hamilton–Jacobi–Bellman (HJB) equation:

$$\text{(4.1)}\quad \begin{cases} 0 = \dfrac{\partial v(t,i,x)}{\partial t} + \min\limits_{u\in\Gamma}\{b(t,i,x,u)\nabla_x v(t,i,x) + L(t,i,x,u)\} \\ \qquad\quad + \dfrac{1}{2}\mathrm{tr}\left(F(t,i)H'(t,i)H(t,i)F(t,i)\nabla_x^2 v(t,i,x)\right) + Q(t)v(t,\cdot,x)(i), \\ v(T,i,x) = 0, \end{cases}$$

where $Q(t)v(t,\cdot,x)(i) = \sum_{j\ne i} q_{ij}(t)(v(t,j,x) - v(t,i,x))$.

Let $u^*(t,i,x)$ denote the feedback control policy minimizing the right-hand side of the HJB equation. The following results can be obtained similarly as in Fleming and Rishel [8] (or Krylov [16]).

LEMMA 4.1. *Assume* (A3) *and* (A7). *Then*

(a) *the value function* $v(t,i,x) \in C^{1,2}([0,T]\times R^p)$, $i \in \mathcal{M}$, *and is the unique solution to the HJB equation* (4.1).

(b) *there exist constants* $\kappa > 0$ *and* K *such that*

$$\left|\frac{\partial v(t,i,x)}{\partial t}\right| + |\nabla_x v(t,i,x)| + |\nabla_x^2 v(t,i,x)| \le K(1+|x|^\kappa).$$

(c) $u^*(t,i,x)$ *is an optimal feedback control and uniformly Lipschitz in* x.

We turn to consider the partially observed system. Let $\mathcal{U}_y$ denote a class of controls $u(\cdot)$ that are $\mathcal{Y}_t$ progressively measurable and $u(t) \in \Gamma$, $t \ge 0$. Given $u(\cdot) \in \mathcal{U}_y$, let $(\tilde{x}(\cdot), \widetilde{R}(\cdot))$ be the corresponding filter given by the equations

$$\begin{aligned} d\tilde{x}(t) =\ & b(t,\tilde{\alpha}(t),\tilde{x}(t),u(t))dt \\ & +\frac{1}{\varepsilon^2}\widetilde{R}(t)H'(t,\tilde{\alpha}(t))\left(dy(t)-h(t,\tilde{\alpha}(t),\tilde{x}(t))dt\right), \\ \frac{d\widetilde{R}(t)}{dt} =\ & \nabla_x b(t,\tilde{\alpha}(t),\tilde{x}(t),u(t))\widetilde{R}(t)+\widetilde{R}(t)(\nabla_x b(t,\tilde{\alpha}(t),\tilde{x}(t),u(t)))' \\ & +F(t,\tilde{\alpha}(t))H'(t,\tilde{\alpha}(t))H(t,\tilde{\alpha}(t))F(t,\tilde{\alpha}(t)) \\ & -\frac{1}{\varepsilon^2}\widetilde{R}(t)H'(t,\tilde{\alpha}(t))H(t,\tilde{\alpha}(t))\widetilde{R}(t), \end{aligned} \tag{4.2}$$

with $\tilde{x}(0) = Ex_0$ and $\widetilde{R}(0) = \mathrm{Cov}(x_0)$. Then the results in Theorems 3.3 and 3.5 hold uniformly with respect to $u(\cdot) \in \mathcal{U}_y$.

Regarding $(\tilde{\alpha}(\cdot), \tilde{x}(\cdot))$ as the "state" and using the feedback control $u^*(t,\alpha,x)$, we define $\tilde{u}(t) = u^*(t,\tilde{\alpha}(t),\tilde{x}(t))$.

Note that $\tilde{\alpha}(\cdot)$ changes values only at $t = j\varepsilon$, $j = 1,2,\ldots$, and for $t \in [j\varepsilon,(j+1)\varepsilon)$, $\tilde{\alpha}(t)$ is $\mathcal{Y}_{j\varepsilon}$ measurable. Therefore, a unique solution $(x(\cdot),y(\cdot),\tilde{x}(\cdot),\widetilde{R}(\cdot))$ to the equations (1.1) and (4.2) with $u(\cdot) = \tilde{u}(\cdot)$ can be obtained piecewisely over the intervals $[j\varepsilon,(j+1)\varepsilon)$, $j = 0,1,\ldots$. In view of these, $\tilde{u}(\cdot)$ is admissible, i.e., $\tilde{u}(\cdot) \in \mathcal{U}_y$. The next theorem concerns the performance of $\tilde{u}(\cdot)$.

THEOREM 4.2. *Assume the conditions of Theorem* 3.3 *and* (A7). *Then for each* $0 < \delta < 1$, *there exists* $\varepsilon_0 > 0$ *such that for* $0 < \varepsilon < \varepsilon_0$,

$$\begin{aligned} &\text{(a)}\ \inf_{u(\cdot)\in\mathcal{U}_y} J(u(\cdot)) = J(\tilde{u}(\cdot)) + O(\varepsilon^{1-\delta}); \\ &\text{(b)}\ \inf_{u(\cdot)\in\mathcal{U}_y} J(u(\cdot)) = Ev(0,\alpha(0),Ex_0) + O(\varepsilon^{1-\delta}). \end{aligned}$$

Proof. We first show part (a) and divide the proof into several steps.

Step 1. We show that

$$\inf_{u(\cdot)\in\mathcal{U}_y} J(u(\cdot)) = \inf_{u(\cdot)\in\mathcal{U}_y} E\int_0^T L(t,\tilde{\alpha}(t),\tilde{x}(t),u(t))dt + O(\varepsilon^{1-\delta}).$$

For each $u(\cdot) \in \mathcal{U}_y$, let $x(\cdot)$ and $\tilde{x}(\cdot)$ denote the the corresponding state and filter processes under $u(\cdot)$. Then, by noticing that $I_{\{\tilde{\alpha}(t)\neq\alpha(t)\}} \le |\tilde{\alpha}(t)-\alpha(t)|/m$ and recalling Theorem 3.3, we have

$$\begin{aligned} & E\int_0^T |L(t,\alpha(t),x(t),u(t)) - L(t,\tilde{\alpha}(t),x(t),u(t))|dt \\ & \le KE\int_0^T I_{\{\tilde{\alpha}(t)\neq\alpha(t)\}}dt \\ & \le \frac{K}{m}E\int_0^T |\tilde{\alpha}(t)-\alpha(t)|dt = O(\varepsilon^{1-\delta}). \end{aligned} \tag{4.3}$$

Moreover, in view of Taylor's expansion, we write

$$\begin{aligned} L(t,\tilde{\alpha}(t),x(t),u(t)) &= L(t,\tilde{\alpha}(t),\hat{x}(t),u(t)) \\ &\quad +\nabla_x L(t,\tilde{\alpha}(t),\hat{x}(t),u(t))(x(t)-\hat{x}(t)) + O(|x(t)-\hat{x}(t)|^2). \end{aligned} \tag{4.4}$$

Since $\tilde{\alpha}(t)$ and $u(t)$ are $\mathcal{Y}_t$ measurable, it follows that

$$\begin{aligned} &E\nabla_x L(t,\tilde{\alpha}(t),\hat{x}(t),u(t))(x(t)-\hat{x}(t)) \\ &\quad = E\left(E[\nabla_x L(t,\tilde{\alpha}(t),\hat{x}(t),u(t))(x(t)-\hat{x}(t))|\mathcal{Y}_t]\right) = 0. \end{aligned} \tag{4.5}$$

Combining (4.4) and (4.5) leads to

$$\begin{aligned} &EL(t,\tilde{\alpha}(t),x(t),u(t)) = EL(t,\tilde{\alpha}(t),\hat{x}(t),u(t)) + O(E|x(t)-\hat{x}(t)|^2) \\ &\quad \le EL(t,\tilde{\alpha}(t),\hat{x}(t),u(t)) + O(E|x(t)-\tilde{x}(t)|^2) + O(E|\tilde{x}(t)-\hat{x}(t)|^2). \end{aligned}$$

Hence, in view of Lemmas A.2 and A.3 and (3.4), for $0<\sigma<\delta$, we can show similarly as in (3.19) that

$$E\int_0^T L(t,\tilde{\alpha}(t),x(t),u(t))dt = E\int_0^T L(t,\tilde{\alpha}(t),\hat{x}(t),u(t))dt + O(\varepsilon^{1-\delta}). \tag{4.6}$$

Furthermore, note that

$$|L(t,\tilde{\alpha}(t),\hat{x}(t),u(t)) - L(t,\tilde{\alpha}(t),\tilde{x}(t),u(t))| \le K|\hat{x}(t)-\tilde{x}(t)|.$$

Thus, in view of Lemma A.3 and by conditioning on $\alpha(\cdot)\in\Theta^{\varepsilon}_{\sigma,\gamma_0}$, we have

$$E\int_0^T L(t,\tilde{\alpha}(t),\hat{x}(t),u(t))dt = E\int_0^T L(t,\tilde{\alpha}(t),\tilde{x}(t),u(t))dt + O(\varepsilon^{1-\delta}). \tag{4.7}$$

Combining (4.3), (4.6), and (4.7), we obtain

$$J(u(\cdot)) = E\int_0^T L(t,\tilde{\alpha}(t),\tilde{x}(t),u(t))dt + O(\varepsilon^{1-\delta})$$

uniformly with respect to $u(\cdot)\in\mathcal{U}_y$.

Step 2. In this step we show that

$$\begin{aligned} &\inf_{u(\cdot)\in\mathcal{U}_y} E\int_0^T L(t,\tilde{\alpha}(t),\tilde{x}(t),u(t))dt \\ &\qquad = \inf_{u(\cdot)\in\mathcal{U}_y} E\int_0^T L(t,\alpha(t),\check{x}(t),u(t))dt + O(\varepsilon^{1-\delta}). \end{aligned} \tag{4.8}$$

Note that (3.21) holds uniformly with respect to $u(\cdot)\in\mathcal{U}_y$ when $\check{x}(\cdot)$ is defined in (3.16) with $b=b(t,\alpha(t),\check{x}(t),u(t))$. In view of this, (4.8) follows from

$$E\int_0^T |L(t,\tilde{\alpha}(t),\tilde{x}(t),u(t))dt - L(t,\alpha(t),\tilde{x}(t),u(t))|dt = O(\varepsilon^{1-\delta})$$

and the Lipschitz property of L, which yields

$$\begin{aligned} &E\int_0^T |L(t,\alpha(t),\tilde{x}(t),u(t))dt - L(t,\alpha(t),\check{x}(t),u(t))|dt \\ &\qquad \le KE\int_0^T |\tilde{x}(t)-\check{x}(t)|dt = O(\varepsilon^{1-\delta}). \end{aligned}$$

Step 3. Let $\mathcal{U}_{y,\alpha} = \{u(\cdot) :\ u(t) \text{ is } \sigma\{y(s), \alpha(s) :\ 0 \le s \le t\}\text{- progressively measurable }\}$. Then, noticing the fact that $\mathcal{U}_y \subset \mathcal{U}_{y,\alpha}$, we obtain

$$\inf_{u(\cdot)\in\mathcal{U}_y} E\int_0^T L(t,\alpha(t),\check{x}(t),u(t))dt \ge \inf_{u(\cdot)\in\mathcal{U}_{y,\alpha}} E\int_0^T L(t,\alpha(t),\check{x}(t),u(t))dt,$$

Step 4. We prove the following estimate based on the dynamic programming approach:

$$\begin{aligned}
&\inf_{u(\cdot)\in\mathcal{U}_{y,\alpha}} E\int_0^T L(t,\alpha(t),\check{x}(t),u(t))dt\\
&\qquad = Ev(0,\alpha(0),\check{x}(0)) + O(\varepsilon^{1-\delta})\\
&\qquad = E\int_0^T L(t,\alpha(t),\check{x}(t),u^*(t,\alpha(t),\check{x}(t)))dt + O(\varepsilon^{1-\delta}).
\end{aligned}$$

For all $u(\cdot) \in \mathcal{U}_{y,\alpha}$, using Dynkin's formula and noticing that $v(T,i,x) = 0$, we have

$$\begin{aligned}
Ev(0,\alpha(0),\check{x}(0)) &= -E\int_0^T \Big\{ \frac{\partial v(t,\alpha(t),\check{x}(t))}{\partial t} + \nabla_x v(t,\alpha(t),\check{x}(t)) b(t,\alpha(t),\check{x}(t),u(t))\\
&+Q(t)v(t,\cdot,\check{x}(t))(\alpha(t)) + \frac{1}{2\varepsilon^2}\mathrm{tr}\left(\check{R}(t)H'(t,\alpha(t))H(t,\alpha(t))\check{R}(t)\nabla_x^2 v(t,\alpha(t),\check{x}(t))\right)\Big\}dt\\
&-E\int_0^T \nabla_x v(t,\alpha(t),\check{x}(t))\left(dy(t) - H(t,\alpha(t))\check{x}(t)dt\right).
\end{aligned}$$

Similarly as in Step 3 of the proof of Theorem 3.5, by conditioning on $\alpha(\cdot) = \theta$, we can show that

$$E\int_0^T \nabla_x v(t,\alpha(t),\check{x}(t))\left(E(H(t,\alpha(t))x(t)|\mathcal{Y}_t) - H(t,\alpha(t))\check{x}(t)\right)dt = O(\varepsilon^{1-\delta})$$

and

$$E\int_0^T \nabla_x v(t,\alpha(t),\check{x}(t))d\hat{v}(t) = 0.$$

Hence,

$$E\int_0^T \nabla_x v(t,\alpha(t),\check{x}(t))\left(dy(t) - H(t,\alpha(t))\check{x}(t)dt\right) = O(\varepsilon^{1-\delta}).$$

Moreover, in view of Lemma 4.1, for $0 < \sigma < \delta$, we have

$$\begin{aligned}
&E\int_0^T \mathrm{tr}\Big[\Big(\frac{\check{R}(t)H'(t,\alpha(t))H(t,\alpha(t))\check{R}(t)}{\varepsilon^2}\\
&\qquad -F(t,\alpha(t))H'(t,\alpha(t))H(t,\alpha(t))F(t,\alpha(t))\Big)\nabla_x^2 v(t,\alpha(t),\check{x}(t))\Big]dt\\
&\le KE\int_0^T \left|\frac{\check{R}(t)}{\varepsilon} - F(t,\alpha(t))\right|\cdot|\nabla_x^2 v(t,\alpha(t),\check{x}(t))|dt\\
&= KE\int_0^T \left|\frac{\check{R}(t)}{\varepsilon} - F(t,\alpha(t))\right|\cdot|\nabla_x^2 v(t,\alpha(t),\check{x}(t))|dtI_{\{\alpha(\cdot)\in\Theta^\varepsilon_{\sigma,\gamma_0}\}}\\
&\quad +KE\int_0^T \left|\frac{\check{R}(t)}{\varepsilon} - F(t,\alpha(t))\right|\cdot|\nabla_x^2 v(t,\alpha(t),\check{x}(t))|dtI_{\{\alpha(\cdot)\notin\Theta^\varepsilon_{\sigma,\gamma_0}\}}\\
&= KE\int_0^T \left|\frac{\check{R}(t)}{\varepsilon} - F(t,\alpha(t))\right|\cdot|\nabla_x^2 v(t,\alpha(t),\check{x}(t))|dtI_{\{\alpha(\cdot)\in\Theta^\varepsilon_{\sigma,\gamma_0}\}} + O(\varepsilon^{1-\sigma}).
\end{aligned}$$

For each $\theta \in \Theta^{\varepsilon}_{\sigma,\gamma_0}$, let $\{t_k\}$ denote the jump times of $\theta = \theta(\cdot)$. Then we obtain

$$\begin{aligned}
&\sum_{j=0}^{[1/\varepsilon^\sigma]} E^\theta \int_{t_j}^{t_{j+1}} \left| \frac{\check{R}(t)}{\varepsilon} - F(t,\alpha(t)) \right| \cdot |\nabla_x^2 v(t,\alpha(t),\check{x}(t))| dt \\
&\quad \le K \sum_{j=0}^{[1/\varepsilon^\sigma]} E^\theta \int_{t_j}^{t_{j+1}} \left(\varepsilon + \exp\left(-\frac{\kappa(t-t_j)}{\varepsilon} \right) \right) \cdot |\nabla_x^2 v(t,\alpha(t),\check{x}(t))| dt \\
&\quad = K \sum_{j=0}^{[1/\varepsilon^\sigma]} \int_{t_j}^{t_{j+1}} \left(\varepsilon + \exp\left(-\frac{\kappa(t-t_j)}{\varepsilon} \right) \right) E^\theta |\nabla_x^2 v(t,\alpha(t),\check{x}(t))| dt \\
&\quad \le K \sum_{j=0}^{[1/\varepsilon^\sigma]} K\varepsilon = O(\varepsilon^{1-\sigma}).
\end{aligned}$$

In view of these, we have

$$\begin{aligned}
Ev(0,\alpha(0),\check{x}(0)) &= -E\int_0^T \Big\{ \frac{\partial v(t,\alpha(t),\check{x}(t))}{\partial t} + b(t,\alpha(t),\check{x}(t),u(t)) \nabla_x v(t,\alpha(t),\check{x}(t)) \\
&\quad + Q(t) v(t,\cdot,\check{x}(t))(\alpha(t)) + L(t,\alpha(t),\check{x}(t),u(t)) \\
&\quad + \frac{1}{2} \mathrm{tr}\left(F(t,\alpha(t)) H'(t,\alpha(t)) H(t,\alpha(t)) F(t,\alpha(t)) \nabla_x^2 v(t,\alpha(t),\check{x}(t)) \right) \Big\} dt + O(\varepsilon^{1-\sigma}) \\
&\quad + E\int_0^T L(t,\alpha(t),\check{x}(t),u(t)) dt \\
&\le E\int_0^T L(t,\alpha(t),\check{x}(t),u(t)) dt + O(\varepsilon^{1-\sigma}),
\end{aligned}$$

where the last inequality is due to the HJB equation and the equality holds if $u(t) = u^*(t,\alpha(t),\check{x}(t))$.

Step 5. Finally, note that $\tilde{u}(\cdot) \in \mathcal{U}_y$. Thus,

$$J(\tilde{u}(\cdot)) \ge \inf_{u(\cdot)\in\mathcal{U}_y} J(u(\cdot)).$$

Moreover, using the Lipschitz property of $u^*(t,i,\cdot)$, we have

$$\begin{aligned}
&|L(t,\alpha(t),\check{x}(t),u^*(t,\alpha(t),\check{x}(t))) - L(t,\tilde{\alpha}(t),\tilde{x}(t),u^*(t,\tilde{\alpha}(t),\tilde{x}(t)))| \\
&\qquad \le K(|\tilde{\alpha}(t) - \alpha(t)| + |\check{x}(t) - \tilde{x}(t)|).
\end{aligned}$$

Thus, it follows that by using Theorem 3.3 and Step 5 in the proof of Theorem 3.5,

$$\begin{aligned}
&E\int_0^T L(t,\alpha(t),\check{x}(t),u^*(t,\alpha(t),\check{x}(t))) dt \\
&\qquad = E\int_0^T L(t,\tilde{\alpha}(t),\tilde{x}(t),u^*(t,\tilde{\alpha}(t),\tilde{x}(t))) dt + O(\varepsilon^{1-\sigma}).
\end{aligned}$$

To show (b), recall that $\check{x}(0) = Ex_0$. Following Steps 1–4, we have

$$\inf_{u(\cdot)\in\mathcal{U}_y} J(u(\cdot)) \ge Ev(0,\alpha(0),Ex_0) + O(\varepsilon^{1-\delta}).$$

Then Steps 4–5 yield

$$Ev(0,\alpha(0),Ex_0) \ge J(\tilde{u}(\cdot)) + O(\varepsilon^{1-\delta}) \ge \inf_{u(\cdot)\in\mathcal{U}_y} J(u(\cdot)) + O(\varepsilon^{1-\delta}).$$

Combine these two inequalities to obtain the result. The proof is now complete. □

5. Hybrid linear quadratic control. In this section we show that the compactness of Γ and boundedness of L in (A7) are not necessary and can be relaxed. We study a simple and useful case in which the system is linear in x and u with a quadratic running cost function. Consider the state $x(t) \in \mathbb{R}^p$, observation $y(t) \in \mathbb{R}^p$, and control $u(t) \in \mathbb{R}^{p_1}$ satisfying the differential equations

$$\left\{\begin{array}{ll} dx(t) = (B_1(t,\alpha(t))x(t) + B_2(t,\alpha(t))u(t))dt & \\ \qquad\qquad +\sigma(t,\alpha(t))dw(t), & x(0) = x_0 \\ dy(t) = H(t,\alpha(t))x(t)dt + \varepsilon dv(t), & y(0) = 0, \end{array}\right. \tag{5.1}$$

where $B_1(t,i)$, $B_1(t,i)$, $\sigma(t,i)$, and $H(t,i)$ are matrices of appropriate dimensions.

The cost function is given by

$$J(u(\cdot)) = E_{\alpha,x}\int_0^T \Big(x'(t)M_1(t,\alpha(t))x(t) + u'(t)M_2(t,\alpha(t))u(t)\Big)dt,$$

where $M_1(t,i)$ and $M_2(t,i)$ are positive definite matrices of appropriate dimensions and $E_{\alpha,x}$ is the conditional expectation given $\alpha(0) = \alpha$ and $x(0) = x$.

Assume all the conditions in the previous sections hold except the conditions on the running cost function L and the control set Γ. In this section we consider $\Gamma = \mathbb{R}^{p_1}$. First of all, consider the completely observable case. Let $\mathcal{U}_{\alpha,x}$ denote the set of admissible controls $u(\cdot)$, which is $\sigma\{(\alpha(r), x(r)) : r \le t\}$ progressively measurable, $E\int_0^T |u(t)|^k dt \le C_k$, $E|x(t)|^k \le C_k$, for each $k = 1, 2, \ldots$, and some constant C_k. In this case, the value function

$$\begin{aligned} v(s,i,x) &= \inf_{u(\cdot)\in\mathcal{U}_{\alpha,x}} E_{s,\alpha,x}\int_s^T \Big(x'(t)M_1(t,\alpha(t))x(t) + u'(t)M_2(t,\alpha(t))u(t)\Big)dt \\ &= x'\Phi(s,i)x + \phi(s,i), \end{aligned}$$

where $E_{s,\alpha,x}$ is the conditional expectation given $(\alpha(s), x(s)) = (\alpha, x)$ and the functions $\Phi(t,i)$ and $\phi(t,i)$ are determined by the following differential equations:

$$\left\{\begin{array}{l} \dfrac{d\Phi(t,i)}{dt} = -\Big\{\Phi(t,i)B_1(t,i) + B_1'(t,i)\Phi(t,i) \\ \quad -\Phi(t,i)B_2(t,i)M_2^{-1}(t,i)B_2'(t,i)\Phi(t,i) + M_1(t,i) + Q(t)\Phi(t,\cdot)(i)\Big\} \\ \Phi(T,i) = 0, \quad \text{for } i \in \mathcal{M} \end{array}\right. \tag{5.2}$$

and

$$\left\{\begin{array}{l} \dfrac{d\phi(t,i)}{dt} = -\Big\{\mathrm{tr}\Big(\sigma(t,i)\sigma'(t,i)\Phi(t,i)\Big) + Q(t)\phi(t,\cdot)(i)\Big\}, \\ \phi(T,i) = 0, \quad \text{for } i \in \mathcal{M}. \end{array}\right.$$

The optimal control for the completely observable case is given by (see Fleming and Rishel [8])

$$u^*(t,i,x) = -M_2^{-1}(t,i)B_2'(t,i)\Phi(t,i)x. \tag{5.3}$$

In the partially observable case, we consider the control

$$\tilde{u}(t) = u^*(t,\tilde{\alpha}(t),\tilde{x}(t)). \tag{5.4}$$

Let $\mathcal{U}_y$ denote the set of controls which are $\sigma\{y(r) : r \le t\}$ progressively measurable and for each $k = 1, 2, \ldots$, there exists C_k such that $E\int_s^T |u(t)|^k dt \le C_k$ and $E|x(t)|^k \le C_k$.

THEOREM 5.1. *For each $0 < \delta < 1$, there exists $\varepsilon_0 > 0$ such that for $0 < \varepsilon < \varepsilon_0$,*

$$\text{(a)} \inf_{u(\cdot)\in\mathcal{U}_y} J(u(\cdot)) = J(\tilde{u}(\cdot)) + O(\varepsilon^{1-\delta});$$
$$\text{(b)} \inf_{u(\cdot)\in\mathcal{U}_y} J(u(\cdot)) = Ev(0, \alpha(0), Ex_0) + O(\varepsilon^{1-\delta}).$$

Proof. The proof can be given following Steps 1-5 in the proof of Theorem 4.2 by using the Holder's inequality as in (3.19) and the quadratic property of the running function $L(t, i, x)$. □

6. Extensions to Picard filter. Note that in Lemma A.4, $R^\theta(t)/\varepsilon$ can be approximated by $F(t, \theta(t))$ for each $\theta \in \Theta$. If we replace $R^\theta(t)/\varepsilon$ by $F(t, \theta(t))$ in the EKF, then we obtain the PF; see Picard [20]. Let $\tilde{\alpha}(\cdot)$ be a filter of $\alpha(\cdot)$ obtained either by the QVT or by the LRT. Let $u(\cdot) \in \mathcal{U}_y$ and define the PF as follows:

$$\begin{aligned} dm(t) &= b(t, \tilde{\alpha}(t), m(t))dt \\ &\quad + \frac{1}{\varepsilon} F(t, \tilde{\alpha}(t)) H'(t, \tilde{\alpha}(t)) \left(dy(t) - h(t, \tilde{\alpha}(t), m(t))dt\right), \end{aligned} \tag{6.1}$$

with $m(0) = Ex_0$.

THEOREM 6.1. *Assume the condition of Theorem 4.2. Then for each $0 < \delta < 1$, there exists $\varepsilon_0 > 0$ such that for $0 < \varepsilon < \varepsilon_0$,*

$$\text{(a)} \ |\hat{x}(\cdot) - m(\cdot)|_T = O(\varepsilon^{\frac{1}{2}-\delta}), \ \textit{uniformly with respect to } u(\cdot) \in \mathcal{U}_y$$
$$\text{(b)} \inf_{u(\cdot)\in\mathcal{U}_y} J(u(\cdot)) = J(\bar{u}(\cdot)) + O(\varepsilon^{\frac{1}{2}-\delta}),$$

where $\bar{u}(t) = u^*(t, \tilde{\alpha}(t), m(t))$.

Proof. We define $\check{m}(t)$, $t \ge 0$, as in (6.1) with $\tilde{\alpha}(\cdot)$ replaced by $\alpha(\cdot)$. Let $0 < \sigma < \delta$. Then we can show as in [11, Theorem 2.6], Lemma A.2, and Step 1 in Theorem 3.5, that

$$E|\check{x}(t) - \check{m}(t)| I_{\{\alpha(\cdot)\in\Theta^\varepsilon_{\sigma,\gamma_0}\}} = O(\varepsilon^{\frac{3}{2}-\sigma}).$$

Then following the proof in Steps 3-5 of Theorem 3.5 we can show

$$E|\check{m}(t) - m(t)| I_{\{\alpha(\cdot)\in\Theta^\varepsilon_{\sigma,\gamma_0}\}} = O(\varepsilon^{\frac{1}{2}-\sigma}).$$

Thus in view of Step 5 in Theorem 3.5, we have

$$E|\tilde{x}(t) - \check{x}(t)| I_{\{\alpha(\cdot)\in\Theta^\varepsilon_{\sigma,\gamma_0}\}} = O(\varepsilon^{1-\sigma}). \tag{6.2}$$

Combining these estimates, we obtain part (a).

Part (b) follows from the fact that

$$\inf_{u(\cdot)\in\mathcal{U}_y} J(u(\cdot)) = E\int_0^T L(t, \alpha(t), \check{x}(t), u^*(t, \alpha(t), \check{x}(t)))dt + O(\varepsilon^{1-\delta}),$$

as in Step 4 of Theorem 4.2, the Lipschitz property of the function $L(t, \alpha, x, u^*(t, \alpha, x))$, part (a), and (6.2). □

As for the hybrid linear quadratic case, we can obtain similarly that

$$\inf_{u(\cdot)\in\mathcal{U}_y} J(u(\cdot)) = J(\bar{u}(\cdot)) + O(\varepsilon^{\frac{1}{2}-\delta}),$$

where $\bar{u}(t) = u^*(t, \tilde{\alpha}(t), m(t))$ with $u^*(t,\alpha,x)$ given by (5.3).

Remark 6.1. In general the EKF provides a better approximation than the PF. However, since the PF does not require computing $\widetilde{R}(\cdot)$, which reduces much of the computation effort, especially when the dimension of the system is large.

In the LRT, we used the EKF to define the test statistics $L^{(i)}(\mathcal{I}(t))$. An alternative way is to use the outcome of the PF to replace the EKF. The results in Lemma 3.2 follows in a similar way.

7. An example and numerical simulations. In this section we consider a simple example and report related computational experiments. We consider the following one-dimensional model:

$$\begin{cases} dx(t) = \Big(B_1(\alpha(t))x(t) + u(t)\Big)dt + \sigma(\alpha(t))dw(t), & x(0) = x_0 \\ dy(t) = H(\alpha(t))x(t)dt + \varepsilon dv(t), & y(0) = 0, \end{cases} \tag{7.1}$$

where $\alpha(t) \in \mathcal{M} = \{1,2\}$, $t \geq 0$, is a Markov chain generated by $Q = \left(\begin{smallmatrix} -\lambda & \lambda \\ \mu & -\mu \end{smallmatrix}\right)$ with $\lambda > 0$ and $\mu > 0$.

We discretize the equation in (7.1) with step size ε. The time horizons in the continuous-time model is $T = 10$ and in the corresponding discrete-time setting is $T_\varepsilon = 10/\varepsilon$. All of our results are based on computations with 100 sample paths.

We consider the model with the following specifications:

$$B_1(1) = -0.01, \quad B_1(2) = -0.02, \quad H(1) = 1, \quad H(2) = 2, \quad \sigma(1) = \sigma(2) = 1,$$

$$\lambda = \mu = 0.02, \quad x_0 = 0, \quad \alpha(0) = 2.$$

Note that the detectability condition in (A5) is satisfied because

$$|H^2(1) - H^2(2)| = 3 > 0.$$

We only consider the QVT because it performs better than the LRT; see Remark 3.3.

Filtering. We compare our results with the well-known interactive multiple models (IMM) algorithm given in Blom and Bar-Shalom [2]. Let $|\tilde{x} - x|_{\text{IMM}}$ denote the norm $\varepsilon \sum_{k=0}^{T_\varepsilon} E|\tilde{x}(k\varepsilon) - x(k\varepsilon)|$ with $\tilde{x}(k\varepsilon)$ obtained by using the IMM algorithm. Similarly, let $|\tilde{x} - x|_{\text{QVT}}$ and $|\tilde{\alpha} - \alpha|_{\text{QVT}}$ denote the corresponding norm when using the QVT algorithm. We take the control $u(t) = 0$. Our numerical results are illustrated by a sample path of $x(k\varepsilon)$, its estimates $\tilde{x}(k\varepsilon)$ using both the IMM and the QVT, and the corresponding errors with parameters $\varepsilon = 0.1$ and $T_\varepsilon = 100$. Their graphs are given in Fig. 7.1.

We also vary the value of ε and obtain upper bounds on estimates of $\tilde{x}(k\varepsilon)$ using the IMM and $(\tilde{\alpha}(k\varepsilon), \tilde{x}(k\varepsilon))$ using the QVT. These are given in Table 7.1.

Remark 7.1. The major advantages of the QVT algorithm is that it does not require the process $\alpha(\cdot)$ to be Markovian. So there is no need to require the generator matrix. Numerically, the QVT works better when ε is small and when the parameter process $\alpha(\cdot)$ does not jump too rapidly. On the other hand, the IMM algorithm works as "an average" device because it tends to average out the fluctuation of the signal

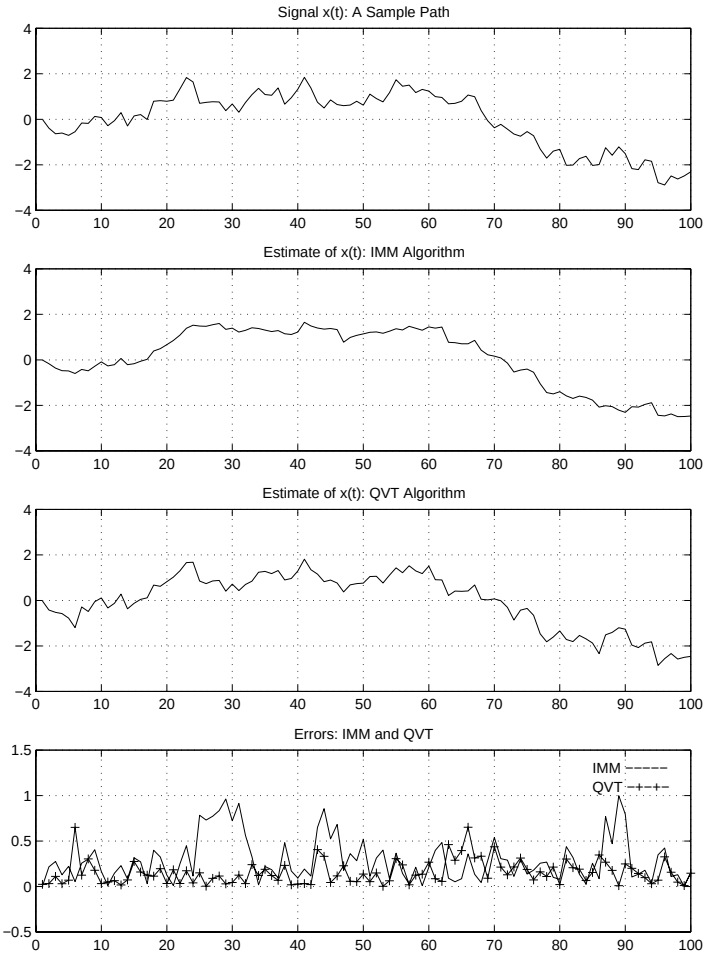


FIG. 7.1. *A sample path of $x(t)$, its estimates, and errors.*

TABLE 7.1
Upper error bounds: filtering.

ε	T_ε	$\lvert\tilde{x}-x\rvert_{\rm IMM}/\sqrt{\varepsilon}$	$\lvert\tilde{\alpha}-\alpha\rvert_{\rm QVT}/\sqrt{\varepsilon}$	$\lvert\tilde{x}-x\rvert_{\rm QVT}/\sqrt{\varepsilon}$
0.1	100	14.17	7.44	9.87
0.05	200	17.71	11.26	13.40
0.033	300	18.93	13.87	15.30
0.025	400	19.15	16.61	16.72
0.02	500	19.16	17.92	18.07

process. So it seems the IMM works better when $\alpha(\cdot)$ fluctuates more frequently. The QVT is a quite promising filtering device in target tracking. It complements the IMM in a number of ways. It could also be used in combination with the IMM algorithm to improve the performance.

Control. We consider the cost function

$$J(u(\cdot)) = E\int_0^T \Big(x^2(t) + u^2(t)\Big)dt.$$

TABLE 7.2
Upper error bounds: control.

ε	T_ε	$\widetilde{J}$	v	$\lvert\widetilde{J}/v-1\rvert/\sqrt{\varepsilon}$	$\lvert\widetilde{J}-v\rvert/\sqrt{\varepsilon}$
0.1	100	10.70	9.09	0.56	5.11
0.05	200	9.73	9.11	0.30	2.75
0.033	300	9.28	9.12	0.09	0.87
0.025	400	9.26	9.12	0.09	0.85
0.02	500	8.97	9.13	0.12	1.13

Then the associated Riccati equation is given by

$$\frac{d\Phi(t,1)}{dt} = -\Big\{2B_1(1)\Phi(t,1) - \Phi^2(t,1) + 1 + \lambda(\Phi(t,2) - \Phi(t,1))\Big\},$$
$$\frac{d\Phi(t,2)}{dt} = -\Big\{2B_1(2)\Phi(t,2) - \Phi^2(t,2) + 1 + \mu(\Phi(t,1) - \Phi(t,2))\Big\},$$

with $\Phi(T,1) = \Phi(T,2) = 0$. The control defined in (5.4) becomes

$$\tilde{u}(t) = -\Phi(t,\tilde{\alpha}(t))\tilde{x}(t).$$

Given $x(s) = x_0$ and $\alpha(s) = i$, the value function

$$v(s,i,x) = x_0^2\Phi(s,i) + \phi(s,i),$$

where $\phi(t,i)$ is determined by the equations

$$\frac{d\phi(t,1)}{dt} = -\Big\{\sigma^2(1)\Phi(t,1) + \lambda(\phi(t,2) - \phi(t,1))\Big\},$$
$$\frac{d\phi(t,2)}{dt} = -\Big\{\sigma^2(2)\phi(t,2) + \mu(\phi(t,1) - \phi(t,2))\Big\},$$

with $\phi(T,1) = \phi(T,2) = 0$.

Let

$$\widetilde{J} = J(\tilde{u}(\cdot)) = E\int_0^T (x^2(t) + \tilde{u}^2(t))dt$$

and $v = v(0,\alpha(0),x(0))$. Then for various ε we have the upper bounds given in Table 7.2.

It can be seen from the numerical simulations that our algorithm gives a quite good approximation to exact optimal solutions.

8. Conclusions. In this paper, we constructed asymptotic filters $(\tilde{x}(\cdot), \widetilde{R}(\cdot))$ and $m(\cdot)$. Using these filters, we constructed nearly optimal controls for the partially observed stochastic system. The information flow is illustrated in Fig. 8.1.

A key assumption in this paper is that the observation noise has to be small. To apply these results in a practical scenario, it is important to determine if the noise in a given problem is small enough to fit the requirement in the paper. In fact, as in general singular perturbation theory, the small parameter ε does not have to be very small in order to have decent numerical results. Typically, it works well when ε is less than 0.1 when all other elements in the coefficients of the system are of order 1.

This paper considers the case when the unknown $\alpha(\cdot)$ does not fluctuate too rapidly. Naturally, it would be interesting to consider the case when $\alpha(\cdot)$ jumps

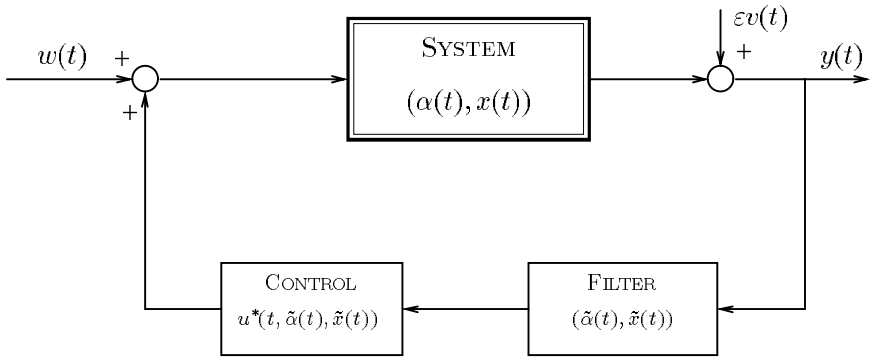


FIG. 8.1. *System diagram.*

rapidly from time to time. Quite often such situation can be formulated as a singular perturbed Markov chain with weak and strong interactions. In this connection, the asymptotic results in Yin and Zhang [21] appear to be useful for constructing filters and the ensuing optimality analysis.

Appendix. In this section we give six technical lemmas used in the paper.

LEMMA A.1. *Assume* (A1)–(A4). *Then there exist constants* ε_0 *and* K *such that for each* $0 < \varepsilon < \varepsilon_0$ *and* $\theta = \theta(\cdot) \in \Theta$,

$$E^\theta \int_0^T \left|\hat{x}^\theta(t) - \tilde{x}^\theta(t)\right|^2 dt \le K\varepsilon^2 E^\theta \int_0^T \left|x(t) - \tilde{x}^\theta(t)\right|^4 dt,$$

where $\hat{x}^\theta(t) = E^\theta[x(t)|\mathcal{Y}_t]$ *and* E^θ *is the conditional expectation given* $\alpha(\cdot) = \theta$.

Proof. The proof of this lemma can be the same as that given as in Haussmann and Zhang [11, Theorem 2.1]. □

LEMMA A.2. *Assume* (A1)–(A3). *Then there exist* γ_0 *and* K *such that for each* $0 < \sigma < 1$ *and for all* $\theta = \theta(\cdot) \in \Theta^\varepsilon_{\sigma,\gamma_0}$,

$$E^\theta \int_0^T \left|x(t) - \tilde{x}^\theta(t)\right|^4 dt \le K\varepsilon^2.$$

Proof. Let $\{t_j\}$ denote the set of jump times of θ. For each t_j, let

$$\begin{aligned} \phi_j &= E^\theta |x(t_j) - \tilde{x}^\theta(t_j)|^2, \\ \psi_j &= E^\theta |x(t_j) - \tilde{x}^\theta(t_j)|^4. \end{aligned}$$

Note that on interval $[t_j, t_{j+1})$, $\theta(\cdot)$ is a constant and therefore the function $H(t, \theta(t))$ is differentiable on this interval.

By considering the differential $d|H(t,\theta(t))x(t) - H(t,\theta(t))\tilde{x}^\theta(t)|^2$ on $[t_j, t_{j+1})$, we can show, by using Gronwall's inequality (as in [11, Lemma 2.3]), that

$$E^\theta |x(t) - \tilde{x}^\theta(t)|^2 \le K_1 \phi_j \exp\left(-\frac{\kappa(t - t_j)}{2\varepsilon}\right) + O(\varepsilon), \tag{A.1}$$

where $\kappa > 0$, $K_1 > 0$, and $O(\cdot)$ are independent of $\{t_j\}$ and σ.

Similarly, considering $d|H(t,\theta(t))x(t) - H(t,\theta(t))\tilde{x}^\theta(t)|^4$, we obtain

$$\begin{aligned} E^\theta|x(t)-\tilde{x}^\theta(t)|^4 \le\ & K_1\psi_j \exp\left(-\frac{\kappa(t-t_j)}{2\varepsilon}\right) \\ & + K_1\phi_j(t-t_j)\exp\left(-\frac{\kappa(t-t_j)}{2\varepsilon}\right) + O(\varepsilon^2). \end{aligned} \tag{A.2}$$

Setting $t = t_{j+1}$ in (9.1) and (9.2), respectively, we have

$$\begin{aligned} \phi_{j+1} \le\ & K_1\phi_j \exp\left(-\frac{\kappa(t_{j+1}-t_j)}{2\varepsilon}\right) + O(\varepsilon), \\ \psi_{j+1} \le\ & K_1\psi_j \exp\left(-\frac{\kappa(t_{j+1}-t_j)}{2\varepsilon}\right) \\ & + K_1\phi_j(t_{j+1}-t_j)\exp\left(-\frac{\kappa(t_{j+1}-t_j)}{2\varepsilon}\right) + O(\varepsilon^2), \end{aligned} \tag{A.3}$$

with $\phi_0 = O(\varepsilon)$ and $\psi_0 = O(\varepsilon^2)$.

Choose γ_0 large enough such that $K_1 e^{-\gamma_0/2} < 1/2$. Then for $\kappa(t_{j+1}-t_j) \ge \gamma_0\varepsilon$,

$$\begin{aligned} \phi_{j+1} &\le \frac{1}{2}\phi_j + O(\varepsilon), \\ \psi_{j+1} &\le \frac{1}{2}\psi_j + K_2\varepsilon\phi_j + O(\varepsilon^2). \end{aligned}$$

It follows by iteration that, for $j = 0, 1, \ldots,$

$$\begin{aligned} \phi_j &\le O(\varepsilon), \\ \psi_j &\le O(\varepsilon), \end{aligned}$$

with $O(\cdot)$ independent of $\{t_j\}$ and σ. In view of (9.3), we obtain

$$\begin{aligned} E^\theta|x(t)-\tilde{x}^\theta(t)|^2 &\le (K_1+1)O(\varepsilon) = O(\varepsilon), \\ E^\theta|x(t)-\tilde{x}^\theta(t)|^4 &\le (K_1+1)O(\varepsilon^2) = O(\varepsilon^2), \end{aligned}$$

with $O(\cdot)$ independent of $\{t_j\}$ and σ. □

Combining Lemmas A.1 and A.2, we have the following lemma.

LEMMA A.3. *Assume* (A1)–(A4). *Then there exist* ε_0, γ_0, *and* K *such that for each* $0 < \sigma < 1$, $0 < \varepsilon < \varepsilon_0$, *and for all* $\theta = \theta(\cdot) \in \Theta^\varepsilon_{\sigma,\gamma_0}$,

$$E^\theta \int_0^T \left|\hat{x}^\theta(t) - \tilde{x}^\theta(t)\right|^2 dt \le K\varepsilon^4.$$

The next lemma is concerned with the bounds and asymptotic estimate of $R^\theta(\cdot)$.

LEMMA A.4. *Assume* (A1)–(A4). *Then the following hold:*

(a) *There exist positive constants* c_1 *and* c_2, *independent of* $0 < \sigma < 1$, $\theta \in \Theta$, *and* $t \in [0,T]$, *such that*

$$c_1 I \le \frac{R^\theta(t)}{\varepsilon} \le c_2 I.$$

(b) *There exist* γ_0, K, *and* $\kappa > 0$ *such that, for each* $0 < \sigma < 1$ *and* $\theta \in \Theta^\varepsilon_{\sigma,\gamma_0}$, *we have*

$$\left|\frac{R^\theta(t)}{\varepsilon} - F(t,\theta(t))\right| \le K\left(\exp\left(-\frac{\kappa(t-t_j)}{\varepsilon}\right) + \varepsilon\right), \tag{A.4}$$

for $t \in [t_j, t_{j+1})$, *where* $\{t_j\}$ *is the set of jump times of* θ.

Proof. Part (a) can be shown as in [11, Lemma 2.5]. To show (b), note that on a given interval $[t_j, t_{j+1})$, $\theta(t)$ is a constant. Let

$$\eta(t) := \operatorname{tr}\left(\frac{R^\theta(t)}{\varepsilon} - F(t,\theta(t))\right)^2.$$

Recall that both $R^\theta(t)$ and $F(t,\theta(t))$ are symmetric matrices. To estimate (9.4), it suffices to obtain a similar upper bound for $\eta(\cdot)$.

Using the conditions in (A1) and (A3), the second equation in (3.2), and the inequality $a_1a_2 \le a_1^2/\varepsilon + \varepsilon a_2^2$ for any numbers a_1 and a_2, we obtain, by considering the derivative of $(R^\theta(t)/\varepsilon - F(t,\theta(t)))^2$,

$$\frac{d\eta(t)}{dt} \le -\frac{\kappa_1}{\varepsilon}\eta(t) + O(\varepsilon),$$

where $\kappa_1 > 0$ and $O(\cdot)$ are independent of the choice of $\{t_j\}$ and σ.

Using Gronwall's inequality, we obtain

$$\eta(t) \le \eta(t_j)\exp\left(-\frac{\kappa_1(t-t_j)}{\varepsilon}\right) + O(\varepsilon^2).$$

As in Lemma A.2, choose γ_0 large enough such that

$$\exp\left(-\frac{\kappa_1(t_{j+1}-t_j)}{\varepsilon}\right) \le \exp\left(-\kappa_1\gamma_0\right) \le \frac{1}{2}.$$

Then

$$\eta(t_{j+1}) \le \frac{\eta(t_j)}{2} + O(\varepsilon).$$

Note that $\eta(t_0) = \eta(0)$ is bounded. It follows that $\eta(t_j)$ is bounded for $j = 0, 1, \ldots$. Hence there exists a constant K_1 such that

$$\eta(t) \le K_1\exp\left(-\frac{\kappa_1(t-t_j)}{\varepsilon}\right) + O(\varepsilon^2).$$

Thus, taking square roots on both sides, we obtain (9.4) for some positive constants κ and K. □

LEMMA A.5. *Given $u(\cdot) \in \mathcal{U}_y$, let $x(t)$, $\tilde{x}(t)$, $\check{x}(t)$, and $\hat{x}(t)$, $t \ge 0$, denote the state, filter, intermediate filter, and conditional mean, respectively. Then, for each $n = 1, 2, \ldots$, there exists a constant K such that*

$$E\left(|x(t)|^n + |\tilde{x}(t)|^n + |\check{x}(t)|^n + |\hat{x}(t)|^n\right) \le K,$$

uniformly with respect to $u(\cdot) \in \mathcal{U}_y$.

Proof. The proof can be given similarly as in [12, Theorem 4.1]. □

LEMMA A.6. *Let $N(T)$ denote the number of jumps of $\alpha(\cdot)$ in $[0,T]$. Then, for each $\sigma > 0$ and $j = 1, 2, \ldots$, there exist positive constants ε_0 and K such that for $0 < \varepsilon < \varepsilon_0$,*

$$P\left(N(T) \ge \frac{1}{\varepsilon^\sigma}\right) \le K\varepsilon^j.$$

Proof. First of all, note that in view of the construction of Markov chains as in Davis [4], there exists a Poisson process $N_0(\cdot)$ with parameter a/ε for some $a > 0$ such that $N(t) \le N_0(t)$, $t \ge 0$. We may assume $a = 1$ for simplicity. Let n_ε denotes the integer part of $[1/\varepsilon^\sigma]$. By using the Poisson distribution of $N_0(\cdot)$, it follows

$$\begin{aligned} P\left(N(T) \ge \frac{1}{\varepsilon^\sigma}\right) &\le P\left(N_0(T) \ge \frac{1}{\varepsilon^\sigma}\right) \\ &\le e^{-T}\left(\frac{T^{n_\varepsilon}}{n_\varepsilon!} + \frac{T^{n_\varepsilon+1}}{(n_\varepsilon+1)!} + \cdots\right) \\ &\le e^{-T}\left(\frac{T^{n_\varepsilon}}{n_\varepsilon!}\right) e^T = \frac{T^{n_\varepsilon}}{n_\varepsilon!}. \end{aligned}$$

Using Stirling's formula, we have

$$\frac{T^{n_\varepsilon}}{n_\varepsilon!} \sim \frac{(Te)^{n_\varepsilon}}{n_\varepsilon^{n_\varepsilon}\sqrt{2\pi n_\varepsilon}} \le \left(\frac{Te}{n_e}\right)^{n_\varepsilon}.$$

Now choose ε_0 such that $Te/n_{\varepsilon_0} \le 1/2$. Then, for $0 < \varepsilon < \varepsilon_0$, we have

$$\left(\frac{Te}{n_\varepsilon}\right)^{n_\varepsilon} \le K\varepsilon^j.$$

The result follows. □

Acknowledgments. Comments and suggestions from the referees that lead to improvement of the paper are greatly appreciated.

REFERENCES

[1] A. Bensoussan, *Stochastic Control of Partially Observed Systems*, Cambridge University Press, Cambridge, 1992.

[2] H. A. P. Blom and Y. Bar-Shalom, *The interacting multiple model algorithm for systems with Markovian switching coefficients*, IEEE Trans. Automat. Control, AC-33 (1988), pp. 780–783.

[3] P. E. Caines and H. F. Chen, *Optimal adaptive LQG control for systems with finite state process parameters*, IEEE Trans. Automat. Control, AC-30 (1985), pp 185–189.

[4] M. H. A. Davis, *Markov Models and Optimization*, Chapman & Hall, New York, 1993.

[5] R. J. Elliott, L. Aggoun, and J. B. Moore, *Hidden Markov Models: Estimation and Control*, Springer-Verlag, New York, 1995.

[6] W. H. Fleming, D. Ji, P. Salame, and Q. Zhang, *Piecewise monotone filtering in discrete time with small observation noise*, IEEE Trans. Automat. Control, AC-36 (1991), pp. 1181–1186.

[7] W. H. Fleming and E. Pardoux, *Piecewise monotone filtering with small observation noise*, SIAM J. Control Optim., 27 (1989), pp. 1156–1181.

[8] W. H. Fleming and R. W. Rishel, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.

[9] W. H. Fleming and Q. Zhang, *Nonlinear filtering with small observation noise: Piecewise monotone observations*, Stochastic Analysis: Liber Amicorum for Moshe Zakai, E. Merzbach, A. Shwartz, and E. Mayer-Wolf, eds., Academic Press, Boston, 1991, pp. 153–168.

[10] W. H. Fleming and Q. Zhang, *Piecewise filtering with small observation noise: Numerical simulations*, in Applied Stochastic Analysis Lecture Notes in Control and Inform. Sci. 177, I. Karatzas and D. L. Ocone, eds., Springer-Verlag, New York, 1992, pp. 108–120.

[11] U. G. Haussmann and Q. Zhang, *Optimal control of diffusions with small observation noise*, in Proc. Imperial College Workshop on Applied Stochastic Analysis, Stochastics Monographs, M. H. Davis and R. J. Elliott, eds., Gordon and Breach, Yverdon, Switzerland, 1989, pp. 237–263.

[12] U. G. Haussmann and Q. Zhang, *Stochastic adaptive control with small observation noise*, Stochastics Stochastics Rep., 32, (1990), pp. 109–144.
[13] U. G. Haussmann and Q. Zhang, *Discrete time stochastic adaptive control with small observation noise*, Appl. Math. Optim., 25 (1992), pp. 303–330.
[14] O. Hijab, *The adaptive LQG problem - Part* I, IEEE Trans. Automat. Control, AC-28 (1983), pp. 171–178.
[15] G. Kallianpur, *Stochastic Filtering Theory*, Springer-Verlag, New York, 1980.
[16] N. V. Krylov, *Controlled Diffusion Processes*, Springer-Verlag, New York, 1980.
[17] H. J. Kushner, *Weak Convergence Methods and Singularly Perturbed Stochastic Control and Filtering Problems*, Birkhäuser, Boston, 1990.
[18] X. R. Li, *Hybrid estimation techniques*, in Control and Dynamic Systems, Vol. 76, C. T. Leondes, ed., Academic Press, New York, 1996.
[19] R. S. Liptser and A. N. Shiryayev, *Statistics of Random Processes,* Volume I and II, Springer-Verlag, New York, 1977.
[20] J. Picard, *Filtrage de diffiusions vectorielles faiblement bruitees*, in Proc. 7th International Conference on Analysis and Optimization of Systems (Antibes 1986), Lecture Notes in Control and Inform. Sci. 83, Springer-Verlag, New York, 1986.
[21] G. Yin and Q. Zhang, *Continuous-Time Markov Chains and Applications: A Singular Perturbation Approach*, Springer-Verlag, New York, 1997.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 5, pp. 1669–1684, September 1998

008

FEEDBACK STABILIZATION OF BILINEAR CONTROL SYSTEMS*

HUALIN WANG†

Abstract. In this paper, we study the region in which a bilinear control system is feedback stabilizable. In particular, we find a necessary and sufficient condition for feedback stabilization in terms of the Lyapunov spectrum.

Key words. Lyapunov exponent, Lyapunov spectrum, Floquet spectrum, control set, feedback stabilization

AMS subject classifications. 93B05, 93D15, 93D20

PII. S0363012996305498

1. Introduction.

In this paper, we consider the following bilinear control systems:

$$\dot{x}(t) = \left(A_0 + \sum_{i=1}^{m} u_i(t) A_i \right) x(t), \text{ in } \mathbb{R}^d$$

$u := (u_1, \ldots, u_m) \in \mathcal{U} := \{u : \mathbb{R} \to \mathbb{R}^m | u(\cdot)$ locally integrable, and $u(t) \in U$ almost everywhere (a.e.)$\}$. Here $A_0, A_1, \ldots, A_m$ are $d \times d$ matrices and $U \subset \mathbb{R}^m$ is compact and convex with $0 \in \text{int}(U)$, the interior of U.

These kinds of systems are obtained, for example, by linearizing nonlinear systems at a common fixed point with respect to the state x only. The purpose of this paper is to characterize the region in which the systems are asymptotically stabilizable using measurable or piecewise analytic feedback laws, which are defined in this paper. The methods used here are based on the Lyapunov spectrum of families of time varying matrices [CK6], in other words, on the collection of Lyapunov exponents of a class of linear differential equations, and on the construction of feedback rank controllers in [Li] which originates from [Su]. Under the accessibility rank condition in the projective space $\mathbb{P}^{d-1}$ (which is weaker than the accessibility rank condition in $\mathbb{R}^d$), the methods allow us not only to find a necessary and sufficient condition for asymptotic feedback stabilization, but also to characterize the region in which the systems are exponentially stabilizable. In particular, we prove the following result: For bilinear control systems with the control ranges satisfying the conditions stated above, exponential stabilization is equivalent to asymptotic stabilization using measurable feedbacks and also to (open loop) asymptotic null controllability.

Our paper is organized as follows: In section 2 we describe the general setup of this paper, in particular, we define the concepts used here, such as piecewise analytic feedback. We also mention some known results on projective systems of the bilinear control systems in $\mathbb{P}^{d-1}$, the projective space. In section 3 we state and prove our main results on the existence of stabilizing feedbacks for cones in the state space $\mathbb{R}^d$. In section 4 we give an example which exhibits the basic idea in this paper. The example can be understood without knowing the construction of the feedback in the proof of our main result. In the appendix, an outline of the proof of Lemma 2.8 is given.

*Received by the editors June 24, 1996; accepted for publication (in revised form) July 1, 1997; published electronically June 22, 1998. This research was partially supported by ONR grants N00014-93-1-0868 and N00014-96-1-0279.

http://www.siam.org/journals/sicon/36-5/30549.html

†Department of Mathematics, Iowa State University, Ames, IA 50011 (hlwang@iastate.edu).

A paper by Clarke et al. [CLSS] shows the equivalence of asymptotic controllability and feedback stabilization. This paper deals with nonlinear systems and their global stabilization. The feedback employed by them is discretized (in time) and their concept of stability is that of "practical stability," i.e., stabilization into arbitrarily small neighborhoods of the fixed point. In contrast, we deal with bilinear systems. Our feedback concept is the classical one (measurable functions on the state space that are not discretized) and stability is global asymptotic (or exponential) stability of the fixed point. Using spectral methods, Gruene constructs discretized controls for nonlinear systems that yield classical asymptotic stability. For semilinear control systems the result can be found in [Gr].

2. Setup and preliminaries. Let U, the control range, be compact, convex with $0 \in \text{int}U \subset \mathbb{R}^m$ and $\mathcal{U}$ be the space consisting of locally integrable open loop control functions taking values in U. Consider

$$\dot{x} = \left(A_0 + \sum_{i=1}^{m} u_i(t) A_i \right) x, \ x \in \mathbb{R}^d, \tag{B}$$

where $u(\cdot) := (u_1(\cdot), \dots, u_m(\cdot)) \in \mathcal{U}$ and A_i are $d \times d$ matrices for $i = 0, 1, \dots, m$. By $A(x, u)$ we denote the right-hand side of (B).

There are many papers devoted to feedback stabilization of (B) (cf., e.g., [AG] and the references listed therein). For two-dimensional systems, a Lyapunov function approach for systems with unconstrained control range is presented in [CSV], while the properties of the Lyapunov spectrum for $d = 2$ are exploited in [CK5] to yield characterizations of feedback stabilizability. Equation (B) can be studied via the associated (angular) system on the projective space $\mathbb{P}^{d-1}$ obtained by identifying opposite points on the sphere in $\mathbb{R}^d$ [CK2]:

$$\dot{s} = h_0(s) + \sum_{i=1}^{m} u_i(t) h_i(s), \ s \in \mathbb{P}^{d-1}, \tag{PB}$$

where $s = \frac{x}{|x|} \in \mathbb{P}^{d-1}$ and $h_i(s) = [A_i - s^T A_i s \cdot I]s, i = 0, 1, \dots m$. Here $|\cdot|$ is the 2-norm on $\mathbb{R}^d$, I is the $d \times d$ identity matrix, and ^{T} denotes transposition. By $h(s, u)$ we denote the right-hand side of (PB).

Assume the accessibility rank condition for (PB), i.e.,

$$\dim Lie\ [h(\cdot, u), u \in U](s) = d - 1 \tag{H}$$

for all $s \in \mathbb{P}^{d-1}$. (Here $\dim Lie\ (X)(s)$ denotes for a set X of vector fields, the dimension of the distribution generated by the Lie algebra $Lie\ (X)$ in the tangent space at the point s.) In order to study the system (PB) we introduce some notations and concepts.

For any point $p \in \mathbb{P}^{d-1}$ and $u \in \mathcal{U}$, let $s(t, p, u)$ denote the solution of (PB) with $s(0, p, u) = p$. For example,

$$\frac{d}{dt}(s(t, p, u)) = h_0(s(t, p, u)) + \sum_{i=1}^{m} u_i(t) h_i(s(t, p, u))$$

holds for all $t \in \mathbb{R}$ except on a set of Lebesque measure zero, where $u = (u_1(\cdot), \dots, u_m(\cdot)) \in \mathcal{U}$.

Define

$$\mathcal{O}^+(p) = \{q \in \mathbb{P}^{d-1} | s(t,p,u) = q \text{ for some } u \in \mathcal{U} \text{ and } t \geq 0\}$$

and

$$\mathcal{O}^-(p) = \{q \in \mathbb{P}^{d-1} | s(t,q,u) = p \text{ for some } u \in \mathcal{U} \text{ and } t \geq 0\}.$$

DEFINITION 2.1. *A set $D \subset \mathbb{P}^{d-1}$ is called a control set of the control system* (PB) *if*

1) *$D \subset c\ell\mathcal{O}^+(p)$ for every $p \in D$ where $c\ell\mathcal{O}^+(p)$ denotes the closure of $\mathcal{O}^+(p)$;*

2) *for every $p \in D$ there is $u \in \mathcal{U}$ such that the corresponding solution, $s(t,p,u)$, of* (PB) *satisfies $s(t,p,u) \in D$ for all $t \in \mathbb{R}$;*

3) *D is maximal (with respect to set inclusion) with the properties* 1) *and* 2).

A main control set is a control set with nonempty interior.

In [CK4] *it is proved that under assumption* (H) *the control system* (PB) *has $k (1 \leq k \leq d)$ main control sets which are linearly ordered, say, $D_1 \prec D_2 \prec \cdots \prec D_k$, where the order is defined by*

$$D_i \prec D_j \quad \text{if and only if there exist } p \in D_i, q \in D_j, \text{ and } u \in \mathcal{U} \text{ such that } s(t,p,u) = q \text{ for some } t \geq 0.$$

Our results will be closely related to various spectral concepts for the bilinear system (B). Here we include the following definition.

DEFINITION 2.2. *Let $u \in \mathcal{U}$, and let $\psi(t,x,u)$ solve* (B) *for all $t \geq 0$ except on a set of Lebesgue measure zero with $\psi(0,x,u) = x \neq 0$. Let*

$$\lambda(u,x) := \limsup_{t \to \infty} \frac{1}{t} \log |\psi(t,x,u)|.$$

Let D be a main control set of (PB), *and let $c\ell(D)$ denote the closure of D. The following set is called the Lyapunov spectrum over D:*

$$\Sigma_{LY}(D) := \left\{ \lambda(u,x) | (u,x) \in \mathcal{U} \times \mathbb{R}^d \text{ s.t. } \frac{\psi(t,x,u)}{|\psi(t,x,u)|} \in c\ell(D) \text{ for all } t \geq 0 \right\}.$$

The following set is called the Floquet spectrum over D:

$$\Sigma_{FL}(D) := \left\{ \lambda(u,x) | (u,x) \in \mathcal{U} \times \mathbb{R}^d, u \text{ piecewise constant periodic with period } T \text{ s.t. } s\left(t, \frac{x}{|x|}, u\right) := \frac{\psi(t,x,u)}{|\psi(t,x,u)|} \in \operatorname{int}(D) \text{ for all } t \geq 0 \text{ and } s\left(T, \frac{x}{|x|}, u\right) = \frac{x}{|x|} \right\}.$$

PROPOSITION 2.3 (see [CK6]). *Under assumption* (H) *we have the following:*

1) *$c\ell\left(\Sigma_{FL}(D_i)\right) =: I_i$ are bounded intervals for $i = 1, \ldots, k$.*

2) *If $D_i \prec D_j$, then $\inf I_i \leq \inf I_j$ and $\sup I_i \leq \sup I_j$.*

An outline of the proof is included here. (The detailed proof is given in [CK6], the proof of Theorem 4.4.)

Proof.

1) Abbreviate $D := D_i$ and let for $j = 1, 2$ $\lambda_j \in \Sigma_{FL}(D)$ with corresponding points $(u_j, p_j) \in \mathcal{U} \times \text{int}(D)$ and periods $T_j \geq 0$. Since $\{s(t, p_j, u_j) | t \geq 0\}$ are compact subsets of $\text{int}(D)$, there exist $T > 0$ and $v_j \in \mathcal{U}$, with $s(t_1, p_1, v_1) = p_2$ and $s(t_2, p_2, v_2) = p_1$ for some $t_1, t_2 \leq T$. For $m, n \in \mathbb{N}$ define a control $u^{m,n}$ via concatenation on the time interval $[0, t^{m,n})$ with $t^{m,n} = mT_1 + t_1 + nT_2 + t_2$ as

$$u^{m,n} = v_2 \cdot \underbrace{n \text{ -times}} \to u_2 \cdot \cdots \cdot u_2 \cdot v_1 \cdot \underbrace{m \text{ -times}} \to u_1 \cdot \cdots \cdot u_1,$$

and on $\mathbb{R}_+$ as the $t^{m,n}$-periodic continuation. By this construction and by virtue of assumption (H), it can be shown that Σ_{FL} is dense in the interval between λ_1 and λ_2.

2) The inequalities follow directly from the construction of the main control sets in the proof of Theorem 3.10 (ii) in [CK4]. □

Based on the following assumption we are going to investigate the region in which the system (B) is stabilizable with state feedback laws. There exists the index $i_0 \in \{1, \ldots, k\}$ such that

$$i_0 = \max\{i | \ \inf I_i < 0\}. \tag{A}$$

Only under some specific circumstances can the whole space or a whole neighborhood of the origin be stabilized. In other words, in general, only part of the space is feedback stabilizable with respect to a given equilibrium of a given control system. We introduce the following definition which allows us to study feedback stabilization in part of the space, specifically, in some directions from the equilibrium (the origin in our paper). So typically, the set of all directions in which (B) is feedback stabilizable is a cone.

DEFINITION 2.4. *For the control system* (B)*, we say that the system is feedback stabilizable in a cone K with the vertex at the origin, if there is a measurable function $u(x)$ defined on K such that*

1) *for any $x \in K$ there is a unique function $\varphi(t, x)(t \geq 0)$ such that*

$$\frac{d\varphi(t,x)}{dt} = A(\varphi(t,x), u(\varphi(t,x))) \text{ and } \varphi(0,x) = x$$

for all $t \geq 0$ except on a set of Lebesgue measure zero;

2) *all solutions starting at any point in K remain in K for all $t \geq 0$;*

3) *any solution starting at any point in K approaches 0 as $t \to \infty$;*

4) *for any cone $\widetilde{K} \subset K$ such that $\widetilde{K} \cap \mathbb{S}^{d-1}$ is compact, ($\mathbb{S}^{d-1}$ is the $d-1$ dimensional unit sphere in $\mathbb{R}^d$), and the following holds: For all $\varepsilon > 0$ $\exists \delta > 0$ s.t. all solutions stay in $K \cap B(\varepsilon)$ for all $t \geq 0$ whenever they start in $\widetilde{K} \cap B(\delta)$, where $B(r)$ denotes the open ball in $\mathbb{R}^d$ centered at the origin, with radius r.*

The system is said to be exponentially feedback stabilizable if 1), 2), 4)*, and the following condition are satisfied:*

3′) *There is $\gamma > 0$ s.t. any solution $\varphi(t, x)$ with $x \in K$ satisfies* $\lim_{t\to\infty} \varphi(t, x)e^{\gamma t} = 0$.

If K is maximal with respect to set inclusion, then we call K the asymptotically (exponentially, respectively) stabilizable region. In the first case we may simply say the stabilizable region. The reason for introducing the cone $\widetilde{K}$ in Definition 2.4.4 *is explained in section* 4.

Remark 2.5. Notice that if $\sup I_k < 0$, then $\lambda(u,x) < 0$ for all $(u,x) \in \mathcal{U} \times \mathbb{R}^d$ (cf. [CK2]); hence any constant u will stabilize (B) in $\mathbb{R}^d$. And if $\inf I_1 > 0$, then $\lambda(u,x) > 0$ for all $(u,x) \in \mathcal{U} \times \mathbb{R}^d$ (cf. [CK2]); hence, no (measurable) $u(x)$ defined on any subset of $\mathbb{R}^d$ will stabilize (B).

So we would like to know what happens if $i_0 \geq 1$ and $\sup I_k \geq 0$. For this we introduce the following definitions.

DEFINITION 2.6. *Let F be a subset of $\mathbb{P}^{d-1}$. The domain of attraction of F is defined as*

$$\mathcal{A}(F) := \{q \in \mathbb{P}^{d-1} | s(t,q,u) \in F \text{ for some } u(\cdot) \in \mathcal{U} \text{ and for some } t \geq 0\},$$

where $s(\cdot,q,u)$ denotes the solution of (PB) *with control $u(\cdot) \in \mathcal{U}$ and $s(0,q,u) = q$.*

In particular, if $F = \{p\}$ for $p \in \mathbb{P}^{d-1}$, we simply write $\mathcal{A}(F)$ as $\mathcal{A}(p)$, which is $\mathcal{O}^-(p)$. If $F = D$ a main control set contained in $\mathbb{P}^{d-1}$, then $\mathcal{A}(D) = \mathcal{A}(p)$ for any $p \in \operatorname{int}(D) : \mathcal{A}(p) \subset \mathcal{A}(D)$ is trivial. For the other direction, let $q \in \mathcal{A}(D)$. Then there exist $u_1 \in \mathcal{U}$ and $t \geq 0$ such that $z := s(t,q,u_1) \in D$. Since $p \in \operatorname{int}(D) \subset \mathcal{O}^+(z), p$ can be reached from z by some $u \in \mathcal{U}$ within finite time. Therefore, $\mathcal{A}(D) \subset \mathcal{A}(p)$. Briefly, we have $\mathcal{A}(D) = \mathcal{A}(p) = \mathcal{O}^-(p)$ for any $p \in \operatorname{int}(D)$.

DEFINITION 2.7 (see [Su]). *A piecewise analytic vectorfield on a real analytic manifold M is a quadruple $(\mathcal{L}, (\mathcal{L}_1, \mathcal{L}_2), \{V_S\}_{S \in \mathcal{L}_1}, E)$ where*

1) *$\mathcal{L}$ is an analytic stratification of M;*

2) *$(\mathcal{L}_1, \mathcal{L}_2)$ is a partition of $\mathcal{L}$ into two classes, i.e., $\mathcal{L} = \mathcal{L}_1 \cup \mathcal{L}_2, \mathcal{L}_1 \cap \mathcal{L}_2 = \emptyset$;*

3) *for each $S \in \mathcal{L}_1, V_S$ is an analytic vectorfield on S;*

4) *E is a map which assigns to each point p in a stratum $S \in \mathcal{L}_2$, a stratum $E(p) \in \mathcal{L}_1$;*

5) *for each $p \in S \in \mathcal{L}_1$, if we let γ denote the integral curve of V_S through p, then either $\gamma(t)$ is defined for all $t \geq 0$, or else, if γ is defined up to a time $T > 0$, and if $\gamma(t), 0 \leq t < T$, remains in a compact subset of M, then $\lim_{t \to T, t<T} \gamma(t)$ exists;*

6) *for each $p \in S \in \mathcal{L}_2$, there is a unique integral curve γ of $V_{E(p)}$ such that $\lim_{t \to 0, t>0} \gamma(t) = p$.*

In our situation we have $M = \mathbb{R}^d \setminus \{0\}$ or $M = \mathbb{P}^{d-1}$, and V is called (*piecewise analytic*) *feedback controller* if for every $p \in S \in \mathcal{L}_1$, there is $u \in U$ such that $V_S(x) = A(x,u), x \in \mathbb{R}^d$ for the system (B) or $V_S(p) = h(p,u), p \in \mathbb{P}^{d-1}$ for system (PB).

The basic idea of this definition is to partition M into two classes of connected real analytic submanifolds (called the strata). On each stratum in one class (i.e., in $\mathcal{L}_1$) an analytic vector field is well defined; hence, the integral curve of the vector field through each point in the stratum is uniquely defined, and in each point of a stratum in the other class an exit rule E is specified in such a way that there is a unique trajectory of $V_{E(p)}$ that starts at p.

The following lemma is one of the basic ingredients of our feedback construction. It extends the results of Sussmann [Su] on the existence of piecewise analytic feedback controllers to the case of systems that are not completely controllable. An outline of the proof is given in the Appendix.

LEMMA 2.8 (see [Li]). *Let $p \in \operatorname{int}(D_j)$, where D_j is a main control set. Then there exists a piecewise analytic feedback controller V_p for the system* (PB) *in $\mathbb{P}^{d-1}$ such that p can be reached from any point of $\mathcal{A}(p)$ in finite time.*

For any subset E of $\mathbb{P}^{d-1}$ let $\mathcal{E}$ be the largest subset of $\mathbb{S}^{d-1}$ whose identification in $\mathbb{P}^{d-1}$ is E, and let K_E be the corresponding cone in $\mathbb{R}^d$, i.e., $K_E = \{x \in \mathbb{R}^d | x = r \cdot e, r \in \mathbb{R} \setminus \{0\}, e \in \mathcal{E}\}$.

Before giving the next definition, let us recall what piecewise analytic feedback

controller means. Sussmann's definition of (piecewise analytic) feedback controller guarantees two things:

1) The vector field is almost surely (a.s.) analytic;

2) There is a unique trajectory of the vector field through any point.

In the next definition we maintain these two aspects with uniquely defined trajectories for all $t \geq 0$. Nevertheless, we relax the exit rule in the following sense: If a trajectory reaches a stratum $S \in \mathcal{L}_2$, it leaves immediately according to a rule (which will be clear from the construction), or it never gets back into any stratum $S \in \mathcal{L}_1$.

DEFINITION 2.9. *For the control system* (B)*, let K be a cone in $\mathbb{R}^d, u : K \to \mathbb{R}^m$ with $u(x) \in U$ for all $x \in K$. The function u is called a piecewise analytic feedback on K if*

1) *K has an analytic stratification $\mathcal{L}$ such that $\mathcal{L}$ is partitioned into two disjoint classes, say, $(\mathcal{L}_1, \mathcal{L}_2)$ with $\dim S = d$ for any $S \in \mathcal{L}_1$ and $\dim S < d$ for any $S \in \mathcal{L}_2$;*

2) *a $(x, u(x))$ is analytic on any $S \in \mathcal{L}_2$;*

3) *there is a rule so that, if a trajectory reaches a stratum $S \in \mathcal{L}_2$ in finite time, then it either leaves S immediately (i.e., for $p \in S$ we have the existence of an open time interval $(0, \tau)$ with $\varphi(t, p) \in S_1 \in \mathcal{L}_1$ for all $t \in (0, \tau)$ and some $S_1 \in \mathcal{L}_1$), or it stays in a lower dimensional submanifold such that for any $x_0 \in K$ the equation $\dot{x} = A(x, u(x))$ has a unique solution $\varphi(t, x_0)$ with $\varphi(0, x_0) = x_0$ and $\varphi(t, x_0) \in K$ for $t \geq 0$.*

Thus we may have different rules for different strata $S \in \mathcal{L}_2$, but for each point $p \in S \in \mathcal{L}_2$ we either assign a stratum $E(p) \in \mathcal{L}_1$ (see Definition 2.7), or $E(p) \in \mathcal{L}_2$ is the stratum S of which p is an element. In the latter case we require that $\varphi(t, p) \in S$ for all $t \geq 0$. Therefore, a piecewise analytic feedback u in the sense of Definition 2.9 guarantees global (in K) existence and uniqueness of solutions of $\dot{x} = A(x, u(x))$ for $t \geq 0$.

3. The existence of stabilizing feedbacks. In this section we prove the existence of stabilizing feedbacks under assumption (H), the accessibility condition on $\mathbb{P}^{d-1}$. Recall the definition of the index $i_0 = \max\{i | \inf \Sigma_{FL}(D_i) < 0\}$, and assume that $i_0 \in \{1, \ldots, k\}$.

THEOREM 3.1. *Under assumptions* (H) *and* (A) *there exists a piecewise analytic feedback $u(x)$ defined in $K_{A(D_{i_0})}$ such that*

$$\dot{x} = A(x, u(x)) \tag{FB}$$

is exponentially stable with respect to the origin in $K_{A(D_{i_0})}$.

Proof. Under assumption (A) we have $\inf \Sigma_{FL}(D_i) < 0$ for $1 \leq i \leq i_0$ and $\inf \Sigma_{FL}(D_i) \geq 0$ for $i_0 < i \leq k$. By the definition of $\Sigma_{FL}(D_{i_o})$, there exist a piecewise constant and periodic control $u_0 \in \mathcal{U}$ and $p \in \text{int}(D_{i_0})$ such that the solution of $\dot{s} = h(s, u_0)$, say, $s_p(t, p, u_0) \in \text{int}(D_{i_0})$ for all $t \geq 0$ and $\lambda(u_0, x_0) < 0$ where $\frac{x_0}{|x_0|} = p$ (or we can simply write $\lambda(u_0, p) < 0$ since $\lambda(u_0, x) = \lambda(u_0, y)$ if $x/|x| = y/|y|$). Let T be the period of both $u_0(\cdot)$ and $s_p(\cdot, p, u_0)$. So we have $s_p(0, p, u_0) = s_p(T, p, u_0) = p$. Let $\psi_0(t, x, u_0)$ be the solution of $\dot{x} = A(x, u_0)$ with $\psi_0(0, x_0, u_0) = x_0$. So $s_p(t, p, u_0)$ is the projection of $\psi_0(t, x_0, u_0)$ onto the projective space $\mathbb{P}^{d-1}$. Since $\lambda(u_0, x_0) < 0, \psi_0(t, x_0, u_0) \to 0$ exponentially as $t \to \infty$. From now on, p, x_0, s_p, and ψ_0 all are fixed throughout the proof.

By Lemma 2.8 there exists a piecewise analytic feedback controller V_p which is well defined on $\mathcal{A}(D_{i_o}) = \mathcal{A}(p)$. This implies the following:

1) There is an analytic stratification of $\mathcal{A}(p)$, say, $\mathcal{L}$;

2) $\mathcal{L}$ is partitioned into two disjoint classes of strata, say, $(\mathcal{L}_1, \mathcal{L}_2)$ with $\dim S = d-1$ for any $S \in \mathcal{L}_1$ and $\dim S < d-1$ for any $S \in \mathcal{L}_2$;

3) $V_p(q) = h(q, u_q)$ for some $u_q \in U$;

4) $V_p(\cdot)$ is analytic in $S \in \mathcal{L}_1$;

5) Instruction for motion for any point in $S \in \mathcal{L}_2$ is given so that p can be reached from any point of $\mathcal{A}(p)$ in finite time via a unique trajectory.

Now we can define $u(x)$ for any $x \in K_{\mathcal{A}(p)}$ (briefly written as K) as follows:

$$u(x) = \begin{cases} u_0(t), \ t = \min\left\{\tau | s_p(\tau, p, u_0) = \dfrac{x}{|x|}\right\}, \\ u_q, \quad \text{if } s_p(t, p, u_0) \neq \dfrac{x}{|x|} \text{ for all } t \geq 0, \end{cases}$$

where $q = \frac{x}{|x|}$ and $u_q \in U$ is specified by (3) above, i.e., by $V_p(q) = h(q, u_q)$. It is clear that $u(x)$ is well defined in the cone K. Note that $u_0(\cdot) \in \mathcal{U}$ is piecewise constant and for any constant c the vector field $h(s, c)$ is analytic in x; hence,

$$\mathcal{N} := \{s_p(t, p, u_0) | 0 \leq t \leq T\}$$

is a closed analytic subset. Therefore it is a subanalytic set. Let $\mathcal{L}$ be the (sub-) analytic stratification specified in 1) above. Let $\mathcal{B}$ the set of all strata belonging to $\mathcal{L}$, i.e.,

$$\mathcal{B} := \{S | S \in \mathcal{L}_1 \text{ or } S \in \mathcal{L}_2\}.$$

Now we are in the situation that $\mathcal{N}$ intersects some strata in $\mathcal{B}$, which may partition some strata in $\mathcal{B}$ into subanalytic sets. Specifically, if $S \in \mathcal{B}$ and $S \cap \mathcal{N} \neq \emptyset$, then $S \cap \mathcal{N}$ = the union of some subanalytic sets, say, $S \cap \mathcal{N} = \bigcup_{i=1}^{I_S} S_i$. Let $\mathcal{T}$ be the family of subanalytic subsets such that $\mathcal{T}$ contains $S \in \mathcal{B}$ if $S \cap \mathcal{N} = \emptyset$ and all $S_i (i = 1, \ldots, I_S)$ if $S \cap \mathcal{N} = \bigcup_{i=1}^{I_S} S_i \neq \emptyset$. Thus $\mathcal{T}$ is a locally finite family of subanalytic subsets of $\mathcal{A}(p)$. By Theorem 4.2 in [Ha] there exists a subanalytic stratification, say, $\mathcal{L}^*$ of $\mathcal{A}(p)$, compatible with $\mathcal{T}$, i.e., every $S \in \mathcal{T}$ is a union of strata of $\mathcal{L}^*$.

Let $\widetilde{\mathcal{A}}$ be the largest subset of $\mathbb{S}^{d-1}$ whose identification in $\mathbb{P}^{d-1}$ is $\mathcal{A}(p)$. Since $\{x = (x_1, \ldots, x_d) \in R^d | \ |x| = 1, x_d = 0\}$ is subanalytic in $\mathbb{S}^{d-1}$, $\mathcal{N}, \mathcal{B}, \mathcal{L}, \mathcal{T}$, and $\mathcal{L}^*$ in $\mathbb{P}^{d-1}$ all have their counterparts in $\mathbb{S}^{d-1}$. (Recall $\mathbb{P}^{d-1}$ is obtained by identifying the opposite points in $\mathbb{S}^{d-1}$). In particular, let $\overline{\mathcal{L}^*}$ be the subanalytic stratification of $\widetilde{\mathcal{A}}$. Notice K is the cone in $\mathbb{R}^d$ containing all x with $\frac{x}{|x|} \in \widetilde{\mathcal{A}}$. Hence, K has a subanalytic stratification whose strata are of the form: $\{r \cdot S | r > 0, S \in \overline{\mathcal{L}^*}\}$.

We need to prove that the feedback function $u(x)$ defined above does exponentially stabilize the system (B) in K (in the sense of Definitions 2.4 and 2.9).

Note that $u(x)$ is constant on each ray from the origin so it makes sense to consider the equation $\dot{s} = h(s, u(s))$ in $\mathcal{A}(p) \subset \mathbb{P}^{d-1}$. First, let us look at how trajectories evolve in $\mathcal{A}(p)$. For any $q \in \mathcal{A}(p)$ the trajectory of $h(s, u(s))$ from q first follows the vectorfield $V_p(s)$. In finite time it hits $\mathcal{N} = \{s_p(t, p, u_0) | 0 \leq t \leq T\}$, and then it follows the periodic solution $s_p(t, p, u_0)$ forever. Since $V_p(s)$ guarantees a unique trajectory from any point $q \in \mathcal{A}(p)$, there is a unique trajectory of $h(s, u(s))$ from q for positive time. Earlier we obtained the subanalytic stratification $\mathcal{L}^*$ of $\mathcal{A}(p)$. Let $\mathcal{L}^* = (\mathcal{L}_1^*, \mathcal{L}_2^*)$. If the trajectory from q reaches any stratum in $\mathcal{L}_2^*$ at r, then the exit rule is given in Lemma 2.8 if $r \notin \mathcal{N}$, or the trajectory follows $\mathcal{N}$ forever if $r \in \mathcal{N}$.

Since $V_p(s)$ steers all points in $\mathcal{A}(p)$ into p in finite time, any trajectory of $h(s,u(s))$ will reach $\mathcal{N}$ in finite time (since $p \in \mathcal{N}$). Since $u(r\cdot x) = u(x)$ for $r \in \mathbb{R}\setminus\{0\}$, (FB) can be projected into the projective space $\mathbb{P}^{d-1}$. The projection yields

$$\text{(FP)} \qquad \dot{s} = h(s,u(s)),\ u(s) = u\left(\frac{x}{|x|}\right),\ s = \frac{x}{|x|} \in \mathbb{P}^{d-1}$$

in $\mathcal{A}(p)$. Let $\varphi(t,x)(t \geq 0)$ be a solution of (FB) with $\varphi(0,x) = x$, i.e., $d\varphi/dt = A(\varphi, u(\varphi))$ a.e. For any $r \neq 0 \quad d(r\cdot\varphi)/dt = A(r\cdot\varphi, u(r\cdot\varphi))$ a.e. Hence, $r\cdot\varphi(t,x)$ is also a solution of (FB) with $r\cdot\varphi(0,x) = r\cdot x$. From this and the fact that (FP) has a unique solution for any initial point $q \in \mathcal{A}(p)$, it follows that (FB) has a unique solution for any initial point $x \in K$ for $t \geq 0$. Since $s_p(t,p,u_0)$ is the solution of $\dot{s} = h(s,u_0(t))$, the projection of $r\cdot\psi_0(t,x_0,u_0)$ for $r \in \mathbb{R}\setminus\{0\}$ is $s_p(t,p,u_0)$. Thus any solution in K will reach one of $r\cdot\psi_0(\cdot,x_0,u_0)$ for $r \in \mathbb{R}$ in finite time and then follow it. Since $r\cdot\psi_0(t,x_0,u_0) \to 0$ as $t \to \infty$ (with the same rate), conditions 1), 2), and 3) in Definition 2.4 are satisfied. To prove 4) let $\widetilde{K}$ be a subset of K considered in the definition. For $\alpha > 0$ all solutions starting at any points in $\widetilde{K} \cap B(\alpha)$ can be uniformly bounded by a constant $M(\alpha, \widetilde{K})$. This can be proved as follows: Consider $B_K := \widetilde{K} \cap \{x \in \mathbb{R}^d || x| = \alpha\}$, which is compact. B_K intersects only a finite number of strata of $\overline{\mathcal{L}^*}$. Let S be a stratum of the first kind and assume $S \cap B_K \neq \emptyset$. Since $A(x,u(x))$ is analytic in S, all solutions starting in $S\cap B_K$ leave S in times bounded by some constant $T(S)$, and the solutions are uniformly bounded by a constant $M(S)$. From the construction of the feedback controller in Theorem 9 [Su] or Lemma 2.8, any trajectory passes through only a finite number of strata before it hits $\{r\cdot\psi_0(t,x_0,u_0)|r \neq 0\}$. Hence all solutions starting in B_K are uniformly bounded by a constant, say, $M(\alpha,\widetilde{K})$, and they hit $\{r\cdot\psi_0|(t,x_0,u_0)|r \neq 0\}$ in times bounded by a constant. Since all solutions of $\dot{x} = A(x,u(x))$ satisfy $r\varphi(t,x) = \varphi(t,rx)$, all solutions starting in $\widetilde{K} \cap B(\alpha)$ are also uniformly bounded by $M(\alpha,\widetilde{K})$. Again in virtue of $r\varphi(t,x) = \varphi(t,rx)$ for $r \in \mathbb{R}$ and $x \in K, M(\alpha,\widetilde{K})$ can be chosen so that $M(r\alpha,\widetilde{K}) = rM(\alpha,\widetilde{K})$. This completes the proof of 4). □

In Theorem 3.1 we used assumption (A) which is related to the Floquet spectrum. An assumption related to the Lyapunov spectrum can be similarly stated as: There exists $i_0 \in \{1,\ldots,k\}$ s.t.

$$\text{(A}') \qquad i_0 = \max\{i|\ \inf \Sigma_{LY}(D_i) < 0\}.$$

There is a similar result using this assumption. Before we state the result, let us recall the following definition.

DEFINITION 3.2. *Let D be a main control set of* (PB). *Define the following subsets of the boundary ∂D:*

$$\begin{aligned}
\Gamma(D) = \{p \in \partial D|\ &\textit{there exist } q \in \operatorname{int}(D) \textit{ and } u \in \mathcal{U} \textit{ with} \\
&p = s(t,q,u) \textit{ for some } t > 0\}, \\
\Gamma^*(D) = \{p \in \partial D|\ &\textit{there exist } q \in \operatorname{int}(D) \textit{ and } u \in \mathcal{U} \textit{ with} \\
&q = s(t,p,u) \textit{ for some } t > 0\}, \\
\widetilde{\Gamma}(D) = \{p \in \partial D|&\mathcal{O}^+(p) \cap \operatorname{int}(D) = \emptyset \textit{ and } \mathcal{O}^-(p) \cap \operatorname{int}(D) = \emptyset\}.
\end{aligned}$$

$\Gamma(D), \Gamma^*(D)$, and $\widetilde{\Gamma}(D)$ are called exit, entrance, and tangential boundary, respectively.

In other words, the boundary of a main control set can be classified into three disjoint classes. From an exit boundary point at least one trajectory leaves $c\ell(D)$ *immediately; from an entrance boundary point at least one trajectory enters into* $\mathrm{int}(D)$ *immediately and the rest of the boundary is the tangential boundary.*

There is the following fact about these concepts.

PROPOSITION 3.3. 1) *The sets* $\Gamma(D)$ *and* $\Gamma^*(D)$ *are open in* ∂D *and* $\widetilde{\Gamma}(D)$ *is closed in* ∂D.

2) *Under* (H), $\widetilde{\Gamma}(D) \subset c\ell(\Gamma^*(D)) \cap c\ell(\Gamma(D))$, *in particular,* $\mathrm{int}_{\partial D}(\widetilde{\Gamma}(D)) = \emptyset$.

3) *For the main control set* D_1 *(the minimal one in the linear ordering), it holds that* $\partial D_1 = \Gamma(D_1)$, *and for the main control set* D_k *(the maximal one in the linear ordering), it holds that* $\partial D_k = \Gamma^*(D_k)$.

The proof is given in [CK4].

THEOREM 3.4. *Under assumptions* (H) *and* (A′), *there exists a measurable feedback law defined in a cone* K, *which exponentially stabilizes* (B) *in* K *if at least one negative Lyapunov exponent can be obtained from a point* $p \in \mathrm{int}(D_{i_o}) \cup \Gamma(D_{i_0})$, *in particular, if* $i_0 = 1$ *or* $\sup \Sigma_{LY}(D_{i_0}) < 0$. *In this case* $K = K_{\mathcal{A}(p)}$.

Proof. The proof is similar to that of Theorem 3.1 except that the periodic solution $s_p(t, p, u_0)$ is replaced by a solution, say, $s_p(t) \subset \mathrm{int}(D_{i_o}) \cup \Gamma(D_{i_0})$ whose corresponding solution of (B) exponentially approaches 0 as $t \to \infty$. This is because if there exists $\overline{p} \in \mathrm{int}(D_{i_o}) \cup \Gamma(D_{i_0})$ with $\lambda(\overline{u}_0, \overline{x}_0) < 0$ for some $\overline{u}_0 \in \mathcal{U}$ and $\frac{\overline{x}_0}{|\overline{x}_0|} = \overline{p}$, then the solution of $\dot{x} = A(x, \overline{u}_0)$, say, $\psi_{\overline{x}_0}(t) \to 0$ as $t \to \infty$. The projection of $\psi_{\overline{x}_0}(t)$, $s_{\overline{p}}(t)$, is a solution of $\dot{s} = h(s, \overline{u}_0)$. By the definition of $\lambda(\overline{u}_0, \overline{x}_0)$ we know $s_{\overline{p}}(t) \in c\ell(D_{i_0})$ for $t \geq 0$. Notice that $\Gamma(D_{i_0})$ is the exit boundary. Hence there exist $p \in \mathrm{int}(D_{i_0})$ and $\overline{u}_1 \in \mathcal{U}$ which steers p into $\overline{p}$ in finite time τ. Let u_0 be the concatenation of $\overline{u}_1(t)$ for $0 \leq t \leq \tau$ and $\overline{u}_0(t + \tau)$ for $t \geq 0$. Then the solution of $\dot{s} = h(s, u_0)$, say, $s_p(t) \in c\ell(D_{i_0})$ and $s_p(0) = p$. Let $\psi(t)$ be the solution of $\dot{x} = A(x, u_0)$, whose projection into $\mathbb{P}^{d-1}$ is $s_p(t)$. So $\psi(t) \to 0$ exponentially as $t \to \infty$ since $\lambda(u_0, x_0) = \lambda(\overline{u}_0, \overline{x}_0) < 0$, where $x_0/|x_0| = p$ and $\psi(0) = x_0$. Once we have these two solutions, $s_p(t)$ and $\psi(t)$, the rest of the proof follows the same lines in the proof of Theorem 3.1. Since $s_p(t)$ for $t \geq 0$ may not be piecewise analytic, the resulting vectorfield $A(x, u(x))$ may not be piecewise analytic. But $u(x)$ is measurable, so is $A(x, u(x))$. □

Remark 3.5. The assumptions (A) and (A′) agree if $c\ell\Sigma_{FL}(D_{i_0}) = \Sigma_{LY}(D_{i_0})$. Conditions for equality of the Floquet and the Lyapunov spectrum can be found in [CK6].

The following corollary leads to the main result (stated in Corollary 3.8). It gives a necessary and sufficient criterion for exponential feedback stabilization of bilinear control systems in $\mathbb{R}^d$.

COROLLARY 3.6. *Under assumption* (H), *there exists a measurable feedback law which exponentially stabilizes the system* (B) *in* $\mathbb{R}^d$ *if and only if at least one negative Lyapunov exponent can be obtained form* $\mathrm{int}(D_k)$, *here* D_k *is the maximal main control set.*

Proof. Let $p \in \mathrm{int}(D_k)$ and $\lambda(u, p) < 0$ for some $u \in \mathcal{U}$. Since D_k is the maximal main control set, $\mathcal{A}(D_k) = \mathcal{A}(p) = \mathbb{P}^{d-1}$. So $K = \mathbb{R}^d \setminus \{0\}$ in Theorem 3.3 and $u(x)$ is well defined in K. Thus $u(0) = 0$ completes the definition of $u(x)$ in $\mathbb{R}^d$.

Conversely, assume $\lambda(u, p) \geq 0$ for all $p \in \mathrm{int}(D_k)$ and $u \in \mathcal{U}$. By 3) of Proposition 3.3, $\partial(D_k) = \Gamma^*(D_k)$. That is to say that any trajectory starting in D_k never leaves D_k. Furthermore, any trajectory starting in $\mathrm{int}(D_k)$ never reaches $\partial(D_k)$ because, if u steered $p \in \mathrm{int}(D_k)$ into $q \in \partial(D_k)$, then u would steer a point near p into a point

not contained in D_k by continuous dependence of solution $\dot{s} = h(s,u)$ on the initial value. Hence $\text{int}(D_k)$ is positively invariant. From this together with $\lambda(u,p) \geq 0$ for any $u \in \mathcal{U}$ and $p \in \text{int}(D_k)$, we see that it is impossible to exponentially stabilize the system in any region containing $\text{int}(D_k)$. □

The next result explains why exponential feedback stability is appropriate stabilization concept for bilinear control systems with compact control ranges.

THEOREM 3.7. *Under* (H), (B) *is asymptotically feedback stabilizable in* $\mathbb{R}^d$ *if and only if it is exponentially feedback stabilizable in* $\mathbb{R}^d$ *(with measurable feedback laws).*

Proof. One direction is obvious. Now assume (B) is not exponentially feedback stabilizable. By Corollary 3.6 $\lambda(u,p) \geq 0$ for all $p \in \text{int}(D_k)$ and $u \in \mathcal{U}$. If (B) were asymptotically feedback stabilizable in $\mathbb{R}^d$ then there would exist a measurable $u : \mathbb{R}^d \to \mathbb{R}^m$ with $u(x) \in U$ for any $x \in \mathbb{R}^d$, such that $\dot{x} = A(x,u(x))$ is asymptotically stable with respect to the origin.

Choose any point $0 \neq x_0 \in \mathbb{R}^d$ with $\frac{x_0}{|x_0|} \in \text{int}(D_k)$ and look at the trajectory of $\dot{x} = A(x,u(x))$ from the point x_0. Since $\dot{x} = A(x,u(x))$ is asymptotically stable, we can choose $T > 0$ such that $|\varphi(T,x_0)| < |x_0|$, here $\varphi(t,x_0)$ is the solution of $\dot{x} = A(x,u(x))$ with $\varphi(0,x_0) = x_0$.

If $\varphi(T,x_0) = rx_0$ with $r \in (-1,1) \setminus \{0\}$, let $v(t) = u(\varphi(t,x_0))$ for $t \in [0,T]$, and extend $v(t)$ periodically with T for $t \in \mathbb{R}$. Consider $\dot{x} = A(x,v(t))$ in $\mathbb{R}^d$. Since $\varphi(t,x_0)$ satisfies $\dot{x} = A(x,u(x))$ for $t \in [0,T], \varphi(t,x_0)$ is a solution of $\dot{x} = A(x,v(t))$ with $\varphi(0,x_0) = x_0$ for $t \in [0,T]$. Notice that $A(x,v(t)) = [A_0 + \sum_{i=1}^m v_i(t)A_i]x$ and $\varphi(T,x_0) = rx_0 = r\varphi(0,x_0)$; hence $r\varphi(t-T,x_0)$ also satisfies $\dot{x} = A(x,v(t))$ with $r\varphi(2T-T,x_0) = r^2x_0$ for $t \in [T,2T]$ (since $v(t)$ is periodic). So, in general, $r^n\varphi(t-nT,x_0)$ for $t \in [nT,(n+1)T], n = 0,1,\ldots$ is a solution of $\dot{x} = A(x,v(t))$. Since for $t \in [0,T]$ $0 < m \leq |\varphi(t,x_0)| \leq M < \infty$ for some m and M, $\lim_{t\to\infty} \frac{1}{t} \log |r^n\varphi(t - nT,x_0)| = \frac{1}{T} \log |r| < 0$ is a Lyapunov exponent. But by assumption all Lyapunov exponents $\lambda(u,p)$ obtained from $\text{int}(D_k)$ are not less than 0.

Thus we have shown that if (B) is not exponentially feedback stabilizable and $\varphi(T,x_0) = rx_0$ for some $x_0 \in \text{int}(D_k)$ with $r \in (-1,1) \setminus \{0\}$, then (B) is not asymptotically feedback stabilizable.

Now pick any point $p_0 \in \text{int}(D_k)$ in $\mathbb{P}^{d-1}$. Let

$$T := \max_{p \in D_k} \min_{u \in \mathcal{U}} \{t | s(t,p,u) = p_0 \text{ and } u \in \mathcal{U}\},$$

where $s(\cdot,p,u)$ is a solution of (PB) with $s(0,p,u) = p$. Since D_k is a closed subset of the compact space $\mathbb{P}^{d-1}, D_k$ is compact. Hence, T is finite [CK1].

For any $\delta > 0$, define

$$L(\delta) := \max\{|\psi(t,x,u)| \mid 0 \leq t \leq T, x \in c\ell(B(\delta)) \text{ and } u \in \mathcal{U}\},$$

where $B(\delta)$ is the open ball in $\mathbb{R}^d$ with radius δ, and $\psi(t,x,u)$ is the solution of $\dot{x} = A(x,u)$ with $\psi(0,x,u) = x$.

Consider the set $\mathcal{U}$, consisting of the open loop control functions. In [CK3] it is proved that $\mathcal{U}$ is compact and metrizable in the week*-topology of $L^\infty(\mathbb{R},\mathbb{R}^m) = (L^1(\mathbb{R},\mathbb{R}^m))^*$. The metric for u and $v \in \mathcal{U}$ is given by

$$d(u,v) = \sum_{n=1}^{\infty} \frac{1}{2^n} \frac{\left| \int_{\mathbb{R}} < u(t) - v(t), x_n(t) > dt \right|}{1 + \left| \int_{\mathbb{R}} < u(t) - v(t), x_n(t) > dt \right|},$$

where $\{x_n | n \in \mathbb{N}\}$ is a countable dense subset of $L^1(\mathbb{R}, \mathbb{R}^m)$, and $\langle \cdot, \cdot \rangle$ denotes an inner product in $\mathbb{R}^m$.

Since $\{x \in \mathbb{R}^d |\ |x| = \delta\}$ and $(\mathcal{U}, d)$ are compact,

$$N(\delta) := \max\{|\psi(t, x, u)| \mid 0 \le t \le T, |x| = \delta \text{ and } u \in \mathcal{U}\}$$

is finite. Since $A(x, u(t)) = (A_0 + \sum_{i=1}^m u_i(t)A_i)x, N(r\delta) = rN(\delta)$ for any $r > 0$. Hence, $L(r\delta) = rL(\delta)$ and $\lim_{\delta \to 0} L(\delta) = 0$. Fix $\delta_0 > 0$ so that $L(\delta_0) < 1$.

Pick $x_0 \in \mathbb{R}^d$ with $|x_0| = 1$ and $\frac{x_0}{|x_0|} = x_0 \in \text{int}(D_k)$. Let $\varphi(\cdot, x_0)$ be the solution of $\dot{x} = A(x, u(x))$ with $\varphi(0, x_0) = x_0$. By assumption $\varphi(t, x_0) \to 0$ as $t \to \infty$ and $\varphi(t, x_0)/|\varphi(t, x_0)| \in \text{int}(D_k)$ for $t \ge 0$ since $\text{int}(D_k)$ is positively invariant.

Let $T_0 > 0$ such that $|\varphi(T_0, x_0)| < \delta_0$. Since (PB) is completely controllable in $\text{int}(D_k)$, there is $u_1 \in \mathcal{U}$ such that

$$s\left(\tau, \frac{\varphi(T_0, x_0)}{|\varphi(T_0, x_0)|}, u_1\right) = \frac{x_0}{|x_0|} \tag{1}$$

for some $\tau \in [0, T]$ (by the definition of T). Let ψ be the corresponding solution of $\dot{x} = A(x, u_1)$ with $\psi(0, \varphi(T_0, x_0), u_1) = \varphi(T_0, x_0)$ for $t \in [0, \tau]$. So $|\psi(\tau, \varphi(T_0, x_0), u_1)| \le L(\delta_0) < 1$. Define

$$v(t) = \begin{cases} u(\varphi(t, x_0)), & 0 \le t \le T_0, \\ u_1(t - T_0), & T_0 < t \le T_0 + \tau. \end{cases}$$

Let $\widetilde{T} = T_0 + \tau$. Then

$$\widetilde{\psi}(t, x_0, v(t)) := \begin{cases} \varphi(t, x_0), & 0 \le t \le T_0, \\ \psi(t - T_0, \varphi(T_0, x_0), v), & T_0 < t \le \widetilde{T} \end{cases}$$

is a solution of $\dot{x} = A(x, v)$ with $\widetilde{\psi}(\widetilde{T}, x_0, v) = rx_0$ for some $r \in (-1, 1) \setminus \{0\}$ (since $0 < |\psi(\tau, \varphi(T_0, x_0), u_1)| < 1$ and (1) holds). Thus we have constructed a solution $\widetilde{\psi}$ in $\mathbb{R}^d$ which satisfies $\widetilde{\psi}(\widetilde{T}, x_0, v) = rx_0$. This is the case we discussed before. So, if (B) is not exponentially feedback stabilizable in $\mathbb{R}^d$ using measurable feedback laws, then (B) is not asymptotically feedback stabilizable in $\mathbb{R}^d$ using measurable feedback laws. □

Combining Corollary 3.6 and Theorem 3.7 we get the following.

COROLLARY 3.8. *Under assumption* (H)*, the following are equivalent:*

1) (B) *is asymptotically feedback stabilizable in* $\mathbb{R}^d$*;*
2) (B) *is exponentially feedback stabilizable in* $\mathbb{R}^d$*;*
3) $\lambda(u, x) < 0$ *for some* $u \in \mathcal{U}$ *and* $x \in \mathbb{R}^d$ *with* $\frac{x}{|x|} \in \text{int}(D_k)$. □

Remark 3.9. Statement 3) of Corollary 3.8 can also be expressed in the following equivalent ways, involving the open loop system (B):

4) For all $x \in \mathbb{R}^d \setminus \{0\}$ there exists $u \in \mathcal{U}$ with $\lambda(u, x) < 0$;
5) The system (B) is asymptotically null-controllable for all $x \in \mathbb{R}^d \setminus \{0\}$;
6) The system (B) is exponentially null-controllable for all $x \in \mathbb{R}^d \setminus \{0\}$.

The proof of 3) $\Leftrightarrow$ 4) follows directly from the construction of the main control sets, for 4) $\Leftrightarrow$ 6) see [CK5], and 5) $\Leftrightarrow$ 6) follows along the same lines as the proof of Theorem 3.7.

4. Example. Consider the controlled linear oscillator $\ddot{y} + 2b\dot{y} + (1+u)y = 0$. With $x = (x_1, x_2)^T = (y, \dot{y})^T$, the equation becomes

$$\dot{x}(t) = \begin{pmatrix} 0 & 1 \\ -1 & -2b \end{pmatrix} x(t) + u(t) \begin{pmatrix} 0 & 0 \\ -1 & 0 \end{pmatrix} x(t) =: A(u)x, \tag{E1}$$

where $u(t) \in U = [A, B]$. This equation is studied in [CK2].

Projection of the equation onto the projective space $\mathbb{P}^1$ yields with $p = (\cos\theta, \sin\theta), \theta \in [0, \pi)$

$$\theta = -\sin^2\theta(t) - (1+u(t))\cos^2\theta(t) - b\sin(2\theta(t)). \tag{E2}$$

Now we consider the case: $b = -2, U = [-2, 2]$. Since the state space of (E2) is one-dimensional, the control sets of (E2) can be calculated simply by checking the monotonicity of $f(\cdot, u)$ where u is constant. The control sets are

$$D_1 = \pi_{\mathbb{P}^1}\left\{\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2 \setminus \{0\} | x_2 = \alpha x_1, \alpha \in (2-\sqrt{5}, 1)\right\} \text{ and}$$
$$D_2 = \pi_{p^1}\left\{\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2 \setminus \{0\} | x_2 = \beta x_1, \beta \in [3, 2+\sqrt{5}]\right\},$$

where $\pi_{\mathbb{P}^1}$ denotes the projection onto $\mathbb{P}^1$. If $\mathbb{P}^1 = [0, \pi)$, then D_1 and D_2 can also be written as

$$D_1 = \left[0, \frac{\pi}{4}\right) \cup \left(\pi + \arctan(2-\sqrt{5}),\ \pi\right) \text{ and}$$
$$D_2 = [\arctan 3,\ \arctan(2+\sqrt{5})],$$

where D_1 is open and D_2 is closed in $\mathbb{P}^1$. By the method provided in [CK2], we can calculate the spectral intervals which are

$$c\ell\Sigma_{FL}(D_1) = [2-\sqrt{5},\ 1] \text{ and}$$
$$c\ell\Sigma_{FL}(D_2) = [3,\ 2+\sqrt{5}].$$

(For this specific case we just need to compute all eigenvalues of $A(u)$ for constant $u \in U$.) It is easy to see that $\mathcal{A}(D_1)$, the domain of attraction of D_1, is D_1 itself (notice that D_1 is the minimal main control set). Hence, the cone generated by $\mathcal{A}(D_1)$ is $K := K_{\mathcal{A}(D_1)}$, which is equal to

$$\left\{\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2 \setminus \{0\} | x_2 = \alpha x_1,\ \alpha \in (2-\sqrt{5}, 1)\right\}.$$

Now we want to find a piecewise analytic feedback law $u(x)$ defined in K to stabilize (E1) in K. First of all we define $u(\theta)$ in $\mathcal{A}(D_1) = D_1$ such that $u(\theta)$ steers any point in D_1 into a specific point p in finite time. Choose $p = \pi + \arctan(-0.05) \in [0, \pi) = \mathbb{P}^1$. Now $u(\theta)$ can be defined as

$$u(\theta) = \begin{cases} 2, & \theta \in \left[0, \frac{\pi}{4}\right) \cup (\pi + \arctan(-0.05),\ \pi), \\ -1.2025, & \theta = \arctan(-0.05), \\ -2, & \theta \in \left(\pi + \arctan(-0.05),\ \pi + \arctan(2-\sqrt{5})\right). \end{cases}$$

By checking the monotonicity of $f(\theta, u(\theta))$, we see that all trajectories starting in $\mathcal{A}(D_1) = D_1$ reach p in finite time. Now we can define $u(x)$ in K as

$$u(x) = \begin{cases} 2, & x \in \left\{ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2 \setminus \{0\} |\ x_2 = \alpha x_1, \alpha \in (-0.05,\ 1) \right\}, \\ -1.2025, & x_2 = -0.05 x_1, \\ -2, & x \in \left\{ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2 \setminus \{0\} |\ x_2 = \beta x, \beta \in (2-\sqrt{5},\ -0.05) \right\}. \end{cases}$$

The proof of stability of $\dot{x} = A(u(x))x$ is easy if we notice the following facts. When $u = -1.2025, A(u) = A(-1.2025)$ has one eigenvector $(x_1, -0.05x_1)^T$ associated with the eigenvalue $\lambda_1 = -0.05$. Let $L := \{(x_1, -0.05x_1)^T \in \mathbb{R}^2 | x_1 \in \mathbb{R}\}$. If $u = 2, A(2)$ has eigenvalues $\lambda_1 = 1$ and $\lambda_2 = 3$ whose corresponding eigenvectors are $(x_1, x_1)^T$ and $(x_1, 3x_1)^T$, respectively. So any trajectory starting in $\{(x_1, x_2)^T \in \mathbb{R}^2 \setminus \{0\} | x_2 = \alpha x_1, \alpha \in (-0.05, 1)\}$ will reach the line L in finite time. If $u = -2, A(-2)$ has eigenvalues $\lambda_1 = 2 - \sqrt{5}$ and $\lambda_2 = 2 + \sqrt{5}$, whose corresponding eigenvectors are $(x_1, (2-\sqrt{5})x_1)^T$ and $(x_1, (2+\sqrt{5})x_1)^T$, respectively. The line $x_2 = (2-\sqrt{5})x_1$ is in the boundary of K, and the line $x_2 = (2+\sqrt{5})x_1$ is outside of $c\ell(K)$. Hence, any trajectory starting in $\{(x_1, x_2)^T \in \mathbb{R}^2 \setminus \{0\} | x_2 = \beta x_1, \beta \in (2-\sqrt{5}, -0.05)\}$ will reach the line L in finite time. So with this feedback law, all trajectories of $\dot{x} = A(u(x))x$ will reach the line L in finite time, and then follow it and exponentially approach to zero. The exponential rate for this feedback is 0.05. (It is easy to see that we can construct feedbacks so that the rates belong to $(2-\sqrt{5}, 0)$.)

This example also shows the following two facts. One is that the set

$$\{t_x | t_x \text{ is the time that the trajectory of } \dot{x} = A(u(x))x \text{ first reaches } L \text{ from } x \in K\}$$

is unbounded on K, but it is bounded on any closed $\widetilde{K} \subset K$. This is the reason why we formulated Definition 2.4.4) using subcones $\widetilde{K} \subset K$ such that $\widetilde{K} \cap \mathbb{S}^{d-1}$ is compact. The other observation is that K may not be maximal. For this example, (E1) is exponentially stabilizable in $K \cup \{(x_1, x_2)^T \in \mathbb{R}^2 \setminus \{0\} | x_2 = (2-\sqrt{5})x_1\}$ (i.e., the union of K and part of its boundary).

Since $\mathbb{P}^1$ is one dimensional and the system (E2) has two control sets on $\mathbb{P}^1$, the periodic solution $s_p(t, p, u_0)$ constructed in the proof of Theorem 3.1 is a single point in this example, i.e., the point $p = \pi + \arctan(-0.05)$. If $d > 2$, or if the projected system has only one control set for $d = 2$, then the periodic solution $s_p(t, p, u_0)$ may not be a constant on $\mathbb{P}^{d-1}$.

Appendix. With the consent of Shan Lin, the author extracts part of the results from his Ph.D. dissertation [Li], from which Lemma 2.8 follows. Most of the notations follow [Su] and results extend those of [Su] to the not completely controllable case.

Consider an affine control system

$$\dot{x} = X_0(x) + \sum_{i=1}^{m} u_i X_i(x) \tag{C}$$

on a paracompact, connected, and real analytic manifold M of dimension n. The vector fields X_0 and X_i, $i = 1, \ldots, m$ are assumed to be real analytic. All admissible controls $u := (u_1, \ldots, u_m)$ are in $\mathcal{U} = \{u : \mathbb{R} \to U \text{ locally integrable}\}$, where $U \subset \mathbb{R}^m$ is compact. Let $\mathcal{U}_0 = \{\text{all piecewise constant controls in } \mathcal{U}\}$, $\mathcal{F} = \{X_0 + \sum_{i=1}^m u_i X_i(x) | u = (u_1, \ldots, u_m) \text{ constant in } U\}$.

We assume that the control system (C) satisfies the Lie algebra rank condition

$$\text{(H)} \qquad \dim Lie\left[X_0 + \sum_{i=1}^{m} u_i X_i,\ u := (u_1, \dots, u_m) \in U\right](x) = n$$

for all $x \in M$.

If X is a smooth vector field on M, Φ^X denotes the flow of X. Let $\xi = (Y_1, \dots, Y_k)$ be a finite sequence of smooth vector fields, and let $\tau = (t_1, \dots, t_k) \in \mathbb{R}^k$, where k is a positive integer. Set $|\xi| = k$ and $\|\tau\| = t_1 + \cdots + t_k$. Denote

$$\Phi_\tau^\xi := \Phi_{t_1}^{Y_1} \Phi_{t_2}^{Y_2} \dots \Phi_{t_k}^{Y_K},$$

where $\Phi_t^X(\cdot) = \Phi^X(t, \cdot)$. Also write $\Phi^\xi(\tau, x) = \Phi_\tau^\xi(x)$ for $\tau \in \mathbb{R}^{|\xi|}$. Define

$$\mathbb{R}_+^{k,T} := \{\tau \in \mathbb{R}_t^k \,|\, \|\tau\| \le T\} \text{ and}$$

for a given $\tau \in \mathbb{R}^k$ $C_\varepsilon^{|\xi|}(\tau) := \{(t_1', \dots, t_k') \in \mathbb{R}^k |\ |t_i' - t_i| \le \varepsilon \text{ for } i = 1, \dots, k\}$.

For $\tau = (t_1, \dots, t_k) \in \mathbb{R}^k$, define a map $\eta_\tau : [0, \|\tau\|] \to \mathbb{R}^k$ by

$$\begin{aligned}
\eta_\tau(t) &= (0, 0, \dots, 0, t), && \text{for } 0 \le t \le t_k \\
&= (0, 0, \dots, t - t_k, t), && \text{for } t_k \le t \le t_k + t_{k-1} \\
&\ \ \vdots && \ \ \vdots \\
&= (t - t_2 - \cdots - t_k, t_2, \dots, t_k), && \text{for } t_2 + t_3 + \cdots + t_k \le t \le \|\tau\|.
\end{aligned}$$

LEMMA A.1. *Assume* (H)*, the accessibility condition for the control system* (C)*, and let $x \in M$ and $y \in \operatorname{int}(D)$ where D is a main control set of control system* (C)*. If there is a $u \in \mathcal{U}$ such that $y = \psi(t, x, u)$ for some $t \in \mathbb{R}_+$, then there is $v \in \mathcal{U}_0$ such that $y = \psi(t', x, v)$ for some $t' \in \mathbb{R}_+$, where $\psi(t, x, u)$ is the trajectory of* (C) *with control u and $\psi(0, x, u) = x$.*

With this lemma, together with a similar argument as in [Su], we can prove the following.

LEMMA A.2. *Let $p \in \operatorname{int}(D)$ and $q \in \mathcal{A}(p)$, where $\mathcal{A}(p)$ is the domain of attraction of p. Then there exist* (1) *a finite subset $\mathcal{F}_p$ of $-\mathcal{F}$,* (2) *a finite sequence ξ of elements of $\mathcal{F}_p$,* (3) *a $\tau \in R_+^{|\xi|}$, and* (4) *an $\varepsilon > 0$ such that*

$$\begin{aligned}
&1)\ \Phi^\xi(\tau, p) = q; && 2)\ C_\varepsilon^{|\xi|}(\tau) \subset \mathbb{R}_+^{|\xi|}; \\
&3)\ \Phi^\xi(\cdot, p)(C_\varepsilon^{|\xi|}(\tau)) \text{ is a neighborhood of } q.
\end{aligned}$$

THEOREM A.3. *Assume* (H) *and let $p \in \operatorname{int}(D)$. Then there exists a piecewise analytic feedback controller V for system* (C) *such that p can be reached from any point of $\mathcal{A}(p)$ in finite time for the system $\dot{x} = X_0(x) + \sum_{i=1}^m u_i(x) X_i(x)$.*

Proof. For each $q \in \mathcal{A}(p)$, pick ξ_q, τ_q and ε_q so that they satisfy all the three properties in Lemma A.2. Let us use F_q to denote the map $\Phi^{\xi_q}(\cdot, p)$. Let A_q be the set of all points of $\mathbb{R}_+^{|\xi_q|}$ that are of the form $\eta_\tau(t)$ for some $\tau \in C_{\xi_q}^{|\xi_1|}(\tau_q)$ and some $t \in [0, \|\tau\|]$. For $i = 1, \dots, |\xi_q|$, put $a_q^i = \tau_{q,i} - \varepsilon_q$ and $b_q^i = \tau_{q,i} + \varepsilon_q$, where $\tau_q = (\tau_{q,1} \dots, \tau_{q,|\xi_q|})$.

Let A_q^i be the set of all points $(t_1, \ldots, t_{|\xi_q|})$ that satisfy (i) $t_j = 0$ for $j \leq |\xi_q| - i$; (ii) $0 \leq t_j \leq b_q^j$ for $j = |\xi_q| + 1 - i$; and (iii) $a_q^j \leq t_j \leq b_q^j$ for $j > |\xi_q| + 1 - i$. Then $A_q = A_q' \cup \cdots \cup A_q^{|\xi_q|}$ and $A_q^i \cap A_q^j = \emptyset$ if $i \neq j$.

Put $B_q = F_q(A_q), B_q^i = F_q(A_q^i)$ for $i = 1, \ldots, |\xi_q|$. Because A_q and A_q^i are compact semianalytic sets, and the map F_q is analytic, it follows that B_q are B_q^i are compact subanalytic subsets of M. Moreover, since $F_q(C_{\varepsilon_q}^{|\xi_q|}) \subset B_q$, the set B_q contains a neighborhood of q.

Let $\{K_j | j \in \mathbb{N}\}$ be a sequence of compact sets such that $K_j \subset \mathrm{int}(K_{j+1})$ and that $\mathcal{A}(p) = \bigcup_{j=1}^{\infty} K_j$. For each j, pick a finite set Q_j of points in $\mathcal{A}(p)$ in such a way that

$$K_j \setminus \mathrm{int}(K_{j-1}) \subset \bigcup_{q \in Q_j} B_q.$$

Let $q_1, q_2, q_3, \ldots$ be a sequence consisting of the points of Q_1, followed by the points of Q_2, followed by the points of Q_3, etc. Form a sequence of sets $D_j = B_{q_k}^i$, where $j \in \mathbb{N}$ and i, k are the unique numbers such that $|\xi_{q_1}| + \cdots + |\xi_{q_{k-1}}| + i = j, 1 \leq i \leq |\xi_{q_k}|$. Let $D_0 = \{p\}$. For $j \geq 0$, put $E_j = D_0 \cup \cdots \cup D_j$. Then the E_j's form an increasing sequence of compact subanalytic sets. Moreover, for every i there is j such that $K_i \subset E_j$. Hence, if we let $H_j = E_j \setminus E_{j-1}$, we find that the H_j's constitute a locally finite partition of $\mathcal{A}(p)$ into relatively compact subanalytic sets. Let $\mathcal{T}$ be the family of sets consisting of H_j for $j \in \mathbb{N}$.

For each set H_j, we have by construction that $j = |\xi_{q_1}| + \cdots + |\xi_{q_{k-1}}| + i$, with $1 \leq i \leq |\xi_{q_k}|$. Let $l = |\xi_{q_k}|, \xi_{q_k} = (Y_1, \ldots, Y_l)$. Then define an analytic vector field Z_j on H_j as $Z_j = -Y_{l+1-i}$. From Theorem 3 and Corollary 7 in [Su] (by letting $F(H_j) = \{Z_j\}$ in the notation of [Su]), we conclude that there is a subanalytic stratification $\mathcal{L}$ compatible with $\mathcal{T}$ such that every $S \in \mathcal{L}$ is a subset of some H_j, and that, if $S \in \mathcal{L}, S \subset H_j$, then either$Z_j$ is everywhere tangent to S, or it is nowhere tangent to S.

Now we can define a piecewise analytic vector field

$$V = (\mathcal{L}, (\mathcal{L}_1, \mathcal{L}_2), \{V_S\}_{S \in \mathcal{L}_1}, E)$$

as follows. We take the stratification to be $\mathcal{L}$. A stratum $S \in \mathcal{L}$ is in $\mathcal{L}_1$ if $S \subset H_j$ and Z_j is tangent to S for some j. Otherwise, S is in $\mathcal{L}_2$. If $S \in \mathcal{L}_1$ and $S \subset H_j$, then define $V_S = Z_j$.

If $S \in \mathcal{L}_2$ and $q \in S$, there is a unique j such that $S \subset H_j$. It is shown in Theorem 9 in [Su] that there is a unique $E(q) \in \mathcal{L}_1$ such that the integral curve γ of Z_j through q satisfies $\gamma(s) \in E(q)$ for small positive s.

Finally, it remains to prove that any point $q \in \mathcal{A}(q)$ can be steered into p in finite time by the feedback controller V defined above. Let $q \in H_j$. The integral curve γ_1 of Z_j through q is such that $\gamma_1(s) \in H_j$ for $0 \leq s < T_1$ and $\gamma_1(T_1) \in H_i$ for some $i < j$. Let γ_2 be the integral curve of Z_i through $\gamma_1(T_1)$. Then $\gamma_2(s) \in H_i$ for $0 \leq s < T_2$ and $\gamma_2(T_2) \in H_k$ for some $k < i$. Define γ_3 to be the integral curve of Z_k through $\gamma(T_2)$, etc. Then the curve γ obtained by following γ_1, then γ_2, then γ_3, and so on, reaches p in time $T \leq \sum_{i=1}^{j_q} T_i$ for some integer $j_q \leq j$, which means at most $j < \infty$ integral curves are needed. So T is finite because each of the T_i's is finite from the construction of Z_i and H_i. □

Acknowledgment. The author is grateful to Professor Wolfgang Kliemann for his constant help and support for years.

REFERENCES

[AG] Z. Aganović and Z. Gajić, *Linear Optimal Control of Bilinear Systems*, Springer-Verlag, Berlin, New York, 1995.

[CSV] R. Chabour, G. Sallet, and J.C Vivalda, *Stabilization of nonlinear systems: A bilinear approach*, Math. Control Signals Systems, 6 (1993), pp. 224–246.

[CK1] F. Colonius and W. Kliemann, *Infinite time optimal control and periodicity*, Appl. Math. Optim., 20 (1989), pp. 113–130.

[CK2] F. Colonius and W. Kliemann, *Minimal and maximal Lyapunov exponents of bilinear control systems*, J. Differential Equations, 101 (1993), pp. 232–275.

[CK3] F. Colonius and W. Kliemann, *Some aspects of control systems as dynamical systems*, J. Dynamic Differential Equations, 5 (1994), pp. 469–494.

[CK4] F. Colonius and W. Kliemann, *Linear control semigroups acting on projective space*, J. Dynamic Differential Equations, 5 (1994), pp. 495–528.

[CK5] F. Colonius and W. Kliemann, *Asymptotic null controllability of bilinear systems*, Banach Center Publications, Vol. 32, Institute of Mathematics, Polish Academy of Sciences, Warsgana, 1995, pp. 139–148.

[CK6] F. Colonius and W. Kliemann, *The Lyapunov spectrum of families of time-varying matrices*, Trans. Amer. Math. Soc., Vol. 348, 11 (1996), pp. 4389–4408.

[CLSS] F.H. Clarke, Yu. S. Ledyaev, E.D. Sontag and A.I. Subbotin, *Asymptotic controllability implies feedback stabilization*, in Proc. Conference on Information Sciences and Systems, Princeton, NJ, 1996.

[Gr] L. Gruene, *Discrete feedback stabilizatin of semilinear control systems*, ESAIM: Control, Optimization and Calculus of Variations, 1 (1996) pp. 207–224.

[Ha] R.M. Hardt, *Stratifications of real analytic mappings and images*, Invent. Math., 28 (1975) pp. 193–208.

[Hi] H. Hironaka, *Subanalytic sets*, Lecture Notes of Istituto Matematico "Leonida Tonelli," Pisa, 1973.

[Li] S. Lin, *Analysis and Synthesis of Nonlinear Control Systems*, Ph.D. dissertation, Department of Mathematics, Iowa State University, Ames, IA, 1996.

[Su] H.J. Sussmann, *Subanalytic sets and feedback control*, J. Differential Equations, 31 (1979), pp. 31–52.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 5, pp. 1685–1702, September 1998

009

STOCHASTIC LINEAR QUADRATIC REGULATORS WITH INDEFINITE CONTROL WEIGHT COSTS*

SHUPING CHEN†, XUNJING LI‡, AND XUN YU ZHOU§

Abstract. This paper considers optimal (minimizing) control of stochastic linear quadratic regulators (LQRs). The assumption that the control weight costs must be positive definite, inherited from the deterministic case, has been taken for granted in the literature. It is, however, shown in this paper that some stochastic LQR problems with indefinite (in particular, negative) control weight costs may still be sensible and well-posed due to the deep nature of stochastic systems. New stochastic Riccati equations, which are backward stochastic differential equations involving complicated nonlinear terms, are presented and their solvability is proved to be sufficient for the well-posedness and the solutions of the optimal LQR problems. Existence and uniqueness of solutions to the Riccati equation for a special case are obtained. Finally, it is argued that, quite contrary to the deterministic systems, the stochastic maximum principle cannot fully characterize the optimality of the stochastic LQR problems.

Key words. stochastic linear quadratic regulator, well-posedness, stochastic Riccati equation, backward stochastic differential equation, maximum principle

AMS subject classifications. 93E, 49K

PII. S0363012996310478

1. Introduction.

Consider the following stochastic linear quadratic regulator (LQR) problem:

$$\text{Minimize} \quad J = E\{\int_0^T \tfrac{1}{2}[x'(t)Q(t)x(t) + u'(t)R(t)u(t)]dt + \tfrac{1}{2}x'(T)Hx(T)\}$$

$$\text{Subject to} \quad \begin{cases} dx(t) = [A(t)x(t) + B(t)u(t)]dt + [C(t)x(t) + D(t)u(t)]dW(t), \\ x(0) = y. \end{cases}$$

Here $W(t)$ is a Brownian motion and the control variable $u(t)$ takes value in some Euclidean space. In the deterministic case (i.e., $C = D = 0$), it is well known that the matrix $R(t)$, the so-called control weight, must be positive definite (for almost all t); otherwise the optimization problem would not be well-posed (or would become trivial) [8, 1]. To be precise, if $R(t)$ is negative (which means a benefit rather than a cost), then the optimal control u can be shown to be such that $|u(t)| = +\infty$, namely, "the larger the better." Stochastic LQR problems have been first studied by Wonham [13] and by many researchers later (cf., e.g., [2, 5]), but the assumption that $R(t) > 0$ has been taken for granted in all of these works. Recently, we observed that some stochastic LQR problems with $D \neq 0$ are nontrivial even when $R(t) < 0$, i.e., the

*Received by the editors October 14, 1996; accepted for publication (in revised form) January 8, 1998; published electronically June 22, 1998.

http://www.siam.org/journals/sicon/36-5/31047.html

†Center for Mathematical Sciences, Zhejiang University, Hangzhou, China (amaschen@dial.zju.edu.ch). The research of this author was partially supported by the National Natural Science Foundation of China and the State Education Commission of China.

‡Laboratory of Mathematics for Nonlinear Sciences and Department of Mathematics, Fudan University, Shanghai 200433, China (xjli@ms.fudan.edu.ch). The research of this author was partially supported by the Climbing Project of China and the Chinese State Education Commission Science Foundation.

§Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, Hong Kong (xyzhou@se.cuhk.edu.hk). The research of this author was supported by RGC earmarked grants CUHK 4125/97E and CUHK 249/94E.

"the-larger-the-better" policy no longer applies. Let us look at a very simple example in one dimension. The following deterministic LQR problem

$$\begin{array}{ll} \text{Minimize} & J = \int_0^1 \frac{1}{2}[x^2(t) + r(t)u^2(t)]dt + \frac{1}{2}x^2(1) \\ \text{Subject to} & \begin{cases} dx(t) = 0, \\ x(0) = 0, \end{cases} \end{array} \tag{1.1}$$

where $r(t) < 0$, is not well-posed. In fact, $J = \int_0^1 \frac{1}{2} r(t)u^2(t)dt \to -\infty$ as $|u(t)| \to +\infty$. Now, consider a stochastic version of (1.1):

$$\begin{array}{ll} \text{Minimize} & J = E\Big\{ \int_0^1 \frac{1}{2}[x^2(t) + r(t)u^2(t)]dt + \frac{1}{2}x^2(1) \Big\} \\ \text{Subject to} & \begin{cases} dx(t) = u(t)dW(t), \\ x(0) = 0. \end{cases} \end{array} \tag{1.2}$$

Substituting $x(t) = \int_0^t u(s)dW(s)$ into the cost function, we obtain, via a simple calculation

$$J = \frac{1}{2} E \int_0^1 [r(t) + (2 - t)]u^2(t)dt. \tag{1.3}$$

Hence, when $r(t)$ is a deterministic function with $r(t) > t-2$, the optimization problem is sensible (with the optimal control $u^*(t) = 0$). In this case, the control weight cost, $r(t)$, could be *negative* as long as, say, $r(t) > -1$. Certainly, $r(t)$ cannot be *too* negative. For example, the problem would obviously become ill-posed if $r(t) < -2$.

The above seemingly surprising observation indeed makes perfect sense when we think a little deeper: the gain due to a larger control size may *not* outweigh the loss due to a greater uncertainty (because $D \neq 0$). It is emphasized that $D \neq 0$, which means that the controller can control the uncertainty or the decision made is going to affect the scale of the uncertainty in the system, plays a key role here. This kind of situation happens in real-world systems. In a stock market, for example, the trading made by the so-called "large investors" is going to influence the fluctuations of the stock prices. If $D = 0$, which is assumed in most of the previous work (see [5] and the references therein), then the control weight R must be positive definite for the well-posedness and the stochastic LQR problem can be treated almost completely parallel to the deterministic case. However, if $D \neq 0$, then R, even being negative, could be compensated by a quadratic term (which is related to D) by taking advantage of the underlying uncertainty. This observation reveals a fundamental difference between deterministic and stochastic systems.

Let us take a more concrete example to illustrate the above idea. Suppose an oil company is investing in an oil prospecting project. This project will cause a certain degree of pollution and suppose the pollution level $x(t)$ during a period of time $[0, T]$ is described by

$$\begin{cases} dx(t) = (\alpha x(t) + \beta u(t))dt + \delta u(t)dW(t), \\ x(0) = x_0, \end{cases} \tag{1.4}$$

where $u(t)$ represents the investment level of the company at time t, x_0 is the initial pollution level, and α, β, and δ are given constants. Suppose that the investment is expected to be very profitable and the return in the time period $[t, t+\Delta t]$ is $r|u(t)|^2\Delta t$

with a constant $r > 0$, and the company has sufficient funds to make the investment so that $u(t) \in (0, +\infty)$. On the other hand, the environmental impact of the project is supervised and monitored by the government so that the pollution level $x(t)$ cannot deviate too much from an allowable level $\bar{x}(t)$ at any time. The objective of the company is on one hand to maximize the total expected return, $E\int_0^T r|u(t)|^2dt$, and on the other hand to minimize the expected negative environmental impact, which in this case is measured by $E\int_0^T |x(t) - \bar{x}(t)|^2dt$. This is a multi-objective optimization problem and it may be converted into a single-objective problem by putting weights on the different objectives. Thus the following function is to be minimized:

$$J = E\int_0^T (\lambda_1|x(t) - \bar{x}(t)|^2 - \lambda_2 r|u(t)|^2)dt, \tag{1.5}$$

where $\lambda_1, \lambda_2 \in (0, 1)$ with $\lambda_1 + \lambda_2 = 1$ represent the weights. Note that this is a stochastic minimizing LQR problem with a negative control cost. If the problem were deterministic (i.e., there was no risk), where a positive return is guaranteed, then by the deterministic LQR theory that the control cost $-\lambda_2 r|u(t)|^2$ would be overwhelming in the overall cost, when λ_1 is small enough. In this case, the optimal policy would be $u(t) = +\infty$ (i.e., the larger the investment size the better) and the problem would become trivial. However, the problem is actually stochastic where the diffusion coefficient depends on the control (i.e., the risk of pollution increases as the investment level increases); then there is a trade-off (no matter how small λ_1 is) between the return (or the investment size) and the risk which makes the optimization problem sensible.

More generally, such a phenomenon can happen in the following situation. Suppose, in a deterministic (minimizing) optimization problem, that the cost *decreases* as the level of activity the decision maker carries out *increases* (a typical example of such situations is an investment that would be "guaranteed" to be profitable if the risk were to be excluded from consideration). Then it is not really an optimization problem because there is no *trade-off* in it, and the optimal decision is simply to take the maximum possible activity level. So the problem is trivial or ill-posed. However, in a stochastic environment, suppose that the uncertainty *increases* with *increasing* magnitude of the activity level and that the uncertainty results in certain additional cost (called *risk adjustment* in the terminology of economics); then there is a trade-off between the activity level and the uncertainty, and the decision maker has to carefully balance the two to achieve an optimal solution. The problem therefore becomes meaningful. Needless to say, such phenomena may occur in a much wider class of optimization problems that can go beyond linear systems and optimal control problems.

For LQR problems, it is natural to study an associated Riccati equation. The Riccati equation presented in this paper for the stochastic LQR problem with an indefinite control weight cost is a backward stochastic differential equation of the Pardoux and Peng type [11] and involves a term $(R + D'PD)^{-1}$ (P is the unknown of the Riccati equation). In the present paper, we show that the stochastic LQR problem is well-posed if there are solutions to the Riccati equation, and an optimal feedback control can then be obtained. However, the existence and uniqueness of the solutions to the Riccati equation, in general, seem to be very difficult problems due to the presence of the complicated nonlinear term. In this paper, we shall solve the existence and uniqueness for a special case where $C = 0$ and all of the matrices A, B, D, Q, R, H are deterministic functions.

For the deterministic LQR problems (with $R > 0$), the Pontryagin maximum prin-

ciple can completely characterize the optimality [1]. More precisely, the solvability of the so-called Hamiltonian system, which consists of the original state equation, the adjoint equation, and the maximum condition, is equivalent to the well-posedness of the LQR problem, and the solution to the Hamiltonian systems gives rise to an optimal feedback control. The stochastic maximum principle has been investigated since the 1960s [6, 9, 2, 7]. However, almost all of the results assume that the diffusion term does not depend on the control variable. Under this assumption, the statements of maximum principle (i.e., an optimal control should maximize pointwisely the usual Hamiltonian, which is *linear* in the drift term and *independent* of the diffusion term) and their proofs are very similar to those of the deterministic case. One does not see much difference between stochastic and deterministic systems from those results. The stochastic maximum principle for systems with control-dependent diffusion coefficients had long been an outstanding open problem until 1988 when Peng [12] first solved it (the proof of Peng was simplified by Zhou [15]). It is observed in [12, 15] that, in addition to the usual (first-order) adjoint equation, one has to introduce an *additional* adjoint equation—called the *second-order adjoint equation*—to represent the *risk factor* due to the underlying uncertainty. The maximum principle is to maximize an extended Hamiltonian, which includes an additional term that is *quadratic* in the diffusion coefficient, to reflect the risk-averse or risk-seeking attitudes of the decision makers. Moreover, it is shown by Zhou [16] that Peng's maximum principle is sufficient under certain convex conditions. However, while Peng's maximum principle has been widely recognized as a significant new result as well as the best result so far, it *cannot*, as will be shown via an example in this paper, lead to an optimal solution for some stochastic LQR problems that are well-posed but with $R < 0$. This is quite different to the deterministic case, which suggests that Peng's maximum principle could be further improved to give *tighter* necessary conditions of optimality which in particular would be sufficient for the stochastic LQR model.

The rest of the paper is organized as follows. In section 2 the optimal control problem of stochastic LQR models with indefinite control weight costs is formulated. In section 3 the corresponding stochastic Riccati equation is introduced and the existence of its solutions is shown to be sufficient for the LQR problem to be well-posed. Section 4 is devoted to the study of the Riccati equation for a special case. In section 5, the gap between Peng's maximum principle and the stochastic LQR problems is demonstrated. Finally, section 6 gives some concluding remarks.

2. Problem formulation and preliminaries. We consider in this paper a stochastic optimal control problem. The system is governed by the following linear Ito's stochastic differential equation (SDE)

$$\begin{cases} dx(t) = [A(t)x(t) + B(t)u(t)]dt + [C(t)x(t) + D(t)u(t)]dW(t), \\ x(s) = y, \end{cases} \tag{2.1}$$

where $(s, y) \in [0, T) \times R^n$ are the initial time and initial state, respectively, $W(t)$ is a given one-dimensional Brownian motion on $[0, T]$, and $u(\cdot)$, the control, is a U-valued $\mathcal{F}_t$-adapted measurable process with

$$\mathcal{F}_t = \sigma\{W(r) : 0 \le r \le t\}. \tag{2.2}$$

Here $U = R^m$. The set of all such admissible controls is denoted by U_{ad}. Note that we assumed the Brownian motion to be one-dimensional just for simplicity. There is no essential difficulty in the analysis below for the multidimensional case.

For each (s, y) and $u(\cdot) \in U_{ad}$, the associated cost is

$$J(s,y;u(\cdot)) = E^s\Big\{\int_s^T \frac{1}{2}[x'(t)Q(t)x(t) + u'(t)R(t)u(t)]dt + \frac{1}{2}x'(T)Hx(T)\Big\}, \tag{2.3}$$

where $E^s \equiv E(\cdot|\mathcal{F}_s)$. The solution $x(\cdot)$ of the SDE (2.1) is called the response of the control $u(\cdot) \in U_{ad}$, and $(x(\cdot), u(\cdot))$ is called an *admissible pair*. The objective of the optimal control problem is to minimize the cost function $J(s, y; u(\cdot))$, for a given $(s, y) \in [0, T) \times R^n$, over all $u(\cdot) \in U_{ad}$. We denote the above problem by $C_{s,y}$ to recall the dependence on the initial time s and the initial state y. The value function is defined as

$$V(s,y) = \inf_{u(\cdot)\in U_{ad}} J(s,y;u(\cdot)). \tag{2.4}$$

Note that V is an $\mathcal{F}_s$-adapted process for each fixed y. An admissible pair $(x^*(\cdot), u^*(\cdot))$ is called *optimal* for $C_{s,y}$ if $u^*(\cdot)$ achieves the infimum of $J(s, y; u(\cdot))$. The optimization problem (2.1)–(2.3) is called *well-posed* if $V(s, y) > -\infty$, $P - a.s.$, for *all* $(s, y) \in [0, T) \times R^n$.

Notation. We make use of the following notation in this paper:

M'	:	the transpose of any vector or matrix M;
M^j	:	the jth entry of any vector M;
$\|M\|$	:	$= \sqrt{\sum_{i,j} m_{ij}^2}$ for any matrix or vector $M = (m_{ij})$;
S^n	:	the space of all $n \times n$ symmetric matrices;
S_+^n	:	the subspace of all nonnegative definite matrices of S^n;
$\hat{S}_+^n$	:	the subspace of all positive definite matrices of S^n;
$C(0, T; X)$	:	the Banach space of X-valued continuous functions on $[0, T]$ endowed with the maximum norm $\|\cdot\|$ for a given Hilbert space X;
ρ_x	:	the gradient or Jacobian of a function ρ with respect to the variable x;
ρ_{xx}	:	the Hessian of a scalar function ρ with respect to the variable x .

Given a probability space $(\Omega, \mathcal{F}, P)$ with a filtration $\{\mathcal{F}_t : a \le t \le b\}$ $(-\infty \le a < b \le +\infty)$, a Hilbert space X with the norm $\|\cdot\|_X$, and p $(1 \le p \le +\infty)$, define the Banach space

$$L^p_{\mathcal{F}}(a,b;X) = \Big\{\phi(\cdot) = \{\phi(t,\omega) : a \le t \le b\}|\ \phi(\cdot) \text{ is an } \mathcal{F}_t - \text{adapted, } X\text{-valued measurable process on } [a,b], \text{ and } E\textstyle\int_a^b \|\phi(t,\omega)\|_X^p\, dt < +\infty\Big\},$$

with the norm

$$\|\phi(\cdot)\|_{\mathcal{F},p} = \left(E\int_a^b \|\phi(t,\omega)\|_X^p\, dt\right)^{\frac{1}{p}}.$$

In the rest of this paper, we shall employ the usual convention of suppressing the ω-dependence of all random functions. Sometimes we even write A for a (deterministic or stochastic) process $A(t)$, omitting the variable t, whenever no confusion arises. Under this convention, when $A \in C(0, T; S^n)$, $A \ge (>)0$ means $A(t) \ge (>)0$, $\forall t \in [0, T]$.

The following basic assumption will be in force throughout this paper:

(A) The data appearing in the LQR problem satisfy

$$\begin{aligned}
&A, C \in L^{\infty}_{\mathcal{F}}(0,T;R^{n\times n}) \cap L^2(\Omega; C(0,T;R^{n\times n})),\\
&B, D \in L^{\infty}_{\mathcal{F}}(0,T;R^{n\times m}) \cap L^2(\Omega; C(0,T;R^{n\times m})),\\
&Q \in L^2_{\mathcal{F}}(0,T;S^n_+) \cap L^2(\Omega; C(0,T;S^n_+)),\\
&R \in L^2_{\mathcal{F}}(0,T;S^m) \cap L^2(\Omega; C(0,T;S^m)),\\
&H \in L^2(\Omega, \mathcal{F}_T; S^n_+).
\end{aligned}$$

3. Stochastic Riccati equation. We introduce the following stochastic Riccati equation:

(3.1)

$$\begin{cases}
dP(t) = \Big\{ -\Big(P(t)A(t) + A'(t)P(t) + C'(t)P(t)C(t) + \Lambda(t)C(t) + C'(t)\Lambda(t) + Q(t)\Big)\\
\qquad +\Big(P(t)B(t) + C'(t)P(t)D(t) + \Lambda(t)D(t)\Big)\Big(R(t) + D'(t)P(t)D(t)\Big)^{-1}\Big(B'(t)P(t)\\
\qquad +D'(t)P(t)C(t) + D'(t)\Lambda(t)\Big)\Big\}dt + \Lambda(t)dW(t),\\
P(T) = H,\\
K(t) \equiv R(t) + D'(t)P(t)D(t) > 0, \quad P-a.s., \ \forall t \in [0,T].
\end{cases}$$

An $\mathcal{F}_t$-adapted *pair* $(P,\Lambda) \in [L^2_{\mathcal{F}}(0,T;S^n) \cap L^2(\Omega; C(0,T;S^n))] \times L^2_{\mathcal{F}}(0,T;S^n)$ is called a solution of the Riccati equation (3.1) if it satisfies all the constraints in (3.1). Note that it is a *backward stochastic differential equation* (BSDE). This type of equation was originally proposed by Bismut [3, 4] for the linear case, then extended to the nonlinear case by Pardoux and Peng [11], and has been further developed extensively in recent years. A distinctive feature of this type of equations is that their solutions are pairs (P,Λ), and the presence of Λ, which is derived by the martingale representation theorem, is necessary to reflect the uncertainty during the period between now and the given terminal time. Note that Λ itself may not satisfy any SDE. For details see [3, 4, 2, 12, 15, 11]. The Riccati equation (3.1) is nonlinear, and the nonlinearity does not satisfy the Lipschitz condition usually imposed in the literature due to the presence of the term $(R + D'PD)^{-1}$.

THEOREM 3.1. *If the stochastic Riccati equation* (3.1) *admits a solution, then the stochastic LQR problem* (2.1)–(2.3) *is well-posed.*

Proof. Let $P \in L^2_{\mathcal{F}}(0,T;S^n) \cap L^2(\Omega; C(0,T;S^n))$ be any semimartingale with the following decomposition:

$$dP(t) = \Gamma(t)dt + \Lambda(t)dW(t), \quad t \in [0,T], \tag{3.2}$$

and let $(x(\cdot), u(\cdot))$ be any admissible pair. Applying Ito's formula, we obtain

$$\begin{aligned}
d(x'Px) = \ & \Big\{x'(\Gamma + PA + A'P + C'PC + \Lambda C + C'\Lambda)x\\
& +2u'(B'P + D'PC + D'\Lambda)x + u'D'PDu\Big\}dt\\
& +\{...\}dW(t).
\end{aligned} \tag{3.3}$$

Integrating from s to T, taking expectations E^s on both sides, and dividing by 2, one gets

$$\begin{aligned}
&\tfrac{1}{2}E^s[x'(T)P(T)x(T)] - \tfrac{1}{2}y'P(s)y\\
&= \tfrac{1}{2}E^s\int_s^T \Big\{x'(\Gamma + PA + A'P + C'PC + \Lambda C + C'\Lambda)x\\
&\quad +2u'(B'P + D'PC + D'\Lambda)x + u'D'PDu\Big\}dt.
\end{aligned}$$

Adding this to (2.3) and, provided $K \equiv R + D'PD > 0$, using the square completion technique, we have

$$\begin{aligned}
&J(s,y;u(\cdot))\\
&= \tfrac{1}{2}E^s\int_s^T \Big\{x'(\Gamma + PA + A'P + C'PC + \Lambda C + C'\Lambda + Q)x\\
&\quad + 2u'(B'P + D'PC + D'\Lambda)x\\
&\quad + u'(R + D'PD)u\Big\}dt + \tfrac{1}{2}E^s[x'(T)(H - P(T))x(T)] + \tfrac{1}{2}y'P(s)y\\
&= \tfrac{1}{2}E^s\int_s^T \Big\{x'(\Gamma + PA + A'P + C'PC + \Lambda C + C'\Lambda + Q - L'K^{-1}L)x\\
&\quad + (u + K^{-1}Lx)'K(u + K^{-1}Lx)\Big\}dt\\
&\quad + \tfrac{1}{2}E^s[x'(T)(H - P(T))x(T)] + \tfrac{1}{2}y'P(s)y,
\end{aligned} \tag{3.4}$$

where $L = B'P + D'PC + D'\Lambda$. Now, if (P,Λ) satisfies the Riccati equation (see (3.2)), i.e.,

$$\Gamma = -(PA + A'P + C'PC + \Lambda C + C'\Lambda + Q - L'K^{-1}L), \tag{3.5}$$

with $K = R + D'PD > 0$ and $P(T) = H$, then

$$\begin{aligned}
&J(s,y;u(\cdot))\\
&= \tfrac{1}{2}E^s\int_s^T (u + K^{-1}Lx)'K(u + K^{-1}Lx)dt + \tfrac{1}{2}y'P(s)y\\
&\geq \tfrac{1}{2}y'P(s)y > -\infty,\ P-a.s.
\end{aligned} \tag{3.6}$$

Therefore, the stochastic LQR problem is well-posed. □

Remark 3.1. We see from the above proof that if the Riccati equation (3.1) admits a solution (P,Λ), then the optimal feedback control would be

$$\begin{aligned}
u(t) &= -K^{-1}(t)L(t)x(t)\\
&= -\Big(R(t) + D'(t)P(t)D(t)\Big)^{-1}\Big(B'(t)P(t) + D'(t)P(t)C(t) + D'(t)\Lambda(t)\Big)x(t)
\end{aligned} \tag{3.7}$$

if the corresponding solutions to the system equation exist. In this case, the value function is $V(t,x) = \frac{1}{2}x'P(t)x$. Note that under (3.7), the system (2.1) reduces to

$$\begin{cases} dx(t) = [A(t) - B(t)K^{-1}(t)L(t)]x(t)dt + [C(t) - D(t)K^{-1}(t)L(t)]x(t)dW(t),\\ x(s) = y. \end{cases} \tag{3.8}$$

This is a linear stochastic differential equation. The existence and uniqueness of its solutions depend on some moment estimates of the coefficients $A - BK^{-1}L$ and $C - DK^{-1}L$ and, in particular, K^{-1}. While existence and uniqueness results are hard to obtain in general, they are indeed available in some special cases; see Theorem 3.2 below.

We see that the solutions to the Riccati equation are pairs (P,Λ). As mentioned, the presence of Λ is necessary when the coefficients A, B, C, D, Q, R, H of the equation are random so as to get an $\mathcal{F}_t$-adapted solution. However, if all the coefficients are deterministic, then we may have a *deterministic* Riccati equation as follows:

$$\begin{cases} \dot{P} + PA + A'P + C'PC - (PB + C'PD)(R + D'PD)^{-1}(B'P + D'PC) + Q = 0,\\ P(T) = H,\\ K = R + D'PD > 0. \end{cases} \tag{3.9}$$

THEOREM 3.2. *Assume that all the coefficients A, B, C, D, Q, R, H are deterministic. Then the statement of Theorem* 3.1 *remains valid with the Riccati equation* (3.1) *replaced by* (3.9). *Moreover, the following feedback control*

$$u(t) = -\Big(R(t) + D'(t)P(t)D(t)\Big)^{-1}\Big(B'(t)P(t) + D'(t)P(t)C(t)\Big)x(t), \tag{3.10}$$

which results in a unique solution of the state equation (3.8), *is optimal.*

Proof. The first assertion comes directly from Theorem 3.1 since, if there exists a solution to (3.9), then there exists a solution to (3.1) with $\Lambda = 0$. As for the second assertion, note that by the conventional Riccati equation theory we have $P \in C(0,T;S^n_+)$. Moreover, $K^{-1} \in C(0,T;S^n_+)$. Hence, all the coefficients in (3.8) are uniformly bounded, which implies the existence and uniqueness of its solutions. This completes the proof. □

By virtue of the newly introduced Riccati equation (3.1) or (3.9), one may now understand why the control weight cost R may be allowed to be indefinite. Indeed, even when $R < 0$, the presence of the term $D'PD$ may offer compensation if it is positive enough so that $R + D'PD > 0$. This is possible, as we will see from the examples below. Note that $D \neq 0$ (namely, the diffusion term depends on the control) is vital for such phenomena to occur, which intuitively means that the controller must be able to control the variance of the uncertainty in dynamics. If $D = 0$, then R must be positive definite in order for the problem to be sensible.

Example 3.1. Consider the example (1.2) presented in the introduction. The corresponding Riccati equation (3.9) is

$$\dot{P}(t) = -1, \qquad P(1) = 1.$$

Hence, $P(t) = 2 - t$. The problem is then well-posed if and only if $r(t) + 2 - t > 0$, which is consistent with the conclusion obtained from the direct computation.

Example 3.2. Consider the following:

$$\begin{array}{ll} \text{Minimize} & J = E^s\Big\{\int_0^1 \frac{1}{2}ru^2(t)dt + \frac{1}{2}x^2(1)\Big\} \\ \text{Subject to} & \begin{cases} dx(t) = u(t)dt + u(t)dW(t), \\ x(s) = y. \end{cases} \end{array} \tag{3.11}$$

Here r is a given (deterministic) constant. We are going to show that the problem is well-posed if

$$-1 \le r < 0, \qquad \ln(-r) + 2 + r < 0 \text{ (or } -0.1586 < r < 0 \text{ approximately)}. \tag{3.12}$$

To this end, we first see that the corresponding Riccati equation (3.9) reads

$$\begin{cases} \dot{P}(t) = \frac{P^2(t)}{r+P(t)}, \\ P(1) = 1, \\ r + P(t) > 0. \end{cases} \tag{3.13}$$

It should be noted that if $r < -1$, then the above equation is not solvable. Indeed, in this case, $P(t) > -r > 1$, so the terminal condition of (3.13) is violated.

The equation (3.13) is equivalent to

$$\begin{cases} \ln P(t) - \frac{r}{P(t)} = t - 1 - r, \\ r + P(t) > 0. \end{cases} \tag{3.14}$$

Define $f^t(p) = \ln p - \frac{r}{p} - t + 1 + r$, $p \in (0, +\infty)$. Since

$$\begin{aligned} &f^t(-r) = \ln(-r) + 2 - t + r < 0, \\ &f^t(1) = 1 - t > 0, \quad \forall t \in [0,1), \end{aligned} \tag{3.15}$$

we conclude that there is $P(t) \in (-r, 1)$ (i.e., $r + P(t) > 0$) such that $f^t(P(t)) = 0$ for $t \in [0,1)$. Moreover,

$$\dot{f}^t(p) = \frac{p+r}{p^2} \begin{cases} < 0, & \text{if } p < -r, \\ > 0, & \text{if } p > -r, \\ = 0, & \text{if } p = -r. \end{cases} \tag{3.16}$$

Thus the $P(t)$ satisfying $f^t(P(t)) = 0$ and $r + P(t) > 0$ is unique. Finally, $f^1(1) = 0$ implies $P(1) = 1$. Hence, (3.14) or (3.13) does admit a solution. In this case, the optimal feedback control is (noting (3.7))

$$u(t) = -\frac{P(t)}{r + P(t)} x(t).$$

4. Existence and uniqueness: A special case. We conclude from the previous section that the study of the stochastic LQR problem may be reduced to that of the Riccati equation (3.1). However, (3.1) is so complicated that we are not able to prove the existence and uniqueness of its solutions at this moment. We can only prove the existence and uniqueness for a special case, where $C(t) \equiv 0$ and all the other coefficients A, B, D, Q, R, H are deterministic functions, which is the objective of this section.

First of all, when $C(t) \equiv 0$ and all the other coefficients are deterministic, the Riccati equation (3.9) is further reduced to

$$\begin{cases} \dot{P} + PA + A'P - PB(R + D'PD)^{-1}B'P + Q = 0, \\ P(T) = H, \\ K = R + D'PD > 0. \end{cases} \tag{4.1}$$

THEOREM 4.1. *If P is a solution to the Riccati equation* (4.1), *then $P \in C(0,T;S_+^n)$ and it is the only solution.*

Proof. That $P \in C(0,T;S_+^n)$ is clear from the conventional Riccati equation theory. Now suppose $\tilde{P}$ is another solution of (4.1). Set $\hat{P} = P - \tilde{P}$. Then $\hat{P}$ satisfies

$$\begin{cases} \dot{\hat{P}} + \hat{P}A + A'\hat{P} - \hat{P}B\tilde{K}^{-1}B'\tilde{P} + PBK^{-1}D'\hat{P}D\tilde{K}^{-1}B'\tilde{P} - PBK^{-1}B'\hat{P} = 0, \\ \hat{P}(T) = 0, \end{cases}$$

where $K = R + D'PD > 0$ and $\tilde{K} = R + D'\tilde{P}D > 0$. Since $|K^{-1}(t)|$ and $|\tilde{K}^{-1}(t)|$ are uniformly bounded due to their continuity, we can apply Gronwall's inequality to get $\hat{P}(t) \equiv 0$. This proves the uniqueness. □

Now let us turn to the existence. We consider the conventional Riccati equation

$$\begin{cases} \dot{P} + PA + A'P - PBK^{-1}B'P + Q = 0, \\ P(T) = H. \end{cases} \tag{4.2}$$

Denote $\mathcal{K} = \{K \in L^\infty(0,T;\hat{S}^m_+) \mid K^{-1} \in L^\infty(0,T;\hat{S}^m_+)\}$. It can be checked that $C(0,T;\hat{S}^m_+) \subset \mathcal{K}$. Fix $Q \in C(0,T;S^n_+)$. For each $K \in \mathcal{K}$, we know from the classical Riccati theory that (4.2) admits a unique solution $P \in C(0,T;S^n_+)$. Thus we can define a mapping $\Psi : \mathcal{K} \to C(0,T;S^n)$ as $P = \Psi(K)$.

LEMMA 4.2. *The operator Ψ is monotonely increasing and continuous.*

Proof. Let $K, \tilde{K} \in \mathcal{K}, P = \Psi(K)$ and $\tilde{P} = \Psi(\tilde{K})$. Denote $\hat{P} = P - \tilde{P}$. Then $\hat{P}$ satisfies

$$\begin{cases} \dot{\hat{P}} + \hat{P}\hat{A} + \hat{A}'\hat{P} - \hat{P}B\tilde{K}^{-1}B'\hat{P} + \hat{Q} = 0, \\ \hat{P}(T) = 0, \end{cases} \tag{4.3}$$

with $\hat{A} = A - B\tilde{K}^{-1}B'\tilde{P}$ and $\hat{Q} = PB(\tilde{K}^{-1} - K^{-1})B'P$. Now if $K \geq \tilde{K} (> 0)$, then $\tilde{K}^{-1} \geq K^{-1} > 0$ which results in $\hat{Q} \geq 0$. Hence, the solution of the conventional Riccati equation (4.3) is nonnegative definite, namely, $\hat{P} \geq 0$. This proves the monotonicity. On the other hand, if $\tilde{K} \to K$, then by (4.3) and Gronwall's inequality it is easily seen that $P - \tilde{P} = \hat{P} \to 0$. This yields the desired continuity, and the proof is complete. ∎

LEMMA 4.3. *The Riccati equation* (4.1) *admits a solution if and only if there is a $K \in C(0,T;\hat{S}^m_+)$ such that*

$$R = K - D'\Psi(K)D. \tag{4.4}$$

Proof. The proof is obvious. ∎

LEMMA 4.4. *The Riccati equation* (4.1) *admits a solution if and only if there exist $K^+, K^- \in C(0,T;\hat{S}^m_+)$ such that*

$$K^+ \geq R + D'\Psi(K^+)D \geq R + D'\Psi(K^-)D \geq K^-. \tag{4.5}$$

Proof. Necessity. If (4.1) admits a solution P, then (4.5) trivially holds by letting $K^+ = K^- = R + D'PD$.

Sufficiency. Let K^+, K^- be given with (4.5) satisfied. Define sequences $\{K^+_i\}^\infty_0$, $\{K^-_i\}^\infty_0$, $\{P^+_i\}^\infty_0$, and $\{P^-_i\}^\infty_0$ iteratively as follows:

$$\begin{cases} K^+_0 = K^+,\ K^-_0 = K^-,\ P^+_0 = \Psi(K^+_0),\ P^-_0 = \Psi(K^-_0); \\ K^+_{i+1} = R + D'P^+_iD,\ K^-_{i+1} = R + D'P^-_iD, \\ P^+_{i+1} = \Psi(K^+_{i+1}),\ P^-_{i+1} = \Psi(K^-_{i+1}),\ \ i = 0,1,2,\cdots \end{cases} \tag{4.6}$$

By (4.5), we have

$$K^+_0 \geq K^+_1 \geq K^-_1 \geq K^-_0 > 0.$$

Since Ψ is increasing (Lemma 4.2), we also have

$$P^+_0 \geq P^+_1 \geq P^-_1 \geq P^-_0 \geq 0.$$

By induction, we obtain

$$\begin{cases} P^+_0 \geq P^+_i \geq P^+_{i+1} \geq P^-_{i+1} \geq P^-_i \geq P^-_0 \geq 0, \\ K^+_0 \geq K^+_i \geq K^+_{i+1} \geq K^-_{i+1} \geq K^-_i \geq K^-_0 > 0, \end{cases} \tag{4.7}$$

for $i = 1,2,\cdots$.

From (4.10), we see that $K_i^+ \in \mathcal{K}$ and there exist $K^+ \in \mathcal{K}$ and P^+ such that

$$\lim_{i\to\infty} K_i^+ = K^+, \quad \lim_{i\to\infty} P_i^+ = P^+. \tag{4.8}$$

By Lemma 4.2, we have

$$P^+ = \lim_{i\to\infty} P_i^+ = \lim_{i\to\infty} \Psi(K_i^+) = \Psi(\lim_{i\to\infty} K_i^+) = \Psi(K^+). \tag{4.9}$$

From (4.9) we see $P^+ \in C(0,T;S^n)$, since it is a solution of Riccati equation (4.2) corresponding to $K = K^+$. This in turn yields $K^+ = R + D'P^+D \in C(0,T;\hat{S}_+^m)$. By Lemma 4.3, we conclude that P^+ is a solution to (4.1). □

Remark 4.1. By the same argument, the following limits exist

$$P^- = \lim_{i\to\infty} P_i^-, \qquad K^- = \lim_{i\to\infty} K_i^-$$

and P^- is also a solution of (4.1). In view of the uniqueness of the solutions, $P^- = P^+$.

Remark 4.2. If (4.5) holds, then we have the following algorithms to compute the solution of the stochastic Riccati equation:

$$K_0 = K^+, \quad P_i = \Psi(K_i), \quad K_{i+1} = R + D'P_iD, \quad i = 0,1,2,\cdots, \tag{4.10}$$

or

$$K_0 = K^-, \quad P_i = \Psi(K_i), \quad K_{i+1} = R + D'P_iD, \quad i = 0,1,2,\cdots. \tag{4.11}$$

The following proposition gives an estimate for the convergence speed of the algorithms (4.10) and (4.11).

PROPOSITION 4.5. *Suppose* (4.5) *holds. Let the sequence* $\{P_i\} \subset C(0,T;S^n)$ *be constructed by either algorithm* (4.10) *or* (4.11) *and let* P *be the solution to the Riccati equation* (4.1). *Then*

$$|P_i(t) - P(t)| \le C\sum_{j=i}^{\infty} \frac{c^{j-2}}{(j-2)!}(T-t)^{j-2}, \qquad i = 2,3,\cdots, \tag{4.12}$$

where C *is a constant that depends only on the coefficients of* (4.1).

Proof. We only prove the estimate (4.12) for the algorithm (4.10). The other one is the same. By definition,

$$P_i(t) = H - \int_t^T [P_iA + A'P_i - P_iBK_i^{-1}B'P_i + Q](s)ds.$$

Set $\hat{P}_i = P_{i+1} - P_i$. Then

$$\begin{aligned}\hat{P}_i(t) = &-\int_t^T [\hat{P}_i(A - BK_{i+1}^{-1}B'P_{i+1}) + (A - BK_i^{-1}B'P_i)'\hat{P}_i \\ &+P_iBK_{i+1}^{-1}D'\hat{P}_{i-1}DK_i^{-1}B'P_{i+1}](s)ds.\end{aligned} \tag{4.13}$$

In view of (4.7), the sequences $\{|P_i|\}, \{|K_i|\}$, and $\{|K_i^{-1}|\}$ are bounded. Hence,

$$|\hat{P}_i(t)| \le C\int_t^T [|\hat{P}_{i-1}(s)| + |\hat{P}_i(s)|]ds. \tag{4.14}$$

Denote $v_i(t) = \int_t^T |\hat{P}_i(s)|ds$. Then (4.14) reads

$$\dot{v}_i(t) + Cv_i(t) + Cv_{i-1}(t) \geq 0,$$

which implies

$$v_i(t) \leq Ce^{CT} \int_t^T v_{i-1}(s)ds \equiv c \int_t^T v_{i-1}(s)ds.$$

By induction, we deduce that

$$v_{i+1}(t) \leq \frac{c^i}{i!}(T-t)^i v_1(0).$$

It then follows from (4.14) that

$$|\hat{P}_i(t)| \leq C\Big\{\frac{c^{i-1}}{(i-1)!}(T-t)^{i-1} + \frac{c^{i-2}}{(i-2)!}(T-t)^{i-2}\Big\}v_1(0). \tag{4.15}$$

This easily yields (4.12). □

Let us examine the condition (4.5) again. Indeed, the existence of K^+ is guaranteed. To see this, put $K^+ = zI$ with $z > 0$. Since $\Psi(K)$ is bounded uniformly for all $K \in C(0,T;\hat{S}_+^m)$, we see that

$$K^+ \geq R + D'\Psi(K^+)D$$

for z sufficiently large. Hence, we have the following result.

THEOREM 4.6. *The Riccati equation* (4.1) *admits a solution if and only if there exist* $K \in C(0,T;\hat{S}_+^m)$ *such that*

$$R + D'\Psi(K)D \geq K. \tag{4.16}$$

This theorem says that while R can be indefinite (or negative definite) for the Riccati equation to have solutions, it cannot be *too* negative. Indeed, in any case, R cannot be smaller than $\inf_{K\in C(0,T;\hat{S}_+^m)} \big[K - D'\Psi(K)D\big]$!

In particular, the condition (4.16) is satisfied automatically if R is positive definite, i.e., $R \in C(0,T;\hat{S}_+^m)$, so we have the following corollary.

COROLLARY 4.7. *If* $R \in C(0,T;\hat{S}_+^m)$, *then the stochastic Riccati equation admits a unique solution* $P \in C(0,T;S_+^n)$.

Theorem 4.6 is very useful in estimating intervals where the Riccati equation (4.1) is solvable without directly dealing with (4.1) itself. Let us use an example to demonstrate.

Example 4.1. Consider the stochastic LQR problem in Example 3.2, where $-1 < r < 0$. By Theorem 4.6, the Riccati equation (3.13) admits a solution on some interval $[\tau,1]$ $(0 \leq \tau \leq 1)$ if and only if there is a positive continuous function $K(\cdot)$ such that

$$r + \Psi(K)(t) \geq K(t), \quad \forall t \in [\tau,1], \tag{4.17}$$

where $\Psi(K)$ is the solution of the conventional Riccati equation

$$\dot{x}(t) = \frac{x^2(t)}{K(t)}, \qquad x(1) = 1.$$

So $\Psi(K)(t) = (1 + \int_t^1 \frac{1}{K(s)} ds)^{-1}$. Setting $f(t) = \frac{1}{K(1-t)}$, one can rewrite (4.17) as

$$1 + \int_0^{1-t} f(s)ds \leq \frac{f(1-t)}{1 - rf(1-t)}, \ \ \forall t \in [\tau, 1],$$

or

$$\int_0^t f(s)ds \leq \frac{(1+r)f(t) - 1}{1 - rf(t)}, \ \ \forall t \in [0, 1-\tau]. \tag{4.18}$$

Since $f(t) > 0$, it is then necessary that $f(t) > \frac{1}{1+r}$ for $t \in [0, 1-\tau]$. Putting

$$f(t) = \frac{1}{1+r} + f_1(t) \ \text{ with } f_1(t) > 0, \ \forall t \in [0, 1-\tau],$$

and substituting into (4.18), we obtain

$$\int_0^t f_1(s)ds \leq \frac{(1+r)^2 f_1(t)}{1 - r(1+r)f_1(t)} - \frac{t}{1+r}, \ \ \forall t \in [0, 1-\tau]. \tag{4.19}$$

Hence, f_1 must satisfy

$$(1+r)[(1+r)^2 + rt]f_1(t) > t, \ \ \forall t \in [0, 1-\tau]. \tag{4.20}$$

It follows that

$$(1+r)^2 + rt > 0, \ \ \forall t \in [0, 1-\tau]. \tag{4.21}$$

The above inequality gives an estimate on the interval where the Riccati equation (3.13) admits a solution, namely,

$$1 - \tau < \left(\frac{1}{\sqrt{-r}} - \sqrt{-r} \right)^2.$$

In particular, in order for (3.13) to be solvable on the whole interval [0,1], it must hold that

$$1 < \left(\frac{1}{\sqrt{-r}} - \sqrt{-r} \right)^2, \ \text{or } r > \frac{\sqrt{5} - 3}{2}.$$

Therefore, while r is allowed to be negative, it cannot be *too* negative.

5. The gap between stochastic maximum principle and LQR problem. This section is going to reveal a gap between Peng's stochastic maximum principle [12, 15], which is regarded as the best result so far in terms of necessary conditions for stochastic optimality, and the solvability of the stochastic LQR problem.

For the reader's convenience, let us state here Peng's maximum principle for general nonlinear systems, and then specialize to the LQR model.

Given $(s, y) \in [0, T) \times R^n$, we are to

$$\text{Minimize } J(s, y; u(\cdot)) = E\left[\int_s^T l(t, x(t), u(t))dt + h(x(1)) \right], \tag{5.1}$$

$$\text{Subject to } \begin{cases} dx(t) = f(t, x(t), u(t))dt + \sigma(t, x(t), u(t))dW(t), \\ x(s) = y. \end{cases} \tag{5.2}$$

The set of admissible controls U_{ad} is defined similarly with U being a given closed set in R^m. Peng's maximum principle asserts that if $(x^*(\cdot), u^*(\cdot))$ is optimal, then it must satisfy

$$\begin{aligned} &\tfrac{1}{2}\mathrm{tr}\Big\{[\sigma(t, x^*(t), u) - \sigma(t, x^*(t), u^*(t))]' P_0(t)[\sigma(t, x^*(t), u) - \sigma(t, x^*(t), u^*(t))]\Big\} \\ &+p'(t)[f(t, x^*(t), u) - f(t, x^*(t), u^*(t))] + q'(t)[\sigma(t, x^*(t), u) - \sigma(t, x^*(t), u^*(t))] \\ &+l(t, x^*(t), u) - l(t, x^*(t), u^*(t)) \geq 0, \ \forall u \in U, \ P-a.s., \ a.e.\ t \in [s, T], \end{aligned} \tag{5.3}$$

where $(p(\cdot), q(\cdot))$ is the $\mathcal{F}_t$-adapted solution to the *first-order* adjoint equation

$$\begin{cases} dp(t) = -[f_x^*(t)' p(t) + \sigma_x^*(t)' q(t) + l_x^*(t)]dt + q(t)dB(t), \\ p(T) = h_x(x^*(T)), \end{cases} \tag{5.4}$$

and $(P_0(t), \Lambda_0(t))$ is the $\mathcal{F}_t$-adapted solution to the *second-order* adjoint equation

$$\begin{cases} dP_0(t) = -[f_x^*(t)' P_0(t) + P_0(t) f_x^*(t) + \sigma_x^*(t)' P_0(t) \sigma_x^*(t) \\ \qquad\quad +\sigma_x^*(t)' \Lambda_0(t) + \Lambda_0(t)\sigma_x^*(t) + \Phi(t)]dt + \Lambda_0(t)dW(t), \\ P_0(T) = h_{xx}(x^*(T)), \end{cases} \tag{5.5}$$

with $\Phi(t) = l_{xx}^*(t) + \sum_{i=1}^n \{p^i(t) f_{xx}^{i*}(t) + q^i(t)\sigma_{xx}^{i*}(t)\}$. In the above, we used the notation $f^*(t) = f(t, x^*(t), u^*(t))$, etc. for simplicity.

Define the generalized Hamiltonian

$$G(t, x, u, p, S) = -\frac{1}{2}\mathrm{tr}\Big(\sigma(t, x, u)' S \sigma(t, x, u)\Big) - p' f(t, x, u) - l(t, x, u), \tag{5.6}$$

for $(t, x, u, p, S) \in [0, T] \times R^n \times R^m \times R^n \times S^n$, and an $\mathcal{H}$-function *corresponding to* the optimal pair $(x^*(\cdot), u^*(\cdot))$ as follows:

$$\mathcal{H}(t, x, u) = G(t, x, u, p(t), P_0(t)) - \sigma(t, x, u)'[q(t) - P_0(t)\sigma(t, x^*(t), u^*(t))], \tag{5.7}$$

for $(t, x, u) \in [s, T] \times R^n \times R^m$, where $p(t), q(t)$, and $P_0(t)$ are determined by adjoint equations (5.4) and (5.5). Then (5.3) is equivalent to the following maximum condition:

$$\mathcal{H}(t, x^*(t), u^*(t)) = \max_{u \in U} \mathcal{H}(t, x^*(t), u), \ \ P-a.s.; \ \ a.e.\ t \in [s, T]. \tag{5.8}$$

In stochastic optimal control theory, the system consisting of the state equation (5.2), adjoint equations (5.4), (5.5), and the maximum condition (5.8) is called a *Hamiltonian system* [14].

Applying the above maximum principle to the LQR model (2.1)–(2.3), we obtain the following result.

THEOREM 5.1. *Let $(x^*(\cdot), u^*(\cdot))$ be an optimal pair for $C_{s,y}$. Then there exist $\mathcal{F}_t$-adapted (p, q) and (P_0, Λ_0) satisfying*

$$\begin{cases} dp(t) = -(A'(t)p(t) + Q(t)x^*(t) + C'(t)q(t))dt + q(t)dW(t), \\ p(T) = Hx^*(T) \end{cases} \tag{5.9}$$

and

(5.10)
$$\begin{cases} dP_0(t) = -(A'(t)P_0(t) + P_0(t)A(t) + C'(t)P_0(t)C(t) + \Lambda_0(t)C(t) + C'(t)\Lambda_0(t) + Q(t))dt \\ \qquad\quad +\Lambda_0(t)dW(t), \\ P_0(T) = H \end{cases}$$

such that

$$R(t)u^*(t) + B'(t)p(t) + D'(t)q(t) = 0, \tag{5.11}$$

$$R(t) + D'(t)P_0(t)D(t) \geq 0, \quad P-a.s., \quad a.e.\ t \in [s,T]. \tag{5.12}$$

Proof. First of all, it is clear that the first- and second-order adjoint equations in the present LQR case are (5.9) and (5.10), respectively. Moreover, $\mathcal{H}(t, x^*(t), u)$ is a quadratic function in u, which attains its maximum at $u^*(t)$ by the maximum condition (5.8). Therefore, it is easily verified that (5.11) and (5.12) are nothing but the first- and second-order conditions, respectively, of the maximum point $u^*(t)$ for the quadratic function $\mathcal{H}(t, x^*(t), \cdot)$. □

We note that the second-order adjoint equation (together with (5.12)) is similar in form to the stochastic Riccati equation (3.1), except that the latter has an additional nonlinear term in its drift coefficient. The relationship between $P_0(t)$ and $P(t)$ is stated in the following proposition.

PROPOSITION 5.2. *If the LQR problem* (2.1)–(2.3) *is well-posed, then*

$$P_0(t) \geq P(t), \ P-a.s., \ \forall t \in [0,T]. \tag{5.13}$$

Proof. Let $(x^*(\cdot), u^*(\cdot))$ be an optimal pair for the problem $C_{0,y}$. Then by [15, Theorem 3.1],

$$P_0(t) \geq V_{xx}(t, x^*(t)), \ P-a.s., \ \forall t \in [0,T], \tag{5.14}$$

provided $V \in C^{1,2}([0,T] \times R^n)$. Now that the LQR problem is well-posed, $V(s,y) = \frac{1}{2}x'P(t)x$, which is smooth enough with $V_{xx}(t, x^*(t)) = P(t)$. The desired result follows. □

The inequality in (5.14) (and, therefore, (5.13) for the LQR problem) could be strict. An example is given in [15, p. 159], where the terminal cost is nonlinear. Another example is the LQR problem studied in Example 3.2. For that problem, it has been shown that $P(t) < 1$ for $t \in [0,1)$. However, The solution to (5.10) is $P_0(t) \equiv 1$.

For deterministic LQR problems, it is well known that the Hamiltonian system completely characterizes the optimal control, namely, a solution of the Hamiltonian system is an optimal pair of the LQR problem and vice versa. In this sense, the maximum principle and the well-posedness of the LQR problem are actually equivalent to each other. It is then natural to expect that in the stochastic case, the solvability of the Hamiltonian system (5.9)–(5.12) would yield the well-posedness of the LQR problem. Unfortunately, it is not true. Indeed, due to Proposition 5.2 and the fact that the inequality therein could be strict, the condition (5.12) is weaker than the inequality involved in the Riccati equation (3.1). Now let us look at an example to make it precise.

Example 5.1. Consider the following

$$\begin{array}{ll} \text{Minimize} & J = E^s\Big\{\int_s^1 -\frac{1}{2}u^2(t)dt + \frac{1}{2}x^2(1)\Big\} \\ \text{Subject to} & \begin{cases} dx(t) = (C(t)x(t) + D(t)u(t))dW(t), \\ x(s) = y, \end{cases} \end{array} \tag{5.15}$$

where C and D are bounded deterministic functions satisfying

$$\exp\left[-\int_0^1 C^2(r)dr\right] < 1, \qquad D(t) = \exp\left[-\frac{1}{2}\int_t^1 C^2(r)dr\right]. \tag{5.16}$$

The system (5.2), (5.9)–(5.12) in this case is

$$\begin{cases} dx^*(t) = (C(t)x^*(t) + D(t)u^*(t))dW(t), & x(s) = y, \\ dp(t) = -C(t)q(t)dt + q(t)dW(t), & p(1) = x^*(1), \\ dP_0(t) = -C^2(t)P_0(t)dt, & P_0(1) = 1, \\ -u^*(t) + D(t)q(t) = 0, \\ -1 + D^2(t)P_0(t) \geq 0. \end{cases} \tag{5.17}$$

Since $P_0(t) = \exp[\int_t^1 C^2(r)dr]$, the last inequality in the above system is satisfied by virtue of (5.16). Hence (5.17) reduces to

$$\begin{cases} dx^*(t) = (C(t)x^*(t) + D^2(t)q(t))dW(t), & x^*(s) = y, \\ dp(t) = -C(t)q(t)dt + q(t)dW(t), & p(1) = x^*(1). \end{cases} \tag{5.18}$$

This is a forward-backward stochastic differential equation, and it can be shown by using a standard contraction mapping theorem that it admits a unique solution $(x^*(\cdot), p(\cdot), q(\cdot))$ if $1 - s > 0$ is small enough (cf. Ma and Yong [10, Theorem 1.5.1]). Now we are going to show that the original LQR problem is *not* well-posed. Fix s and $y \neq 0$. Using the Ito formula, we have

$$dx^2(t) = 2x(t)(C(t)x(t) + D(t)u(t))dW(t) + (C(t)x(t) + D(t)u(t))^2dt. \tag{5.19}$$

Thus

$$J(s, y; u(\cdot)) = \frac{1}{2}E^s\int_s^1 [-u^2(t) + (C(t)x(t) + D(t)u(t))^2]dt + \frac{1}{2}y^2. \tag{5.20}$$

By (5.16), for a sufficiently small $\varepsilon_0 > 0$, one can find a $\delta_0 > 0$ such that

$$1 - D^2(t) \geq \varepsilon_0, \ \ \forall t \in [s, 1 - \delta_0]. \tag{5.21}$$

For an integer $k > 0$, define a feedback control

$$u_k(t) = \begin{cases} kC(t)x(t), & s \leq t < 1 - \delta_0, \\ -D^{-1}(t)C(t)x(t), & 1 - \delta_0 \leq t \leq 1. \end{cases} \tag{5.22}$$

It follows from (5.20) that

$$\begin{aligned} J(s, y; u_k(\cdot)) &= \frac{1}{2}E^s\int_s^{1-\delta_0} [-k^2 + (1 + kD(t))^2]|C(t)x(t)|^2dt - \frac{1}{2}E^s\int_{1-\delta_0}^1 u_k^2dt + \frac{1}{2}y^2 \\ &= -\frac{1}{2}E^s\int_s^{1-\delta_0} \left[(1 - D^2(t))(k - \frac{D(t)}{1 - D^2(t)})^2 - \frac{1}{1 - D^2(t)}\right]|C(t)x(t)|^2dt \\ &\quad -\frac{1}{2}E^s\int_{1-\delta_0}^1 u_k^2(t)dt + \frac{1}{2}y^2. \end{aligned}$$

Noting that

$$E^s x^2(t) = \exp \int_s^t (1 + D(t)k)^2 C^2(t) dt \cdot y^2 > 0,$$

we conclude

$$J(s, y; u_k(\cdot)) \to -\infty$$

as $k \to +\infty$. So the LQR problem is not well-posed.

6. Concluding remarks. In this paper, we have studied a general class of stochastic linear quadratic regulators with the diffusion coefficients dependent on the control variables. It is observed that the optimal control problem may be well-posed even when the control weight costs are indefinite by virtue of the uncertainty involved. A new stochastic Riccati equation is introduced, and the existence of solutions to it is shown to be sufficient for the well-posedness of the LQR problem. It is also found that the stochastic maximum principle cannot fully characterize the optimality. These distinctive features reveal some fundamental differences between the deterministic and stochastic situations.

Many interesting and challenging problems remain open. The first problem is the existence and uniqueness of solutions to the stochastic Riccati equation in a general situation. The resolution of this problem might have to involve some delicate analysis on the nonlinear backward SDEs. Second, how do we numerically solve the stochastic Riccati equation? For the special case (4.1), this paper gives an algorithm (Remark 4.2) along with its convergence speed (Proposition 4.5). Note that the algorithm involves computing the operator Ψ (i.e., the solution to the conventional Riccati equation), and numerical schemes have been widely available for solving Ψ. However, numerical solutions to the general Riccati equation (3.1) remain an interesting but perhaps difficult problem; to the best of our knowledge there has been few works on numerically solving nonlinear backward stochastic differential equations. Last but not least, what is a "better" stochastic maximum principle that can fully solve the LQR problem as in the deterministic case? These topics will be studied in forthcoming papers.

Acknowledgments. The authors wish to thank S. Peng and J. Yong for inspiring discussions. Thanks are also due to the referees for their constructive comments that led to an improved version of the paper.

REFERENCES

[1] B.D.O. Anderson and J.B. Moore, *Optimal Control–Linear Quadratic Methods*, Prentice–Hall, Englewood Cliffs, NJ, 1989.
[2] A. Bensoussan, *Lecture on stochastic control, part* I, in Nonlinear Filtering and Stochastic Control, Lecture Notes in Math. 972, Springer-Verlag, Berlin, 1983, pp. 1–39.
[3] J. M. Bismut, *Analyse Convexe et Probabilites*, These, Faculte des Sciences de Paris, 1973.
[4] J. M. Bismut, *An introductory approach to duality in stochastic control*, SIAM Rev., 20 (1978), pp. 62–78.
[5] M.H.A. Davis, *Linear Estimation and Stochastic Control*, Chapman and Hall, London, 1977.
[6] W. H. Fleming, *Optimal control of partially observable diffusions,* SIAM J. Control Optim., 6 (1968), pp. 194–215.
[7] U. G. Haussmann, *A Stochastic Maximum Principle for Optimal Control of Diffusions*, Pitman, Boston, 1986.
[8] R. E. Kalman, *Contributions to the theory of optimal control,* Bol. Soc. Math. Mexicana, 5 (1960), pp. 102–119.

[9] H. J. Kushner, *Necessary conditions for continuous parameter stochastic optimization problems*, SIAM J. Control Optim., 10 (1972), pp. 550–565.
[10] J. Ma and J. Yong, *Forward-Backward Stochastic Differential Equations and Their Applications*, in preparation.
[11] E. Pardoux and S. Peng, *Adapted solutions of backward stochastic equations*, System Control Lett., 14 (1990), pp. 55–61.
[12] S. Peng, *A general stochastic maximum principle for optimal control problems*, SIAM J. Control Optim., 28 (1990), pp. 966–979.
[13] W. M. Wonham, *On a matrix Riccati equation of stochastic control,* SIAM J. Control Optim., 6 (1968), pp. 312–326.
[14] J. Yong and X. Y. Zhou, *Stochastic Controls: Hamiltonian Systems and HJB Equations*, Springer-Verlag, New York, to appear.
[15] X. Y. Zhou, *A unified treatment of maximum principle and dynamic programming in stochastic controls*, Stochastics Stochastics Rep., 36 (1991), pp. 137–161.
[16] X. Y. Zhou, *Sufficient conditions of optimality for stochastic systems with controllable diffusions*, IEEE Trans. Automat. Control, 41 (1996), pp. 1176–1179.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 5, pp. 1703–1749, September 1998

010

ON QUADRATIC DIFFERENTIAL FORMS*

J. C. WILLEMS† AND H. L. TRENTELMAN†

Abstract. This paper develops a theory around the notion of quadratic differential forms in the context of linear differential systems. In many applications, we need to not only understand the behavior of the system variables but also the behavior of certain functionals of these variables. The obvious cases where such functionals are important are in Lyapunov theory and in LQ and H_∞ optimal control. With some exceptions, these theories have almost invariably concentrated on first order models and state representations. In this paper, we develop a theory for linear time-invariant differential systems and quadratic functionals. We argue that in the context of systems described by *one-variable* polynomial matrices, the appropriate tool to express quadratic functionals of the system variables are *two-variable* polynomial matrices. The main achievement of this paper is a description of the interaction of one- and two-variable polynomial matrices for the analysis of functionals and for the application of higher order Lyapunov functionals.

Key words. quadratic differential forms, linear systems, polynomial matrices, two-variable polynomial matrices, Lyapunov theory, positivity, spectral factorization, dissipativeness, storage functions

AMS subject classifications. 93A10, 93A30, 93D05, 93D20, 93D30, 93C05, 93C45

PII. S0363012996303062

1. Introduction.

In the theory of models for dynamical systems, it has been customary to consider both external input/output as well as state space models. Also, there is a well developed theory for passing from one type of model to another. Thus, there are efficient algorithms for passing from a convolution, to a transfer function, to a state model, and back. Even for stochastic and nonlinear systems, there are methods for associating a first order state representation to a high order model.

However, in addition to understanding the interaction between system variables, we need in many applications to understand also the behavior of certain functionals of these variables. The obvious cases where such functionals are crucial are in Lyapunov theory, in the theory of dissipative systems, and in optimal control. In these contexts it is remarkable to observe that the theory of dynamics has almost invariably concentrated on first order models and state representations. Thus, in studying system stability using Lyapunov methods, we are constrained to consider state representations, and optimal control problems invariably assume that the cost is an integral of a function of the state and the input. The question thus occurs of whether it is possible to develop an external theory—for example, Lyapunov theory—for systems and functionals so that analysis of stability and passivity, for instance, could proceed on the basis of a *first principles* model instead of first having to find a state representation. In this paper, we consider models that are not in state form (even though some proofs use state representations). Our models are externally specified yet they are not completely general *first principles* models in that we concentrate on models in kernel or in image representation.

It is the purpose of this paper to develop such a theory. We do not, however, set our aims too high and start with a very well-understood class of systems and functionals: linear time-invariant differential systems and quadratic functionals in the

*Received by the editors May 6, 1996; accepted for publication (in revised form) September 9, 1997; published electronically June 22, 1998.
http://www.siam.org/journals/sicon/36-5/30306.html
†Research Institute for Mathematics and Computing Science, P.O. Box 800, 9700 AV Groningen, The Netherlands (j.c.willems@math.rug.nl, h.l.trentelman@math.rug.nl).

system variables and their derivatives. We shall see that *one-variable* polynomials are the appropriate tool in which to parametrize the model (see also, among others, [16], [17]) and *two-variable* polynomials are the appropriate tool for parametrizing the functionals. Thus, the paper presents an interesting interplay between one- and two-variable polynomial matrices. Two-variable polynomials turn out to be a very effective tool for analyzing linear systems with quadratic functionals.

This paper consists of a series of general concepts and questions, combined with some specific results concerned with Lyapunov stability and with dissipativity, i.e., with positivity of (integrals of) quadratic differential forms. As such, the paper aims at making a contribution to the development of the very useful and subtle notions of dissipative and lossless (conservative) systems.

In companion papers, these ideas will be applied to LQ and H_∞ problems. The main achievement of this paper—the interaction of one- and two-variable polynomial matrices for the analysis of functionals and application in higher order Lyapunov functions—appears to be new. However, seeds of this have appeared previously in the literature. We mention especially Brockett's early work on *path integrals* [7], [8] in addition to classical work on Routh–Hurwitz-type conditions (see, for example, [6]), and early work by Kalman [13], [14].

2. Review. In order to make this paper reasonably self-contained, we first introduce some notation and some basic facts from the behavioral approach to linear dynamical systems. References in which more details can be found include [31], [32], and [33].

We will deal exclusively with continuous-time real linear time-invariant differential dynamical systems. Thus, the time axis is $\mathbb{R}$, the signal space is $\mathbb{R}^q$ (the number of variables q, of course, depends on the case at hand), and the behavior $\mathfrak{B}$ is the solution set of a system of linear constant coefficient differential equations

$$R\left(\frac{d}{dt}\right)w = 0 \tag{2.1}$$

in the real variables $w_1, w_2, \cdots, w_q$, arranged as the column vector w; R is a real polynomial matrix with, of course, q columns. The number of rows of R depends, as do its coefficients, on the particular dynamical system described by (2.1). Hence we denote this as $R \in \mathbb{R}^{\bullet\times q}[\xi]$, where ξ denotes the indeterminate. Thus, if $R(\xi) = R_0 + R_1\xi + \cdots + R_N\xi^N$, then (2.1) denotes the system of differential equations

$$R_0 w + R_1\frac{dw}{dt} + \cdots + R_N\frac{d^N w}{dt^N} = 0. \tag{2.2}$$

For the behavior, i.e., for the solution set of (2.1) or (2.2), it is usually advisable to consider locally integrable w's as candidate solutions and to interpret the differential equation in the sense of distributions. However, it is our explicit intention to avoid mathematical technicalities as much as possible in this paper. In keeping with this, we assume that the solution set consists of infinitely differentiable functions, even though many of the results are valid without this assumption. Hence the behavior of (2.1) is defined as

$$\mathfrak{B} = \left\{ w \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^q) \mid R\left(\frac{d}{dt}\right)w = 0 \right\}. \tag{2.3}$$

We denote the family of dynamical systems obtained this way by $\mathfrak{L}^q$. Hence elements of $\mathfrak{L}^q$ are dynamical systems $\Sigma = (\mathbb{R}, \mathbb{R}^q, \mathfrak{B})$ with time axis $\mathbb{R}$, signal space $\mathbb{R}^q$, and

behavior $\mathfrak{B}$ described through some $R \in \mathbb{R}^{\bullet\times q}[\xi]$ by (2.3). Note that instead of writing $\Sigma \in \mathfrak{L}^q$ we may as well write $\mathfrak{B} \in \mathfrak{L}^q$, and we prefer to use this notation in this paper.

As explained in the previous paragraphs, each $R \in \mathbb{R}^{\bullet\times q}[\xi]$ unambiguously defines a system $\mathfrak{B} \in \mathfrak{L}^q$. However, there are always many R's defining the same $\mathfrak{B} \in \mathfrak{L}^q$. For example, if U is any unimodular polynomial matrix such that the product UR makes sense, then R and UR induce the same element of $\mathfrak{L}^q$. Also, there are many other ways of specifying a given $\mathfrak{B} \in \mathfrak{L}^q$. Note that (2.1) describes $\mathfrak{B}$ as $\mathfrak{B} = \ker(R(\frac{d}{dt}))$ with $R(\frac{d}{dt})$ viewed as a map from $\mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^q)$ into $\mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^{\mathrm{rowdim}(R)})$. For obvious reasons we hence refer to (2.1) as *a kernel representation* of $\mathfrak{B} \in \mathfrak{L}^q$. We will meet other representations, in particular image, latent variable, input/output, state, and input/state/output representations. These are now briefly introduced.

A system $\mathfrak{B} \in \mathfrak{L}^q$ is said to be *controllable* if for each $w_1, w_2 \in \mathfrak{B}$ there exists a $w \in \mathfrak{B}$ and a $t' \geq 0$ such that $w(t) = w_1(t)$ for $t < 0$ and $w(t) = w_2(t - t')$ for $t \geq t'$. It can be shown that $\mathfrak{B}$ is controllable iff its kernel representation satisfies $\mathrm{rank}(R(\lambda)) = \mathrm{rank}(R)$ for all $\lambda \in \mathbb{C}$. Here, $\mathrm{rank}(R)$ is defined as the rank of R considered as a matrix with elements in the field $\mathbb{R}(\xi)$ of real rational functions. On the other hand, for a given $\lambda \in \mathbb{C}$, $R(\lambda)$ is a matrix with elements in $\mathbb{C}$. Accordingly, $\mathrm{rank}(R(\lambda))$ denotes the rank of the complex matrix $R(\lambda)$. It is easy to see that $\mathrm{rank}(R) = \max_{\lambda\in\mathbb{C}} \mathrm{rank}(R(\lambda))$.

Controllable systems are exactly those that admit image representations. More concretely, $\mathfrak{B} \in \mathfrak{L}^q$ is controllable iff there exists an $M \in \mathbb{R}^{q\times\bullet}[\xi]$ such that $\mathfrak{B} = \mathrm{im}(M(\frac{d}{dt}))$, with $M(\frac{d}{dt})$ viewed as a mapping from $\mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^{\mathrm{coldim}(M)})$ into $\mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^q)$. The resulting representation

$$w = M\left(\frac{d}{dt}\right)\ell \tag{2.4}$$

is called an *image representation* of $\mathfrak{B}$.

An image representation is a special case of what we call a latent variable representation of $\mathfrak{B}$. The system of differential equations

$$R\left(\frac{d}{dt}\right)w = M\left(\frac{d}{dt}\right)\ell \tag{2.5}$$

is said to be a *latent variable representation* of $\mathfrak{B} \in \mathfrak{L}^q$ if

$$\mathfrak{B} = \{w \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^q) \mid \exists\, \ell \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^\bullet) \text{ such that (2.5) holds}\}.$$

A latent variable representation is said to be *observable* if $(R(\frac{d}{dt})w = M(\frac{d}{dt})\ell_1$ and $R(\frac{d}{dt})w = M(\frac{d}{dt})\ell_2)$ implies $(\ell_1 = \ell_2)$. Observability is equivalent to the condition that $M(\lambda)$ is of full column rank for all $\lambda \in \mathbb{C}$. A controllable system, it turns out, always allows an observable image representation.

Of special interest in section 4 will be the observability of the system

$$A\left(\frac{d}{dt}\right)\ell = 0,\ w = C\left(\frac{d}{dt}\right)\ell. \tag{2.6}$$

Of course, the definition of observability applies to (2.6). If this is the case, then we call the pair of polynomial matrices (A, C) with the same number of columns an *observable pair*. Hence (A, C) is an observable pair iff

$$\begin{bmatrix} A(\lambda) \\ C(\lambda) \end{bmatrix} \tag{2.7}$$

is of full column rank for all $\lambda \in \mathbb{C}$.

Systems in $\mathfrak{L}^q$ admit many other useful representations. We already encountered kernel and image representations. Next, we introduce state and input/output representations.

In [22] the notion of state models and their construction has been discussed in detail. Here we limit ourselves to the bare essentials. Let $\mathfrak{B} \in \mathfrak{L}^q$. A latent variable representation (with the latent variable denoted by x this time) of the form (2.5) is said to be a *state model* if, whenever (w_1, x_1) and (w_2, x_2) are $\mathfrak{C}^\infty$-solutions of (2.5) with $x_1(0) = x_2(0)$, then the concatenation $(w_1, x_1) \wedge (w_2, x_2)$ also satisfies (2.5). Since this concatenation need not be in $\mathfrak{C}^\infty$, it need only be a weak solution of (2.5), that is, a solution in the sense of distributions. State models are governed by equations of the form (2.5) with special structure. In fact, (2.5) is a state model iff there exist matrices E, F, and G such that

$$Gw + Fx + E\frac{dx}{dt} = 0 \tag{2.8}$$

is equivalent to (2.5) in the case of state models. Thus (2.8) is called a *state representation* of the behavior $\mathfrak{B}$ if

$$\mathfrak{B} = \{w \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^q) \mid \exists\, x \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^n) \text{ such that (2.8) is satisfied}\}.$$

Here n denotes the dimension of the vector x. The important feature of (2.8) is that it is an (implicit) differential equation containing derivatives of order at most one in x and zero in w. We call a state representation *state minimal* if among all state representations of $\mathfrak{B}$, n is as small as possible. It is possible to prove that (2.8) is state minimal iff it is *state trim* (meaning that for all $a \in \mathbb{R}^n$ there exists $(w, x) \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^{q+n})$ such that $x(0) = a$) and observable. The dimension of the state space of a state minimal representation of $\mathfrak{B} \in \mathfrak{L}^q$ is called the *McMillan degree* of $\mathfrak{B}$. The notion of McMillan degree usually refers to properties of polynomial matrices. Actually, for the case at hand this correspondence holds in terms of full row rank kernel representation matrices R or observable image representation matrices M, but we do not need this correspondence in this paper.

Every system $\mathfrak{B} \in \mathfrak{L}^q$ also admits an input/output representation. By reordering the components of the vector w, if need be, we can decompose w into

$$w = \begin{bmatrix} u \\ y \end{bmatrix} \tag{2.9}$$

with, in terms of R, $\operatorname{rank}(R)$ components for y and $q - \operatorname{rank}(R)$ components for u, such that $\mathfrak{B} \in \mathfrak{L}^q$ admits the special kernel representation

$$P\left(\frac{d}{dt}\right) y = Q\left(\frac{d}{dt}\right) u, \tag{2.10}$$

with P square, $\det P \neq 0$, and $P^{-1}Q$ a matrix of proper rational functions. Thus, in (2.10) u has the usual properties of input and y those of output. Therefore (2.10) is called an *input/output representation*.

Actually, for controllable systems, we can also recover the input/output structure in terms of the image representation. Thus the image representation

$$\begin{bmatrix} u \\ y \end{bmatrix} = \begin{bmatrix} U(\frac{d}{dt}) \\ Y(\frac{d}{dt}) \end{bmatrix} \ell \tag{2.11}$$

is an input/output representation if U is square, $\det U \neq 0$, and YU^{-1} is a matrix of proper rational functions. The number of input components of a system in image representation (2.4) equals $\operatorname{rank}(M)$.

It is possible to combine the above, leading to the familiar input/state/output representation

$$\frac{dx}{dt} = Ax + Bu, \quad y = Cx + Du. \tag{2.12}$$

This representation is state minimal iff it is observable, i.e., iff (A, C) is an observable pair of matrices (not to be confused with an observable pair of polynomial matrices).

Summarizing, given any $w \in \mathfrak{B}$, we may partition the components of w into inputs and outputs. Also, there exists an $X \in \mathbb{R}^{\bullet \times q}[\xi]$ such that

$$x = X\left(\frac{d}{dt}\right) w \tag{2.13}$$

is a (minimal) state map for $\mathfrak{B}$. For a system in image representation (2.4) this leads to a state representation of the form

$$x = X'\left(\frac{d}{dt}\right) \ell. \tag{2.14}$$

The resulting relation between u and y is as in (2.10); that between w and x is as in (2.8); and that between u, y, and x is as in (2.12).

We need a few more details about the state construction for systems in image representation (2.4). Assume that M is of full column rank. Then after permutation of the components of w (i.e., of the rows of M), if need be, M is of the form

$$M = \begin{bmatrix} U \\ Y \end{bmatrix},$$

with U square, $\det(U) \neq 0$, and YU^{-1} a matrix of proper rational functions. The resulting system

$$\begin{bmatrix} u \\ y \end{bmatrix} = \begin{bmatrix} U(\frac{d}{dt}) \\ Y(\frac{d}{dt}) \end{bmatrix} \ell$$

is then an input/output representation. Consider all polynomial row vectors $F \in \mathbb{R}^{1 \times \bullet}[\xi]$ such that FU^{-1} is strictly proper. It can be shown (see [22]) that this set is a vector space. Now

$$x = X\left(\frac{d}{dt}\right) \ell \tag{2.15}$$

is a state map for (2.4) iff the rows of X span this vector space. It is a minimal state map iff the rows of X form a basis for this vector space.

Next, consider associated with (2.4) the variable v governed by

$$v = L\left(\frac{d}{dt}\right) \ell.$$

Then it follows from the above that there exist matrices P and Q such that

$$v = Px + Qu$$

iff LU^{-1} is proper. Moreover, Q is zero iff LU^{-1} is strictly proper, and Q is invertible iff LU^{-1} is biproper.

3. Quadratic differential forms. Differential equations and one-variable polynomial matrices play an essential role in describing the dynamics of systems, as we have seen in section 2 and the references given therein. When studying functions of the dynamical variables, as in Lyapunov theory, studying dissipation and passivity, or specifying performance criteria in optimal control, we invariably encounter quadratic expressions in the variables and their derivatives. As we shall see, two-variable polynomial matrices are the proper mathematical tool to express these quadratic functionals. We aim to illustrate throughout this paper that linear dynamical equations expressed through one-variable polynomial matrices, and quadratic functionals expressed through two-variable polynomial matrices fit as a glove fits a hand.

Let $\mathbb{R}^{q_1\times q_2}[\zeta,\eta]$ denote the set of real polynomial matrices in the (commuting) indeterminates ζ and η. Explicitly, an element $\Phi\in\mathbb{R}^{q_1\times q_2}[\zeta,\eta]$ is thus given by

$$\Phi(\zeta,\eta)=\sum_{k,\ell}\Phi_{k\ell}\zeta^k\eta^\ell. \tag{3.1}$$

The sum in (3.1) ranges over the nonnegative integers and is assumed to be finite, and $\Phi_{k\ell}\in\mathbb{R}^{q_1\times q_2}$. Such a Φ induces a *bilinear differential form* (BLDF), that is, the map

$$L_\Phi:\mathfrak{C}^\infty(\mathbb{R},\mathbb{R}^{q_1})\times\mathfrak{C}^\infty(\mathbb{R},\mathbb{R}^{q_2})\to\mathfrak{C}^\infty(\mathbb{R},\mathbb{R}) \tag{3.2}$$

defined by

$$(L_\Phi(v,w))(t):=\sum_{k,\ell}\left(\frac{d^kv}{dt^k}(t)\right)^T\Phi_{k\ell}\left(\frac{d^\ell w}{dt^\ell}(t)\right). \tag{3.3}$$

If $q_1=q_2$ $(=:q)$, then Φ induces a *quadratic differential form* (QDF)

$$Q_\Phi:\mathfrak{C}^\infty(\mathbb{R},\mathbb{R}^q)\to\mathfrak{C}^\infty(\mathbb{R},\mathbb{R}) \tag{3.4}$$

defined by

$$Q_\Phi(w):=L_\Phi(w,w). \tag{3.5}$$

Define the asterisk operator * by

$$^*:\mathbb{R}^{q_1\times q_2}[\zeta,\eta]\to\mathbb{R}^{q_2\times q_1}[\zeta,\eta];\quad \Phi^*(\zeta,\eta):=\Phi^T(\eta,\zeta), \tag{3.6}$$

where ^{T} denotes transposition. Obviously $L_\Phi(v,w)=L_{\Phi^*}(w,v)$. If $\Phi\in\mathbb{R}^{q\times q}[\zeta,\eta]$ satisfies $\Phi=\Phi^*$, then Φ is called *symmetric*. The symmetric elements of $\mathbb{R}^{q\times q}[\zeta,\eta]$ are denoted by $\mathbb{R}^{q\times q}_s[\zeta,\eta]$. Clearly

$$Q_\Phi=Q_{\Phi^*}=Q_{\frac{1}{2}(\Phi+\Phi^*)} \tag{3.7}$$

This shows that when considering quadratic differential forms, we can hence in principle restrict our attention to Φ's in $\mathbb{R}^{q\times q}_s[\zeta,\eta]$. However, both bilinear and quadratic forms are of interest to us.

Associated with $\Phi\in\mathbb{R}^{q_1\times q_2}[\zeta,\eta]$, we can form the matrix

$$\tilde\Phi=\begin{bmatrix}\Phi_{00}&\Phi_{01}&\cdots&\cdot&\cdots\\ \Phi_{10}&\Phi_{11}&\cdots&\cdot&\cdots\\ \vdots&\vdots&&\vdots&\\ \cdot&\cdot&\cdots&\Phi_{k\ell}&\cdots\\ \vdots&\vdots&&\vdots&\end{bmatrix}. \tag{3.8}$$

Note that, although $\tilde{\Phi}$ is an infinite matrix, all but a finite number of its elements are zero. We can factor $\tilde{\Phi}$ as $\tilde{\Phi} = \tilde{N}^T \tilde{M}$, with $\tilde{N}$ and $\tilde{M}$ infinite matrices having a finite number of rows and all but a finite number of elements equal to zero. This decomposition leads, after premultiplication by $[I_{q_1} \;\; I_{q_1}\zeta \;\; I_{q_1}\zeta^2 \cdots]$ and postmultiplication by $\mathrm{col}[I_{q_2} \;\; I_{q_2}\eta \;\; I_{q_2}\eta^2 \cdots]$, to the following factorization of Φ:

$$\Phi(\zeta, \eta) = N^T(\zeta) M(\eta). \tag{3.9}$$

This decomposition is not unique, but if we take $\tilde{N}$ and $\tilde{M}$ surjective, then their number of rows is equal to the rank of $\tilde{\Phi}$. The factorization (3.9) is then called a *canonical factorization* of Φ. Associated with (3.9), we obtain the following expression for the BLDF L_Φ:

$$L_\Phi(w_1, w_2) = \left(N\left(\frac{d}{dt}\right) w_1 \right)^T M\left(\frac{d}{dt}\right) w_2. \tag{3.10}$$

Next we discuss the case that Φ is symmetric. Clearly $\Phi = \Phi^*$ iff $\tilde{\Phi}$ is symmetric. In that case, it can be factored as $\tilde{\Phi} = \tilde{M}^T \Sigma_M \tilde{M}$ with $\tilde{M}$ an infinite matrix having a finite number of rows and all but a finite number of elements equal to zero, and Σ_M a signature matrix, i.e., a matrix of the form

$$\begin{bmatrix} I_{r_+} & 0 \\ 0 & -I_{r_-} \end{bmatrix}.$$

This decomposition leads to the following decomposition of Φ:

$$\Phi(\zeta, \eta) = M^T(\zeta) \Sigma_M M(\eta). \tag{3.11}$$

Also, this decomposition is not unique but if we take $\tilde{M}$ surjective, then Σ_M is unique. We denote this Σ_M as Σ_Φ and the resulting pair (r_-, r_+) by (ϕ_-, ϕ_+). This pair is called the *inertia* of Φ. The resulting factorization

$$\Phi(\zeta, \eta) = M^T(\zeta) \Sigma_\Phi M(\eta) \tag{3.12}$$

is called a *symmetric canonical factorization* of Φ. Of course, a symmetric canonical factorization is not unique. However, they can all be obtained from one by replacing $M(\xi)$ by $UM(\xi)$ with $U \in \mathbb{R}^{\mathrm{rank}(\tilde{\Phi}) \times \mathrm{rank}(\tilde{\Phi})}$ such that $U^T \Sigma_\Phi U = \Sigma_\Phi$.

Associated with (3.11), we obtain the following decomposition of Q_Φ into a sum of positive and negative squares:

$$Q_\Phi(w) = \left\| P\left(\frac{d}{dt}\right) w \right\|^2 - \left\| N\left(\frac{d}{dt}\right) w \right\|^2, \tag{3.13}$$

where $N, P \in \mathbb{R}^{\bullet \times q}[\xi]$ are obtained by partitioning $\tilde{M}$ conform Σ_M as:

$$\tilde{M} = \begin{bmatrix} \tilde{P} \\ \tilde{N} \end{bmatrix}. \tag{3.14}$$

For a given symmetric $\Phi(\zeta, \eta)$ we are also interested in the symmetric two-variable polynomial matrix $|\Phi|(\zeta, \eta)$, the *absolute value* of Φ, which we define as follows. For a given real symmetric matrix $A \in \mathbb{R}^{n \times n}$ define its absolute value, $|A| \in \mathbb{R}^{n \times n}$, as the unique symmetric nonnegative definite matrix $X \in \mathbb{R}^{n \times n}$ such that $X^2 = A^2$. This

matrix $|A|$ can be computed as follows. Factor $A = \tilde{U}^T \Lambda \tilde{U}$, where Λ is the diagonal matrix with, on its diagonal, the nonzero eigenvalues of A in decreasing order, and where $\tilde{U}\tilde{U}^T = I$, and define $|A| := \tilde{U}^T |\Lambda| \tilde{U}$ with $|\Lambda|$ defined in the obvious way. Let $\tilde{\Phi}$ be the symmetric matrix associated with $\Phi(\zeta, \eta)$. Let $|\tilde{\Phi}|$ be the absolute value of $\tilde{\Phi}$. Next, define $|\Phi|(\zeta, \eta)$ as the symmetric two-variable polynomial matrix associated with $|\tilde{\Phi}|$:

$$|\Phi|(\zeta, \eta) := \begin{bmatrix} I \\ \zeta I \\ \zeta^2 I \\ \vdots \end{bmatrix}^T |\tilde{\Phi}| \begin{bmatrix} I \\ \eta I \\ \eta^2 I \\ \vdots \end{bmatrix}.$$

Note that a factorization $\tilde{\Phi} = \tilde{U}^T \Lambda \tilde{U}$ immediately yields a symmetric canonical factorization of $\Phi(\zeta, \eta)$. Indeed, define $\tilde{M}_c := \sqrt{|\Lambda|}\tilde{U}$. We then have

$$\tilde{\Phi} = \tilde{U}^T \sqrt{|\Lambda|} \Sigma_\Phi \sqrt{|\Lambda|} \tilde{U} = \tilde{M}_c^T \Sigma_\Phi \tilde{M}_c, \tag{3.15}$$

with $\tilde{M}_c$ surjective. The corresponding $M_c(\xi) := \tilde{M}_c \operatorname{col}[I \quad \xi I \quad \xi^2 I \quad \ldots]$ then yields a canonical factorization $\Phi(\zeta, \eta) = M_c^T(\zeta) \Sigma_\Phi M_c(\eta)$. This particular canonical factorization has the property that

$$|\Phi|(\zeta, \eta) = M_c^T(\zeta) M_c(\eta). \tag{3.16}$$

In general, if $M(\xi)$ is any canonical factor of Φ, then we have $UM(\xi) = M_c(\xi)$ with U satisfying $U^T \Sigma_\Phi U = \Sigma_\Phi$, and hence $|\Phi|(\zeta, \eta) = M^T(\zeta) U^T U M(\eta)$.

One of the conveniences of identifying BLDFs and QDFs with two-variable polynomial matrices is that they allow a very convenient calculus. One instance of this is differentiation. Obviously if L_Φ is a BLDF, so is $\frac{d}{dt} L_\Phi$, and if Q_Φ is a QDF, so is $\frac{d}{dt} Q_\Phi$. The result of differentiation is easily expressed in terms of the two-variable polynomial matrices and leads to the dot operator $\bullet$ defined as

$$\bullet : \mathbb{R}^{q_1 \times q_2}[\zeta, \eta] \to \mathbb{R}^{q_1 \times q_2}[\zeta, \eta]; \quad \overset{\bullet}{\Phi}(\zeta, \eta) := (\zeta + \eta)\Phi(\zeta, \eta). \tag{3.17}$$

It is easily calculated that

$$\frac{d}{dt} L_\Phi = L_{\overset{\bullet}{\Phi}} \text{ and } \frac{d}{dt} Q_\Phi = Q_{\overset{\bullet}{\Phi}} \tag{3.18}$$

In the following, an important role is played by certain one-variable polynomial matrices obtained from two-variable polynomial matrices by means of the delta operator ∂, defined as

$$\partial : \mathbb{R}^{q_1 \times q_2}[\zeta, \eta] \to \mathbb{R}^{q_1 \times q_2}[\xi]; \quad \partial\Phi(\xi) := \Phi(-\xi, \xi).$$

Note that, among other things, this allows one to associate a differential operator $\Phi(-\frac{d}{dt}, \frac{d}{dt})$ with a QDF—this is one of the key ingredients in LQ—and variational problems.

Introduce the star operator * acting on matrix polynomials by

$$^\star : \mathbb{R}^{q_1 \times q_2}[\xi] \to \mathbb{R}^{q_2 \times q_1}[\xi]; \quad R^\star(\xi) := R^T(-\xi).$$

The importance of this operation stems from the fact that $M(\frac{d}{dt})$ and $M^\star(\frac{d}{dt})$ are formal adjoints as differential operators. A polynomial matrix $M \in \mathbb{R}^{q \times q}[\xi]$ is called

para-Hermitian if $M = M^\star$. Note that $(\partial\Phi)^\star = \partial(\Phi^*)$. Hence if $\Phi \in \mathbb{R}_s^{q\times q}[\zeta,\eta]$, then $\partial\Phi$ is para-Hermitian.

In addition to studying BLDFs and QDFs as maps to $\mathfrak{C}^\infty(\mathbb{R},\mathbb{R})$, we are interested in their integrals. In order to make sure that those integrals exist, we assume in this case that the arguments have compact support. As is common, we denote by $\mathfrak{D}(\mathbb{R},\mathbb{R}^q) := \{w \in \mathfrak{C}^\infty(\mathbb{R},\mathbb{R}^q) \mid w \text{ has compact support}\}$. Let $\Phi \in \mathbb{R}^{q_1\times q_2}[\zeta,\eta]$. Then obviously $L_\Phi : \mathfrak{D}(\mathbb{R},\mathbb{R}^{q_1}) \times \mathfrak{D}(\mathbb{R},\mathbb{R}^{q_2}) \to \mathfrak{D}(\mathbb{R},\mathbb{R})$. Consider the integral

$$\int L_\Phi : \mathfrak{D}(\mathbb{R},\mathbb{R}^{q_1}) \times \mathfrak{D}(\mathbb{R},\mathbb{R}^{q_2}) \to \mathbb{R} \tag{3.19}$$

defined as

$$\int L_\Phi(v,w) := \int_{-\infty}^{+\infty} L_\Phi(v,w)dt. \tag{3.20}$$

The notation $\int Q_\Phi$ follows readily from this. Furthermore, consider the same integral over a finite interval $[t_1,t_2]$

$$\int_{t_1}^{t_2} L_\Phi(v,w)dt \tag{3.21}$$

denoted as $\int_{t_1}^{t_2} L_\Phi$. We call this integral *independent of path* if for any t_1 and t_2 the result of the integral (3.21) depends only on the values of v and w and (a finite number of) their derivatives at $t = t_1$ and $t = t_2$ but not on the intermediate path used to connect these endpoints, assuming, of course that $v \in \mathfrak{C}^\infty(\mathbb{R},\mathbb{R}^{q_1})$ and $w \in \mathfrak{C}^\infty(\mathbb{R},\mathbb{R}^{q_2})$.

The questions of when the map $\int L_\Phi$ is zero and when path independence holds are studied next.

THEOREM 3.1. *Let $\Phi \in \mathbb{R}^{q_1\times q_2}[\zeta,\eta]$. Then the following statements are equivalent:*

1. *$\int L_\Phi = 0$, equivalently $\int_{t_1}^{t_2} L_\Phi$ is independent of path.*
2. *There exists a $\Psi \in \mathbb{R}^{q_1\times q_2}[\zeta,\eta]$ such that $\Phi = \dot{\Psi}$, equivalently, such that $L_\Phi = \frac{d}{dt}L_\Psi$. Obviously Ψ is given by*

$$\Psi(\zeta,\eta) = \frac{\Phi(\zeta,\eta)}{\zeta+\eta}. \tag{3.22}$$

3. *$\partial\Phi = 0$, i.e., $\Phi(-\xi,\xi) = 0$.*

The same equivalence holds for QDFs. Simply assume $\Phi \in \mathbb{R}_s^{q\times q}[\zeta,\eta]$ and replace the L's by Q's in 1 *and* 2.

Proof. For the proof, see the appendix.

The importance of this theorem is that condition (3) gives a very convenient way of checking (1) or (2). Path integrals and path independence featured prominently in Brockett's work in the sixties (see [7], [8]), and indeed some of our results can be viewed as streamlined versions of this work. Another potentially interesting connection of the above theorem and our paper with the existing literature is [3] where, in our notation, $\Phi(\zeta,\eta) = R(\zeta)M(-\eta)$ is studied, with R and M associated with a kernel representation (2.1) and an image representation (2.4) of a controllable system. This Φ defines an intriguing path independent BLDF that can be associated with any controllable $\mathfrak{B}$.

In this paper we also study the behavior of QDFs evaluated along a differential behavior $\mathfrak{B} \in \mathfrak{L}^q$. In order to do so, it is convenient to introduce an equivalence relation on both the one- and two-variable polynomial matrices modulo a given $\mathfrak{B} \in \mathfrak{L}^q$.

Let $D_1, D_2 \in \mathbb{R}^{\bullet \times q}[\xi]$. Define $(D_1 \overset{\mathfrak{B}}{=} D_2) :\Leftrightarrow (D_1(\frac{d}{dt}) - D_2(\frac{d}{dt}))\mathfrak{B} = 0$. Let $\Phi_1, \Phi_2 \in \mathbb{R}_s^{q \times q}[\zeta, \eta]$. Define $(\Phi_1 \overset{\mathfrak{B}}{=} \Phi_2) :\Leftrightarrow (Q_{\Phi_1}(w) = Q_{\Phi_2}(w)$ for all $w \in \mathfrak{B})$. These equivalencies are easily expressed in terms of a kernel or an image representation of $\mathfrak{B}$.

PROPOSITION 3.2. *Let $R \in \mathbb{R}^{\bullet \times q}[\xi]$ define a kernel representation of $\mathfrak{B} \in \mathfrak{L}^q$. Then $D_1 \overset{\mathfrak{B}}{=} D_2$ iff*

$$D_1 - D_2 = FR \tag{3.23}$$

for some $F \in \mathbb{R}^{\bullet \times \bullet}[\xi]$ and $\Phi_1 \overset{\mathfrak{B}}{=} \Phi_2$ iff

$$\Phi_2(\zeta, \eta) = \Phi_1(\zeta, \eta) + R^T(\zeta)F(\zeta, \eta) + F^*(\zeta, \eta)R(\eta) \tag{3.24}$$

for some $F \in \mathbb{R}^{\bullet \times q}[\zeta, \eta]$. Let $M \in \mathbb{R}^{q \times \bullet}[\zeta, \eta]$ define an image representation of $\mathfrak{B} \in \mathfrak{L}^q$. Then $D_1 \overset{\mathfrak{B}}{=} D_2$ iff

$$D_1 M = D_2 M$$

and $\Phi_1 \overset{\mathfrak{B}}{=} \Phi_2$ iff

$$M^T(\zeta)\Phi_1(\zeta, \eta)M(\eta) = M^T(\zeta)\Phi_2(\zeta, \eta)M(\eta)$$

Proof. For the proof, see the appendix.

The first equivalence in the above proposition was already proven in [23], with an account of the history of the result, which goes back to 1895. We will return to the second equivalence at the end of section 4.

We now briefly discuss positivity of QDFs. This will be a major issue in the following; here we restrict our attention to the basic definitions.

DEFINITION 3.3. *Let $\Phi \in \mathbb{R}_s^{q \times q}[\zeta, \eta]$. We call the QDF Q_Φ* nonnegative, *denoted $\Phi \geq 0$, if $Q_\Phi(w) \geq 0$ for all $w \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^q)$, and* positive, *denoted $\Phi > 0$, if $\Phi \geq 0$ and if the only $w \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^q)$ for which $Q_\Phi(w) = 0$ is $w = 0$.*

Using the matrix representation of Φ, it is easy to see that $\Phi \geq 0$ iff there exists $D \in \mathbb{R}^{\bullet \times q}[\xi]$ such that $\Phi(\zeta, \eta) = D^T(\zeta)D(\eta)$. Simply factor $\tilde{\Phi}$ as $\tilde{\Phi} = \tilde{D}^T\tilde{D}$ and take $D(\xi) = \tilde{D}$ col $[I_q \;\; I_q\xi \;\; I_q\xi^2 \cdots]$. Moreover $\Phi > 0$ iff this D has the property that $D(\lambda)$ is of rank q for all $\lambda \in \mathbb{C}$; in other words, iff the image representation $w = D(\frac{d}{dt})\ell$ defined by D is observable. Note that, for $\Phi \in \mathbb{R}_s^{q \times q}[\zeta, \eta]$, we always have $|\Phi| \geq 0$.

We are also interested in QDFs which are zero or positive along a behavior $\mathfrak{B} \in \mathfrak{L}^q$.

DEFINITION 3.4. *We call* Φ zero along $\mathfrak{B}$, *denoted $\Phi \overset{\mathfrak{B}}{=} 0$, if $Q_\Phi(w) = 0$ for all $w \in \mathfrak{B}$. The notions of* nonnegative $(\overset{\mathfrak{B}}{\geq})$ *and* positive $(\overset{\mathfrak{B}}{>})$ along $\mathfrak{B}$ *follow readily.*

Note that it immediately follows from Proposition 3.2 that, if $R(\frac{d}{dt})w = 0$ is a kernel representation of $\mathfrak{B}$, then $\Phi \overset{\mathfrak{B}}{=} 0$ iff it can be written as $\Phi(\zeta, \eta) = F^*(\zeta, \eta)R(\eta) + R^T(\zeta)F(\zeta, \eta)$. A similar result holds for positivity as follows.

PROPOSITION 3.5. *Let $\Phi \in \mathbb{R}_s^{q \times q}[\zeta, \eta]$, $\mathfrak{B} \in \mathfrak{L}^q$, and $R \in \mathbb{R}^{\bullet \times q}[\xi]$ induce a kernel representation of $\mathfrak{B}$. Then*

(i) $\Phi \overset{\mathfrak{B}}{\geq} 0$ *iff there exists $\Phi' \in \mathbb{R}_s^{q \times q}[\zeta, \eta]$ with $\Phi \overset{\mathfrak{B}}{=} \Phi'$ and $\Phi' \geq 0$;*

(ii) $\Phi \overset{\mathfrak{B}}{>} 0$ *iff there exists* $\Phi' \in \mathbb{R}^{q\times q}_s[\zeta,\eta]$ *with* $\Phi \overset{\mathfrak{B}}{=} \Phi'$ *and* $\Phi'(\zeta,\eta) = D^T(\zeta)D(\eta)$, *with* (R, D) *an observable pair.*

Proof. For the proof, see the appendix.

4. Lyapunov theory. Lyapunov theory is a firmly established and very useful technique for establishing stability. It pertains to systems described by explicit first order differential equations. However, as argued in [33], models obtained from first principles are seldomly in first order form, will contain latent variables, and may contain high order derivatives. Writing them in explicit first order form without introducing spurious solutions may not be an easy matter. Moreover, stability considerations do not require systems to be in first order form. In fact, historically the very first stability questions and results, as Maxwell's statement of the stability problem and the Routh–Hurwitz conditions, pertain to high order differential equations. Oddly enough, to our knowledge, no attempts seem to have been made to establish Lyapunov theory for high order differential equations. This is the purpose of this section. We limit our attention, however, to linear differential systems and to Lyapunov functions that are quadratic differential forms, but we recognize the urgency of generalizing this work to nonlinear systems. We should remark that in this section for stability, we only consider systems in which the latent variables have been eliminated, although latent variables do not cause essential difficulties in the context of stability.

First, we introduce the notion of stability. We say that a system $\mathfrak{B} \in \mathfrak{L}^q$ is *asymptotically stable* if $(w \in \mathfrak{B}) \Rightarrow (w(t) \underset{t\to\infty}{\longrightarrow} 0)$ and *stable* if $(w \in \mathfrak{B}) \Rightarrow (w$ is bounded on the half-line $[0,\infty))$. For a system $\mathfrak{B} \in \mathfrak{L}^q$ to be (asymptotically) stable it has to be autonomous. A system $\mathfrak{B} \in \mathfrak{L}^q$ is said to be *autonomous* if $(w_1, w_2 \in \mathfrak{B})$ and $(w_1(t) = w_2(t)$ for $t < 0)$ imply $(w_1 = w_2)$. It is easy to see that the system with kernel representation $R(\frac{d}{dt})w = 0$ is autonomous iff $\operatorname{rank}(R) = q$; in particular, if R is square and $\det(R) \neq 0$.

DEFINITION 4.1. *Let* $R \in \mathbb{R}^{\bullet\times q}[\xi]$. *The complex number* $\lambda \in \mathbb{C}$ *is said to be a* singularity *of* R *if* $\operatorname{rank}(R(\lambda)) < \operatorname{rank}(R)$; R *is said to be* Hurwitz *if* $\operatorname{rank}(R) = q$ *and if* R *has all its singularities in the open left half of the complex plane.*

Thus a square $R \in \mathbb{R}^{q\times q}[\xi]$ is Hurwitz iff $\det(R)$ is a *Hurwitz polynomial*, i.e., a nonzero polynomial with its roots in the open left half-plane. We record the following classical result for easy reference.

PROPOSITION 4.2. *The system with kernel representation* (2.1) *is asymptotically stable iff* R *is Hurwitz.*

Our most basic Lyapunov theorem regarding high order systems is the following.

THEOREM 4.3. *Let* $\mathfrak{B} \in \mathfrak{L}^q$. *Then* $\mathfrak{B}$ *is asymptotically stable iff there exists* $\Psi \in \mathbb{R}^{q\times q}_s[\zeta,\eta]$ *such that* $\Psi \overset{\mathfrak{B}}{\geq} 0$ *and* $\overset{\bullet}{\Psi} \overset{\mathfrak{B}}{<} 0$.

Proof. For the proof, see the appendix.

Example 4.4. Consider the scalar system described by $w + \frac{dw}{dt} + \frac{d^2w}{dt^2} = 0$. Consider the QDF $w^2 + (\frac{dw}{dt})^2$. Its derivative QDF is $2(w + \frac{d^2w}{dt^2})\frac{dw}{dt}$. Since $w \in \mathfrak{B}$ iff $w + \frac{d^2w}{dt^2} = -\frac{dw}{dt}$, we see that this QDF is $\mathfrak{B}$-equivalent to the QDF $-2(\frac{dw}{dt})^2$. Finally observe that $(1 + \xi + \xi^2, \sqrt{2}\xi)$ is an observable pair. Hence $-2(\frac{dw}{dt})^2$ is negative on $\mathfrak{B}$. Theorem 4.3 establishes asymptotic stability (a rather trivial matter for the case at hand). Our aim was to show the use of Lyapunov theory without getting involved with state representations (admittedly also a trivial matter).

Example 4.5. Consider the multivariable system

$$Kw + D\frac{dw}{dt} + M\frac{d^2w}{dt^2} = 0, \tag{4.1}$$

with $K, D, M \in \mathbb{R}^{q\times q}$, $K = K^T \geq 0$, $D + D^T \geq 0$, and $M = M^T \geq 0$. Such second order equations occur frequently as models of (visco-)elastic mechanical systems. Take $\Psi(\zeta,\eta) = K + M\zeta\eta$. Then $\overset{\bullet}{\Psi}(\zeta,\eta) = K(\zeta+\eta) + M(\zeta^2\eta + \zeta\eta^2)$ which is obviously $\mathfrak{B}$-equivalent to $-(D+D^T)\zeta\eta$. Thus, asymptotic stability follows if

$$\left(K + D\xi + M\xi^2, \sqrt{(D+D^T)}\xi\right) \tag{4.2}$$

is an observable pair. This is the case, for example, if $\{0\} = \ker(K) \subset \ker(D+D^T) \subset \ker(M)$. Indeed, under this condition

$$\begin{bmatrix} K + D\lambda + M\lambda^2 \\ \sqrt{(D+D^T)}\lambda \end{bmatrix}$$

has full column rank for all $\lambda \in \mathbb{C}$.

State representations of autonomous systems take a very special form. Indeed, it is easy to see that $\mathfrak{B} \in \mathfrak{L}^q$ is autonomous iff it admits a state representation of the form $\frac{dx}{dt} = Ax$, $w = Cx$. Such state representations are automatically state trim. If (A, C) is observable, then they are state minimal. It also follows that for *every* $D \in \mathbb{R}^{\bullet\times q}[\xi]$ there exists a matrix $H \in \mathbb{R}^{\bullet\times n}$ such that $D(\frac{d}{dt})w \overset{\mathfrak{B}}{=} Hx$, i.e., every linear differential operator acting on an autonomous $\mathfrak{B} \in \mathfrak{L}^q$ is $\mathfrak{B}$-equivalent to an instantaneous function of the state. An analogous statement holds, of course, for QDFs.

Viewed from this perspective, one can regard Theorem 4.3 as being about state systems and in this sense not very different from classical Lyapunov theorems. The point of Theorem 4.3 is twofold:

1. It avoids the state construction which algorithmically (and conceptually) is not always easy in the multivariable case; and
2. It has the usual Lyapunov theory as a special case by applying it to systems in first order form and using memoryless QDFs. For the sake of completeness, we record this as a corollary.

COROLLARY 4.6. *Let $\mathfrak{B}$ be the behavior of $\frac{dw}{dt} = Aw$. Let $\Psi(\zeta,\eta) = \Psi_0$ with $\Psi_0 \in \mathbb{R}^{q\times q}$, $\Psi_0 = \Psi_0^T \geq 0$. Then $\overset{\bullet}{\Psi}(\zeta,\eta) \overset{\mathfrak{B}}{=} A\Psi_0 + \Psi_0 A^T =: \Delta_0$. Whence, if $\Delta_0 = \Delta_0^T \leq 0$ and if (A, Δ_0) is an observable pair of matrices, $\mathfrak{B}$ is asymptotically stable.*

Proof. For the proof, see the appendix.

In section 3, we discussed $\mathfrak{B}$-positive QDFs. When $\mathfrak{B}$ is autonomous, it is useful to consider also a stronger concept. Let $\mathfrak{B} \in \mathfrak{L}^q$ and $\Phi \in \mathbb{R}_s^{q\times q}[\zeta,\eta]$; we call Φ *strongly $\mathfrak{B}$-positive* (denoted $\Phi \overset{\mathfrak{B}}{\gg} 0$) if $\Phi \overset{\mathfrak{B}}{\geq} 0$ and if ($w \in \mathfrak{B}$ and $Q_\Phi(w)(0) = 0$) imply ($w = 0$). It is easy to see that $\Phi \overset{\mathfrak{B}}{\gg} 0$ implies $\Phi \overset{\mathfrak{B}}{>} 0$ and that in order for $\Phi \overset{\mathfrak{B}}{\gg} 0$, $\mathfrak{B}$ must be autonomous. In fact, $((Q_\phi(w)(0) = 0) \Rightarrow (w = 0))$ by itself already implies ($\Phi \overset{\mathfrak{B}}{\gg} 0$ or $\Phi \overset{\mathfrak{B}}{\ll} 0$). Using this notion we arrive at the following refinement of theorem 4.3.

PROPOSITION 4.7. *If $\Psi \overset{\mathfrak{B}}{\geq} 0$ and $\overset{\bullet}{\Psi} \overset{\mathfrak{B}}{<} 0$, then $\mathfrak{B}$ is asymptotically stable and $\Psi \overset{\mathfrak{B}}{\gg} 0$.*

Proof. For the proof, see the appendix.

We now formulate a stronger version of the "only if" part of Theorem 4.3.

THEOREM 4.8. *Assume that $\mathfrak{B} \in \mathfrak{L}^q$ is asymptotically stable. Then for any $\Phi \in \mathbb{R}^{q\times q}_s[\zeta,\eta]$ there exists a $\Psi \in \mathbb{R}^{q\times q}_s[\zeta,\eta]$ such that $\overset{\bullet}{\Psi} \overset{\mathfrak{B}}{=} \Phi$; Ψ is unique up to $\mathfrak{B}$-equivalence in the sense that, if $\overset{\bullet}{\Psi}_1 \overset{\mathfrak{B}}{=} \Phi$ and $\overset{\bullet}{\Psi}_2 \overset{\mathfrak{B}}{=} \Phi$, then $\Psi_1 \overset{\mathfrak{B}}{=} \Psi_2$. If $\Phi \overset{\mathfrak{B}}{\leq} 0$, then $\Psi \overset{\mathfrak{B}}{\geq} 0$, and if $\Phi \overset{\mathfrak{B}}{<} 0$, then $\Psi \overset{\mathfrak{B}}{\gg} 0$.*

In order to compute Ψ from Φ, the following algorithm may be used. Let $R \in \mathbb{R}^{\bullet\times q}[\xi]$ induce a kernel representation of $\mathfrak{B}$. Consider the polynomial matrix equation

$$X^T(-\xi)R(\xi) + R^T(-\xi)X(\xi) = \Phi(-\xi,\xi) \tag{4.3}$$

in the unknown $X \in \mathbb{R}^{\bullet\times q}[\xi]$. Then (4.3) has a solution. Let X_0 be a solution. If R is square, then all its solutions can be obtained from this one as

$$X(\xi) = X_0(\xi) + F(\xi)R(\xi), \tag{4.4}$$

where F ranges over all polynomial matrices of appropriate size satisfying

$$F^T(-\xi) = -F(\xi). \tag{4.5}$$

Consider any $Y \in \mathbb{R}^{\bullet\times q}[\zeta,\eta]$ such that

$$Y(-\xi,\xi) = X(\xi) \tag{4.6}$$

and compute

$$\Psi(\zeta,\eta) = \frac{\Phi(\zeta,\eta) - Y^*(\zeta,\eta)R(\eta) - R^T(\zeta)Y(\zeta,\eta)}{\zeta+\eta}. \tag{4.7}$$

Then $\overset{\bullet}{\Psi} \overset{\mathfrak{B}}{=} \Phi$. Since any two Ψ_1, Ψ_2 such that $\overset{\bullet}{\Psi}_1 \overset{\mathfrak{B}}{=} \Phi$ and $\overset{\bullet}{\Psi}_2 \overset{\mathfrak{B}}{=} \Phi$ satisfy $\Psi_1 \overset{\mathfrak{B}}{=} \Psi_2$, any other solutions of (4.3) and/or (4.6) yield Ψ's in (4.7) that are $\mathfrak{B}$-equivalent.

Proof. For the proof, see the appendix.

Theorem 4.8 is more than a mouthful and so we illustrate it for ordinary state space systems $\frac{dw}{dt} = Aw$. Let $\Phi = \Phi^T \in \mathbb{R}^{n\times n}$. Then (4.3) becomes

$$X^T(-\xi)(A - I\xi) + (A^T + I\xi)X(\xi) = \Phi. \tag{4.8}$$

This equation has a constant solution which must be symmetric, $X_0 = X_0^T$, the solution of the ordinary Lyapunov equation

$$X_0 A + A_0^T X = \Phi. \tag{4.9}$$

Choose $Y = X_0$ and verify that (4.7) reduces to $\Psi = X_0$, whence the Lyapunov function is $Q_\Psi(w) = w^T X_0 w$ and for $w \in \mathfrak{B}$ its derivative is $Q_\Phi(w) = w^T \Phi w$. Because of this analogy, we refer to (4.3) as the polynomial matrix *Lyapunov equation*.

The above shows that it seems to suffice to consider Lyapunov functions Ψ and their derivatives Φ that are of lower degree than that of R. That is, in fact, a general feature of the equations in Theorem 4.8. However, in order to formalize this, we return first to the notion of $\mathfrak{B}$-equivalence of differential operators in the case that $\mathfrak{B} \in \mathfrak{L}^q$ is autonomous.

Let $\mathfrak{B} \in \mathfrak{L}^q$ be autonomous. Then there always exists a square kernel representation for it. Let $R \in \mathbb{R}^{q\times q}[\xi]$ be such that $\mathfrak{B} = \ker(R(\frac{d}{dt}))$. We assume in the remainder of this section that R is square.

Let $D \in \mathbb{R}^{\bullet\times q}[\xi]$. We call D *R-canonical* if DR^{-1} is a matrix of strictly proper rational functions. Let $\Phi \in \mathbb{R}_s^{q\times q}[\zeta,\eta]$. We call Φ *R-canonical* if $(R^T(\zeta))^{-1}\Phi(\zeta,\eta)$ $(R(\eta))^{-1}$ is a matrix of strictly proper two-variable rational functions. (Note that there is no ambiguity about what "strictly proper" means for these two-variable rational functions.) Since for autonomous systems all differential operators can be seen as instantaneous functions of the state, it is clear that for any D there exists a canonical D' that is R-equivalent to D'. The aim of the next result is to derive this also for QDFs.

PROPOSITION 4.9. *Let $D \in \mathbb{R}^{\bullet\times q}[\xi]$. Among all differential operators $\mathfrak{B}$-equivalent to D, there is exactly one, D', which is R-canonical. This D' can be computed as follows. Compute $DR^{-1} \in \mathbb{R}^{\bullet\times q}(\xi)$ and write it as $DR^{-1} = P + S$, with P the polynomial part and S the strictly proper rational part of DR^{-1} . Then $D' = D - PR$. Let $\Phi \in \mathbb{R}_s^{q\times q}[\zeta,\eta]$. Among all QDFs $\mathfrak{B}$-equivalent to Φ there is exactly one, Φ', which is R-canonical. This Φ' can be computed as follows. Write Φ as $\Phi(\zeta,\eta) = M^T(\zeta)N(\eta)$. Compute the R-canonical representatives M' of M and N' of N. Then $\Phi'(\zeta,\eta) = M'^T(\zeta)N'(\eta)$.*

Proof. For the proof, see the appendix.

The following proposition shows that $\mathfrak{B}$-positivity reduces to positivity of the $\mathfrak{B}$-canonical representative.

PROPOSITION 4.10. *If Ψ is R-canonical, then we have*

(i) $(\Psi \overset{\mathfrak{B}}{=} 0) \Leftrightarrow (\Psi = 0)$,

(ii) $(\Psi \overset{\mathfrak{B}}{\geq} 0) \Leftrightarrow (\Psi \geq 0) \Leftrightarrow (\Psi(\zeta,\eta) = D^T(\zeta)D(\eta)$ *with D R-canonical*$)$,

(iii) $(\Psi \overset{\mathfrak{B}}{>} 0) \Leftrightarrow (\Psi > 0$ *and* $\Psi(\zeta,\eta) = D^T(\zeta)D(\eta)$ *with* (R,D) *observable*$) \Leftrightarrow$ $(\Psi(\zeta,\eta) = D^T(\zeta)D(\eta)$ *with* (R,D) *observable and D R-canonical*$)$.

Proof. For the proof, see the appendix.

We immediately obtain the following consequence of Theorem 4.3.

COROLLARY 4.11. *$\mathfrak{B} \in \mathfrak{L}^q$ is asymptotically stable iff there exists a $\Psi \in \mathbb{R}_s^{q\times q}[\zeta,\eta]$, $\Psi \geq 0$, such that the R-canonical representative of $(\zeta+\eta)\Psi(\zeta,\eta)$, computed as in Proposition* 4.9, *is ≤ 0 and factors as $-D^T(\zeta)D(\eta)$ with (R,D) observable.*

Our next result is perhaps the most useful of all. It shows how to walk through the algorithm of Theorem 4.8 and preserve canonicity.

THEOREM 4.12. *Assume that $\mathfrak{B} \in \mathfrak{L}^q$ is asymptotically stable and has kernel representation* (2.1) *with R square. Assume that Φ is R-canonical. Then the polynomial matrix Lyapunov equation* (4.3) *has a unique R-canonical solution. Denote it by X'. Then*

$$\Psi(\zeta,\eta) = \frac{\Phi(\zeta,\eta) - X'^T(\zeta)R(\eta) - R^T(\zeta)X'(\eta)}{\zeta+\eta} \tag{4.10}$$

is the unique R-canonical Ψ such that $\overset{\bullet}{\Psi} \overset{\mathfrak{B}}{=} \Phi$. Hence if $\Phi \leq 0$, then $\Psi \geq 0$, and if, in addition, $\Phi(\zeta,\eta) = -D^T(\zeta)D(\eta)$ with (R,D) observable, then $\Psi \overset{\mathfrak{B}}{\gg} 0$.

Proof. For the proof, see the appendix.

We make a short comment relating these results to state representations. The state maps (2.13) associating a minimal state to $\mathfrak{B}$ are uniquely defined up to $\mathfrak{B}$-equivalence. There is, consequently, a minimal state map (unique up to premultiplication by a nonsingular matrix that is R-canonical, say, $x = X(\frac{d}{dt})w$). An R-canonical

$\Phi \in \mathbb{R}^{q\times q}_s[\zeta,\eta]$ is of the form $Q_\Phi(w) = x^T\Gamma x$, i.e., $\Phi(\zeta,\eta) = X^T(\zeta)\Gamma X(\eta)$, with $\Gamma = \Gamma^T$ an $(n\times n)$ matrix. Of course for this Φ there holds $(\Phi \overset{\mathfrak{B}}{\geq} 0) \Leftrightarrow (\Phi \geq 0) \Leftrightarrow (\Gamma \geq 0)$; furthermore, $(\Phi \overset{\mathfrak{B}}{>} 0) \Leftrightarrow (\Gamma \geq 0$ and observability of the pair of matrices (A,Γ) (with A associated with the state x)), and finally $(\Phi \overset{\mathfrak{B}}{\gg} 0) \Leftrightarrow (\Gamma > 0)$.

The above results allow generalizations to unstable systems. Let us briefly mention a few. We have seen that ($\mathfrak{B}$ asymptotically stable) $\Leftrightarrow$ ($\exists\Psi(\zeta,\eta)$ such that ($\Psi \overset{\mathfrak{B}}{\geq} 0$ and $\overset{\bullet}{\Psi} \overset{\mathfrak{B}}{<} 0$)). There also holds ($\mathfrak{B}$ stable) $\Leftrightarrow$ ($\exists\Psi(\zeta,\eta)$ such that ($\Psi \overset{\mathfrak{B}}{>} 0$ and $\overset{\bullet}{\Psi} \overset{\mathfrak{B}}{\leq} 0$)) and (an autonomous $\mathfrak{B}$ is not stable) $\Leftrightarrow$ ($\exists\Psi(\zeta,\eta)$ such that ($\Psi \overset{\mathfrak{B}}{\not\geq} 0$ and $\overset{\bullet}{\Psi} \overset{\mathfrak{B}}{<} 0$)). Furthermore, the result that a Lyapunov function Ψ can be constructed so that it has a given derivative Φ (Theorem 4.8) can be generalized to autonomous systems, as long as they have the property that if λ is a singularity of R then $-\lambda$ will not be a singularity of R. As such this theorem extends in this sense to a large class of unstable systems.

We close this section with two extensive examples.

Example 4.13. In this first example we use Theorem 4.8 in order to give a Lyapunov proof of the Routh–Hurwitz test for stability of scalar systems. Let $R \in \mathbb{R}[\xi]$ be a Hurwitz polynomial. Hence $R(\frac{d}{dt})w = 0$ defines an asymptotically stable scalar system. Take, for the derivative of the Lyapunov function,

$$\Phi(\zeta,\eta) = -\frac{1}{2}R(-\zeta)R(-\eta) = -\frac{1}{2}R^\star(\zeta)R^\star(\eta). \tag{4.11}$$

Then obviously, since R has no imaginary axis roots, $(R,R^\star)$ is an observable (i.e., a coprime) pair. The polynomial matrix Lyapunov equation (4.3) yields $X(\xi) = -\frac{1}{4}R(\xi)$ as a solution. Take $Y(\zeta,\eta) = X(\eta)$. Then (4.7) yields

$$B(\zeta,\eta) = \frac{1}{2}\frac{R(\zeta)R(\eta) - R(-\zeta)R(-\eta)}{\zeta+\eta} \tag{4.12}$$

as a Lyapunov function. Note that this Lyapunov function can be written directly from the system parameters, without having to solve linear equations! This fact is actually well known, even though it is not presented in the vein of providing a higher order Lyapunov function ([6], [13], [14]).

The two-variable polynomial B defined by (4.12) is called the *Bezoutian* of R. Note that $\overset{\bullet}{B}(\zeta,\eta) \overset{\mathfrak{B}}{=} -\frac{1}{2}R(-\zeta)R(-\eta)$; B is R-canonical, but $\overset{\bullet}{B}$ is not. However, $\overset{\bullet}{B}$ is $\mathfrak{B}$-equivalent to $-\frac{1}{2}R(-\zeta)R(-\eta) + \frac{1}{2}R(\zeta)R(\eta)$, which is. If we take this for the Φ in Theorem 4.8, then the Lyapunov equation yields $X = 0$. Taking $Y = 0$ then also yields the Bezoutian (4.12) as the corresponding (hence R-canonical) Lyapunov function B.

A close examination of the arguments involved yields the equivalence of the following three conditions on a polynomial R of degree n and the corresponding $B \in \mathbb{R}[\zeta,\eta]$ given by (4.12):

1. R is Hurwitz,
2. $B \geq 0$ and $(R,R^\star)$ is coprime,
3. $\tilde{B}$ (the constant matrix associated with B) has rank n and is ≥ 0.

The Lyapunov function (4.12), the Bezoutian, is a very useful one for deriving various stability tests. It is a classical concept in stability (see [11] for a recent reference). Let us illustrate its usefulness by deriving the Routh stability test from it.

Let $R \in \mathbb{R}[\xi]$ be a polynomial of degree n. Decompose R in its even and odd parts as

$$R(\xi) = E_0(\xi^2) + \xi E_1(\xi^2).$$

Form the *Routh table* by computing the polynomials $E_2, E_3, \ldots, E_n$ as

$$E_k(\xi) = \xi^{-1}(E_{k-1}(0)E_{k-2}(\xi) - E_{k-2}(0)E_{k-1}(\xi)).$$

Assume for simplicity that $R(0) = E_0(0) \geq 0$. Routh's stability criterion states that R is Hurwitz iff all elements of the *Routh array* $E_0(0), E_1(0), \ldots, E_n(0)$ are positive. Define $R_k(\xi) = E_{k-1}(\xi^2) + \xi E_k(\xi^2)$ for $k = 1, \ldots, n$, and let B_k be the Bezoutian associated with R_k. Examining expression (4.12) yields, after a simple calculation

$$E_k(0)B_k(\zeta, \eta) = \zeta\eta B_{k+1}(\zeta, \eta) + E_{k-1}(0)E_k(\zeta^2)E_k(\eta^2) \tag{4.13}$$

for $k = 1, \ldots, n$ (define $B_{n+1} = 0$). Assume that $E_0(0), E_1(0), \ldots, E_n(0)$ are all positive. Then we obtain (note that $B = B_1$)

$$B(\zeta, \eta) = \sum_{k=1}^{n} \alpha_k \zeta^{k-1}\eta^{k-1}E_k(\zeta^2)E_k(\eta^2),$$

where $\alpha_k = E_{k-1}(0)/E_1(0)E_2(0)\ldots E_k(0)$. Obviously $\tilde{B} \geq 0$ and has rank n. Therefore R is Hurwitz. To show the converse, assume that R is Hurwitz. Then $E_0(0) > 0$. Also, $\tilde{B} \geq 0$ and has rank n. Therefore, by (4.13), $\tilde{B}_2 \geq 0$ and has rank $n-1$. Hence R_2 is Hurwitz, and $E_1(0) > 0$. Now proceed by induction.

The key point thus is that

$$\sum_{k=1}^{n} \alpha_k \left(E_k\left(\frac{d^2}{dt^2}\right) \frac{d^{k-1}}{dt^{k-1}} w \right)^2$$

is a QDF which is well defined and nonnegative definite when the Routh conditions are satisfied. It has derivative

$$\frac{1}{2}\left(\left(R\left(-\frac{d}{dt}\right) w \right)^2 - \left(R\left(\frac{d}{dt}\right) w \right)^2 \right),$$

which is obviously nonnegative definite along solutions of $R(\frac{d}{dt})w = 0$.

Example 4.14. Let $E_1, E_2, \ldots, E_N$ and $O_1, O_2, \ldots, O_{N'}$ be two sets of real polynomials, and assume that $R_{k,\ell}(\xi) := E_k(\xi^2) + \xi O_\ell(\xi^2)$ is Hurwitz for all $k = 1, 2, \ldots, N$ and $\ell = 1, 2, \ldots, N'$. Then any combination

$$R(\xi) = \sum_{k=1}^{N} \alpha_k E_k(\xi^2) + \sum_{\ell=1}^{N'} \beta_\ell \xi O_\ell(\xi^2)$$

is also Hurwitz whenever all the α_k's and β_ℓ's are positive. In order to see this, simply observe that, in the obvious notation, (4.12) yields

$$B(\zeta, \eta) = \sum_{k=1}^{N} \sum_{\ell=1}^{N'} \alpha_k \beta_\ell B_{k,\ell}(\zeta, \eta)$$

and the conclusion follows.

This may be applied to interval polynomials. Assume that $R \in R[\xi]$ is given by $R(\xi) = R_0 + R_1\xi + \cdots + R_n\xi^n$, with $R_k \in [a_k, A_k]$. The question arises under what conditions all these polynomials are Hurwitz. The *weak Kharitonov* test states that this is the case iff the 2^n extreme polynomials, that is, those obtained by replacing each R_k by a_k or A_k, are all Hurwitz. This result is an immediate consequence of the above. With a little bit of extra work, we can also obtain the *strong Kharitonov* test [15] which states that the interval polynomials are Hurwitz iff the four *Kharitonov* polynomials obtained by taking the initial sequences $a_0, a_1, A_2, \ldots$, or $a_0, A_1, A_2, \ldots$, or $A_0, A_1, a_2, \ldots$, or $A_0, a_1, a_2, \ldots$, and continuing by alternating between two consecutive maxima and minima, are all Hurwitz. Indeed (see [20]), observe that for all $\omega, R(i\omega)$ lies in the rectangle in the complex plane spanned by the four points obtained by taking for R the Kharitonov polynomials. This rectangle does not contain the origin since, by the above, the convex hull of the Kharitonov polynomials contains only Hurwitz polynomials if the Kharitonov polynomials are themselves Hurwitz.

5. Average positivity. Up to now, we have considered positivity of QDFs and its use in establishing stability through Lyapunov functions. However, in many applications, especially in control theory, we are interested in an average type of positivity. In section 3, we already discussed when $\int Q_\Phi$ is zero. We now study when it is positive. With an eye towards applications in LQ and H_∞ control we have to distinguish several (unfortunately not less than three) types of average positivity. All of them have quite logical definitions.

DEFINITION 5.1. *Let $\Phi \in \mathbb{R}^{q\times q}_s[\zeta, \eta]$. The QDF Q_Φ (or simply Φ) is said to be*

1. average nonnegative*, denoted $\int Q_\Phi \geq 0$, if $\int_{-\infty}^{+\infty} Q_\Phi(w)dt \geq 0$ for all $w \in \mathfrak{D}(\mathbb{R}, \mathbb{R}^q)$,*
2. average positive*, denoted by $\int Q_\Phi > 0$, if $\int Q_\Phi \geq 0$ and if $\int_{-\infty}^{+\infty} Q_\Phi(w)dt = 0$ implies $w = 0$,*
3. strongly average positive*, denoted $\int Q_\Phi \overset{\text{per}}{>} 0$, if for all nonzero periodic $w \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^q)$ there holds $\frac{1}{T}\int_0^T Q_\Phi(w)dt > 0$, where T denotes the period of w.*

Note that (3) looks somewhat different from the other definitions in this paper since, for the first time, periodic functions are involved. Actually, wherever in the paper definitions refer to compact support functions, they could have been written just as well in terms of periodic functions. However, strong average positivity is the only instance where the converse is not true.

PROPOSITION 5.2. *Let $\Phi \in \mathbb{R}^{q\times q}_s[\zeta, \eta]$. Then*

(i) $(\int Q_\Phi \geq 0) \Longleftrightarrow (\partial\Phi(i\omega) \geq 0 \ \ \forall\omega \in \mathbb{R})$.

(ii) $(\int Q_\Phi > 0) \Longleftrightarrow (\partial\Phi(i\omega) \geq 0 \ \ \forall\omega \in \mathbb{R}$ *and* $\det(\partial\Phi) \neq 0)$.

(iii) $(\int Q_\Phi \overset{\text{per}}{>} 0) \Longleftrightarrow (\partial\Phi(i\omega) > 0 \ \ \forall\omega \in \mathbb{R})$.

Proof. For the proof, see the appendix.

Concerning the equivalence (ii), note that $\partial\Phi(i\omega) \geq 0 \ \forall\omega \in \mathbb{R}$ and $\det(\partial\Phi) \neq 0$ is equivalent to: $\partial\Phi(i\omega) > 0$ for all but finitely many $\omega \in \mathbb{R}$.

Intuitively, we think of $Q_\Phi(w)$ as the power going into a physical system. In many applications, the power is indeed a quadratic differential form of some system variables. For example, in mechanical systems, it is $\sum_k F_k \frac{dq_k}{dt}$ with F_k the external force acting on the system, and q_k the position of the kth pointmass; in electrical circuits it is $\sum_k V_k I_k$, with V_k the potential and I_k the current going into the circuit at the kth terminal. Note that in these examples the variables are themselves also related. When this relation is expressed as an image representation, then we obtain a general QDF in terms of latent variables for the power delivered to a system.

Average nonnegativity states that the net flow of energy going into the system is nonnegative: the system dissipates energy. Of course, sometimes energy flows into the system, while at other times it flows out of it. This outflow is due to the fact that energy is stored. However, because of dissipation, the rate of increase of storage cannot exceed the supply. This interaction between supply, storage, and dissipation is now formalized.

DEFINITION 5.3. *Let $\Phi \in \mathbb{R}^{q\times q}_s[\zeta,\eta]$ induce the QDF Q_Φ. The QDF Q_Ψ induced by $\Psi \in \mathbb{R}^{q\times q}_s[\zeta,\eta]$ is said to be a* storage function *for Φ if*

$$\frac{d}{dt}Q_\Psi \leq Q_\Phi. \tag{5.1}$$

A QDF Q_Δ induced by $\Delta \in \mathbb{R}^{q\times q}_s[\zeta,\eta]$ is said to be a dissipation function *for Φ if*

$$\Delta \geq 0 \text{ and } \int Q_\Phi = \int Q_\Delta. \tag{5.2}$$

The next proposition shows that one can always interpret average positivity by an instantaneous positivity condition involving the difference between the rate of change of storage function and the supply rate.

PROPOSITION 5.4. *The following conditions are equivalent:*

1. $\int Q_\Phi \geq 0$,
2. Φ *admits a storage function,*
3. Φ *admits a dissipation function.*

Moreover, there is a one-one relation between storage and dissipation functions, Ψ and Δ, respectively, defined by

$$\frac{d}{dt}Q_\Psi(w) = Q_\Phi(w) - Q_\Delta(w)$$

equivalently, $\overset{\bullet}{\Psi} = \Phi - \Delta$, i.e.,

$$\Psi(\zeta,\eta) = \frac{\Phi(\zeta,\eta) - \Delta(\zeta,\eta)}{\zeta+\eta}. \tag{5.3}$$

Proof. For the proof, see the appendix.

Of course, we should expect that a storage function is related to memory, to state. The question, however, is: *the state of which system*? After all, we are considering a QDF, not a dynamical system. However, the factorization of Φ as

$$\Phi(\zeta,\eta) = M^T(\zeta)\Sigma_M M(\eta) \tag{5.4}$$

discussed earlier in section 3 allows us to introduce a state for the QDF Q_Φ. Indeed, (5.4) induces the dynamical system in image representation

$$v = M\left(\frac{d}{dt}\right)w. \tag{5.5}$$

Note that in (5.5) we are considering w as the latent variable and v as the manifest one. This is in keeping with the idea that $v^T\Sigma_M v$, the supply rate, is the variable of interest and that w is a latent variable that explains it. We are hence considering the behavior of the possible trajectories v. Assume that M has r rows, i.e., that

$M \in \mathbb{R}^{r\times q}[\xi]$. Thus, (5.5) defines a system $\mathfrak{B} \in \mathfrak{L}^r$ with $\mathfrak{B} = \mathrm{im}(M(\frac{d}{dt}))$ and $M(\frac{d}{dt})$ viewed as a map from $\mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^q)$ to $\mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^r)$. Hence this system has a state representation. Assume that

$$x = X\left(\frac{d}{dt}\right) w \tag{5.6}$$

induces such a state representation. Thus $X \in \mathbb{R}^{\bullet\times q}[\xi]$ is a polynomial matrix defining a state map for $\mathfrak{B} = \mathrm{im}(M(\frac{d}{dt}))$. Let $\Psi \in \mathbb{R}^{q\times q}_s[\zeta, \eta]$. Then the QDF Q_Ψ is said to be a *state function* (relative to the state of Φ) if there exists a real (symmetric) matrix P such that

$$Q_\Psi(w) = \|X\left(\frac{d}{dt}\right) w\|^2_P. \tag{5.7}$$

It is said to be a *state/supply function* if there exists a real (symmetric) matrix E such that

$$Q_\Psi(w) = \| \begin{bmatrix} M(\frac{d}{dt}) \\ X(\frac{d}{dt}) \end{bmatrix} w\|^2_E \tag{5.8}$$

where, as always, $\|a\|^2_A$ denotes $a^T Aa$. Note that the factorization (5.4) is not unique. However, any such factorization is related in a simple way to a canonical one, say, to

$$\Phi(\zeta, \eta) = \hat{M}^T(\zeta)\Sigma_\Phi \hat{M}(\eta) \tag{5.9}$$

by the existence of a matrix $F \in \mathbb{R}^{\bullet\times\bullet}$ such that $\hat{M}(\xi) = FM(\xi)$. This relation has as a consequence that any (possibly nonminimal) state map X for the system in image representation (5.5) is related in a static way to a minimal state map $\hat{X}$ associated with the system in image representation

$$\hat{v} = \hat{M}\left(\frac{d}{dt}\right) w \tag{5.10}$$

based on a canonical factorization. Indeed, there exists a matrix $L \in \mathbb{R}^{\bullet\times\bullet}$ such that

$$\hat{X}(\xi) = LX(\xi). \tag{5.11}$$

Thus, considering arbitrary (i.e., not necessarily canonical) factorizations and arbitrary (i.e., not necessarily minimal) state representations yields *a* (rather than *the*) state of Q_Φ. Thus, the situation with the state is similar to the situation with the state of a system $\mathfrak{B} \in \mathfrak{L}^q$.

We have the following important result.

THEOREM 5.5. *Let $\int Q_\Phi \geq 0$, and let $\Psi \in \mathbb{R}^{q\times q}_s[\zeta, \eta]$ be a storage function for Φ, i.e., $\overset{\bullet}{\Psi} \leq \Phi$. Then Ψ is a state function. Let $\Delta \in \mathbb{R}^{q\times q}_s[\zeta, \eta]$ be a dissipation function for Φ. Then Δ is a state/supply function. In fact, if $X \in \mathbb{R}^{\bullet\times q}[\xi]$ is a state map for Φ, then there exist real symmetric matrices P and E such that*

$$\Psi(\zeta, \eta) = X^T(\zeta)PX(\eta), \tag{5.12}$$

$$\Delta(\zeta, \eta) = \begin{bmatrix} M(\zeta) \\ X(\zeta) \end{bmatrix}^T E \begin{bmatrix} M(\eta) \\ X(\eta) \end{bmatrix}. \tag{5.13}$$

Equivalently

$$L_\Psi(w_1, w_2) = \left(X\left(\frac{d}{dt}\right) w_1\right)^T PX\left(\frac{d}{dt}\right) w_2,$$

$$L_\Delta(w_1, w_2) = \begin{bmatrix} M(\frac{d}{dt})w_1 \\ X(\frac{d}{dt})w_1 \end{bmatrix}^T E \begin{bmatrix} M(\frac{d}{dt})w_2 \\ X(\frac{d}{dt})w_2 \end{bmatrix}$$

for all $w_1, w_2 \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^q)$.

Proof. For the proof, see the appendix.

Let $\Gamma \in \mathbb{R}^{q\times q}[\xi]$ be para-Hermitian: $\Gamma^\star = \Gamma$. An $F \in \mathbb{R}^{q\times q}[\xi]$ is said to induce a *symmetric factorization* of Γ if $\Gamma(\xi) = F^T(-\xi)F(\xi)$. It is said to be a *symmetric Hurwitz factorization* if F is square and Hurwitz and a *symmetric anti-Hurwitz factorization* if $F^\star$ is square and Hurwitz. It is easy to see that for a symmetric factorization to exist we need to have $\Gamma(i\omega) \geq 0\ \forall \omega \in \mathbb{R}$ and for an (anti-)Hurwitz one to exist we must have $\Gamma(i\omega) > 0\ \forall \omega \in \mathbb{R}$. The converses are also true but not at all trivial in the matrix case. This result is well known (see, e.g., [9], [10], [18], [21]), and we state it for easy reference.

PROPOSITION 5.6. *Let* $\Gamma \in \mathbb{R}^{q\times q}[\xi]$ *be para-Hermitian. Then*

(i) Γ *allows a symmetric factorization iff* $\Gamma(i\omega) \geq 0$ *for all* $\omega \in \mathbb{R}$.

(ii) Γ *allows a symmetric Hurwitz factorization iff* $\Gamma(i\omega) > 0$ *for all* $\omega \in \mathbb{R}$. *Such a factorization* $\Gamma(\xi) = F^T(-\xi)F(\xi)$ *is unique up to premultiplication of* $F(\xi)$ *by an orthogonal matrix.*

(iii) Γ *allows a symmetric anti-Hurwitz factorization iff* $\Gamma(i\omega) > 0$ *for all* $\omega \in \mathbb{R}$. *Such a factorization* $\Gamma(\xi) = F^T(-\xi)F(\xi)$ *is unique up to premultiplication of* $F(\xi)$ *by an orthogonal matrix.*

An important issue of concern is the uniqueness of the storage function, and therefore of the dissipation function, because of the one-to-one relation between the two. When $\int Q_\Phi = 0$, then the associated storage function is unique ($\Psi(\zeta, \eta) = \frac{\Phi(\zeta,\eta)}{\zeta+\eta}$) and the dissipation function is zero. However, in general there are many possibilities.

THEOREM 5.7. *Let* $\int Q_\Phi \geq 0$. *Then there exist storage functions* Ψ_- *and* Ψ_+ *for* Φ *such that any other storage function* Ψ *for* Φ *satisfies*

$$\Psi_- \leq \Psi \leq \Psi_+. \tag{5.14}$$

If $\int Q_\Phi \overset{per}{>} 0$ *then* Ψ^- *and* Ψ^+ *may be constructed as follows. Let* $\partial\Phi(\xi) = H^T(-\xi)H(\xi)$ *and* $\partial\Phi(\xi) = A^T(-\xi)A(\xi)$ *be, respectively, Hurwitz and anti-Hurwitz factorizations of* $\partial\Phi$. *Then*

$$\Psi_+(\zeta, \eta) = \frac{\Phi(\zeta, \eta) - A^T(\zeta)A(\eta)}{\zeta + \eta} \tag{5.15}$$

and

$$\Psi_-(\zeta, \eta) = \frac{\Phi(\zeta, \eta) - H^T(\zeta)H(\eta)}{\zeta + \eta}. \tag{5.16}$$

Proof. For the proof, see the appendix.

We close this section with a few remarks.

Remark 5.8. In this section we have studied average positivity with, in $Q_\Phi(w)$, $w \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^q)$ or $\mathfrak{D}(\mathbb{R}, \mathbb{R}^q)$, but otherwise free. It is of interest to generalize these

concepts to the case that $w \in \mathfrak{B}$ with $\mathfrak{B}$ a given element of $\mathfrak{L}^q$. Of course, in section 4 we have considered precisely such situations for $\mathfrak{B}$'s that are *autonomous*. Actually, it turns out that the theory of section 5 is immediately applicable to systems $\mathfrak{B} \in \mathfrak{L}^q$ that are *controllable*. Indeed, let $\mathfrak{B} \in \mathfrak{L}^q$ be controllable and assume that we want to study when

$$\int_{-\infty}^{+\infty} Q_\Phi(w)dt \geq 0 \quad \text{or} \quad \int_{-\infty}^{+\infty} Q_\Phi(w)dt = 0 \tag{5.17}$$

holds for all $w \in \mathfrak{B} \cap \mathfrak{D}(\mathbb{R}, \mathbb{R}^q)$. Simply construct an image representation for $\mathfrak{B}$, say,

$$w = M\left(\frac{d}{dt}\right)\ell. \tag{5.18}$$

Upon substituting (5.18) in (5.17), we see that the issue then becomes one of studying when

$$\int_{-\infty}^{+\infty} Q_\Phi\left(M\left(\frac{d}{dt}\right)\ell\right)dt \geq 0 \quad \text{or} \quad \int_{-\infty}^{+\infty} Q_\Phi\left(M\left(\frac{d}{dt}\right)\ell\right)dt = 0$$

for all $\ell \in \mathfrak{D}(\mathbb{R}, \mathbb{R}^\bullet)$. Since obviously $Q_\Phi(M(\frac{d}{dt})\ell) = Q_{\Phi'}(\ell)$ with

$$\Phi'(\zeta, \eta) := M^T(\zeta)\Phi(\zeta, \eta)M(\eta),$$

the problem reduces to studying Φ'. For example, the existence of a storage function is established as follows. Without loss of generality, take (5.18) to be an observable image representation. Then M has a polynomial left inverse $M^\dagger$. By Proposition 5.4, there exists Ψ' such that $\frac{d}{dt}Q_{\Psi'}(\ell) \leq Q_{\Phi'}(\ell)$ for all $\ell \in \mathfrak{D}(\mathbb{R}, \mathbb{R}^\bullet)$. Now define $\Psi(\zeta, \eta) := M^{\dagger T}(\zeta)\Psi'(\zeta, \eta)M^\dagger(\eta)$. Then for $w = M(\frac{d}{dt})\ell$ we have $Q_\Psi(w) = Q_{\Psi'}(M^\dagger(\frac{d}{dt})w) = Q_{\Psi'}(\ell)$ and $Q_\Phi(w) = Q_{\Phi'}(\ell)$, so we obtain $\frac{d}{dt}Q_\Psi(w) \leq Q_\Phi(w)$.

The case that $\mathfrak{B} \in \mathfrak{L}^q$ is neither controllable nor autonomous will be studied in a later publication. The next comment is relevant to the question of what the appropriate definition of dissipativity is in that case.

Remark 5.9. Finding an appropriate definition of a dissipative system is an issue that has attracted considerable attention (see [29], [12], [24], [27]). Of course, this is at the root of the issues discussed in the present article. Let $\mathfrak{B} \in \mathfrak{L}^q$. There are many examples where the instantaneous rate of supply (say, of energy) into the system is given, not by a static function of the external variables, but by a QDF, $Q_\Phi(w)$. The study of supply rates that are themselves dynamic is one of the novel aspects of the present paper. When would one want to call $\mathfrak{B}$ dissipative with respect to $Q_\Phi(w)$? Lossless? Conservative? When would one want to say that $\mathfrak{B}$ absorbs some of the supply? The definitions of average nonnegativity for dissipativeness, and $\int Q_\Phi = 0$ for losslessness (= conservativeness), are fully adequate *provided* that $\mathfrak{B}$ *is controllable* (see Remark 5.8). However, Proposition 5.4 points to another definition which does not need controllability and which, in the controllable case, reduces to it. Thus, we arrive at the following definition as the most general: $\mathfrak{B} \in \mathfrak{L}^q$ is said to be *dissipative* with respect to the supply rate Q_Φ if there exists a Q_Ψ such that $\frac{d}{dt}Q_\Psi(w) \leq Q_\Phi(w)$ for all $w \in \mathfrak{B}$, and *lossless, or conservative,* if this holds with equality. The unfortunate aspect of this definition is its *existential* nature — it shares this notorious feature with the first and second law of thermodynamics. It does not seem an easy matter in the noncontrollable case to reduce this to a statement involving only Q_Φ, and without

invoking a to-be-constructed Q_Ψ. In Theorem 5.5 we have unraveled this existence question a bit by proving that this Q_Ψ will be a state function.

Note that the proposed definition of dissipativity and losslessness is an interesting generalization of the notion of a Lyapunov function since, for autonomous systems, it is natural to take the external supply $\Phi = 0$. Also note that this definition holds for any $\mathfrak{B}$ and Φ and does not require the introduction of the notion of state. In other words, dynamical systems with free variables that allow interaction with the environment relate to flows on manifolds, just as dissipative systems relate to Lyapunov functions.

Remark 5.10. Let $\mathfrak{B}$ be controllable and assume that it is dissipative with respect to the supply rate $Q_\Phi(w)$ (see Remark 5.9). Also in this general case every storage function is a state function, and every dissipation function is a state/supply function. However, this time, not simply the state of Φ is involved, but the state of a system obtained by combining the dynamics of Φ and $\mathfrak{B}$. This is elaborated in [26].

Remark 5.11. Let $\Phi(\zeta, \eta) = M^T(\zeta)\Sigma_M M(\eta)$. Consider the system in image representation (5.5). Then it can be shown that for Φ to be average nonnegative, there must be an input/output partition for this system so that all the input components correspond to +1's in Σ_Φ. In other words, the supply rate $v^T\Sigma_M v$ is always of the form $\|u\|^2 + \|y_1\|^2 - \|y_2\|^2$, with u an input, and y_1, y_2 outputs.

Remark 5.12. It follows from Theorem 5.5 that a factorization of the polynomial matrix $\Phi(-\xi, \xi) = M^T(-\xi)\Sigma_M M(\xi)$ into $F^T(-\xi)F(\xi)$ always leads to a situation in which the McMillan degree of M is equal to that of $\mathrm{col}(M, F)$. This means that the factorization is a *regular* factorization (as this property is called). In the H_∞-problem factorization, questions are encountered in which the existence of a regular factorization poses a serious problem.

Remark 5.13. It is easy to see that the set of storage functions corresponding to a given supply rate is convex. Moreover, in the case of average positivity, $\Psi_- \neq \Psi_+$ and hence, in this case, there are an infinite number of possible storage functions. Actually, in this respect it is worth mentioning the following refinement of Theorem 5.7, which follows immediately from our proof of this theorem. If $\Phi(-\xi, \xi)$ satisfies $\Phi(-i\omega, i\omega) \geq 0$ (but not $\Phi(-i\omega, i\omega) > 0$) for all $\omega \in \mathbb{R}$, then a symmetric Hurwitz factorization does not exist. In this case, there are two possibilities: either $\det(\partial\Phi) \neq 0$ or $\det(\partial\Phi) = 0$. In the former case, $\Phi(-\xi, \xi)$ allows a factorization $\Phi(-\xi, \xi) = H^T(-\xi)H(\xi)$ with H "almost Hurwitz" (i.e., H has all its singularities in $\Re e(\lambda) \leq 0$). In the latter case, there exists a unimodular matrix U such that

$$\Phi(\zeta, \eta) = U^T(\zeta)\Phi'(\zeta, \eta)U(\eta),$$

with Φ' of the form

$$\Phi' = \begin{bmatrix} \Phi_1 & \Phi_2 \\ \Phi_2^* & \Phi_3 \end{bmatrix},$$

with $\det(\partial\Phi_1) \neq 0$ and $\partial\Phi_2 = 0$, $\partial\Phi_3 = 0$. Factor $\Phi_1(-\xi, \xi)$ as before as $H_1^T(-\xi)H_1(\xi)$. Then

$$H = \begin{bmatrix} H_1 & 0 \\ 0 & 0 \end{bmatrix} U_1$$

yields an almost Hurwitz-like factorization of $\partial\Phi$. Similarly, we can define an almost anti-Hurwitz-like factorization $\partial\Phi = A^\star A$ of any Φ satisfying $\Phi(-i\omega, i\omega) \geq 0$ for all $\omega \in \mathbb{R}$. The computation of Ψ_+ and Ψ_- given in Theorem 5.7 holds unaltered with

this A and H. Note that in the lossless case ($\partial\Phi = 0$) this yields $\Psi_+ = \Psi_-$, whence the uniqueness of Ψ.

Remark 5.14. It is easy to deduce from the proof of Theorem 5.7 that Ψ_- and Ψ_+ have the following interpretations. Let $x = X(\frac{d}{dt})w$ be the state (see 5.6). Let $a \in \mathbb{R}^n$. Consider all $w \in \mathfrak{D}(\mathbb{R}, \mathbb{R}^q)$ such that $(X(\frac{d}{dt})w)(0) = a$. Denote this set by $\mathfrak{B}_a$. By Theorem 5.5 we know that Q_{Ψ_-} is a state function, say, $Q_{\Psi_-}(w) = \|X(\frac{d}{dt})w\|_{K_-}$, for some symmetric matrix $K_- \in \mathbb{R}^{n\times n}$. Hence for $w \in \mathfrak{B}_a$ we have $Q_{\Psi_-}(w)(0) = a^T K_- a$. Similarly $Q_{\Psi_+}(w)(0) = a^T K_+ a$ for some symmetric matrix K_+. Then it can be shown that

$$a^T K_- a = \sup_{w\in\mathfrak{B}_a} \left(-\int_0^{+\infty} Q_\Phi(w)dt \right) \tag{5.19}$$

and

$$a^T K_+ a = \inf_{w\in\mathfrak{B}_a} \left(\int_{-\infty}^{0} Q_\Phi(w)dt \right). \tag{5.20}$$

For this reason, $Q_{\Psi_-}(w)(0)$ is called the *available storage* and $Q_{\Psi_+}(w)(0)$ the *required supply* at $t = 0$ due to w. In this inf and sup, one keeps the past, respectively, the future of w fixed.

6. Half-line positivity. In section 5, we studied QDFs for which $\int_{-\infty}^{+\infty} Q_\Phi(w)dt \geq 0$. The intuitive idea was that this expresses that the net supply (of "energy") is directed into the system: energy is being absorbed and dissipated in the system. There are, however, situations where at any moment in time the system has absorbed energy, i.e., $\int_{-\infty}^{t} Q_\Phi(w)(\tau)d\tau \geq 0$ for all $t \in \mathbb{R}$. For example, electrical circuits and mechanical devices at rest are in a state of minimum energy, and therefore the energy delivered *up to any time* is nonnegative. This type of positivity is studied in this section. It plays a crucial role in H_∞ problems.

DEFINITION 6.1. *Let $\Phi \in \mathbb{R}^{q\times q}_s[\zeta, \eta]$. The QDF Q_Φ (or simply Φ) is said to be* half-line nonnegative*, denoted by $\int^t Q_\Phi \geq 0$, if $\int_{-\infty}^0 Q_\Phi(w)dt \geq 0$ for all $w \in \mathfrak{D}(\mathbb{R}, \mathbb{R}^q)$, and* half-line positive*, denoted $\int^t Q_\Phi > 0$, if in addition $\int_{-\infty}^0 Q_\Phi(w)dt = 0$ implies $w(t) = 0$ for $t \leq 0$.*

Note that half-line nonnegativity implies average nonnegativity, and that half-line positivity implies average positivity.

Write $\Phi(\zeta, \eta) = M^T(\zeta)\Sigma_M M(\eta)$ and partition M conform Σ_M as

$$M = \begin{bmatrix} P \\ N \end{bmatrix} \tag{6.1}$$

so that $\Phi(\zeta, \eta) = P^T(\zeta)P(\eta) - N^T(\zeta)N(\eta)$ and hence $Q_\Phi(w) = \|P(\frac{d}{dt})w\|^2 - \|N(\frac{d}{dt})w\|^2$. In the following, for $\lambda \in \mathbb{C}$, let $\bar{\lambda}$ denote its complex conjugate.

PROPOSITION 6.2. *Let $\Phi \in \mathbb{R}^{q\times q}_s[\zeta, \eta]$. Then*

(i) $(\int^t Q_\Phi \geq 0) \Rightarrow (\Phi(\bar{\lambda}, \lambda) \geq 0 \ \forall \lambda \in \mathbb{C}, \Re e(\lambda) \geq 0)$

(ii) $(\int^t Q_\Phi > 0) \Rightarrow (\Phi(\bar{\lambda}, \lambda) \geq 0 \ \forall \lambda \in \mathbb{C}, \Re e(\lambda) \geq 0$ *and* $\det(\partial\Phi) \neq 0)$.

Proof. For the proof, see the appendix.

As noted before, it immediately follows from the definitions that half-line nonnegativity implies average nonnegativity, etc. Thus, Proposition 5.4 implies the existence of a storage function. It is the nonnegativity of the storage function that allows us to conclude the *half-line* positivity.

THEOREM 6.3. *Let $\Phi \in \mathbb{R}^{q\times q}_s[\zeta, \eta]$. Then the following statements are equivalent.*

1. $\int^t Q_\Phi \geq 0$,
2. *there exists a storage function* $\Psi \geq 0$ *for* Φ,
3. Φ *admits a storage function, and the storage function* Ψ_+ *defined in Theorem* 5.7 *satisfies* $\Psi_+ \geq 0$.

Proof. For the proof, see the appendix.

In order to check half-line nonnegativity, one could thus in principle proceed as follows. Verify that $\Phi(-i\omega, i\omega) \geq 0$ for all $\omega \in \mathbb{R}$, compute Ψ_+, and check whether $\Psi_+ \geq 0$. In some situations, it is actually possible to verify this condition in a more immediate fashion; for example, when $\Phi(\zeta, \eta) = \Phi_0$, a constant matrix, with $\Phi_0 \geq 0$ (trivial, but that is the case that occurs in standard LQ theory!), or when in (6.1) P is square and $\det(P) \neq 0$. Then, under the assumption that a storage function exists (equivalently: $N^T(-i\omega)N(i\omega) \leq P^T(-i\omega)P(i\omega)$ for all $\omega \in \mathbb{R}$), all storage functions are actually nonnegative if one of them is nonnegative. In fact, in this case the following theorem holds.

THEOREM 6.4. *Let* $\Phi \in \mathbb{R}^{q\times q}_s[\zeta, \eta]$. *Assume it is factored as* $\Phi(\zeta, \eta) = P^T(\zeta)P(\eta) - N^T(\zeta)N(\eta)$ *with* P *square and* $\det(P) \neq 0$. *Let* $X \in \mathbb{R}^{\bullet\times q}[\xi]$ *be a minimal state map for the* $\mathfrak{B}$ *given in image representation by* (6.1). *The following statements are equivalent:*

1. $\int^t Q_\Phi \geq 0$,
2. $\Phi(\bar{\lambda}, \lambda) \geq 0$ *for all* $\lambda \in \mathbb{C}$, $\Re e(\lambda) \geq 0$,
3. NP^{-1} *has no poles in* $\Re e(\lambda) \geq 0$ *and* $\Phi(-i\omega, i\omega) \geq 0$ *for all* $\omega \in \mathbb{R}$,
4. *there exists a storage function* $\Psi \geq 0$ *for* Φ,
5. *there exists a storage function for* Φ *and every storage function* Ψ *for* Φ *satisfies* $\Psi \geq 0$,
6. *there exists a real symmetric matrix* $K > 0$ *such that* $Q_K(w) := \|X(\frac{d}{dt})w\|^2_K$ *is a storage function for* Φ,
7. *there exists a storage function for* Φ *and every real symmetric matrix* K *such that* $Q_K(w) := \|X(\frac{d}{dt})w\|^2_K$ *is a storage function for* Φ *satisfies* $K > 0$.

Furthermore, if $\left[\begin{smallmatrix} P \\ N \end{smallmatrix}\right]$ *is observable, then any of the above statements is equivalent with*

3′. P *is Hurwitz and* $\Phi(-i\omega, i\omega) \geq 0$ *for all* $\omega \in \mathbb{R}$.

Proof. For the proof, see the appendix.

7. Observability. One of the noticeable features of QDFs is that a number of interesting systems theory concepts generalize very nicely to QDFs. We have already seen that the state of a symmetric canonical factorization of Φ functions as the state of the QDF Q_Φ. In this section we introduce observability of a QDF. In a later section we will discuss duality of QDFs.

For $\Phi \in \mathbb{R}^{q_1\times q_2}[\zeta, \eta]$ and $w_1 \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^{q_1})$ fixed, the linear map $w_2 \mapsto L_\Phi(w_1, w_2)$ is denoted by $L_\Phi(w_1, \bullet)$. For $w_2 \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^{q_2})$ fixed, the linear map $w_1 \mapsto L_\Phi(w_1, w_2)$ is denoted by $L_\Phi(\bullet, w_2)$. The BLDF Φ is called observable if $L_\Phi(w_1, \bullet)$ and $L_\Phi(\bullet, w_2)$ determine w_1 and w_2 uniquely. Equivalently we have the following.

DEFINITION 7.1. *Let* $\Phi \in \mathbb{R}^{q_1\times q_2}[\zeta, \eta]$. *We call* Φ observable *if, for all* $w_1 \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^{q_1})$ *and for all* $w_2 \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^{q_2})$, *we have*

$$L_\Phi(w_1, \bullet) = 0 \;\Leftrightarrow\; w_1 = 0$$

and

$$L_\Phi(\bullet, w_2) = 0 \;\Leftrightarrow\; w_2 = 0.$$

The following theorem gives necessary and sufficient conditions for observability purely in terms of the two-variable polynomial matrix Φ and in terms of the (one-

variable) polynomial matrices N and M occurring in any canonical factorization of Φ.

THEOREM 7.2. *Let $\Phi(\zeta,\eta) = N^T(\zeta)M(\eta)$ be a canonical factorization. The following statements are equivalent:*

1. *Φ is observable,*
2. *for every $\lambda \in \mathbb{C}$, the rows of $\Phi(\lambda,\xi) \in \mathbb{R}^{q_1\times q_2}[\xi]$, and the columns of $\Phi(\xi,\lambda) \in \mathbb{R}^{q_1\times q_2}[\xi]$ are linearly independent over $\mathbb{C}$,*
3. *$N(\lambda)$ and $M(\lambda)$ have full column rank for all $\lambda \in \mathbb{C}$; equivalently, the image representations $v_1 = N(\frac{d}{dt})w_1$ and $v_2 = M(\frac{d}{dt})w_2$ are observable,*

Proof. For the proof, see the appendix.

If Φ is symmetric, then we have $\Phi^T(\lambda,\xi) = \Phi(\xi,\lambda)$, so condition (ii) above can be replaced by a single statement on the independence of the rows of $\Phi(\lambda,\xi)$. Also, in this case the maps $L_\Phi(w,\bullet)$ and $L_\Phi(\bullet,w)$ coincide. Furthermore, for the particular symmetric canonical factorization $\Phi(\zeta,\eta) = M_c^T(\zeta)\Sigma_\Phi M_c(\eta)$ obtained from (3.15), we have $|\Phi|(\zeta,\eta) = M_c^T(\zeta)M_c(\eta)$. Hence observability of Φ is also equivalent with $|\Phi| > 0$ and with the condition $|\Phi|(\bar\lambda,\lambda) > 0$ for all $\lambda \in \mathbb{C}$. This immediately yields the following.

COROLLARY 7.3. *Let $\Phi \in \mathbb{R}_s^{q\times q}[\zeta,\eta]$ and let $\Phi(\zeta,\eta) = M^T(\zeta)\Sigma_\Phi M(\eta)$ be a symmetric canonical factorization. Then the following statements are equivalent:*

1. *Φ is observable,*
2. *$L_\Phi(w,\bullet) = 0 \;\Leftrightarrow\; w = 0$,*
3. *for every $\lambda \in \mathbb{C}$, the rows of $\Phi(\lambda,\xi) \in \mathbb{R}^{q_1\times q_2}[\xi]$ are linearly independent over $\mathbb{C}$,*
4. *$M(\lambda)$ has full column rank for all $\lambda \in \mathbb{C}$, equivalently, the image representation $v = M(\frac{d}{dt})w$ is observable,*
5. *$|\Phi| > 0$,*
6. *$|\Phi|(\bar\lambda,\lambda) > 0$ for all $\lambda \in \mathbb{C}$.*

8. Strict positivity. Throughout this section we assume that $\Phi \in \mathbb{R}_s^{q\times q}[\zeta,\eta]$ is observable. We now introduce and develop the notion of strict positivity. The concept of strict half-line positivity given here is very analogous to that used by Meinsma [19].

DEFINITION 8.1. *Let $\Phi \in \mathbb{R}_s^{q\times q}[\zeta,\eta]$ be observable. We call the QDF Q_Φ* strictly positive*, denoted $\Phi \gg 0$, if there exists $\epsilon > 0$ such that $\Phi - \epsilon|\Phi| \geq 0$. We call it* strictly average positive*, denoted by $\int Q_\Phi \gg 0$, if there exists $\epsilon > 0$ such that*

$$\int_{-\infty}^{+\infty} Q_\Phi(w)dt \geq \epsilon \int_{-\infty}^{+\infty} Q_{|\Phi|}(w)dt \tag{8.1}$$

for all $w \in \mathfrak{D}(\mathbb{R},\mathbb{R}^q)$. We call it strictly half-line positive*, denoted $\int^t Q_\Phi \gg 0$, if there exists an $\epsilon > 0$ such that*

$$\int_{-\infty}^{0} Q_\Phi(w)dt \geq \epsilon \int_{-\infty}^{0} Q_{|\Phi|}(w)dt \tag{8.2}$$

for all $w \in \mathfrak{D}(\mathbb{R},\mathbb{R}^q)$. Note that (because of observability) strict positivity implies positivity, and similarly for the other cases.

These notions of strict positivity involve $|\Phi|$ which may be difficult to evaluate. However, it is possible to relate it to any canonical factorization of Φ. This is stated in the next proposition. For simplicity we state only the case of strict average positivity. However, completely analogous statements hold for simple strict positivity or for strict half-line positivity.

PROPOSITION 8.2. *Let* $\Phi \in \mathbb{R}^{q\times q}_s[\zeta,\eta]$ *be observable and let* $\Phi(\zeta,\eta)=P^T(\zeta)P(\eta)-N^T(\zeta)N(\eta)$ *be a symmetric canonical factorization of* Φ (*see* (3.12)). *Denote* $M=\left[\begin{smallmatrix}P\\N\end{smallmatrix}\right]$ *The following are equivalent:*

1. Φ *is strictly average positive.*
2. *There exists an* $\epsilon>0$ *such that*

$$\int_{-\infty}^{+\infty} \|M\left(\frac{d}{dt}\right)w\|^2_{\Sigma_\Phi}dt \geq \epsilon \int_{-\infty}^{+\infty} \|M\left(\frac{d}{dt}\right)w\|^2 dt \tag{8.3}$$

for all $w\in\mathfrak{D}(\mathbb{R},\mathbb{R}^q)$. *Here* $\|a\|^2_{\Sigma_\Phi}$ *denotes* $a^T\Sigma_\Phi a$.

3. *There exists an* $\alpha<1$ *such that*

$$\int_{-\infty}^{+\infty} \|N\left(\frac{d}{dt}\right)w\|^2 dt \leq \alpha \int_{-\infty}^{+\infty} \|P\left(\frac{d}{dt}\right)w\| dt \tag{8.4}$$

for all $w\in\mathfrak{D}(\mathbb{R},\mathbb{R}^q)$.

Moreover, also for a noncanonical factorization (3.11), (2) *and* (3) *(with* Σ_Φ *replaced by* Σ_M) *are equivalent and imply* (1).

Proof. For the proof, see the appendix.

9. A Pick matrix condition for half-line positivity. It is surprisingly difficult to establish some type of analogue of Proposition 5.2 for half-line positivity, and earlier attempts [28], [30], [1]) turned out to be flawed. In Proposition 6.4 such an analogue of Proposition 5.2 was given but only in the special case where $\Phi(\zeta,\eta)=P^T(\zeta)P(\eta)-N^T(\zeta)N(\eta)$ with $\det(P)\neq 0$. In this section we give a necessary and sufficient condition for strict half-line positivity in terms of Φ.

As is well known, the Pick matrix plays an important role in system and circuit theory, in particular in connection with passivity properties of linear dynamical systems; see [34], [4], [5]. We derive a Pick-matrix-type test for nonnegativity of Ψ_+. This test is perhaps the most original specific result of this paper. For simplicity we consider only the case of strict half-line positivity. First, however, we need to define the Pick-type matrix which may be computed effectively from a $\Phi\in\mathbb{R}^{q\times q}_s[\zeta,\eta]$. Let $F\in\mathbb{R}^{q\times q}[\xi]$, and assume that $\det(F)\neq 0$. We call F *semisimple* if for all $\lambda\in\mathbb{C}$ the dimension of the kernel of $F(\lambda)$ is equal to the multiplicity of λ as a root of $\det(F)$. Note that F is certainly semisimple if $\det(F)$ has distinct roots. We now define the matrix T_Φ. Since the expression is much simpler in the semisimple case, we explain that case first.

DEFINITION 9.1 (semisimple case). *Let* $\Phi\in\mathbb{R}^{q\times q}_s[\zeta,\eta]$ *be observable, and assume that* $\det(\partial\Phi)$ *has no roots on the imaginary axis. Let* $\lambda_1,\lambda_2,\dots,\lambda_n\in\mathbb{C}$ *be the roots of* $\det(\partial\Phi)$ *with positive real part and let* $a_1,a_2,\dots,a_n\in\mathbb{C}^q$ *be such that* $\partial\Phi(\lambda_i)a_i=0$, *and such that the* a_k*'s associated with the same* λ_i *form a basis of* $\ker(\partial\Phi(\lambda_i))$. *Then the* Pick matrix *of* Φ *is defined as*

$$T_\Phi := \left[\frac{\bar a_i^T\Phi(\bar\lambda_i,\lambda_j)a_j}{\bar\lambda_i+\lambda_j}\right]_{i,j=1,\dots,n}. \tag{9.1}$$

In order to define the matrix T_Φ in the general case, we need to take into account the algebraic multiplicities of the roots λ_i.

DEFINITION 9.2 (general case). *Let* $\Phi\in\mathbb{R}^{q\times q}_s[\zeta,\eta]$ *be observable,* $\det(\partial\Phi)\neq 0$, *and assume that* $\det(\partial\Phi)$ *has no roots on the imaginary axis. Let* $\lambda_1,\lambda_2,\dots,\lambda_k\in\mathbb{C}$ *be the* distinct *roots of* $\det(\partial\Phi)$ *with positive real part, and denote by* n_i *the multiplicity*

of λ_i as a root of $\det(\partial\Phi)$. *For $i = 1, 2, \ldots, k$, there are n_i linearly independent vectors $a_{i,0}, a_{i,1}, \ldots, a_{i,n_i-1}$ determined by the (n_i) linear equations*

$$\sum_{j=\ell}^{n_i-1} \begin{pmatrix} j \\ \ell \end{pmatrix} (\partial\Phi)^{(j-\ell)}(\lambda_i) a_{i,j} = 0, \quad (\ell = 0, 1, \ldots, n_i - 1).$$

Here, $(\partial\Phi)^{(k)}(\xi)$ denotes the kth derivative of the polynomial matrix $\partial\Phi$.

For $i = 1, 2, \ldots, k$, define

$$A_i := \begin{bmatrix} a_{i,0} & 0 & \cdots & 0 \\ a_{i,0} & a_{i,1} & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ a_{i,0} & a_{i,1} & \ldots & a_{i,n_i-1} \end{bmatrix} \in \mathbb{C}^{n_i q \times n_i}$$

Also, define $\Phi_{i,j} \in \mathbb{C}^{n_i q \times n_j q}$ by defining its (r,s)th block to be the $q \times q$ matrix

$$(\Phi_{i,j})_{r,s} := \Phi^{(r,s)}(\bar{\lambda}_i, \lambda_j), \quad r = 1, 2, \ldots, n_i \; ; \; s = 1, 2, \ldots, n_j,$$

where $\Phi^{(k,\ell)}$ means taking the kth partial derivative with respect to ζ and the ℓth with respect to η.

Then we define the Pick matrix *of Φ as the matrix T_Φ whose (i,j)th block is given by $T_{i,j} \in \mathbb{C}^{n_i \times n_j}$, with*

$$T_{i,j} := \frac{1}{\bar{\lambda}_i + \lambda_j} \bar{A}_i^T \Phi_{i,j} A_j.$$

Note that the sum $\sum_{i=1}^{k} n_i$ of the multiplicities is equal to $n := \frac{1}{2} \deg \det(\partial\Phi)$, and that T_Φ is a complex Hermitian matrix of size $n \times n$.

The next theorem is the most refined result of this paper. It shows, on the one hand, the relation between strict half-line positivity and positivity of a storage function, and, on the other hand, the relation with the positivity of the Pick matrix T_Φ.

We have seen in Theorem 5.5 that a storage function is a quadratic state function, i.e., $Q_\Phi(w)$ is of the form $x^T K x$, $K = K^T$, with $x = X(\frac{d}{dt})w$ a minimal state map for Φ. We call this state function *positive definite* if $K > 0$.

THEOREM 9.3. *Let $\Phi \in \mathbb{R}_s^{q \times q}[\zeta, \eta]$ be observable. The following are equivalent:*

1. $\int^t Q_\Phi \gg 0$,
2. (a) $\int Q_\Phi \gg 0$,
   (b) *there exists a storage function that is a positive definite state function,*
3. (a) $\exists \epsilon > 0$ *such that* $\Phi(-i\omega, i\omega) \geq \epsilon |\Phi|(-i\omega, i\omega)$ *for all* $\omega \in \mathbb{R}$,
   (b) $T_\Phi > 0$.

Proof. For the proof, see the appendix.

Remark 9.4. It follows from the proof of Theorem 9.3 that half-line nonnegativity implies that the Pick-type matrix T_Φ (see (9.1)) is ≥ 0 whenever any set of λ_i's in the right half of the complex plane and any set of a_i's are chosen. It is possible to prove that if T_Φ is nonnegative definite for any such choice for the λ_i's and a_i's, then we have half-line nonnegativity. The remarkable thing about Theorem 9.3 is that it suffices to evaluate T_Φ at the set of special λ_i's and a_i's obtained from the singularities of $\partial\Phi$.

Remark 9.5. It is well known that solvability of a certain Nevanlinna–Pick interpolation problem is equivalent to positive definiteness of a given Pick matrix. In fact, in [34] the necessity of the positive definite Pick matrix is shown using a half-line positivity argument. Theorem 9.3 states that positive definiteness of a given Pick matrix is also *sufficient* for half-line positivity.

Remark 9.6. It can be shown that if Φ is observable, then $\int Q_\Phi \gg 0$ implies that $\Psi_+ - \Psi_-$ is a positive definite state function, with Ψ_+ and Ψ_- as defined in Theorem 5.7.

Remark 9.7. It is possible to generalize the T_Φ-test of Theorem 9.3 to half-line positive (instead of strictly half-line positive QDFs) by including "infinite zeros" of $\det(\partial\Phi)$. However, the notation gets very involved, and therefore we will not do this here.

10. Duality. In the present section we discuss some remarkable relations between positivity of QDFs and their duals. These relations are of interest in their own right and will be of crucial importance in our treatment of the H_∞-problem [25].

Let $\mathfrak{B}_1$ and $\mathfrak{B}_2 \in \mathfrak{L}^q$. We call $\mathfrak{B}_1$ and $\mathfrak{B}_2$ *complementary* if $\mathfrak{B}_1 \oplus \mathfrak{B}_2 = \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^q)$. It is easy to see that this implies that both $\mathfrak{B}_1$ and $\mathfrak{B}_2$ must be controllable: uncontrollable $\mathfrak{B}$'s in $\mathfrak{L}^q$ have no complement in $\mathfrak{L}^q$. We call them *dual* if they are complementary and if $\langle w_1, w_2\rangle = 0$ for all $w_1 \in \mathfrak{B}_1 \cap \mathfrak{D}(\mathbb{R}, \mathbb{R}^q)$ and $w_2 \in \mathfrak{B}_2 \cap \mathfrak{D}(\mathbb{R}, \mathbb{R}^q)$, where $\langle w_1, w_2\rangle$ denotes the usual inner product $\int_{-\infty}^{+\infty} w_1^T w_2 dt$. If this is the case, then we denote $\mathfrak{B}_2$ as $\mathfrak{B}_1^\perp$, since $\mathfrak{B}_1$ defines $\mathfrak{B}_1^\perp$ uniquely. Obviously we also have $(\mathfrak{B}_1^\perp)^\perp = \mathfrak{B}_1$.

It is easy to see that $R(\frac{d}{dt})w = 0$ is a minimal kernel representation of the controllable $\mathfrak{B}$ iff $v = R^T(-\frac{d}{dt})\ell$ is an observable image representation of $\mathfrak{B}^\perp$ (a kernel representation $R(\frac{d}{dt})w = 0$ of $\mathfrak{B}$ is called *minimal* if R has full row rank). Consequently, $w = M(\frac{d}{dt})\ell$ is an observable image representation of $\mathfrak{B}$ iff $M^T(-\frac{d}{dt})v = 0$ is a minimal kernel representation of $\mathfrak{B}^\perp$. This duality can also be extended to state representations, in the following sense. If

$$E\frac{dx}{dt} + Fx + Gw = 0, \tag{10.1}$$

is an n-dimensional minimal state representation of $\mathfrak{B}$, then $\mathfrak{B}^\perp$ admits an also n-dimensional minimal state representation (thus the dimensions of the minimal state representations are the same), say,

$$E'\frac{dz}{dt} + F'z + G'v = 0, \tag{10.2}$$

having the property that for all $(w, x) \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^{q+n})$ satisfying (10.1), and for all $(v, z) \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^{q+n})$ satisfying (10.2) there holds the following kind of duality involving the state

$$\frac{d}{dt}z^T x = v^T w. \tag{10.3}$$

In fact, this is an immediate consequence of the following.

PROPOSITION 10.1. *Let $R(\frac{d}{dt})w = 0$ and $w = M(\frac{d}{dt})\ell$ be a minimal kernel representation and an observable image representation, respectively, of the controllable system $\mathfrak{B} \in \mathfrak{L}^q$. Assume that $X \in \mathbb{R}^{n\times\bullet}[\xi]$ defines a minimal state map for $\mathfrak{B}$, i.e.,*

$x = X(\frac{d}{dt})\ell$ defines a minimal state of $\mathfrak{B}$. Then there exists a $Z \in \mathbb{R}^{n\times\bullet}[\xi]$ defining a minimal state map $Z(\frac{d}{dt})$ for $\mathfrak{B}^\perp$, such that for all $\ell, \ell' \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^\bullet)$, we have

$$\frac{d}{dt}\left(Z\left(\frac{d}{dt}\right)\ell'\right)^T X\left(\frac{d}{dt}\right)\ell = \left(R^T\left(-\frac{d}{dt}\right)\ell'\right)^T M\left(\frac{d}{dt}\right)\ell. \tag{10.4}$$

If we define $\Psi(\zeta,\eta) := Z^T(\zeta)X(\eta)$ and $\Phi(\zeta,\eta) := R(-\zeta)M(\eta)$, then (10.4) is equivalent to $\overset{\bullet}{\Psi} = \Phi$.

Proof. For the proof, see the appendix.

If a pair of minimal state maps (X, Z) of $\mathfrak{B}$ and $\mathfrak{B}^\perp$ satisfies (10.4), then we call it a *matched pair* of state maps.

We now associate with a QDF a dual one and relate their average nonnegativity and average positivity. Let $\Phi \in \mathbb{R}^{q\times q}_s[\zeta,\eta]$ and let $\Phi(\zeta,\eta) = M^T(\zeta)\Sigma_\Phi M(\eta)$ be a symmetric canonical factorization, with

$$\Sigma_\Phi = \begin{bmatrix} I_{r_+} & 0 \\ 0 & -I_{r_-} \end{bmatrix}.$$

Let us assume that $M \in \mathbb{R}^{r\times q}[\xi]$. Partition M conformably to Σ_Φ as

$$M = \begin{bmatrix} P \\ N \end{bmatrix}$$

so that Φ is written as

$$\Phi(\zeta,\eta) = P^T(\zeta)P(\eta) - N^T(\zeta)N(\eta). \tag{10.5}$$

Consider the dynamical system $\mathfrak{B} \in \mathfrak{L}^r$ with image representation

$$w = M\left(\frac{d}{dt}\right)\ell. \tag{10.6}$$

There are a number of integers associated with M that are of interest to us:

$$r_+ = \operatorname{rowdim}(P), \tag{10.7}$$

$$r_- = \operatorname{rowdim}(N), \tag{10.8}$$

$$m = \operatorname{rank}(M)\ . \tag{10.9}$$

The number r_+ corresponds to the number of positive squares in Q_Φ, r_- to the number of negative squares, while m equals the number of inputs in any input/output or input/state/output representation of $\mathfrak{B}$. Since it is defined by an image representation, $\mathfrak{B}$ is a controllable system and, as such, it admits a dual, $\mathfrak{B}^\perp \in \mathfrak{L}^r$. Let $R(\frac{d}{dt})w = 0$ be a minimal kernel representation of $\mathfrak{B}$. Then

$$v = R^T\left(-\frac{d}{dt}\right)\ell' \tag{10.10}$$

is an observable image representation for $\mathfrak{B}^\perp$. Let $\Phi'(\zeta,\eta) := R(-\zeta)\Sigma_\Phi R^T(-\eta)$. Note that the QDFs $Q_\Phi(\ell) = (M(\frac{d}{dt})\ell)^T\Sigma_\Phi M(\frac{d}{dt})\ell$ and

$$Q_{\Phi'}(\ell') = \left(R^T\left(-\frac{d}{dt}\right)\ell'\right)^T \Sigma_\Phi R^T\left(-\frac{d}{dt}\right)\ell'$$

are in a sense also dual. Their positivity properties are very much related, as shown in the following theorem.

THEOREM 10.2. *Assume that $r_+ = m$. Then*

(i) $\int Q_\Phi \geq 0 \Leftrightarrow \int Q_{\Phi'} \leq 0$,
(ii) $\int Q_\Phi > 0 \Leftrightarrow \int Q_{\Phi'} < 0$,
(iii) (*assume* Φ *is observable*) $\int Q_\Phi \gg 0 \Leftrightarrow \int Q_{\Phi'} \ll 0$,
(iv) *Let* (X, Z) *be a matched pair of minimal state maps for* $\mathfrak{B}$ *and* $\mathfrak{B}^\perp$. *Assume that* $\int Q_\Phi \geq 0$, *and let* Ψ *define a storage function for* Φ. *By Theorem* 5.5, Ψ *is a state function, i.e., there exists a real symmetric matrix* K *such that*

$$Q_\Psi(\ell) = \left(X\left(\frac{d}{dt}\right)\ell\right)^T KX\left(\frac{d}{dt}\right)\ell. \tag{10.11}$$

Assume that K *is nonsingular. Then*

$$Q_{\Psi'}(\ell') = -\left(Z\left(\frac{d}{dt}\right)\ell'\right)^T K^{-1}Z\left(\frac{d}{dt}\right)\ell' \tag{10.12}$$

is a storage function for $-Q_{\Phi'}$.
(v) (assume Φ *is observable*) $\int_t Q_\Phi \gg 0 \Leftrightarrow \int_t Q_{\Phi'} \ll 0$. *Here* $\int_t Q_{\Phi'} \ll 0$ *is defined as the property that there exists* $\epsilon > 0$ *such that* $\int_0^\infty Q_{\Phi'}(w)dt \leq -\epsilon \int_0^\infty Q_{\Phi'|}(w)dt$ *for all* $w \in \mathfrak{D}(\mathbb{R}, \mathbb{R}^q)$ *(i.e., half-line positivity over the* positive *half-line).*

Proof. For the proof, see the appendix.

We close this section by pointing out that it is of interest to generalize the notion of duality by using, instead of the usual inner product, an inner product that is itself induced by a QDF. These ramifications are a matter of future research.

11. Conclusions. In this paper we studied two-variable polynomial matrices and their role in a number of problems in linear system theory. The basic premise set forward is the following. Dynamic models lead naturally to the study of one-variable polynomial matrices. By substituting the time derivative for the indeterminate, and by letting the resulting differential operator act on a variable, one arrives at a dynamical system, which may then be in kernel or in image representation. The study of quadratic functionals in a variable and its derivative, on the other hand, leads to two-variable polynomial matrices. Important instances where dynamical systems occur in conjunction with functionals are, for example, Lyapunov theory, the theory of dissipative systems, and LQ and H_∞ control. We developed the former two applications in the present paper. The latter two will be discussed elsewhere.

Appendix.

Proof of Theorem 3.1. We prove the equivalence of the two statements in (1) at the end of the proof and proceed with the first statement by running the circle (1) $\Rightarrow$ (3) $\Rightarrow$ (2) $\Rightarrow$ (1). Assume that $\int L_\Phi = 0$. Then obviously $\int_{-\infty}^\infty L_\Phi(v, w)dt = 0$ for all $v \in \mathfrak{D}(\mathbb{R}, \mathbb{C}^{q_1})$ and $w \in \mathfrak{D}(\mathbb{R}, \mathbb{C}^{q_2})$, with $L_\Phi(v, w)$ in this case (for complex functions) defined by $\sum_{k,\ell}(\frac{d^k\bar{v}}{dt^k})^T\Phi_{k,\ell}(\frac{d^\ell w}{dt^\ell})$. Then

$$\int_{-\infty}^\infty \hat{v}^T(-i\omega)\Phi(-i\omega, i\omega)\hat{w}(i\omega)d\omega = 0$$

for all $\hat{v} \in L_2(\mathbb{C}, \mathbb{C}^{q_1})$, $\hat{w} \in L_2(\mathbb{C}, \mathbb{C}^{q_2})$ that are Fourier transforms of $v \in \mathfrak{D}(\mathbb{R}, \mathbb{C}^{q_1})$, and $w \in \mathfrak{D}(\mathbb{R}, \mathbb{C}^{q_2})$. This implies that $\partial\Phi = 0$. Assume to the contrary that there exist $\omega_0 \in \mathbb{R}$, $a \in \mathbb{C}^{q_1}$, $b \in \mathbb{C}^{q_2}$ such that $\bar{a}^T\Phi(-i\omega_0, i\omega_0)b \neq 0$. Define $v_N \in \mathfrak{D}(\mathbb{R}, \mathbb{C}^{q_1})$

for $N = 1, 2, \ldots$, by

$$v_N(t) = \begin{cases} e^{i\omega_0 t}a & |t| \leq \frac{2\pi N}{\omega_0}, \\ \tilde{v}(t + \frac{2\pi N}{\omega_0}) & t < -\frac{2\pi N}{\omega_0}, \\ \tilde{v}(t - \frac{2\pi N}{\omega_0}) & t > \frac{2\pi N}{\omega_0}. \end{cases} \tag{A.1}$$

Define $w_N \in \mathfrak{D}(\mathbb{R}, \mathbb{C}^{q_2})$ analogously by replacing a by b. Note that $\tilde{v}$ and $\tilde{w}$ can be chosen independent of N, and obtain smoothness for all N: indeed, if v_1 is smooth, then by the periodic nature of v_N for $|t| \leq \frac{2\pi N}{\omega_0}$, v_N will also be smooth.

Next evaluate $\int_{-\infty}^{\infty} L_\Phi(v_N, w_N)dt$ and observe that this integral equals

$$\frac{4\pi N}{\omega_0}\bar{a}^T\Phi(-i\omega_0, i\omega_0)b + E$$

with E independent of N. It follows that $\int_{-\infty}^{\infty} L_\Phi(v_N, w_N)dt \neq 0$ for N sufficiently large. In order to obtain this for real-valued functions, consider the real and imaginary parts of v_N, w_N and the integrals. This establishes the contradiction. Hence (1) implies (3).

To prove (3) $\Rightarrow$ (2), view $\Phi(\zeta, \eta)$ as a one-variable polynomial in ζ and carry out the division by $\zeta + \eta$. This yields $\Phi(\zeta, \eta) = (\zeta + \eta)d(\zeta, \eta) + r(\zeta, \eta)$. Hence $\partial\Phi = 0$ implies $r = 0$. This yields (2).

To show that (2) $\Rightarrow$ (1), observe that $\int_{-\infty}^{\infty} L_\Phi(v, w)dt = \int_{-\infty}^{\infty} \frac{d}{dt}L_\Psi(v, w)dt$. The last term obviously vanishes since v and w have compact support.

To show the equivalence of the two statements in (1), observe that it follows trivially that $\int L_\Phi = 0$ implies path independence. Conversely, if $\partial\Phi = 0$ then, according to (3), there exists $\Psi \in \mathbb{R}^{q_1 \times q_2}[\zeta, \eta]$ such that $L_\Phi = \frac{d}{dt}L_\Psi$. Thus, for any pair of functions $v \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^{q_1})$ and $w \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^{q_2})$, and for any t_1 and t_2, we have

$$\int_{t_1}^{t_2} L_\Phi(v, w)dt = \int_{t_1}^{t_2} \frac{d}{dt}L_\Psi(v, w)dt = L_\Psi(v, w)(t_2) - L_\Psi(v, w)(t_1).$$

Hence the integral depends only on the values taken on by v and w and their derivatives at the endpoints t_1 and t_2. □

Proof of Proposition 3.2. This proposition is proven following the standard proofs used in behavioral theory: reduce the problem to the scalar case using the Smith form. Let $R = U\Delta V$ with U, V unimodular and Δ diagonal. Define $\mathfrak{B}' = V(\frac{d}{dt})\mathfrak{B}$. Then $\mathfrak{B}'$ has $\Delta(\frac{d}{dt})w = 0$ as kernel representation. To prove the proposition, note that the "if" parts are immediate.

To show the first "only if" part, we show that $D(\frac{d}{dt})\mathfrak{B} = 0$ implies that there exists F such that $D = FR$ or equivalently, with $D = D'V$, that $D'(\frac{d}{dt})\mathfrak{B}' = 0$ implies that there exists F' such that $D' = F'\Delta$. Let $\Delta = \mathrm{diag}(d, \Delta')$, let d' be the first column of D', and let w_1 be a solution of $d(\frac{d}{dt})w_1 = 0$. Since $\mathrm{col}[w_1, 0, \ldots, 0] \in \mathfrak{B}'$ and $D'(\frac{d}{dt})\mathfrak{B}' = 0$, it follows that $d'(\frac{d}{dt})w_1 = 0$. It is easily seen that $d(\frac{d}{dt})w_1 = 0$ implies $d'(\frac{d}{dt})w_1 = 0$ iff each element of the polynomial vector d' is a factor of d. Proceeding this way column by column yields $D' = F'\Delta$.

To show the second "only if" part, we prove first the analogous result for BLDFs. This states that with $\Phi \in \mathbb{R}^{q_1 \times q_2}[\zeta, \eta]$, the BLDF $L_\Phi(w_1, w_2) = 0$ for all $w_1 \in \mathfrak{B}_1$ and $w_2 \in \mathfrak{B}_2$ iff there exists F_1, F_2 such that

$$\Phi(\zeta, \eta) = R_1^T(\zeta)F_2(\zeta, \eta) + F_1(\zeta, \eta)R_2(\eta), \tag{A.2}$$

where R_1 and R_2 induce kernel representations of $\mathfrak{B}_1$ and $\mathfrak{B}_2$. The "if" part is once again obvious. To prove the "only if" part, consider first the following lemma, which proves the scalar case $q_1 = q_2 = 1$.

LEMMA A.1. *Let $r_1, r_2 \in \mathbb{R}[\xi]$ and $\Phi \in \mathbb{R}[\zeta, \eta]$. Let $\mathfrak{B}_m \in \mathfrak{L}^1$, $m = 1, 2$ be given in kernel representation by $r_m(\frac{d}{dt})w_m = 0$. Then $Q_\Phi(w_1, w_2) = 0$ for all $w_m \in \mathfrak{B}_m$ iff there exists $f_m \in \mathbb{R}[\zeta, \eta]$ such that*

$$\Phi(\zeta, \eta) = r_1(\zeta) f_2(\zeta, \eta) + f_1(\zeta, \eta) r_2(\eta). \tag{A.3}$$

Proof. The "if" part is obvious. To show the "only if" part, let r_1 have degree n_1 and r_2 have degree n_2, and assume that they are monic. Consider the term $\Phi_{k,\ell}\zeta^k\eta^\ell$ of $\Phi(\zeta, \eta)$. In the quadratic form $Q_\Phi(w_1, w_2)$ this term contributes $\Phi_{k,\ell}\frac{d^k w_1}{dt^k}\frac{d^\ell w_2}{dt^\ell}$. If w_1 satisfies $r_1(\frac{d}{dt})w_1 = 0$, and if $k \geq n_1$, then the contribution in $Q_\Phi(w_1, w_2)$ of $\Phi_{k,\ell}\zeta^k\eta^\ell$ is equivalent to that of $\Phi_{k,\ell}(\zeta^k - \zeta^{k-n_1}r_1(\zeta))\eta^\ell$. Proceeding analogously with the ℓ's and the other terms shows that there exists $\Phi'(\zeta, \eta) = \sum_{k,\ell}\Phi'_{k,\ell}\zeta^k\eta^\ell$ with $\Phi'_{k,\ell} = 0$ for $k \geq n_1$ or $\ell \geq n_2$ such that

$$\Phi(\zeta, \eta) = r_1(\zeta) f_2(\zeta, \eta) + f_1(\zeta, \eta) r_2(\eta) + \Phi'(\zeta, \eta). \tag{A.4}$$

Obviously $Q_{\Phi'}(w_1, w_2) = Q_\Phi(w_1, w_2)$ for $w_m \in \mathfrak{B}_m$. Therefore $Q_{\Phi'}(w_1, w_2) = 0$ for $w_m \in \mathfrak{B}_m$. Consider $Q_{\Phi'}(w_1, w_2)(0)$ and observe that this is a quadratic form in $w_1(0), \frac{dw_1}{dt}(0), \ldots, \frac{d^{n_1-1}w_1}{dt^{n_1-1}}(0)$ and $w_2(0), \frac{dw_2}{dt}(0), \ldots, \frac{d^{n_2-1}w_2}{dt^{n_2-1}}(0)$. These initial conditions can be chosen arbitrarily in the sense that for any values of $w_m(0), \frac{dw_m}{dt}(0), \ldots, \frac{d^{n_m-1}w_m}{dt^{n_m-1}}(0)$ there exist $w_m \in \mathfrak{B}_m$ having these initial values. It follows that $\Phi' = 0$. ∎

Now return to the proof of the case for general q_1, q_2. Bring R_1 and R_2 in Smith form, showing that it suffices to prove (A.2) for $R = \Delta_1$ and $R_2 = \Delta_2$ with Δ_1 and Δ_2 in Smith form. Let d_1 be the (k_1, k_1)th element of Δ_1 and d_2 the (k_2, k_2)th element of Δ_2. Examine (A.2) and observe that we need to show that the (k_1, k_2)th element of Φ, $\Phi_{k_1k_2}$, can be written as

$$\Phi_{k_1,k_2}(\zeta, \eta) = d_1(\zeta) f_2(\zeta, \eta) + f_1(\zeta, \eta) d_2(\eta) \tag{A.5}$$

whenever it holds that $d_1(\frac{d}{dt})v_1 = 0$ and $d_2(\frac{d}{dt})v_2 = 0$ implies that $L_{\Phi_{k_1k_2}}(v_1, v_2) = 0$. Now use the previous lemma.

In order to prove Proposition 3.2 for $\Phi \in \mathbb{R}_s^{n_1 \times n_2}[\zeta, \eta]$, use the *-operator on (A.2), and add.

The image representation part of Proposition 3.2 is proven analogously. ∎

Proof of Proposition 3.5. The proof follows exactly along the same lines as the proof of Proposition 3.2, and we can therefore be very brief. The Smith form once again implies that it suffices to prove the case $q_1 = q_2 = 1$. Denote a kernel representation of $\mathfrak{B}$ by $r(\frac{d}{dt})w = 0$. Using (A.4) with $r_1 = r_2 = r$ shows that $Q_\Phi \geq 0$ on $\mathfrak{B}$ iff $Q_{\Phi'} \geq 0$ on $\mathfrak{B}$. However, again by the arbitrariness of the initial conditions, ($Q_\Phi \geq 0$ on $\mathfrak{B}$) iff the matrix $\tilde{\Phi}'$ associated with Φ' is nonnegative definite. Part (i) of the proposition follows.

To show part (ii), factor Φ' (using $\tilde{\Phi}'$) as $\Phi'(\zeta, \eta) = D^T(\zeta)D(\eta)$ with $D \in \mathbb{R}^{\bullet \times 1}[\xi]$ having elements whose degree is less than that of r. It thus suffices to find conditions for $r(\frac{d}{dt})w = 0$ and $D(\frac{d}{dt})w = 0$ to imply $w = 0$. That, however, is exactly equivalent to the observability of the pair (r, D). ∎

Proof of Theorem 4.3. The "if" part is shown as follows. By Proposition 3.5 we know that $\overset{\bullet}{\Psi} \overset{\mathfrak{B}}{<} 0$ implies that $\overset{\bullet}{\Psi}(\zeta,\eta) \overset{\mathfrak{B}}{=} -D^T(\zeta)D(\eta)$ with $D \in \mathbb{R}^{\bullet\times q}[\xi]$ such that (R,D) is observable, with $R \in \mathbb{R}^{\bullet\times q}[\xi]$ a kernel representation of $\mathfrak{B}$. It also holds that

$$\frac{d}{dt}Q_\Psi(w) = Q_{\overset{\bullet}{\Psi}}(w). \tag{A.6}$$

Integrate this from 0 to T along a $w \in \mathfrak{B}$ and obtain

$$Q_\Psi(w)(T) - Q_\Psi(w)(0) = \int_0^T Q_{\overset{\bullet}{\Psi}}(w)dt = -\int_0^T \|D\left(\frac{d}{dt}\right)(w)\|^2 dt. \tag{A.7}$$

Using $\Psi \overset{\mathfrak{B}}{\geq} 0$, this yields

$$\int_0^T \|D\left(\frac{d}{dt}\right)(w)\|^2 dt \leq Q_\Psi(w)(0). \tag{A.8}$$

Therefore

$$\int_0^\infty \|D\left(\frac{d}{dt}\right)(w)\|^2 dt < \infty. \tag{A.9}$$

This implies the asymptotic stability of $\mathfrak{B}$. Assume that $ae^{\lambda t} \in \mathfrak{B}$, $a \neq 0$. Then $R(\lambda)a = 0$ and by (A.8) there must hold that either $D(\lambda)a = 0$ or $\Re e(\lambda) < 0$. (Note that we silently use the obvious fact that (A.8) also holds for the complexification of $\mathfrak{B}$.) However, by observability of (R,D), $R(\lambda)a = 0$ and $D(\lambda)a = 0$ imply $a = 0$. Hence all exponential solutions $ae^{\lambda t}$ of $R(\frac{d}{dt})w = 0$ must have $\Re e(\lambda) < 0$. It is well known from the theory of differential equations that this implies that all solutions approach zero as $t \to \infty$. The "only if" follows from the stronger Theorem 4.8 and will be proven then. □

Proof of Corollary 4.6. $\overset{\bullet}{\Psi}(\zeta,\eta) = (\zeta+\eta)\Psi_0$. In the case at hand, $R(\xi) = A - \xi I$. Using Proposition 3.2, $(\zeta+\eta)\Psi \overset{\mathfrak{B}}{=} A\Psi_0 + \Psi_0 A^T$. Finally, observe that observability of $(A-\xi I, \sqrt{\Delta_0})$ (as a pair of polynomial matrices) is equivalent to that of $(A, \sqrt{\Delta_0})$ (as a pair of matrices) which is equivalent to that of (A, Δ_0). □

Proof of Proposition 4.7. Examine formula (A.8) in the proof of Theorem 4.3. It implies $Q_\Psi(w)(0) \geq \int_0^\infty \|D(\frac{d}{dt})w\|^2 dt$. Therefore $Q_\Psi(w)(0) = 0$ implies $D(\frac{d}{dt})w = 0$. However, by observability of D, $D(\frac{d}{dt})w = 0$ in turn implies $w = 0$. □

Proof of Theorem 4.8. The proof is organized as follows. First, we prove that (4.3) is solvable; second, that if R is square (4.4), (4.5) gives all Sits solutions; third, that (4.7) yields $\overset{\bullet}{\Psi} \overset{\mathfrak{B}}{=} \Phi$; fourth, that $\overset{\bullet}{\Psi}_1 \overset{\mathfrak{B}}{=} \overset{\bullet}{\Psi}_2$ implies $\Psi_1 \overset{\mathfrak{B}}{=} \Psi_2$; fifth, that $\Phi \overset{\mathfrak{B}}{\leq} 0$ yields $\Psi \overset{\mathfrak{B}}{\geq} 0$; and sixth, that $\Phi \overset{\mathfrak{B}}{<} 0$ yields $\Psi \overset{\mathfrak{B}}{\gg} 0$.

(i) First put R in Smith form: let

$$R = U \begin{bmatrix} D \\ 0 \end{bmatrix} V,$$

with D diagonal and U, V unimodular. Observe that it suffices to prove (4.3) with $R = D$. The (k,ℓ)th component of the matrix equation (4.3) in the obvious notation takes the form

$$x_{\ell k}(-\xi)d_\ell(\xi) + d_k(-\xi)x_{k\ell}(\xi) = \Phi_{k\ell}(-\xi,\xi). \tag{A.10}$$

Since d_k and d_ℓ are Hurwitz, $d_\ell(\xi)$ and $d_k(-\xi)$ are coprime and hence, by Bezout, (A.10) has a solution. This then yields a solution of the matrix version.

(ii) Again use the Smith form. Obtain that the difference of two solutions must satisfy

$$x_{\ell k}(-\xi)d_\ell(\xi) + d_k(-\xi)x_{k\ell}(\xi) = 0. \tag{A.11}$$

Hence again using coprimeness of $d_\ell(\xi)$ and $d_k(-\xi)$, there exists a polynomial $f_{k\ell}$ such that $x_{k\ell}(\xi) = f_{k\ell}(\xi)d_\ell(\xi)$. This yields (4.4). To show (4.5), obtain

$$R^T(-\xi)(F(\xi) + F^T(-\xi))R(\xi) = 0. \tag{A.12}$$

If R is square and $\det(R) \neq 0$, (4.5) follows by pre- and postmultiplying by $(R^T(-\xi))^{-1}$ and $(R(\xi))^{-1}$.

(iii) This proof is obvious.

(iv) Let $w \in \mathfrak{B}$ and assume that $\overset{\bullet}{\Psi}_1 \overset{\mathfrak{B}}{=} \overset{\bullet}{\Psi}_2$, i.e., $\overset{\bullet}{\Delta} \overset{\mathfrak{B}}{=} 0$, where $\Delta = \Psi_1 - \Psi_2$. Then

$$\int_0^t Q_{\overset{\bullet}{\Delta}}(w)dt = \int_0^t \frac{d}{dt}Q_\Delta(w)dt = Q_\Delta(w)(t) - Q_\Delta(w)(0). \tag{A.13}$$

Since $Q_{\overset{\bullet}{\Delta}}(w) = 0$, asymptotic stability of $\mathfrak{B}$ implies $Q_\Delta(w)(t) \to 0$ as $t \to \infty$ and hence that $Q_\Delta(w)(0) = 0$. Therefore $\Delta \overset{\mathfrak{B}}{=} 0$.

(v) This follows immediately from (A.7), and Proposition 4.7 yields (vi). □

Proof of Proposition 4.9. Existence of both D' and Ψ' follows from the algorithm given in the statement of the proposition. To show uniqueness of D' observe that $D' \overset{\mathfrak{B}}{=} D''$, i.e., $D'' - D' = FR$, and $D'R^{-1}, D''R^{-1}$ strictly proper, implies $F = 0$, i.e., $D' = D''$. In the two-variable case assume $\Psi' \overset{\mathfrak{B}}{=} \Psi''$, i.e.,

$$\Psi'(\zeta,\eta) = \Psi''(\zeta,\eta) + F^T(\eta,\zeta)R(\eta) + R^T(\zeta)F(\zeta,\eta).$$

Thus,

$$(R^T(-\zeta))^{-1}(\Psi' - \Psi'')(\zeta,\eta)(R(\eta))^{-1} = (R^T(\zeta))^{-1}F^T(\eta,\zeta) + F(\zeta,\eta)(R(\eta))^{-1}$$

Strict properness again implies $F = 0$. □

Proof of Proposition 4.10. Let $\Psi \overset{\mathfrak{B}}{=} 0$. Then $\Psi(\zeta,\eta) = F^T(\eta,\zeta)R(\eta) + R^T(\zeta)F(\zeta,\eta)$. Pre- and postmultiply by $(R^T(\zeta))^{-1}$ and $(R(\eta))^{-1}$, respectively, and conclude that $F = 0$. The result follows. If $\Psi \overset{\mathfrak{B}}{\geq} 0$, use the same reasoning and Proposition 3.5 on $\Psi(\zeta,\eta) \overset{\mathfrak{B}}{=} D^T(\zeta)D(\eta)$ with D R-canonical. The case $\Psi \overset{\mathfrak{B}}{>} 0$ is similar. □

Proof of Theorem 4.12. We first show that (4.4) has an R-canonical solution. Let X be any solution. Factor XR^{-1} as $XR^{-1} = P + S$ with P polynomial and S strictly proper. First observe that it follows from (4.3) that $P(\xi) + P^T(-\xi) = 0$. Next, show that $X - PR$ is a canonical solution. Uniqueness of this R-canonical solution follows from (4.4).

Next, we show that (4.10) yields an R-canonical Ψ. Simply pre- and postmultiply by $(R^T(\zeta))^{-1}$ and $R^{-1}(\eta)$ and observe properness. Uniqueness follows from ($\overset{\bullet}{\Psi}_1 \overset{\mathfrak{B}}{=} \Phi$ and $\overset{\bullet}{\Psi}_2 \overset{\mathfrak{B}}{=} \Phi$) $\Longrightarrow$ ($\overset{\bullet}{\Psi}_1 \overset{\mathfrak{B}}{=} \overset{\bullet}{\Psi}_2$). Now apply Proposition 4.9.

The remaining statements follow from Proposition 4.10. □

Proof of Proposition 5.2. The proof of all three statements is analogous. Therefore we only give the proof of (i). To prove ($\Leftarrow$), let $w \in \mathfrak{D}(\mathbb{R}, \mathbb{R}^q)$ and let $\hat{w}$ be its Fourier transform. Observe, using Parseval's Theorem, that

$$\int_{-\infty}^{+\infty} Q_\Phi(w)dt = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \hat{w}(-i\omega)^T\Phi(-i\omega, i\omega)\hat{w}(i\omega)d\omega, \tag{A.14}$$

whence ($\Leftarrow$). To show the converse, as in the proof of Theorem 3.1, we silently switch from $\mathbb{R}^q$ as signal space to $\mathbb{C}^q$. Assume that there exists $a \in \mathbb{C}^q$ and $\omega_0 \in \mathbb{R}$ such that $\bar{a}^T\Phi(-i\omega_0, i\omega_0)a < 0$. Consider the function $w_N \in \mathfrak{D}(\mathbb{R}, \mathbb{C}^q)$ for $N = 1, 2, \ldots$, defined exactly as v_N was in the proof of Theorem 3.1. Next evaluate $\int_{-\infty}^{+\infty} Q_\Phi(w_N)dt$ and observe (using the idea in the proof of Theorem 3.1) that this integral can be made negative by taking N sufficiently large. □

Proof of Proposition 5.4. We will run the circle (3) $\Rightarrow$ (2) $\Rightarrow$ (1) $\Rightarrow$ (3). To see that (3) $\Rightarrow$ (2), assume that Δ is a dissipation function. Then $\Phi(-\xi, \xi) = \Delta(-\xi, \xi)$, by Theorem 3.1. Define

$$\Psi(\zeta, \eta) = \frac{\Phi(\zeta, \eta) - \Delta(\zeta, \eta)}{\zeta + \eta}. \tag{A.15}$$

Hence $\overset{\bullet}{\Psi} = \Phi - \Delta$. Use $\Delta \geq 0$ to conclude that Ψ is a storage function. To see that (2) $\Rightarrow$ (1), use $\overset{\bullet}{\Psi} \leq \Phi$ and Theorem 3.1 to conclude (1). To see that (1) $\Rightarrow$ (3), use Propositions 5.2 and 5.6 to construct a D such that $\Phi(-\xi, \xi) = D^T(-\xi)D(\xi)$. Observe that $\Delta(\zeta, \eta) := D^T(\zeta)D(\eta)$ defines a dissipation function. The one-one relation between Ψ and Δ is given by (A.15). □

Proof of Theorem 5.5. By (5.11) it suffices to consider minimal state representations obtained from a canonical factorization of Φ. Let

$$v = M = \left(\frac{d}{dt}\right) w$$

be obtained from such a factorization, and let $x = X(\frac{d}{dt})w$ be a minimal state. There exists a permutation matrix P such that

$$PM = \begin{bmatrix} U \\ Y \end{bmatrix},$$

with $\det(U) \neq 0$ and such that YU^{-1} is a matrix of proper rational functions. Denote $u = U(\frac{d}{dt})w$. Consider $f = F(\frac{d}{dt})w$, where F is an arbitrary polynomial matrix. Then (see section 2) f is a state function, (i.e., there exists a matrix K such that $f = Kx$) iff FU^{-1} is strictly proper and a state/input function (i.e., there exists matrices L, J such that $f = Kx + Ju$) iff FU^{-1} is proper.

We first prove the second part of the theorem, i.e., that every dissipation function is a state/supply function. Let $\Delta(\zeta, \eta) = D^T(\zeta)D(\eta)$ be a dissipation function. Then

$$M^T(-\xi)\Sigma_\Phi M(\xi) = D^T(-\xi)D(\xi). \tag{A.16}$$

Pre- and postmultiply by U^{-1}, to obtain

$$(M(-\xi)U^{-1}(-\xi))^T\Sigma_\Phi M(\xi)U^{-1}(\xi) = (D(-\xi)U^{-1}(-\xi))^T D(\xi)U^{-1}(\xi). \tag{A.17}$$

Since the left-hand side is proper, so is the right-hand side. This obviously implies that $D(\xi)U^{-1}(\xi)$ is proper. Hence $D(\frac{d}{dt})w$ is a state/input function and equivalently,

$D(\xi) = KX(\xi) + JU(\xi)$ for suitable constant matrices K, J. From this it is readily seen that there exists a matrix E such that (5.13) holds.

We now prove the first part of the theorem, i.e., that every storage function is a state function. Let $\Psi(\zeta, \eta)$ define a storage function for Φ. Let $D(\xi)$ be such that

$$(\zeta + \eta)\Psi(\zeta, \eta) = M^T(\zeta)\Sigma_\Phi M(\eta) - D^T(\zeta)D(\eta).$$

By redefining

$$\tilde{M}^T(\zeta)\tilde{\Sigma}_\Phi \tilde{M}(\eta) = \begin{bmatrix} M(\zeta) \\ D(\zeta) \end{bmatrix}^T \begin{bmatrix} \Sigma_\Phi & 0 \\ 0 & -I \end{bmatrix} \begin{bmatrix} M(\eta) \\ D(\eta) \end{bmatrix},$$

and observing that a minimal state for the system with image representation $v = M(\frac{d}{dt})w$ is also a minimal state for the system with image representation

$$v = \begin{bmatrix} M(\frac{d}{dt}) \\ D(\frac{d}{dt}) \end{bmatrix} w,$$

it suffices to prove the claim in the lossless case, i.e., when $\overset{\bullet}{\Psi} = \Phi$.

Assume that $\Phi(\zeta, \eta) = M^T(\zeta)\Sigma_\Phi M(\eta)$ is a symmetric canonical factorization of Φ and that $\Psi(\zeta, \eta) = N^T(\zeta)\Sigma_\Psi N(\eta)$ is a symmetric canonical factorization of Ψ. Postmultiply the identity $\overset{\bullet}{\Psi} = \Phi$ by $U^{-1}(\eta)$ to obtain

$$(\zeta + \eta)N^T(\zeta)\Sigma_\Psi N(\eta)U^{-1}(\eta) = M^T(\zeta)\Sigma_\Phi M(\eta)U^{-1}(\eta). \tag{A.18}$$

Assume that $L_k\eta^k$ is the term of degree k in the polynomial part of the matrix of rational functions $N(\eta)U^{-1}(\eta)$. Using that the right-hand side of (A.18) is proper in η, by equating powers of η yields $N^T(\zeta)\Sigma_\Psi L_k = 0$. Express $N(\zeta)$ as

$$N(\zeta) = \begin{bmatrix} N_0 & N_1 & \dots & N_L \end{bmatrix} \begin{bmatrix} I \\ I\zeta \\ \vdots \\ I\zeta^L \end{bmatrix}$$

and use (A.18) to obtain $\begin{bmatrix} N_0 & N_1 & \dots & N_L \end{bmatrix}^T \Sigma_\Psi L_k = 0$. Since the factorization $\Psi(\zeta, \eta) = N^T(\zeta)\Sigma_\Psi N(\eta)$ is canonical, $\begin{bmatrix} N_0 & N_1 & \dots & N_L \end{bmatrix}$ is surjective. Hence (A.18) yields $L_k = 0$. This shows that $N(\xi)U^{-1}(\xi)$ is strictly proper and hence that $N(\frac{d}{dt})w$ is a state function as desired. Thus there exists a constant matrix K such that $N(\xi) = KX(\xi)$. This shows that there exists a matrix P such that (5.12) holds. This completes the proof of the theorem. □

Proof of Theorem 5.7. We first prove the second part, the part regarding strong average positivity. In this case it follows from Proposition 5.3 that $\Phi(-i\omega, i\omega) > 0$ for all ω. Hence by Proposition 5.6, $\Phi(-\xi, \xi)$ has a Hurwitz and an anti-Hurwitz factorization. The associated storage functions, Ψ_+ and Ψ_-, satisfy

$$\frac{d}{dt}(Q_{\Psi_+}(w) - Q_{\Psi_-}(w)) = \|H\left(\frac{d}{dt}\right)w\|^2 - \|A\left(\frac{d}{dt}\right)w\|^2. \tag{A.19}$$

Let $x = X(\frac{d}{dt})w$ be a minimal state associated with a canonical factorization of Φ. By Theorem 5.5, there exist real symmetric matrices, say K_+ and K_-, such that for

all $w \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^q)$, we have

$$Q_{\Psi_+}(w) = X\left(\frac{d}{dt}w\right)^T K_+ X\left(\frac{d}{dt}\right) w,$$
$$Q_{\Psi_-}(w) = X\left(\frac{d}{dt}w\right)^T K_- X\left(\frac{d}{dt}\right) w.$$

Then if for all a there exists a solution w of $A(\frac{d}{dt})w = 0$ such that $(X(\frac{d}{dt})w)(0) = a$, we obtain, by integrating along this solution,

$$a^T K_+ a - a^T K_- a = Q_{\Psi_+}(w)(0) - Q_{\Psi_-}(w)(0) = \int_{-\infty}^{0} \|H\left(\frac{d}{dt}\right) w\|^2 dt, \tag{A.20}$$

whence $K_+ \geq K_-$, so $\Psi_+ \geq \Psi_-$.

The problem is that there may not be a solution of $A(\frac{d}{dt})w = 0$ for all a such that $(X(\frac{d}{dt})w)(0) = a$. In order to circumvent this difficulty we first prove the statements of the second part under the additional assumption that $\Phi \geq \epsilon|\Phi|$ for some $\epsilon > 0$, in addition to the assumption that $\Phi(-i\omega, i\omega) > 0$ for all ω. Next, we modify Φ to Φ_ϵ such that these conditions hold for $\epsilon > 0$, and, finally, take the limit for $\epsilon \downarrow 0$.

Assume that $\Phi(-i\omega, i\omega) > 0$ for all $\omega \in \mathbb{R}$ and $\Phi \geq \epsilon|\Phi|$ for some $\epsilon > 0$. The system (5.5) allows an I/O representation, in the sense that there exists a permutation matrix P such that

$$Pv = \begin{bmatrix} U(\frac{d}{dt}) \\ Y(\frac{d}{dt}) \end{bmatrix} w, \tag{A.21}$$

with $\det(U) \neq 0$ and $G := YU^{-1}$ proper. Let $u = U(\frac{d}{dt})w$, $y = Y(\frac{d}{dt})w$. There exist constant matrices A, B, C, and D such that u, x, and y are related by $\frac{dx}{dt} = Ax + Bu, y = Cx + Du$. Since AU^{-1} is biproper, $A(\frac{d}{dt})w$ is of the form $Fx + Lu$ with L nonsingular. Using $u = -L^{-1}Fx$ and $x(0) = a$ in these equations then results in a solution of $A(\frac{d}{dt})w = 0$. To show that AU^{-1} is indeed biproper, use Proposition 5.2 to obtain

$$\begin{bmatrix} U(-i\omega) \\ Y(-i\omega) \end{bmatrix}^T P\Sigma_\Phi P^T \begin{bmatrix} U(i\omega) \\ Y(i\omega) \end{bmatrix} =$$
$$A^T(-i\omega)A(i\omega) \geq \epsilon \begin{bmatrix} U(-i\omega) \\ Y(-i\omega) \end{bmatrix}^T \begin{bmatrix} U(i\omega) \\ Y(i\omega) \end{bmatrix}.$$

After pre- and postmultiplying by $(U^{-1}(-i\omega))^T$ and $U^{-1}(i\omega)$, respectively, we obtain that

$$\begin{bmatrix} I \\ G(-i\omega) \end{bmatrix}^T P\Sigma_\Phi P^T \begin{bmatrix} I \\ G(i\omega) \end{bmatrix} = ((AU^{-1})(-i\omega))^T (AU^{-1})(i\omega) \geq \epsilon I. \tag{A.22}$$

Since G is proper, AU^{-1} is proper, by the equality on the left. The inequality on the right gives biproperness.

Consider a general Φ and define Φ_ϵ by $\Phi_\epsilon = \Phi + \epsilon|\Phi| + \epsilon I$. Then Φ_ϵ satisfies the above conditions and hence there exists (in the obvious notation) Ψ_ϵ^- and Ψ_ϵ^+ such that $\Psi_\epsilon^- \leq \Psi_\epsilon \leq \Psi_\epsilon^+$. Observe that for $0 < \epsilon_1 \leq \epsilon_2$ there holds $\Phi_{\epsilon_1} \leq \Phi_{\epsilon_2}$ and deduce

from $\overset{\bullet}{\Psi}{}^+_{\epsilon_1} \leq \Phi_{\epsilon_1} \leq \Phi_{\epsilon_2}$ that $\Psi^+_{\epsilon_2} \geq \Psi^+_{\epsilon_1}$. Similarly, $\Psi^-_{\epsilon_2} \leq \Psi^-_{\epsilon_1}$. Consequently $\Psi^-_{\epsilon_2} \leq \Psi^-_{\epsilon_1} \leq \Psi^+_{\epsilon_1} \leq \Psi^+_{\epsilon_2}$. Prove (using for example the associated matrix representations) that this monotonicity implies the existence of $\lim_{\epsilon \downarrow 0} \Psi^-_\epsilon =: \Psi^-_0$ and $\lim_{\epsilon \downarrow 0} \Psi^+_\epsilon =: \Psi^+_0$.

We now prove that Ψ^-_0 and Ψ^+_0 satisfy $\overset{\bullet}{\Psi}{}^-_0 \leq \Phi$ and $\overset{\bullet}{\Psi}{}^+_0 \leq \Phi$, and subsequently that any storage function Ψ of Φ satisfies $\Psi^-_0 \leq \Psi \leq \Psi^+_0$. To prove the first part, observe that $\overset{\bullet}{\Psi}{}^-_\epsilon \leq \Phi_\epsilon$ and $\overset{\bullet}{\Psi}{}^+_\epsilon \leq \Phi_\epsilon$ for $\epsilon > 0$ and take the limit for $\epsilon \downarrow 0$. To prove the second part, assume that $\overset{\bullet}{\Psi} \leq \Phi$. Then $\overset{\bullet}{\Psi} \leq \Phi \leq \Phi_\epsilon$. Therefore $\Psi^-_\epsilon \leq \Psi \leq \Psi^+_\epsilon$. Now take the limit for $\epsilon \downarrow 0$.

We still have to prove the formulas (5.15) and (5.16) for the computation of Ψ_- and Ψ^+ for the case that we only have $\Phi(i\omega, -i\omega) > 0$ for all $\omega \in \mathbb{R}$ and not necessarily $\Phi \geq \epsilon|\Phi|$ for some $\epsilon > 0$. Let H_ϵ be a symmetric Hurwitz factor of $\Phi_\epsilon(-\xi, \xi)$: $\Phi_\epsilon(-\xi, \xi) = H^T_\epsilon(-\xi)H_\epsilon(\xi)$, as discussed in Proposition 5.6. In order to make it unique, normalize H_ϵ to $\sqrt{\Phi_\epsilon(0)}$. It holds that

$$H^T_\epsilon(\zeta)H_\epsilon(\eta) = \Phi_\epsilon(\zeta, \eta) - (\zeta + \eta)\Psi^-_\epsilon(\zeta, \eta).$$

Since $\Phi_\epsilon \to \Phi$ as $\epsilon \downarrow 0$ and $\Psi^-_\epsilon \to \Psi^-_0$ as $\epsilon \downarrow 0$, we also have that H_ϵ converges. Clearly the limit H_0 satisfies $\Phi(-\xi, \xi) = H^T_0(-\xi)H_0(\xi)$ and must be Hurwitz. The formula for Ψ^-_0 $(= \Psi^-)$ follows. The situation for Ψ^+ is treated analogously. □

Proof of Proposition 6.2: (i) Compute $\int^0_{-\infty} Q_\Phi(w)$ for $w(t) = e^{\lambda t}a$ with $\Re e(\lambda) > 0$ and $a \in \mathbb{C}^m$. This integral equals $\frac{\bar{a}^T \Phi(\bar{\lambda}, \lambda)a}{\bar{\lambda} + \lambda}$. This w is not of compact support, but an approximation argument can be used to complete the proof of (i). For $\Re e(\lambda) = 0$ the result follows from Proposition 5.2. (ii) is proven similarly. □

Proof of Theorem 6.3 : We prove that (3) ⇒ (2) ⇒ (1) ⇒ (3). That (3) ⇒ (2) is trivial. In order to see that (2) ⇒ (1), integrate $\frac{d}{dt}Q_\Psi(w) \leq Q_\Phi(w)$ from $-\infty$ to 0. We now prove that (1) ⇒ (3). Assume first that Φ satisfies the assumptions $\Phi(-i\omega, i\omega) > 0$ for all ω and $\Phi \geq \epsilon|\Phi|$ for some $\epsilon > 0$. By Theorem 5.7 we then have

$$\Psi_+(\zeta, \eta) = \frac{\Phi(\zeta, \eta) - A^T(\zeta)A(\eta)}{\zeta + \eta}. \tag{A.23}$$

This yields $\frac{d}{dt}Q_{\Psi_+}(w) = Q_\Phi(w) - \|A(\frac{d}{dt})w\|^2$ for all w. Let $x = X(\frac{d}{dt})w$ be a minimal state map of Φ. By Theorem 5.5, $Q_{\Psi_+}(w) = \|X(\frac{d}{dt})w\|^2_{K_+}$ for some real symmetric matrix K_+. Using this expression in (A.23) and integrating from $-\infty$ to 0 yields that, for all a such that $X(\frac{d}{dt})w(0) = a$ and $A(\frac{d}{dt})w = 0$, we have $a^T K_+ a = \int^0_{-\infty} Q_\Phi(w)dt$. This integral is ≥ 0, so we must have $a^T K_+ a \geq 0$ (actually, such w does not have compact support but, by an approximation argument, the integral cannot be < 0). As in the proof of Theorem 5.7, it can be shown that for any initial condition a such w exists. This proves that $\Psi_+ \geq 0$. Take a general Φ. As in the proof of Theorem 5.7, first replace Φ by Φ_ϵ. By applying the previous to Φ_ϵ, we can conclude that (in the obvious notation) $\Psi^\epsilon_+ \geq 0$. Then take the limit for $\epsilon \downarrow 0$. □

Proof of Theorem 6.4. We will first run the circle (1) ⇒ (2) ⇒ (3) ⇒ (7) ⇒ (4) ⇒ (1).

(1) ⇒(2). This was proven in Proposition 6.2.

(2) ⇒(3). We have $P^T(\bar{\lambda})P(\lambda) \geq N^T(\bar{\lambda})N(\lambda)$ for $\lambda \in \mathbb{C}$, $\Re e(\lambda) \geq 0$. Assume that NP^{-1} has a pole λ such that $\Re e(\lambda) \geq 0$. Then there exists a vector $v \neq 0$ such that $P(\lambda)v = 0$ while $N(\lambda)v \neq 0$. This, however, contradicts the above inequality.

(3) $\Rightarrow$(7). Let $Q_K(w) = \|X(\frac{d}{dt})w\|_K^2$ be a storage function. We want to show that $K > 0$. First we show that NP^{-1} is a proper rational matrix. We have $P^T(-i\omega)P(i\omega) - N^T(-i\omega)N(i\omega) \geq 0$. Define $G := NP^{-1}$. Then for all ω such that $P(i\omega)$ is nonsingular, we have $G^T(-i\omega)G(i\omega) \leq I$, which shows that G is proper.

There exist matrices A, B, C, and D such that all smooth x, w_1 and w_2 satisfying $\dot{x} = Ax + Bw_2$, $w_1 = Cx + Dw_2$ can be written as $x = X(\frac{d}{dt})w$, $w_1 = N(\frac{d}{dt})w$, and $w_2 = P(\frac{d}{dt})w$ for some $w \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^\bullet)$. Since NP^{-1} has no poles in $\Re e(\lambda) \geq 0$, the matrix A can be chosen such that its eigenvalues are in the open left half of the complex plane. Moreover, we may assume that the pair (C, A) is observable. Let $a \in \mathbb{R}^n$. Choose $w_2 = 0$, let x satisfy $\frac{d}{dt}x = Ax$, $x(0) = a$, and let $w_1 = Cx$. This shows that there exists $w \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^\bullet)$ such that $(X(\frac{d}{dt})w)(0) = a$ and $w_2 = P(\frac{d}{dt})w = 0$. Also, $X(\frac{d}{dt})w \in L_2[0, \infty)$ since A is a Hurwitz matrix. Since $w_1 = N(\frac{d}{dt})w = CX(\frac{d}{dt})w$, we also have that $N(\frac{d}{dt})w \in L_2[0, \infty)$. Thus we can integrate the dissipation inequality from 0 to ∞ to obtain

$$-\left\|\left(X\left(\frac{d}{dt}\right)w\right)(0)\right\|_K^2 \leq -\int_0^\infty \|N(\frac{d}{dt})w\|^2 dt. \tag{A.24}$$

This shows that $a^T K a \geq 0$. Assume that $a^T K a = 0$. Then we must have $w_1 = N(\frac{d}{dt})w = 0$. By observability of the pair (C, A) this implies that $a = 0$.

(7) $\Rightarrow$(4). Let Q_Ψ be any storage function. Since X is a state map, by Theorem 5.5 there exists a real symmetric matrix K such that $Q_\Psi(w) = \|X(\frac{d}{dt})w\|_K^2$ for all w. By assumption, K is positive definite.

(4) $\Rightarrow$(1). This was proven in Theorem 6.3.

The implications (7) $\Rightarrow$(5), (5) $\Rightarrow$(4), (7) $\Rightarrow$(6), and (6) $\Rightarrow$(4) are obvious.

Finally, if we assume observability, then the poles of NP^{-1} coincide with the singularities of the polynomial matrix P. This shows that, under this assumption, (3) and (3′) are equivalent. This completes the proof. □

Proof of Theorem 7.2. Consider the representation (3.10) of L_Φ. From the fact that the factorization is canonical, it is easily seen that the mappings $w_1 \mapsto (N(\frac{d}{dt})w_1)(0)$ and $w_2 \mapsto (M(\frac{d}{dt})w_2)(0)$ are surjective. Thus we have

$$L_\Phi(w_1, \bullet) = 0 \Leftrightarrow \left(N\left(\frac{d}{dt}\right)w_1\right)^T M\left(\frac{d}{dt}\right)w_2 = 0 \text{ for all } w_2 \Leftrightarrow N\left(\frac{d}{dt}\right)w_1 = 0.$$

Similarly, $L_\Phi(\bullet, w_2) = 0$ iff $M(\frac{d}{dt})w_2 = 0$. From this, the equivalence of (1) and (3) is immediate.

To prove (1) $\Rightarrow$ (2), assume that, for some $\lambda \in \mathbb{C}$, $a^T\Phi(\lambda, \xi) = 0$, where a is a complex vector. Define $w_1(t) := e^{\bar{\lambda}t}\bar{a}$. For any w_2 and for all t, we then have

$$L_\Phi(w_1, w_2)(t) = e^{\lambda t}\left(a^T\Phi\left(\lambda, \frac{d}{dt}\right)w_2\right)(t) = 0.$$

This implies $w_1 = 0$, so $a = 0$, which proves that the rows of $\Phi(\lambda, \xi)$ are linearly independent over $\mathbb{C}$. Similarly, we can prove that the columns of $\Phi(\xi, \lambda)$ are linearly independent.

Finally, we prove that (2) implies (3). Let $\lambda \in \mathbb{C}$ and put $M(\lambda)a = 0$ for some complex vector a. We want to prove that $a = 0$. We clearly get $N^T(\xi)M(\lambda)a = 0$ so $\Phi(\xi, \lambda)a = 0$. Since the columns of $\Phi(\xi, \lambda)$ are linearly independent over $\mathbb{C}$, this yields $a = 0$. Likewise we can prove that $N(\lambda)$ has full column rank for all λ. □

Proof of Proposition 8.2. Write (8.3) as

$$(1+\epsilon)\int_{-\infty}^{+\infty}\|N\left(\frac{d}{dt}\right)w\|^2dt \leq (1-\epsilon)\int_{-\infty}^{+\infty}\|P\left(\frac{d}{dt}\right)w\|^2dt. \tag{A.25}$$

Put $\alpha = \frac{1-\epsilon}{1+\epsilon}$ and conclude that (2) and (3) are equivalent (also in the noncanonical case). To show that (1) $\Longleftrightarrow$ (2), observe that $|\Phi|(\zeta,\eta) = M_c^T(\zeta)M_c(\eta)$ for the special symmetric canonical factorization of $\Phi(\zeta,\eta)$ corresponding to (3.15). Hence statement (1) of the theorem is actually statement (2) for this special canonical factorization of Φ. It thus suffices to prove that if (2) holds for one canonical factorization, then it holds for any. From matrix theory, it follows that two canonical factorizations

$$M_1^T(\zeta)\Sigma_\Phi M_1(\eta) = M_2^T(\zeta)\Sigma_\Phi M_2(\eta) \tag{A.26}$$

are related by $M_1(\xi) = SM_2(\xi)$, with S a nonsingular matrix. Hence (2) for M_2 implies

$$\begin{aligned}\int_{-\infty}^{+\infty}\|M_1\left(\frac{d}{dt}\right)w\|^2_{\Sigma_\Phi}dt &= \int_{-\infty}^{+\infty}\|M_2\left(\frac{d}{dt}\right)w\|^2_{\Sigma_\Phi}dt\\ &\geq \epsilon_2\int_{-\infty}^{+\infty}\|M_2\left(\frac{d}{dt}\right)w\|^2dt\\ &\geq \frac{\epsilon_2}{\|S\|^2}\int_{-\infty}^{+\infty}\|M_1\left(\frac{d}{dt}\right)w\|^2dt\end{aligned}$$

and (2) for M_1 follows. Obviously this proof can be reversed with M_1 playing the role of M_2. If M_2 comes from a noncanonical factorization, then S may not be nonsingular and the proof goes through (but cannot be reversed). □

Proof of Theorem 9.3. The proof is structured as follows. We first prove that (1) $\Leftrightarrow$ (2). Subsequently, we show that (1) $\Rightarrow$(3) and finally that (3) $\Rightarrow$(1).

(1) $\Rightarrow$(2). That $\int^t Q_\Phi \gg 0$ implies (2a) is obvious. In order to prove (2b) we need the following lemma. Recall that a quadratic state function Q_Ψ (or simply Ψ), $Q_\Psi(w) = \|X(\frac{d}{dt})w\|^2_K$, is called *positive definite* if $K > 0$.

LEMMA A.2. *Let $M \in \mathbb{R}^{\bullet\times q}[\xi]$ be observable. Then there exists a positive definite state function Ψ such that*

$$\frac{d}{dt}Q_\Psi(w) \leq \|M\left(\frac{d}{dt}\right)w\|^2. \tag{A.27}$$

Proof. We show that Ψ_+, the supremal storage function associated with $M^T(\zeta)M(\eta)$, fits the bill. By Theorem 5.5, Ψ_+ is a state function, say, $Q_{\Psi_+} = \|X(\frac{d}{dt})w\|^2_{K_+}$. Here we take X to be any minimal state map of M. Obviously $\Psi_+ \geq 0$, since $\Psi = 0$ satisfies (A.27) and $\Psi_+ \geq \Psi = 0$. In order to show that $K_+ > 0$, let $a \neq 0$ be arbitrary. We show that $a^TK_+a > 0$. Factor $M^T(-\xi)M(\xi) = A^T(-\xi)A(\xi)$, with $A(\xi)$ anti-Hurwitz. Then

$$\frac{d}{dt}Q_{\Psi_+}(w) = \|M\left(\frac{d}{dt}\right)w\|^2 - \|A\left(\frac{d}{dt}\right)w\|^2 \tag{A.28}$$

for all $w \in \mathfrak{C}^\infty(\mathbb{R},\mathbb{R}^q)$. As in the proof of Theorem 5.7, it is easily seen that there exists $w \neq 0$ such that $A(\frac{d}{dt})w = 0$ and $X(\frac{d}{dt})w(0) = a$ (show that AU^{-1} is biproper,

with $M = \mathrm{col}(U, Y)$ an I/O partitioning). For this w in (A.28) we obtain $a^T K_+ a = \int_{-\infty}^0 \|M(\frac{d}{dt})w\|^2 dt > 0$, where the strict inequality follows from the observability of M. □

We now return to the proof of (1) ⇒(2b) of Theorem 9.3. Let $\Phi(\zeta,\eta) = M^T(\zeta)\Sigma_\Phi M(\eta)$ be the symmetric canonical factorization such that $|\Phi|(\zeta,\eta) = M^T(\zeta)M(\eta)$ (i.e., the one obtained by factoring $\tilde{\Phi} = \tilde{U}^T\Lambda\tilde{U}$, with Λ the diagonal matrix consisting of the nonzero eigenvalues of $\tilde{\Phi}$, and putting $\tilde{M} := \sqrt{|\Lambda|}\tilde{U}$). By the above lemma there exists a positive definite state function $\tilde{\Psi}$ such that $\frac{d}{dt}Q_{\tilde{\Psi}}(w) \leq \|M(\frac{d}{dt})w\|^2$. Now, $\int^t Q_\Psi \gg 0$ implies that there exists $\epsilon > 0$ such that $\int^t Q_{\Phi_\epsilon} \geq 0$, where $\Phi_\epsilon := \Phi - \epsilon|\Phi|$. Let Ψ_+^ϵ be the supremal storage function associated with Φ_ϵ. Clearly $\dot{\Psi}_+^\epsilon \leq \Phi_\epsilon \leq \Phi$, so Ψ_+^ϵ is also a storage function for Φ. This immediately yields $\Psi_+^\epsilon \leq \Psi_+$. Consider the two-variable polynomial matrix $\Psi_+^\epsilon + \epsilon\tilde{\Psi}$. Clearly this defines a storage function for Φ as well, so $\Psi_+^\epsilon \leq \Psi_+^\epsilon + \epsilon\tilde{\Psi} \leq \Psi_+$. According to Theorem 6.3, $\Psi_+^\epsilon \geq 0$. Since $\tilde{\Psi}$ is a positive definite state function, this implies that Ψ_+ is a positive definite state function.

We show that (2) ⇒(1). Let $X \in \mathbb{R}^{n\times q}[\xi]$ define a minimal state map for the system $v = M(\frac{d}{dt})w$. By (2b) Ψ_+ is a positive definite state function. Hence there exists $K_+ = K_+^T > 0$ such that $Q_{\Psi_+}(w) = \|X(\frac{d}{dt})w\|^2_{K_+}$. Factor $M^T(-\xi)\Sigma_M M(\xi) = A^T(-\xi)A(\xi)$ with A anti-Hurwitz. Then we have

$$\int_{-\infty}^0 \|M\left(\frac{d}{dt}\right)\|^2_{\Sigma_\Phi} dt = \|X\left(\frac{d}{dt}\right)w(0)\|^2_{K_+} + \int_{-\infty}^0 \|A\left(\frac{d}{dt}\right)w\|^2 dt \tag{A.29}$$

for all $w \in \mathfrak{D}(\mathbb{R},\mathbb{R}^q)$. There exists a permutation matrix P such that $PM = \mathrm{col}(U, Y)$, with $\det(U) \neq 0$ and YU^{-1} proper. Write $u = U(\frac{d}{dt})w$, $y = Y(\frac{d}{dt})w$, and $x = X(\frac{d}{dt})w$, with w ranging over $\mathfrak{C}^\infty(\mathbb{R},\mathbb{R}^q)$. Write the associated input/state/output representation. Hence there are constant matrices A_1, B_1, C_1, and D_1 such that these u, y, and x are exactly those that are related by the equations

$$\dot{x} = A_1x + B_1u, \quad y = C_1x + D_1u. \tag{A.30}$$

As in the proof of Theorem 5.7, by strict positivity we have that AU^{-1} is biproper. Thus there exist constant matrices F and L, $\det(L) \neq 0$ such that $A(\frac{d}{dt})w = Fx + Lu$. Solving this equation for u and substituting the result in (A.30) yields that the relation between $a := A(\frac{d}{dt})w$, $v = P\mathrm{col}(u, y)$, and $x = X(\frac{d}{dt})w$ is given by linear equations of the form

$$\dot{x} = A_2x + B_2a \quad v = C_2x + D_2a \tag{A.31}$$

with (C_2, A_2) observable and where the eigenvalues of A_2 coincide with the singularities of the spectral factor $A(\xi)$. This shows that for given $a \in L_2((-\infty,0],\mathbb{R}^\bullet)$ and final condition $x(0) = x_0$, the corresponding v is in $L_2((-\infty,0],\mathbb{R}^\bullet)$. In other words (A.31) defines a bounded operator from $L_2((-\infty,0],\mathbb{R}^\bullet) \times \mathbb{R}^n$ to $L_2((-\infty,0],\mathbb{R}^\bullet)$, mapping (a, x_0) to v. Hence there exists a constants C_1 and C_2 such that

$$\int_{-\infty}^0 \|v\|^2 dt \leq C_1 \int_{-\infty}^0 \|a\|^2 dt + C_2\|x_0\|^2.$$

Since $K_+ > 0$, there exists $\epsilon > 0$ such that $\frac{1}{\epsilon}K_+ > C_2I$ and $\frac{1}{\epsilon} > C_1$. For this ϵ we have

$$\int_{-\infty}^0 \|v\|^2 dt \leq \frac{1}{\epsilon}\left(\int_{-\infty}^0 \|a\|^2 dt + x_0^T K_+ x_0\right)$$

which, by (A.29), is equivalent to

$$\int_{-\infty}^{0} \|M\left(\frac{d}{dt}\right)\|^2 dt \leq \frac{1}{\epsilon}\int_{-\infty}^{0} \|M\left(\frac{d}{dt}\right)\|^2_{\Sigma_\Phi} dt.$$

This shows that Φ is strictly half-line positive. Whence, (2) $\Rightarrow$(1).

Next we show that (1) $\Rightarrow$(3). We will only consider the semisimple case. That (1) $\Rightarrow$ (3a) follows from Proposition 5.2. To prove (3b), calculate $\int_{-\infty}^{0} Q_\Phi(a)dt$ for

$$a(t) = \sum_{k=1}^{n} \alpha_k e^{\lambda_k t} a_k$$

and obtain the result

$$\int_{-\infty}^{0} Q_\Phi(a)dt = \begin{bmatrix} \bar{\alpha}_1 \\ \bar{\alpha}_2 \\ \vdots \\ \bar{\alpha}_n \end{bmatrix}^T T_\Phi \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix}. \tag{A.32}$$

We know that, for some $\epsilon > 0$, $\int_{-\infty}^{0} Q_\Phi(w)dt \geq \epsilon \int_{-\infty}^{0} Q_{|\Phi|}(w)dt$ for all w of compact support. An approximation argument yields that this implies $\int_{-\infty}^{0} Q_\Phi(a)dt > 0$ for $a \neq 0$, equivalently for col $(\alpha_1, \alpha_2, \ldots, \alpha_n) \neq 0$. Hence, $T_\Phi > 0$.

Finally, we turn to (3) $\Rightarrow$(2). The implication (3) $\Rightarrow$(2a) follows from Proposition 5.2. To show that (3) $\Rightarrow$(2b), we show that $T_\Phi > 0$ implies that the supremal storage function Ψ_+ defines a positive definite state function. Let $\partial\Phi(\xi) = A^T(-\xi)A(\xi)$ be an anti-Hurwitz factorization. We claim that $\lambda_1, \lambda_2, \ldots, \lambda_n$ are exactly the singularities of $A(\xi)$, with associated vectors $a_1, a_2, \ldots, a_n$ in the kernel of $A(\lambda_k)$, $k = 1, 2, \ldots, n$. Indeed, if λ has $\Re e(\lambda) > 0$, then $A^T(-\lambda)$ is nonsingular. Hence $\Phi(-\lambda_k, \lambda_k)a_k = 0$ and $\Re e(\lambda_k) > 0$ implies $A(\lambda_k)a_k = 0$.

According to Theorem 5.7, it holds that

$$\frac{d}{dt}Q_{\Psi_+}(w) = Q_\Phi(w) - \|A\left(\frac{d}{dt}\right)w\|^2. \tag{A.33}$$

For any solution $w = \sum_{k=1}^{r} \alpha_k e^{\lambda_k t} a_k$ of $A(\frac{d}{dt})w = 0$, we thus have

$$Q_{\Psi_+}(w)(0) = \int_{-\infty}^{0} Q_\Phi(w)dt = \begin{bmatrix} \bar{\alpha}_1 \\ \bar{\alpha}_2 \\ \vdots \\ \bar{\alpha}_n \end{bmatrix}^T T_\Phi \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix}. \tag{A.34}$$

Also, there exists a real symmetric matrix K_+ such that

$$Q_{\Psi_+}(w) = \|X\left(\frac{d}{dt}\right)w\|^2_{K_+},$$

where $X(\frac{d}{dt})$ is a minimal state map. Since $T_\Phi > 0$, (A.34) implies that $K_+ > 0$. Indeed, let $a \neq 0$ be arbitrary. Let $w \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^q)$ be such that $A(\frac{d}{dt})w = 0$, say,

$w = \sum_{k=1}^{r} a_k e^{\lambda_k t}\alpha_k$, and $X(\frac{d}{dt})w(0) = a$. Such w exists by strict positivity (see the proof of Theorem 5.7). Thus we have

$$a^T K_+ a = \left\| X\left(\frac{d}{dt}\right) w(0) \right\|^2_{K_+} = \begin{bmatrix} \bar{\alpha}_1 \\ \bar{\alpha}_2 \\ \vdots \\ \bar{\alpha}_n \end{bmatrix}^T T_\Phi \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix} > 0.$$

This completes the proof of Theorem 9.3. ∎

Proof of Proposition 10.1. Let $R(\frac{d}{dt})w = 0$ and $w = M(\frac{d}{dt})\ell$ be, respectively, a kernel and an observable image representation of $\mathfrak{B}$. Then we have $R(\xi)M(\xi) = 0$. Furthermore, $v = R^T(-\frac{d}{dt})\ell'$is an image representation of $\mathfrak{B}^\perp$. Consider the BLDF $(R^T(-\frac{d}{dt})\ell')^T M(\frac{d}{dt})\ell$. Note that this is the BLDF associated with the two-variable polynomial matrix $R(-\zeta)M(\eta)$. By Theorem 3.1, there exists $\Psi(\zeta,\eta)$ such that $\frac{d}{dt}Q_\Psi(\ell',\ell) = (R^T(-\frac{d}{dt})\ell')^T M(\frac{d}{dt})\ell$, and by Theorem 5.5 $Q_\Psi(\ell',\ell)$ is a state function; in other words, if $X(\frac{d}{dt})$ and $\tilde{Z}(\frac{d}{dt})$ are minimal state maps of $\mathfrak{B}$ and $\mathfrak{B}^\perp$, then $Q_\Psi(\ell,\ell') = (\tilde{Z}(\frac{d}{dt})\ell')^T K X(\frac{d}{dt})\ell$ for some matrix K. The proposition follows by taking $Z := K^T\tilde{Z}$ if we can show that K is nonsingular. To show this, assume to the contrary that $Ka = 0$. Let $\tilde{w} \in \mathfrak{B} \cap \mathfrak{D}(\mathbb{R},\mathbb{R}^q)$ be a trajectory emanating at $t = 0$ from $x(0) = a$. It follows that

$$\int_0^\infty v^T \tilde{w} dt = 0$$

for all $v \in \mathfrak{B}^\perp \cap \mathfrak{D}(\mathbb{R},\mathbb{R}^q)$. Consider the function $\hat{w} : \mathbb{R} \to \mathbb{R}^q$ such that $w(0) = 0$ for $t \le 0$ and $\hat{w}(t) = \tilde{w}(t)$ for $t \ge 0$. Obviously it holds that

$$\int_{-\infty}^{+\infty} v^T \hat{w} dt = 0$$

for all $v \in \mathfrak{B}^\perp \cap \mathfrak{D}(\mathbb{R},\mathbb{R}^q)$. Therefore $\hat{w}$ belongs to the $L_2(\mathbb{R},\mathbb{R}^q)$ closure of $\mathfrak{B}$ (this is the one point in this paper where $\mathfrak{C}^\infty$ solutions are inadequate). Since $\hat{w}(t) = 0$ for $t \le 0$, it must hold that $x(0) = 0$. Hence $Ka = 0$ implies $a = 0$, yielding the result. ∎

In order to prove Theorem 10.2 we use the following lemma. Recall that the inertia of an $n \times n$ complex Hermitian matrix H is the triple (π_-, π_0, π_+), with π_- the number of negative eigenvalues, π_+ the number of positive eigenvalues, and $\pi_0 (= n - \pi_- - \pi_+)$ the multiplicity of the zero eigenvalue.

LEMMA A.3. *Let $\mathcal{L}$ be a linear subspace of $\mathbb{R}^n$. Consider the quadratic form $x^T Q x$ on $\mathbb{R}^n$ with $Q = Q^T$ nonsingular. Let the inertia of Q be $(\pi_-, 0, \pi_+)$ and assume that $\pi_+ = \dim(\mathcal{L})$. Then $a^T Q a > 0$ for all $0 \ne a \in \mathcal{L}$ iff $a^T Q^{-1} a < 0$ for $0 \ne a \in \mathcal{L}^\perp$, and $a^T Q a \ge 0$ for all $a \in \mathcal{L}$ iff $a^T Q^{-1} a \le 0$ for $a \in \mathcal{L}^\perp$.*

Proof. Let $\mathcal{L} = \ker(R) = \mathrm{im}(M)$ with R surjective and M injective. Then $\mathcal{L}^\perp = \mathrm{im}(R^T) = \ker(M^T)$. Furthermore, $a^T Q a > 0$ for $0 \ne a \in \mathcal{L}$ means $M^T Q M > 0$. Consider the relations

$$\begin{bmatrix} M^T \\ RQ^{-1} \end{bmatrix} Q \begin{bmatrix} M & Q^{-1}R^T \end{bmatrix} = \begin{bmatrix} M^T Q M & 0 \\ 0 & RQ^{-1}R^T \end{bmatrix},$$

$$\begin{bmatrix} M^T \\ R \end{bmatrix} Q \begin{bmatrix} M & Q^{-1}R^T \end{bmatrix} = \begin{bmatrix} M^T Q M & 0 \\ RQM & RR^T \end{bmatrix}.$$

The second relation shows that $[M \;\; Q^{-1}R^T]$ is nonsingular. The first shows that

$$\text{in}(M^TQM) + \text{in}(RQ^{-1}R^T) = \text{in}(Q).$$

Hence $M^TQM > 0$ implies $RQ^{-1}R^T < 0$. To get the $\geq$ case, replace Q by $Q + \epsilon I$ and let $\epsilon \downarrow 0$. □

Proof of Theorem 10.2. To prove (i) and (ii), combine Proposition 5.2 and Lemma A.3 in the following way. For $\omega \in \mathbb{R}$ fixed, define $\mathcal{L} := \text{im}(M(i\omega)) = \ker(R(i\omega))$. Define $Q := \Sigma_\Phi$. Note that $Q^{-1} = \Sigma_\Phi$ as well. Using that $M(i\omega)$ and $R^T(-i\omega)$ are injective, we get the equivalence

$$M^T(i\omega)\Sigma_\Phi M(i\omega) > 0 \;\; \Leftrightarrow \;\; R(i\omega)\Sigma_\Phi R^T(-i\omega) < 0$$

which yields statement (ii). Statement (i) follows from the second assertion of Lemma A.3, which yields the same equivalence with nonstrict inequalities.

We now prove (iii). Again this can be proven using Lemma A.3, this time with $Q = \Sigma_\Phi - \epsilon I$. We then get

$$M^T(i\omega)(\Sigma_\Phi - \epsilon I)M(i\omega) \geq 0 \;\; \Leftrightarrow \;\; R(i\omega)(\Sigma_\Phi - \epsilon I)^{-1}R^T(-i\omega) \leq 0.$$

Using the formula $\Sigma_\Phi + \epsilon I = (1 - \epsilon^2)(\Sigma_\Phi - \epsilon I)^{-1}$, the latter is equivalent with

$$R(i\omega)\Sigma_\Phi R^T(-i\omega) \leq -\epsilon R(i\omega)R^T(-i\omega).$$

This shows (iii).

In order to prove (iv), we need the following lemma.

LEMMA A.4. *Let (X, Z) be a matched pair of minimal state maps for $\mathfrak{B}$ and $\mathfrak{B}^\perp$. Define subspaces $\mathcal{L} \subset \mathbb{R}^{r+2n}$, $\mathcal{M} \subset \mathbb{R}^{r+2n}$ by*

$$\text{(A.35)}\; \mathcal{L} := \left\{ \begin{bmatrix} w \\ x \\ a \end{bmatrix} \in \mathbb{R}^{r+2n} \mid \exists \ell \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^\bullet) \;\; \begin{bmatrix} w \\ x \\ a \end{bmatrix} = \begin{bmatrix} M(\frac{d}{dt})\ell \\ X(\frac{d}{dt})\ell \\ \frac{d}{dt}X(\frac{d}{dt})\ell \end{bmatrix} (0) \right\},$$

$$\text{(A.36)}\; \mathcal{M} := \left\{ \begin{bmatrix} v \\ b \\ z \end{bmatrix} \in \mathbb{R}^{r+2n} \mid \exists \ell' \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^\bullet) \;\; \begin{bmatrix} v \\ b \\ z \end{bmatrix} = \begin{bmatrix} R^T(-\frac{d}{dt})\ell' \\ -\frac{d}{dt}Z(\frac{d}{dt})\ell' \\ -Z(\frac{d}{dt})\ell' \end{bmatrix} (0) \right\}.$$

Then $\dim(\mathcal{L}) = n + m$ *and* $\mathcal{L}^\perp = \mathcal{M}$.

Proof. There exists a permutation matrix P such that

$$PM = \begin{bmatrix} U \\ Y \end{bmatrix}$$

with $U \in \mathbb{R}^{m\times m}[\xi]$ and YU^{-1} a proper rational matrix. If we define $u = U(\frac{d}{dt})\ell$ and $y = Y(\frac{d}{dt})\ell$, then u has the usual properties of input and y has the usual properties of output of $\mathfrak{B}$. There exist matrices $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$, and $D \in \mathbb{R}^{p\times m}$ (with $p = r - m$, the number of outputs) such that $x = X(\frac{d}{dt})\ell$, $u = U(\frac{d}{dt})\ell$, and $y = Y(\frac{d}{dt})\ell$ are exactly related by $\frac{dx}{dt} = Ax + Bu$, $y = Cx + Du$. Thus for all $\ell \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^\bullet)$ we have:

$$\begin{bmatrix} P & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} M(\frac{d}{dt})\ell \\ X(\frac{d}{dt})\ell \\ \frac{d}{dt}X(\frac{d}{dt})\ell \end{bmatrix} (0) = \begin{bmatrix} 0 & I_m \\ C & D \\ I_n & 0 \\ A & B \end{bmatrix} \cdot \begin{bmatrix} X(\frac{d}{dt})\ell \\ U(\frac{d}{dt})\ell \end{bmatrix} (0).$$

This implies

$$\begin{bmatrix} P & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix} \mathcal{L} \subset \operatorname{im} \begin{bmatrix} 0 & I_m \\ C & D \\ I_n & 0 \\ A & B \end{bmatrix}.$$

Here, in fact, equality holds. Indeed, given col (x_0, u_0), take any $x \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^n)$ and $u \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^m)$ such that $x(0) = x_0$ and $u(0) = u_0$, and such that $\frac{dx}{dt} = Ax + Bu$, $y = Cx + Du$. There exists $\ell \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^\bullet)$ such that $x = X(\frac{d}{dt})\ell$, $u = U(\frac{d}{dt})\ell$. This shows that equality holds and that $\dim(\mathcal{L}) = n + m$.

We now prove that $\mathcal{L}^\perp = \mathcal{M}$. For all $\ell \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^\bullet)$ and $\ell' \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^\bullet)$ we have that (10.4) holds. By evaluating this for $t = 0$, we immediately obtain that $\mathcal{L} \perp \mathcal{M}$. Thus it suffices to show that $\dim \mathcal{L} = n + (r - m)$. This is, however, an immediate consequence of the fact that the number of inputs of $\mathfrak{B}^\perp$, $m(\mathfrak{B}^\perp)$ is equal to $r - m$. ∎

We now return to the proof of (iv) of Theorem 10.2. Assume that $\int Q_\Phi \geq 0$ and let $\Psi(\zeta, \eta) = X^T(\zeta) K X(\eta)$, with $K = K^T$ a storage function for Φ, i.e., $\overset{\bullet}{\Psi} \leq \Phi$. In terms of $w = M(\frac{d}{dt})\ell$, $x = X(\frac{d}{dt})\ell$, $\frac{dx}{dt} = \frac{d}{dt}X(\frac{d}{dt})\ell$ this inequality yields, in particular,

$$\begin{bmatrix} w(0) \\ x(0) \\ \frac{dx}{dt}(0) \end{bmatrix}^T \begin{bmatrix} \Sigma_\Phi & 0 & 0 \\ 0 & 0 & -K \\ 0 & -K & 0 \end{bmatrix} \begin{bmatrix} w(0) \\ x(0) \\ \frac{dx}{dt}(0) \end{bmatrix} \geq 0. \tag{A.37}$$

Denote the symmetric matrix in (3.16) by Q. Note that (A.37) says that $a^T Q a \geq 0$ for all $a \in \mathcal{L}$, with $\mathcal{L}$ defined by (A.35). Since $\dim(\mathcal{L}) = n + m = n + r_+$, which is exactly the number of positive eigenvalues of Q, it follows from Lemma A.3 that $a^T Q^{-1} a \leq 0$ for all $a \in \mathcal{L}^\perp = \mathcal{M}$. More explicitly,

$$a^T \begin{bmatrix} \Sigma_\Phi & 0 & 0 \\ 0 & 0 & -K^{-1} \\ 0 & -K^{-1} & 0 \end{bmatrix} a \leq 0 \tag{A.38}$$

for $a \in \mathcal{L}^\perp$. A typical element of $\mathcal{M}$ has the form

$$a = \begin{bmatrix} R^T(-\frac{d}{dt})\ell' \\ -\frac{d}{dt}Z(\frac{d}{dt})\ell' \\ -Z(\frac{d}{dt})\ell' \end{bmatrix}(t),$$

where $\ell' \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^\bullet)$. By letting $t \in \mathbb{R}$ be arbitrary, the inequality (A.38) yields exactly the dissipation inequality

$$\frac{d}{dt}\|Z\left(\frac{d}{dt}\right)\ell'\|^2_{-K^{-1}} \leq \|R^T\left(-\frac{d}{dt}\right)\ell'\|^2_{-\Sigma_\Phi}$$

which is the content of (4). To show (v), use (iv) and Theorem 9.3. ∎

REFERENCES

[1] B.D.O. Anderson, *Algebraic properties of minimal degree spectral factors*, Automatica, 9 (1973), pp. 491–500.

[2] B.D.O. Anderson and S. Vongpanitlerd, *Network Analysis and Synthesis*, Prentice–Hall, Englewood Cliffs, NJ, 1973.
[3] B.D.O. Anderson and E.I. Jury, *Generalized Bezoutian and Sylvester matrices in multivariable linear control*, IEEE Trans. Automat. Control, 21 (1976), pp. 551–556.
[4] D.Z. Arov, *Passive linear stationary dynamical systems*, Siberian Math. J., 20 (1979), pp. 149–162.
[5] J.A. Ball, J.W. Helton, and J. William, *Shift invariant subspaces, passivity, reproducing kernels and H_∞-optimization*, in Contributions to Operator Theory and its Applications, Oper. Theory Adv. Appl. 35, I. Gohberg, J.W. Helton, and L. Rodman, eds., Birkhäuser, Basel, 1988, pp. 265–310.
[6] S. Barnett, *Polynomials and Linear Control Systems*, Marcel Dekker, New York, 1983.
[7] R.W. Brockett and J.L. Willems, *Frequency domain stability criteria, parts* 1 *and* 2, IEEE Trans. Automat. Control, 10 (1965), pp. 255–261; 401–413.
[8] R.W. Brockett, *Path integrals, Lyapunov functions, and quadratic minimization*, in Proc. 4th Allerton Conference on Circuit and System Theory, University of Illinois, Monticello, IL, 1966, pp. 685–698.
[9] F.M. Callier, *On polynomial spectral factorization by symmetric extraction*, IEEE Trans. Automat. Control, 30 (1985), pp. 453–464.
[10] W.A. Coppel, *Linear Systems*, Notes in Pure Mathematics 6, Australian National University, Canberra, 1972.
[11] P.A. Fuhrmann, *Algebraic methods in system theory*, in Mathematical System Theory, The Influence of R.E. Kalman, A.C. Antoulas, ed., Springer-Verlag, New York, 1991, pp. 233–265.
[12] D.J. Hill and P.J. Moylan, *Stability of nonlinear dissipative systems*, IEEE Trans. Automat. Control, 21 (1976), pp. 708–711.
[13] R.E. Kalman, *On the Hermite-Fujiwara theorem in stability theory*, Quart. Appl. Math., 23 (1965), pp. 279–282.
[14] R.E. Kalman, *Algebraic characterization of polynomials whose zeros lie in certain algebraic domains*, in Proc. Nat. Acad. of Sci., 64 (1969), pp. 818–823.
[15] V.L. Kharitonov, *Asymptotic stability of an equilibrium position of a family of systems of linear differential equations*, Differentsial'nye Uravneniya, 14 (1978), pp. 2086–2088.
[16] V. Kucera, *Discrete Linear Control, The Polynomial Approach*, John Wiley, Chichester, 1979.
[17] H. Kwakernaak, *The polynomial approach to H_∞ optimal regulation*, in H_∞-Control Theory, Lecture Notes in Math. 1496, E. Mosca and L. Pandolfi, eds., Springer-Verlag, Berlin, 1991, pp. 141–221.
[18] H. Kwakernaak and M. Sebek, *Polynomial J-spectral factorization*, IEEE Trans. Automat. Control, 39 (1994), pp. 315–328.
[19] G. Meinsma, *Frequency Domain Methods in H_∞ Control*, Ph.D. thesis, University of Twente, The Netherlands, 1993.
[20] R.J. Minnichelli, J.J. Anagnost, and C.A. Desoer, *An elementary proof of Kharitonov's stability theorem with extensions*, IEEE Trans. Automat. Control, 34 (1989), pp. 995–998.
[21] A.C.M. Ran and L. Rodman, *Factorization of matrix polynomials with symmetries*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 845–864.
[22] P. Rapisarda and J.C. Willems, *State maps for linear systems*, SIAM J. Control Optim., 35 (1997), pp. 1053–1091.
[23] J.W. Schumacher, *Transformations of linear systems under external equivalence*, Linear Algebra Appl., 102 (1988), pp. 1–33.
[24] H.L. Trentelman and J.C. Willems, *The dissipation inequality and the algebraic Riccati equation*, in The Riccati Equation, S. Bittanti, A.J. Laub, and J.C. Willems, eds., Springer-Verlag, New York, 1991, pp. 197–242.
[25] H.L. Trentelman and J.C. Willems, *H_∞ control in a behavioral context: The full information case*, IEEE Trans. Automat. Control, to appear.
[26] H.L. Trentelman and J.C. Willems, *Every storage function is a state function*, Systems Control Lett., 32 (1997), pp. 249–259.
[27] S. Weiland and J.C. Willems, *Dissipative systems in a behavioral context*, Math. Models Methods Appl. Sci., 1 (1991), pp. 1–25.
[28] J.C. Willems, *Least squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automat. Control, 16 (1971), pp. 621–634.
[29] J.C. Willems, *Dissipative dynamical systems–Part* I: *General theory, Part* II: *Linear systems with quadratic supply rates*, Arch. Rational Mech. Anal., 45 (1972), pp. 321–351; 352–393.
[30] J.C. Willems, *On the existence of a nonpositive solution to the Riccati equation*, IEEE Trans. Automat. Control, 19 (1974), pp. 592–593.

[31] J.C. Willems, *From time series to linear system. Part* I: *Finite dimensional linear time invariant systems; Part* II: *Exact modelling; Part* III: *Approximate modelling*, Automatica, 22 (1986), pp. 561–580; 675–694; 23 (1987), pp. 87–115.
[32] J.C. Willems, *Models for dynamics*, Dynamics Report, 2 (1989), pp. 171–269.
[33] J.C. Willems, *Paradigms and puzzles in the theory of dynamical systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 259–294.
[34] D.C. Youla and M. Saito, *Interpolation with positive-real functions*, J. Franklin Inst., 284 (1967), pp. 77–108.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 5, pp. 1750–1794, September 1998

011

TRUST-REGION INTERIOR-POINT SQP ALGORITHMS FOR A CLASS OF NONLINEAR PROGRAMMING PROBLEMS*

J. E. DENNIS†, MATTHIAS HEINKENSCHLOSS†, AND LUÍS N. VICENTE‡

Abstract. In this paper, a family of trust-region interior-point sequential quadratic programming (SQP) algorithms for the solution of a class of minimization problems with nonlinear equality constraints and simple bounds on some of the variables is described and analyzed. Such nonlinear programs arise, e.g., from the discretization of optimal control problems. The algorithms treat states and controls as independent variables. They are designed to take advantage of the structure of the problem. In particular they do not rely on matrix factorizations of the linearized constraints but use solutions of the linearized state equation and the adjoint equation. They are well suited for large scale problems arising from optimal control problems governed by partial differential equations.

The algorithms keep strict feasibility with respect to the bound constraints by using an affine scaling method proposed, for a different class of problems, by Coleman and Li [*SIAM J. Optim.,* 6 (1996), pp. 418–445] and they exploit trust-region techniques for equality-constrained optimization. Thus, they allow the computation of the steps using a variety of methods, including many iterative techniques.

Global convergence of these algorithms to a first-order Karush–Kuhn–Tucker (KKT) limit point is proved under very mild conditions on the trial steps. Under reasonable, but more stringent, conditions on the quadratic model and on the trial steps, the sequence of iterates generated by the algorithms is shown to have a limit point satisfying the second-order necessary KKT conditions. The local rate of convergence to a nondegenerate strict local minimizer is q-quadratic. The results given here include, as special cases, current results for only equality constraints and for only simple bounds.

Numerical results for the solution of an optimal control problem governed by a nonlinear heat equation are reported.

Key words. nonlinear programming, SQP methods, trust-region methods, interior-point algorithms, Dikin–Karmarkar ellipsoid, Coleman–Li affine scaling, simple bounds, optimal control problems

AMS subject classifications. 49M37, 90C06, 90C30

PII. S0036012995279031

1. Introduction.

In this paper we introduce and analyze a family of algorithms for the solution of an important class of minimization problems which often arise from the discretization of optimal control problems. These problems are specially structured nonlinear programming problems of the following form:

$$\begin{aligned} &\text{minimize } f(y,u) \\ &\text{subject to } C(y,u)=0, \\ &\qquad u\in\mathcal{B}=\{u:\ a\le u\le b\}, \end{aligned} \tag{1}$$

*Received by the editors November 22, 1995; accepted for publication (in revised form) December 18, 1997; published electronically June 25, 1998.

http://www.siam.org/journals/sicon/36-5/27903.html

†Department of Computational and Applied Mathematics, Rice University, Houston, TX 77005–1892 (dennis@rice.edu, heinken@rice.edu). The research of the first author was supported by DOE grant FG03-93ER25178, CRPC grant CCR-9120008, and AFOSR grant F49620-9310212. The research of the second author was supported by NSF grant DMS-9403699, DOE grant DE-FG03-95ER25257, and AFOSR grant F49620-93-1-0280.

‡Departamento de Matemática, Universidade de Coimbra, 3000 Coimbra, Portugal (lvicente@mat.uc.pt). This research was developed while the author was a graduate student at the Department of Computational and Applied Mathematics, Rice University, Houston, TX and was supported by INVOTAN (NATO scholarship), CCLA (Fulbright scholarship), FLAD (Portugal), and DOE grant FG03-93ER25178.

where $y \in \mathbb{R}^m$, $u \in \mathbb{R}^{n-m}$, $a \in (\mathbb{R} \cup \{-\infty\})^{n-m}$, and $b \in (\mathbb{R} \cup \{+\infty\})^{n-m}$. The functions $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ and $C : \mathbb{R}^n \longrightarrow \mathbb{R}^m$, $m < n$, are assumed to be at least continuously differentiable. As indicated above, minimization problems of the form (1) often arise from the discretization of optimal control problems. In this case y is the vector of state variables, u is the vector of control variables, and $C(y, u) = 0$ is the discretized state equation. Other applications, which might be viewed as special optimal control problems, include optimal design and parameter identification problems. Minimization problems (1) originating from optimal control problems governed by large systems of ordinary differential equations (ODEs) or partial differential equations (PDEs) are the targets of the algorithms in this paper.

Although there are algorithms available for the solution of nonlinear programming problems that are more general than (1), the family of algorithms presented in this paper is unique in the consequent use of structure inherent in many optimal control problems, the use of optimization techniques successfully applied in other contexts of nonlinear programming, and the rigorous theoretical justification.

Our algorithms are based on sequential quadratic programming (SQP) methods and use trust-region interior-point techniques to guarantee global convergence and to handle the bound constraints on the controls. SQP methods find a solution of the nonlinear programming problem (1) by solving a sequence of quadratic programming problems. It is known, see, e.g., [37], [38], that the structure of optimal control problems can be used to implement and analyze SQP methods. In particular, to implement SQP methods, it is sufficient to compute quantities of the form $C_y(y, u)v_y$, $C_y(y, u)^T v_y$, $C_u(y, u)v_u$, $C_u(y, u)^T v_y$ and to compute solutions of the linearized state equation $C_y(y, u)v_y = r$ and of the "adjoint equation" $C_y(y, u)^T v_y = r$. Here C_y and C_u denote the derivatives of C with respect to y and u. This is an important observation, because these are tasks that arise naturally in the context of optimal control problems. All of the early SQP algorithms, and many of the recent ones, rely on matrix factorizations, such as sparse LU decompositions, of the Jacobian $J(x)$ of $C(x)$. For the applications we have in mind this is not feasible. Often, the involved matrices are too large to perform such computations and very often these matrices are not even available in explicit form. On the other hand, matrix-vector multiplications $C_y(x)v_y$, $C_y(x)^T v_y$, $C_u(x)v_u$, $C_u(x)^T v_y$ can be performed, and efficient solvers for the linearized state equation $C_y(x)v_y = r$, and the adjoint equation $C_y(x)^T v_y = r$, are often available. For example, the partial Jacobian $C_y(x)$ in the application treated in section 11 has a block bidiagonal structure with diagonal matrices being tridiagonal. Thus, while the Jacobian is large, the solution of the linearized state equation or the adjoint equation can be done by block forward substitution or block backward substitution, respectively. In each substitution step, only a relatively small system with a tridiagonal system has to be solved. This is typical for many applications, in particular those in dynamical systems. Many SQP-based codes for optimal control problems governed by ODEs or DAEs (differential algebraic equations) exploit this structure efficiently in their numerical linear algebra. See, e.g., [1], [2], [42], [58], [62], and the references therein. For many applications, in particular those governed by PDEs, such factorizations of the Jacobian $J(x)$ of $C(x)$ are not feasible from a practical point of view, but solution techniques for $C_y(y, u)v_y = r$ and $C_y(y, u)^T v_y = r$ are available. This has motivated us to require only this information and to design a practicable algorithm that disjoins the particular equation solver from the optimization algorithm. In the presence of bound constraints, this task goes well beyond the mere replacement of matrix factorizations by black-box solvers. The implementation

of our algorithm is given in [16].

A purely local analysis for the case with no bound constraints has been given in [34], [36], [37], and [39]. However, we consider here the much more difficult issue of incorporating this entire structure into an algorithm that converges globally and handles bound constraints on the control variables u.

The global convergence of our algorithms is guaranteed by a trust-region strategy. In our framework the trust region serves a dual purpose. Besides ensuring global convergence, trust regions also introduce a regularization of the subproblems which is related to the Tikhonov regularization. For the solution of optimal control problems, the partitioning of the variables into states y and controls u motivates a partial decoupling of step components that leads to interesting alternatives for the choice of the trust region. In Sections 5.2.1 and 5.2.2 we will introduce a decoupled and a coupled trust-region approach. As indicated by the names, in the decoupled approach the trust region will act on step components separately. This allows a more efficient implementation of algorithms for the computation of these steps. However, for problems with ill-conditioned state equations, this decoupling does not give an accurate estimate of the size of the steps and might lead to poor performance. In this situation the coupled approach is better, and so we include both.

For the treatment of the bound constraints on u we use an affine scaling interior-point method introduced by Coleman and Li [13] for problems with simple bounds. Interior-point approaches are attractive for many optimization problems with a large number of bounds, including the structured problem (1). In our context, the affine scaling interior-point method is also of interest, because it does not interfere with the structure of the problem (1). To apply this method, no information in addition to that needed for the case without bound constraints is required from the user. This or similar interior-point approaches have recently also been used, e.g., in [6], [14], [43], [44], and [50]. The advantage of the approach in [13] is that the scaling matrix is determined by the distance of the iterates to the bounds and by the direction of the gradient. This dependence on the direction of the gradient is important for global convergence and its good effect can be seen in numerical examples; see, e.g., Figures 1 and 2.

Another important issue that is addressed in the implementations of the algorithms presented in this paper is the problem scaling inherent in optimal control problems. As we have pointed out, the problems we are primarily interested in are discretizations of optimal control problems governed by partial differential equations. The infinite-dimensional problem structure greatly influences the finite-dimensional problem. In our implementation, we take this into account by choosing scalar products for the states y, the controls u, and the duality pairing needed to represent $\lambda^T C(y, u)$, products that are discretizations of proper infinite-dimensional ones. It is beyond the scope of this paper to give a comprehensive theoretical study of these issues, but it is important to notice that the formulation of the algorithms discussed here fully supports the use of such scalar products without any changes. This is a great advantage. In some of our numerical experiments [11], [30] this improved the performance of our algorithms significantly, avoided artificial ill conditioning, and enhanced the quality of the solution computed for a given stopping tolerance. Moreover, our numerical experiments also indicate the mesh-independent behavior of our algorithms when this type of scaling is used.

We believe that the features and strong theoretical properties of these algorithms make them very attractive and powerful tools for the solution of optimal control

problems. They have been successfully applied to a boundary control problem (see section 11), a distributed nonlinear elliptic control problem [31], and optimal control problems arising in fluid flow [11], [30]. The software that produced these results is currently being beta-tested with the intent of electronic distribution [16].

Before we give an outline of this paper, it is worth discussing the relationship between the constrained minimization problem (1) and an equivalent reduced problem. Under the assumptions of the implicit function theorem it is possible to solve $C(y,u) = 0$ for y. This defines a smooth function $y(u)$ and allows us to reduce the minimization problem (1). The reduced problem is given by

$$\begin{aligned} &\text{minimize} && \hat{f}(u) \equiv f(y(u),u) \\ &\text{subject to} && u \in \mathcal{B} = \{u :\ a \le u \le b\}. \end{aligned} \tag{2}$$

This leads to the so-called *black-box* approach in which the nonlinear constraint $C(y,u) = 0$ is not visible to the optimizer. Its solution is part of the evaluation of the objective function $\hat{f}(u)$. The reduced problem can be solved by a gradient or a Newton-like method. For optimal control problems, many algorithms follow this approach. Often, projection techniques are used to handle the box constraints; see, e.g., [28], [51].

Recently, so-called *all-at-once* approaches that treat both y and u as independent variables have been proposed to solve optimal control problems; see, e.g., [1], [2], [4], [29], [32], [33], [34], [35], [36], [37], [39], [41], [42], [57], [58], [62].

Since *all-at-once approaches* move towards optimality and feasibility at the same time, they offer significant advantages. SQP methods are of particular interest. They do not require the possibly very expensive solution of the nonlinear state equation in every step, but as indicated above allow use of the structure of optimal control problems. In addition, SQP methods have proven to be very successful for the solution of other nonlinear programming problems. See, e.g., [5], [9], [23], [24], [40], [47], [48], [50], [56].

As outlined before, we use SQP-based methods for the solution of (1), i.e., the all-at-once approach. However, the reduced problem (2) is important to us for two reasons. Firstly, the relation between the full problem (1) and the reduced problem (2) gives important insight into the structure of (1) and allows us to extend techniques successfully applied to problems of the form (2). Secondly, black-box approaches are used very often to solve the problems we have in mind. We want to use this expertise in designing more efficient codes. Specifically, our consequent use of the structure of the optimal control problems leads to our family of trust-region interior-point SQP algorithms. These algorithms only require information that the user has to provide if a black-box approach is used with a Newton-like method for the solution of the nonlinear state equation and adjoint equation techniques for the computation of gradients. Thus, we combine the possible implementational advantages of a black-box approach with the generally more efficient all-at-once approach. It will be seen that in our algorithms the step s is decomposed into two components: $s = s^{\mathsf{n}} + s^{\mathsf{t}}$, where s^{n} is called the quasi-normal component and s^{t} is called the tangential component. The role of quasi-normal component s^{n} is to move towards feasibility. It is of the form $s^{\mathsf{n}} = ((s_y^{\mathsf{n}})^T\ 0^T)^T$, where s_y^{n} is essentially a Newton step for the solution of the nonlinear state equation $C(y,u) = 0$ for given u. For most problems of interest here, the computation of a "true" normal component is not practical. The tangential component s^{t} moves towards optimality. This component is in the null space of the linearized constraints and it is of the form $s^{\mathsf{t}} = ((-C_y(y,u)^{-1}C_u(y,u)s_u)^T\ s_u^T)^T$,

where s_u is essentially a Newton-like step for the reduced problem (2).

This paper is organized as follows: In section 2 we discuss the structure of the problem and motivate our SQP approach. We study the relationship between the all-at-once approach based on (1) and the black-box approach for (2) and the relationship between SQP methods for (1) and Newton methods for (2). For problems without box constraints, these connections are known, but for problems with box constraints, this will reveal useful new information. The first- and second-order Karush–Kuhn–Tucker (KKT) conditions for (1) are stated in section 3. We will state them in a nonstandard form that will lead to the scaling matrix used in the affine scaling interior-point approach. In section 4 we will discuss the application of Newton's method to the system of nonlinear equations arising from the first-order KKT conditions. This will be important for the derivation of our SQP method. In section 5 we describe our trust-region interior-point SQP algorithms. Sections 5.1 and 5.2 contain descriptions of the quasi-normal component and the tangential component. Using the derivations in sections 2 and 4, the connections between the quasi-normal component s^{n} and the Newton step for the solution of the nonlinear state equation $C(y,u) = 0$ for given u, and the relations between the tangential component s^{t} and Newton-like steps for the reduced problem (2), will be made precise. As noticed previously, the partial decoupling of the step components motivated by the partitioning of the variables into states y and controls u, and the roles of the decoupled and coupled trust-region approaches, will be exposed in sections 5.2.1 and 5.2.2. A complete statement of the trust-region interior-point SQP algorithms is given in section 5.4.

The convergence theory for these algorithms is given in sections 6, 7, 8, and 9. Section 6 contains some technical results. In section 7 we establish the existence of an accumulation point of the iterates which satisfies the first-order KKT conditions (Corollary 7.6). This result is established under very mild assumptions on the steps and on the Lagrange multipliers. It simultaneously extends the results presented recently by Coleman and Li [13] for simple bounds and those by Dennis, El-Alem, and Maciel [15] for equality constraints. Under additional conditions on the steps and the quadratic model, we show that the accumulation point satisfying the first-order necessary KKT conditions also solves the second-order necessary KKT conditions (Theorem 8.2). This latter result simultaneously extends those by Coleman and Li [13] for simple bounds and those by Dennis and Vicente [19] for equality constraints (see also [65]). Finally, we prove that if the sequence converges to a nondegenerate point satisfying the sufficient second-order KKT conditions, then the rate of convergence is q-quadratic (Corollary 9.4). Our analysis allows the application of a variety of methods for the computation of the step components s^{n} and s^{t}. In section 10 we discuss practical algorithms for the computation of trial steps and the multiplier estimates that are currently used in our implementation. Numerical results obtained with our implementation of these algorithms, called TRICE (trust-region interior-point SQP algorithms for optimal control and engineering design problems) [16], are reported in section 11. Section 12 contains conclusions and a discussion of future work.

We review the notation used in this paper. The vector x is given by

$$x = \begin{pmatrix} y \\ u \end{pmatrix}.$$

The Jacobian matrix of $C(x)$ is denoted by $J(x)$. We use subscripted indices to represent the evaluation of a function at a particular point of the sequences $\{x_k\}$ and

$\{\lambda_k\}$. For instance, f_k represents $f(x_k)$, and ℓ_k is the same as $\ell(x_k, \lambda_k)$. The vector and matrix norms used are the ℓ_2 norms, and I_l represents the identity matrix of order l. Also, $(z)_y$ and $(z)_u$ represent the subvectors of $z \in \mathbb{R}^n$ corresponding to the y and u components, respectively.

2. The structure of the minimization problem. The purpose of this section is to discuss some of the basic relationships between the problem (1) and its reduction (2). This will introduce fundamental quantities that are subsequently needed, and it will support our claim that the basic quantities needed to implement our SQP approach are already available if one uses a gradient or Newton-like method for the solution of the reduced problem (2).

The Lagrange function $\ell : \mathbb{R}^{n+m} \longrightarrow \mathbb{R}^n$ associated with the objective function $f(x)$ and the equality constraint $C(x) = (c_1(x), \ldots, c_m(x))^T = 0$ is given by

$$\ell(x, \lambda) = f(x) + \lambda^T C(x),$$

where $\lambda \in \mathbb{R}^m$ are the Lagrange multipliers.

The linearized constraints are given by $J(x)s = -C(x)$ or, equivalently, by

$$\begin{pmatrix} C_y(x) & C_u(x) \end{pmatrix} \begin{pmatrix} s_y \\ s_u \end{pmatrix} = -C(x). \tag{3}$$

We say that

$$s = \begin{pmatrix} s_y \\ s_u \end{pmatrix}, \quad s_y \in \mathbb{R}^m,\ s_u \in \mathbb{R}^{n-m}$$

satisfies the linearized state equation if it is a solution to (3). If $C_y(x)$ is invertible, the solutions of the linearized state equation are of the form

$$s = s^{\mathsf{n}} + W(x)s_u, \tag{4}$$

where

$$s^{\mathsf{n}} = \begin{pmatrix} -C_y(x)^{-1}C(x) \\ 0 \end{pmatrix} \tag{5}$$

is a particular solution and

$$W(x) = \begin{pmatrix} -C_y(x)^{-1}C_u(x) \\ I_{n-m} \end{pmatrix}$$

is a matrix whose columns form a basis for the null space $\mathcal{N}(J(x))$ of $J(x)$. One can see that matrix-vector multiplications of the form $W(x)^T s$ and $W(x)s_u$ involve only the solution of linear systems with the matrices $C_y(x)$ and $C_y(x)^T$. Moreover, the y component of the particular solution s^{n} is just the step that one would compute if one would apply Newton's method for the solution of the nonlinear equation $C(y, u) = 0$ for given u.

The point we want to convey in this section has nothing to do with the presence or absence of the bound constraints $a \leq u \leq b$. Therefore, for the remainder of this section, we consider the simpler case where there are no bound constraints, i.e., where $\mathcal{B} = \mathbb{R}^{n-m}$. If we solve (1) with $\mathcal{B} = \mathbb{R}^{n-m}$ by an SQP method, then the quadratic programming subproblem we have to solve at every iteration is of the form

$$\begin{aligned} \text{minimize} \quad & \nabla f(x)^T s + \tfrac{1}{2} s^T \nabla^2_{xx} \ell(x, \lambda)\ s \\ \text{subject to} \quad & C_y(x)s_y + C_u(x)s_u + C(x) = 0. \end{aligned} \tag{6}$$

If the reduced Hessian $W(x)^T\nabla^2_{xx}\ell(x,\lambda)W(x)$ is nonsingular, the solution of (6) is given by (4) with

$$s_u = -\Big(W(x)^T\nabla^2_{xx}\ell(x,\lambda)\,W(x)\Big)^{-1}W(x)^T\Big(\nabla f(x) + \nabla^2_{xx}\ell(x,\lambda)s^{\mathsf{n}}\Big). \tag{7}$$

In practice the Hessian $\nabla^2_{xx}\ell(x,\lambda)$ or the reduced Hessian $W(x)^T\nabla^2_{xx}\ell(x,\lambda)\,W(x)$ are often approximated using quasi-Newton updates. In the latter case, when an approximation to $\nabla^2_{xx}\ell(x,\lambda)$ is not available, then the "cross term" $W(x)^T\nabla^2_{xx}\ell(x,\lambda)s^{\mathsf{n}}$ has also to be approximated. This term can be approximated by zero, by finite differences, or by other quasi-Newton approximations; see, e.g., [3]. In the case where this cross term is approximated by zero, the right-hand side of the linear system (7) defining s_u can be written as

$$W(x)^T\nabla f(x) = -C_u(x)^TC_y(x)^{-T}\nabla_y f(x) + \nabla_u f(x).$$

Thus, if the Lagrange multiplier is computed by the adjoint formula

$$\lambda = -C_y(x)^{-T}\nabla_y f(x), \tag{8}$$

then

$$W(x)^T\nabla f(x) = C_u(x)^T\lambda + \nabla_u f(x) = \nabla_u\ell(x,\lambda).$$

Now we turn to the reduced problem with $\mathcal{B} = \mathbb{R}^{n-m}$. Suppose there exists an open set $\mathcal{U}$ such that for all $u \in \mathcal{U}$ there exists a solution y of $C(y,u) = 0$ and such that the matrix $C_y(x)$ is invertible for all $x = (y,u)$ with $u \in \mathcal{U}$ and $C(y,u) = 0$. Then the implicit function theorem guarantees the existence of a differentiable function

$$y : \mathcal{U} \to \mathbb{R}^m$$

defined by

$$C(y(u),u) = 0,$$

and the problem (1) can be reduced to (2). Since $y(\cdot)$ is differentiable, the function $\hat{f}$ is differentiable and its gradient is given by

$$\nabla\hat{f}(u) = W(y(u),u)^T\nabla f(y(u),u),$$

cf. [29]. Moreover, it can be shown that the Hessian of $\hat{f}$ is equal to the reduced Hessian

$$\nabla^2\hat{f}(u) = W(y(u),u)^T\nabla^2_{xx}\ell(y(u),u,\lambda)\;W(y(u),u),$$

provided that the Lagrange multiplier is computed from (8).

One can see that the gradient and the Hessian information in the SQP method for (1) and in the Newton method for (2) are the same if (y, u) solves $C(y, u) = 0$. Thus, if Newton-like methods are applied for the solution of (2), then one has all the available ingredients necessary to implement an SQP method for the solution of (1). The important difference, of course, is that in the SQP method we do not have to solve the nonlinear constraints $C(y, u) = 0$ at every iteration.

In these considerations we neglected the bound constraints $a \le u \le b$. These will be analyzed in the following sections. We already point out that these relationships between (1) and (2) are basically the same with or without the bound constraints.

3. Optimality conditions. A point x_* satisfies the first-order KKT conditions if there exist $\lambda_* \in \mathbb{R}^m$ and $\mu_*^a, \mu_*^b \in \mathbb{R}^{n-m}$ such that

$$
\begin{aligned}
&C(x_*) = 0,\\
&a \le u_* \le b,\\
&\begin{pmatrix} \nabla_y f(x_*) \\ \nabla_u f(x_*) \end{pmatrix} + \begin{pmatrix} C_y(x_*)^T \lambda_* \\ C_u(x_*)^T \lambda_* \end{pmatrix} - \begin{pmatrix} 0 \\ \mu_*^a \end{pmatrix} + \begin{pmatrix} 0 \\ \mu_*^b \end{pmatrix} = 0,\\
&((u_*)_i - a_i)\,(\mu_*^a)_i = (b_i - (u_*)_i)\,(\mu_*^b)_i = 0, \quad i = 1, \ldots, n-m, \quad \text{and}\\
&\mu_*^a \ge 0,\ \mu_*^b \ge 0.
\end{aligned}
$$

These KKT conditions are necessary conditions for x_* to be a local solution of (1). Note that the constraint qualifications are satisfied, since the invertibility of $C_y(x_*)$ and the form of the bound constraints imply the linear independence of the active constraints. Under the assumption of the invertibility of $C_y(x_*)$, we can rewrite the first-order KKT conditions:

$$
\begin{aligned}
&C(x_*) = 0,\\
&a \le u_* \le b,\\
&\lambda_* = -C_y(x_*)^{-T} \nabla_y f(x_*),\\
&a_i < (u_*)_i < b_i \implies (\nabla_u \ell(x_*, \lambda_*))_i = 0,\\
&(u_*)_i = a_i \implies (\nabla_u \ell(x_*, \lambda_*))_i \ge 0, \quad \text{and}\\
&(u_*)_i = b_i \implies (\nabla_u \ell(x_*, \lambda_*))_i \le 0.
\end{aligned}
$$

One can obtain a useful form of the first-order KKT conditions by noting that

$$
\begin{aligned}
\nabla_u \ell(x_*, \lambda_*) &= \nabla_u f(x_*) + C_u(x_*)^T \lambda_*\\
&= \nabla_u f(x_*) - C_u(x_*)^T C_y(x_*)^{-T} \nabla_y f(x_*)\\
&= W(x_*)^T \nabla f(x_*).
\end{aligned}
$$

In other words, $\nabla_u \ell(x_*, \lambda_*)$ is just the reduced gradient corresponding to the u variables. Hence x_* is a first-order KKT point if

$$
\begin{aligned}
&C(x_*) = 0,\\
&a \le u_* \le b,\\
&a_i < (u_*)_i < b_i \implies \big(W(x_*)^T \nabla f(x_*)\big)_i = 0,\\
&(u_*)_i = a_i \implies \big(W(x_*)^T \nabla f(x_*)\big)_i \ge 0, \quad \text{and}\\
&(u_*)_i = b_i \implies \big(W(x_*)^T \nabla f(x_*)\big)_i \le 0.
\end{aligned}
$$

Furthermore, x_* satisfies the second-order necessary KKT conditions if it satisfies the first-order KKT conditions and if the principal submatrix of the reduced Hessian

$$W(x_*)^T \nabla^2_{xx} \ell(x_*, \lambda_*) W(x_*)$$

corresponding to indices i such that $a_i < (u_*)_i < b_i$ is positive semidefinite, where the multipliers λ_* are given by $\lambda_* = -C_y(x_*)^{-T} \nabla_y f(x_*)$.

Now we adapt the idea of Coleman and Li [12] to this context and define $D(x) \in \mathbb{R}^{(n-m)\times(n-m)}$ to be the diagonal matrix with diagonal elements given by

$$\Big(D(x)\Big)_{ii} = \begin{cases} (b-u)_i^{\frac{1}{2}} & \text{if} \quad \big(W(x)^T \nabla f(x)\big)_i < 0 \text{ and } b_i < +\infty, \\ 1 & \text{if} \quad \big(W(x)^T \nabla f(x)\big)_i < 0 \text{ and } b_i = +\infty, \\ (u-a)_i^{\frac{1}{2}} & \text{if} \quad \big(W(x)^T \nabla f(x)\big)_i \geq 0 \text{ and } a_i > -\infty, \\ 1 & \text{if} \quad \big(W(x)^T \nabla f(x)\big)_i \geq 0 \text{ and } a_i = -\infty, \end{cases} \tag{9}$$

for $i = 1, \ldots, n-m$. In the following proposition we give the form of the first- and second-order necessary KKT conditions that we use in this paper. To us, they indicate the suitability of (9) as a scaling for (1). See also [13], [18], [64], and the remark below for further discussions on the choice of D as a scaling matrix.

PROPOSITION 3.1. *The point x_* satisfies the first-order KKT conditions if and only if*

$$\begin{aligned} &C(x_*) = 0, \quad a \leq u_* \leq b, \quad \textit{and} \\ &D(x_*) W(x_*)^T \nabla f(x_*) = 0. \end{aligned}$$

The point x_* satisfies the second-order necessary KKT conditions if and only if it satisfies the first-order KKT conditions and

$$D(x_*) W(x_*)^T \nabla^2_{xx} \ell(x_*, \lambda_*) W(x_*) D(x_*)$$

is positive semidefinite. The corresponding multiplier is given by

$$\lambda_* = -C_y(x_*)^{-T} \nabla_y f(x_*).$$

Remark 3.1. Proposition 3.1 remains valid for a larger class of diagonal matrices $D(x)$. The scalar 1 in the definition (9) of D can be replaced by any other positive scalar, and Proposition 3.1 also remains valid with $D(x)$ replaced by $D(x)^p$, $p > 0$. Most of our convergence results still hold true if $D(x)$ is replaced by $D(x)^p$, $p \geq 1$. See also Remark 8.1 and, for the case of simple bound constraints, see [18], [64]. However, the square roots in the definition of $D(x)$ will be necessary for the proof of local q-quadratic convergence of our algorithms.

The form of the sufficient optimality conditions used in this paper requires the definition of nondegeneracy or strict complementarity.

DEFINITION 3.2. *A point x in $\mathcal{B}$ is said to be nondegenerate if $(W(x)^T \nabla f(x))_i = 0$ implies $a_i < u_i < b_i$ for all $i \in \{1, \ldots, n-m\}$.*

We now define a diagonal $(n-m)\times(n-m)$ matrix $E(x)$ with diagonal elements given by

$$\Big(E(x)\Big)_{ii} = \begin{cases} \big|\big(W(x)^T\nabla f(x)\big)_i\big| & \text{if } \big(W(x)^T\nabla f(x)\big)_i < 0 \text{ and } b_i < +\infty, \text{ or} \\ & \text{if } \big(W(x)^T\nabla f(x)\big)_i > 0 \text{ and } a_i > -\infty, \\ 0 & \text{in all other cases,} \end{cases}$$

for $i = 1,\ldots,n-m$. The significance of this matrix will become clear in the next section when we apply Newton's method to the system of nonlinear equations arising from the first-order KKT conditions. From the definitions of $D(x)$ and $E(x)$ we have the following property.

PROPOSITION 3.3. *A nondegenerate point x_* satisfies the second-order sufficient KKT conditions if and only if it is a first-order KKT point and*

$$D(x_*)W(x_*)^T\nabla^2_{xx}\ell(x_*,\lambda_*)W(x_*)D(x_*) + E(x_*)$$

is positive definite, where $\lambda_* = -C_y(x_*)^{-T}\nabla_y f(x_*)$.

4. Newton's method. One way to motivate the algorithms described in this paper is to apply Newton's method to the system of nonlinear equations

$$\begin{aligned} &C(x) = 0, \\ &D(x)^2W(x)^T\nabla f(x) = 0, \end{aligned} \tag{10}$$

where x is strictly feasible with respect to the bounds on the variables u, i.e., $a < u < b$. This is related to Goodman's approach [27] for an orthogonal null-space basis and equality constraints. Although $D(x)^2$ is usually discontinuous at points where $(W(x)^T\nabla f(x))_i = 0$, the function $D(x)^2W(x)^T\nabla f(x)$ is continuous (but not differentiable) at such points. The application of Newton's method to this type of nonlinear systems has first been suggested by Coleman and Li [12] in the context of nonlinear minimization problems with simple bounds. They have shown that this type of nondifferentiability still allows the Newton process to achieve local q-quadratic convergence. In order to apply Newton's method we first need to compute some derivatives.

To calculate the Jacobian of the reduced gradient $W(x)^T\nabla f(x)$, we write

$$W(x)^T\nabla f(x) = \nabla_u f(x) + C_u(x)^T\lambda,$$

where λ is given by $C_y(x)^T\lambda = -\nabla_y f(x)$ and has derivatives

$$\begin{aligned} \tfrac{\partial\lambda}{\partial y} &= -C_y(x)^{-T}\left(\textstyle\sum_{i=1}^m \nabla^2_{yy}c_i(x)\lambda_i + \nabla^2_{yy}f(x)\right) \\ &= -C_y(x)^{-T}\nabla^2_{yy}\ell(x,\lambda), \\ \tfrac{\partial\lambda}{\partial u} &= -C_y(x)^{-T}\left(\textstyle\sum_{i=1}^m \nabla^2_{yu}c_i(x)\lambda_i + \nabla^2_{yu}f(x)\right) \\ &= -C_y(x)^{-T}\nabla^2_{yu}\ell(x,\lambda). \end{aligned}$$

This implies the equalities

$$\begin{aligned}\frac{\partial}{\partial y}\left(W(x)^T\nabla f(x)\right) &= C_u(x)^T\frac{\partial\lambda}{\partial y}+\nabla^2_{uy}f(x)+\textstyle\sum_{i=1}^m\nabla^2_{uy}c_i(x)\lambda_i\\ &= W(x)^T\begin{pmatrix}\nabla^2_{yy}\ell(x,\lambda)\\ \nabla^2_{uy}\ell(x,\lambda)\end{pmatrix},\\ \frac{\partial}{\partial u}\left(W(x)^T\nabla f(x)\right) &= C_u(x)^T\frac{\partial\lambda}{\partial u}+\nabla^2_{uu}f(x)+\textstyle\sum_{i=1}^m\nabla^2_{uu}c_i(x)\lambda_i\\ &= W(x)^T\begin{pmatrix}\nabla^2_{yu}\ell(x,\lambda)\\ \nabla^2_{uu}\ell(x,\lambda)\end{pmatrix},\end{aligned}$$

and we can conclude that

$$\frac{\partial}{\partial x}\left(W(x)^T\nabla f(x)\right)=W(x)^T\nabla^2_{xx}\ell(x,\lambda),$$

where $\lambda=-C_y(x)^{-T}\nabla_y f(x)$.

A linearization of (10) gives

$$C_y(x)s_y+C_u(x)s_u=-C(x), \tag{11}$$

$$\left(D(x)^2W(x)^T\nabla^2_{xx}\ell(x,\lambda)+\left[0\mid E(x)\right]\right)\begin{pmatrix}s_y\\ s_u\end{pmatrix}=-D(x)^2W(x)^T\nabla f(x), \tag{12}$$

where 0 denotes the $(n-m)\times m$ matrix with zero entries. Equation (11) is the linearized state equation. The diagonal elements of $E(x)$ are the product of the derivative of the diagonal elements of $D(x)^2$ and the components of the reduced gradient $W(x)^T\nabla f(x)$. The derivative of $(D(x)^2)_{ii}$ does not exist if $(W(x)^T\nabla f(x))_i=0$. In this case we set the corresponding quantities in the Jacobian to zero (see references [12], [13]). This gives the equation (12).

By using (4) we can rewrite the linear system (11)–(12) as

$$\begin{aligned}&s=s^{\mathsf{n}}+W(x)s_u,\\ &\quad\left(D(x)^2W(x)^T\nabla^2_{xx}\ell(x,\lambda)W(x)+E(x)\right)s_u\\ &=-D(x)^2W(x)^T\Big(\nabla^2_{xx}\ell(x,\lambda)s^{\mathsf{n}}+\nabla f(x)\Big).\end{aligned} \tag{13}$$

We define our Newton-like step as the solution of

$$\begin{aligned}&s=s^{\mathsf{n}}+W(x)s_u,\\ &\quad\left(\bar D(x)^2W(x)^T\nabla^2_{xx}\ell(x,\lambda)W(x)+E(x)\right)s_u\end{aligned} \tag{14}$$

$$=-\bar D(x)^2W(x)^T\Big(\nabla^2_{xx}\ell(x,\lambda)s^{\mathsf{n}}+\nabla f(x)\Big), \tag{15}$$

where $\bar D(x)\in\mathbb{R}^{(n-m)\times(n-m)}$ is the diagonal matrix defined by

(16)

$$\left(\bar D(x)\right)_{ii}=\begin{cases}(b-u)_i^{\frac12} & \text{if } \left(W(x)^T\left(\nabla^2_{xx}\ell(x,\lambda)s^{\mathsf{n}}+\nabla f(x)\right)\right)_i<0 \text{ and } b_i<+\infty,\\ 1 & \text{if } \left(W(x)^T\left(\nabla^2_{xx}\ell(x,\lambda)s^{\mathsf{n}}+\nabla f(x)\right)\right)_i<0 \text{ and } b_i=+\infty,\\ (u-a)_i^{\frac12} & \text{if } \left(W(x)^T\left(\nabla^2_{xx}\ell(x,\lambda)s^{\mathsf{n}}+\nabla f(x)\right)\right)_i\geq 0 \text{ and } a_i>-\infty,\\ 1 & \text{if } \left(W(x)^T\left(\nabla^2_{xx}\ell(x,\lambda)s^{\mathsf{n}}+\nabla f(x)\right)\right)_i\geq 0 \text{ and } a_i=-\infty,\end{cases}$$

for $i = 1, \ldots, n-m$. This change of the diagonal scaling matrix is based on the form of the right-hand side of (14). Unlike D, the scaling matrix $\bar{D}$ includes information from the cross term $\nabla^2_{xx}\ell(x,\lambda)s^{\mathsf{n}}$ and is therefore used as the scaling matrix for the computation of s_u in our algorithm, cf. (23). In the subsequent sections we will allow the replacement of the Hessian $\nabla^2_{xx}\ell(x,\lambda)$ to be a suitable matrix H.

If x is close to a nondegenerate point x_* satisfying the second-order sufficient KKT conditions, and if $W(x)^T\nabla^2_{xx}\ell(x,\lambda)s^{\mathsf{n}}$ is sufficiently small, a step s defined in this way is a Newton step on the following system of nonlinear equations:

$$\begin{aligned} &C(x) = 0, \\ &D(x)_u^2 W(x)^T \nabla f(x) = 0, \end{aligned} \tag{17}$$

where $D(x)_u$ depends on x_* as follows:

$$(D(x)_u)_{ii} = \begin{cases} 1 \text{ or } (b-u)_i^{\frac{1}{2}} \text{ or } (u-a)_i^{\frac{1}{2}} & \text{if } \left(W(x_*)^T\nabla f(x_*)\right)_i = 0, \\ (b-u)_i^{\frac{1}{2}} & \text{if } \left(W(x_*)^T\nabla f(x_*)\right)_i < 0, \\ (u-a)_i^{\frac{1}{2}} & \text{if } \left(W(x_*)^T\nabla f(x_*)\right)_i > 0, \end{cases}$$

for $i = 1, \ldots, n-m$. If $(W(x_*)^T\nabla f(x_*))_i = 0$, the ith diagonal element of $D(x)_u$ has to be chosen so that $\bar{D}(x)$ and $D(x)_u$ are the same matrix. Of course, this depends on the sign of $(W(x)^T(\nabla^2_{xx}\ell(x,\lambda)s^{\mathsf{n}} + \nabla f(x)))_i$. As Coleman and Li [12] pointed out, $D(x)_u$ is just for theoretical use since x_* is unknown. One can see that $D(x)_u^2 W(x)^T\nabla f(x)$ is continuously differentiable with Lipschitz continuous derivatives in an open neighborhood of x_*, that $D(x_*)_u^2 W(x_*)^T\nabla f(x_*) = 0$, and that the Jacobian of $D(x)_u^2 W(x)^T\nabla f(x)$ at x_* is nonsingular, for all choices of $D(x)_u$. These conditions are those typically required to get q-quadratic convergence for the Newton iteration (see [17, Thm. 5.2.1]). Thus the sequence of iterates generated by the Newton step (14)–(15) will converge q-quadratically to a nondegenerate point that satisfies the sufficient KKT conditions. The interior-point process damps the Newton step so that it stays strictly feasible, but this does not affect the rate of convergence. The details are provided in Corollary 9.4.

5. Trust-region interior-point SQP algorithms. The algorithms that we propose generate a sequence of iterates $\{x_k\}$ where

$$x_k = \begin{pmatrix} y_k \\ u_k \end{pmatrix}$$

and u_k is strictly feasible with respect to the bounds, i.e., $a < u_k < b$. At iteration k we are given x_k, and we need to compute a trial step s_k. If s_k is accepted, we set $x_{k+1} = x_k + s_k$. Otherwise we set x_{k+1} to x_k, reduce the trust-region radius, and compute a new trial step.

Following the application of Newton's method (14), each trial step s_k is decomposed as

$$s_k = s_k^{\mathsf{n}} + s_k^{\mathsf{t}} = s_k^{\mathsf{n}} + W_k(s_k)_u,$$

where s_k^{n} is called the quasi-normal component and s_k^{t} is the tangential component.

The role of s_k^{n} is to move towards feasibility. It will be seen that s_k^{n} is related to the Newton step for the solution of $C(y, u_k) = 0$ for fixed u_k. The role of s_k^{t} is to move towards optimality. The u component of s_k^{t} is related to the Newton step for the reduced problem (2). However, as made clear previously, we do not require feasibility with respect to the nonlinear equality constraints.

The global convergence is guaranteed by imposing an appropriate trust region on the step and monitoring the progress by a suitable merit function. The definition of the quasi-normal component, the tangential component, and the merit function, as well as the complete formulation of our algorithms, is the content of this section.

5.1. The quasi-normal component. Let δ_k be the trust radius at iteration k. The quasi-normal component s_k^{n} is related to the trust-region subproblem for the linearized constraints

$$\begin{aligned} &\text{minimize } \frac{1}{2}\|J_k s^{\mathsf{n}} + C_k\|^2 \\ &\text{subject to } \|s^{\mathsf{n}}\| \le \delta_k, \end{aligned}$$

and it is required to have the form

$$s_k^{\mathsf{n}} = \begin{pmatrix} (s_k^{\mathsf{n}})_y \\ 0 \end{pmatrix}. \tag{18}$$

Thus, the displacement along s_k^{n} is made only in the y variables, and as a consequence, x_k and $x_k + s_k^{\mathsf{n}}$ have the same u components. Since $(s_k^{\mathsf{n}})_u = 0$, the trust-region subproblem introduced above can be rewritten as

$$\text{minimize } \frac{1}{2}\|C_y(x_k)(s^{\mathsf{n}})_y + C_k\|^2 \tag{19}$$

$$\text{subject to } \|(s^{\mathsf{n}})_y\| \le \delta_k. \tag{20}$$

Thus, the quasi-normal component s_k^{n} is a trust-region globalization of the component s^{n} given in (5) of the Newton step (14). We do not have to solve (19)–(20) exactly; we only have to assume that the quasi-normal component satisfies the conditions

$$\|s_k^{\mathsf{n}}\| \le \kappa_1 \|C_k\| \tag{21}$$

and

$$\|C_k\|^2 - \|C_y(x_k)(s_k^{\mathsf{n}})_y + C_k\|^2 \ge \kappa_2 \|C_k\| \min\{\kappa_3 \|C_k\|, \delta_k\}, \tag{22}$$

where κ_1, κ_2, and κ_3 are positive constants independent of k. In section 10.1, we describe several ways of computing a quasi-normal component that satisfies the requirements (18), (21), and (22). Condition (21) tell us that the quasi-normal component is small close to feasible points. Condition (22) is just a weaker form of Cauchy decrease or simple decrease for the trust-region subproblem (19), (20).

5.2. The tangential component. The computation of the tangential component $(s_k)_u$ follows a trust-region globalization of the Newton step (15). Following Coleman and Li [13] we symmetrize (15) and get

$$\left(\bar{D}_k W_k^T H_k W_k \bar{D}_k + E_k\right) \bar{D}_k^{-1} s_u = -\bar{D}_k W_k^T \Big(H_k s_k^{\mathsf{n}} + \nabla f_k\Big),$$

where $E_k = E(x_k)$ and H_k denotes a symmetric approximation to the Hessian matrix $\nabla^2_{xx}\ell_k$. The scaling matrix $\bar{D}_k$ is equal to $\bar{D}(x_k)$ defined by (16) with $\nabla^2_{xx}\ell_k$ replaced by H_k. This suggests the change of variables $\hat{s}_u = \bar{D}_k^{-1} s_u$ and the consideration in the scaled space $\hat{s}_u$ of the trust-region subproblem

$$\begin{aligned} &\text{minimize } \left(\bar{D}_k W_k^T \left(H_k s_k^{\mathsf{n}} + \nabla f_k\right)\right)^T \hat{s}_u + \frac{1}{2}\hat{s}_u^T \left(\bar{D}_k W_k^T H_k W_k \bar{D}_k + E_k\right) \hat{s}_u \\ &\text{subject to } \|\hat{s}_u\| \le \delta_k. \end{aligned}$$

Now we can rewrite the previous subproblem in the unscaled space s_u as

$$\begin{aligned} &\text{minimize} \quad \left(W_k^T \left(H_k s_k^{\mathsf{n}} + \nabla f_k\right)\right)^T s_u + \tfrac{1}{2} s_u^T \left(W_k^T H_k W_k + E_k \bar{D}_k^{-2}\right) s_u \\ &\text{subject to} \quad \|\bar{D}_k^{-1} s_u\| \le \delta_k. \end{aligned} \tag{23}$$

Of course, we also have to require that the new iterate is in the interior of the box constraints. To ensure that $u_k + s_k$ is strictly feasible with respect to the box constraints, we choose $\sigma_k \in [\sigma, 1)$, $\sigma \in (0, 1)$ and compute s_u with $\sigma_k(a - u_k) \le s_u \le \sigma_k(b - u_k)$. However, one of the strengths of this trust-region approach is that we can allow for approximate solutions of this subproblem. In particular, it is not necessary to solve the full trust-region subproblem including the box constraints. For example, one can compute the solution of the trust-region subproblem without the box constraints and then scale the computed solution back so that the resulting damped s_u obeys $\sigma_k(a - u_k) \le s_u \le \sigma_k(b - u_k)$; see, e.g., section 5.2.4. We will show that under suitable assumptions this strategy guarantees global convergence and local q-quadratic convergence. Another way to compute an approximate u component of the step is to use a modified conjugate-gradient algorithm applied to the trust-region subproblem, without the box constraints, that is truncated if one of the bounds $\sigma_k(a - u_k) \le s_u \le \sigma_k(b - u_k)$ is violated. See section 10.2. More ways to compute the tangential component are possible. The conditions on the tangential component necessary to guarantee global convergence are stated in section 5.2.3.

We now introduce a quadratic model

$$q_k(s) = \ell_k + \nabla_x \ell_k{}^T s + \frac{1}{2} s^T H_k s$$

of $\ell(x_k + s, \lambda_k)$ about (x_k, λ_k). A trivial manipulation shows that

$$q_k(s_k^{\mathsf{n}} + W_k s_u) = q_k(s_k^{\mathsf{n}}) + \bar{g}_k^T s_u + \frac{1}{2} s_u{}^T W_k^T H_k W_k s_u, \tag{24}$$

with

$$\bar{g}_k = W_k^T \nabla q_k(s_k^{\mathsf{n}}) = W_k^T \Big(H_k s_k^{\mathsf{n}} + \nabla f_k\Big).$$

For convenience we define

$$\Psi_k(s_u) = q_k(s_k^{\mathsf{n}} + W_k s_u) + \frac{1}{2} s_u^T \left(E_k \bar{D}_k^{-2}\right) s_u. \tag{25}$$

5.2.1. The decoupled trust-region approach. We can restate the trust-region subproblem (23) as

$$\text{minimize } \Psi_k(s_u) \tag{26}$$

$$\text{subject to } \|\bar{D}_k^{-1} s_u\| \le \delta_k. \tag{27}$$

We refer to the approach based on this subproblem as the decoupled approach. In this decoupled approach, the trust-region constraint is of the form $\|\bar{D}_k^{-1}s_u\| \le \delta_k$ corresponding to the constraint $\|\hat{s}_u\| \le \delta_k$ in the scaled space. One can see from (20) and (27) that we are imposing the trust region separately on the y part of the quasi-normal component and on the u part of the tangential component. Moreover, if the cross term $W_k^T H_k s_k^{\mathsf{n}}$ is set to zero, then the trust-region subproblems for the quasi-normal component and for the tangential component are completely separated.

5.2.2. The coupled trust-region approach. The approach we present now forces the y and u parts of the tangential component $s_k^{\mathsf{t}} = W_k(s_k)_u$ to lie inside the trust region of radius δ_k. The reference trust-region subproblem is given by

$$\text{minimize } \Psi_k(s_u) \tag{28}$$

$$\text{subject to } \left\| \begin{pmatrix} -C_y(x_k)^{-1}C_u(x_k)s_u \\ \bar{D}_k^{-1}s_u \end{pmatrix} \right\| \le \delta_k. \tag{29}$$

In the case where there are no bounds on u, this trust-region constraint is of the form

$$\left\| \begin{pmatrix} -C_y(x_k)^{-1}C_u(x_k)s_u \\ s_u \end{pmatrix} \right\| = \|W_k s_u\| \le \delta_k.$$

As opposed to the decoupled case, one can see that the term $C_y(x_k)^{-1}C_u(x_k)s_u$ is present in the trust-region constraint (29). If W_k^+ denotes the Moore–Penrose pseudoinverse of W_k (see [25, sec. 5.5.4]), then

$$\frac{1}{\|W_k^+\|}\|s_u\| \le \|W_k s_u\| \le \|W_k\|\|s_u\|.$$

Thus, if the condition number $\kappa(W_k) = \|W_k^+\|\,\|W_k\|$ is small, then the decoupled and the coupled approach will generate similar iterates. In this case, the decoupled approach will be more efficient since it uses fewer linear system solvers with the system matrix $C_y(x_k)$. See section 10.2. However, if $\kappa(W_k)$ is large, e.g., if $C_y(x_k)$ is ill conditioned, then the coupled approach will use the true size of the tangential component, whereas the decoupled approach may vastly underestimate the size of this step component. This can lead to poor performance of the decoupled approach when steps are rejected and the trust-region radius is reduced based on the incorrect estimate $\|s_u\|$ of the norm of $s^{\mathsf{t}} = W_k s_u$. This indicates that when $C_y(x)$ is ill conditioned the coupled approach offers a better regularization of the step.

5.2.3. Cauchy decrease for the tangential component. To assure global convergence to a first-order KKT point, we consider analogues for the subproblems (26)–(27) and (28)–(29) of the fraction of Cauchy decrease or simple decrease conditions for the unconstrained minimization problem.

First we consider the decoupled trust-region subproblem (26)–(27). The Cauchy step c_k^{d} is defined for this case as the solution of

$$\begin{aligned} &\text{minimize } \Psi_k(s_u) \\ &\text{subject to } \|\bar{D}_k^{-1}s_u\| \le \delta_k, \quad s_u \in span\{-\bar{D}_k^2\bar{g}_k\}, \\ &\qquad \sigma_k(a - u_k) \le s_u \le \sigma_k(b - u_k), \end{aligned}$$

where $-\bar{D}_k^2\bar{g}_k$ is the steepest-descent direction for $\Psi_k(s_u)$ at $s_u = 0$ in the norm $\|\bar{D}_k^{-1}\cdot\|$. Here $\sigma_k \in [\sigma, 1)$ ensures that the Cauchy step c_k^{d} remains strictly feasible

with respect to the box constraints. The parameter $\sigma \in (0,1)$ is fixed for all k. As in many trust-region algorithms, we require the tangential component $(s_k)_u$ with $\sigma_k(a-u_k) \le (s_k)_u \le \sigma_k(b-u_k)$ to give a decrease on $\Psi_k(s_u)$ smaller than a uniform fraction of the decrease given by c_k^{d} for the same function $\Psi_k(s_u)$. This condition is often called fraction of Cauchy decrease, and in this case is

$$\Psi_k(0) - \Psi_k((s_k)_u) \ge \beta_1^{\mathsf{d}} \left(\Psi_k(0) - \Psi_k(c_k^{\mathsf{d}})\right), \tag{30}$$

where β_1^{d} is positive and fixed across all iterations. It is not difficult to see that dogleg or conjugate-gradient algorithms can conveniently compute components $(s_k)_u$ that satisfy condition (30) with $\beta_1^{\mathsf{d}} = 1$. We leave these issues to section 10.2.

In a similar way, the component $(s_k)_u$ with $\sigma_k(a-u_k) \le (s_k)_u \le \sigma_k(b-u_k)$ satisfies a fraction of Cauchy decrease for the coupled trust-region subproblem (28)–(29) if

$$\Psi_k(0) - \Psi_k((s_k)_u) \ge \beta_1^{\mathsf{c}} \Big(\Psi_k(0) - \Psi_k(c_k^{\mathsf{c}})\Big), \tag{31}$$

for some β_1^{c} independent of k, where the Cauchy step c_k^{c} is the solution of

$$\begin{aligned} &\text{minimize } \Psi_k(s_u) \\ &\text{subject to } \left\| \begin{pmatrix} -C_y(x_k)^{-1} C_u(x_k) s_u \\ \bar{D}_k^{-1} s_u \end{pmatrix} \right\| \le \delta_k, \quad s_u \in span\{-\bar{D}_k^2 \bar{g}_k\}, \\ &\sigma_k(a-u_k) \le s_u \le \sigma_k(b-u_k). \end{aligned}$$

In section 10.2 we show how to use conjugate gradients to compute components $(s_k)_u$ satisfying the condition (31).

One final comment is in order. In the coupled approach, the Cauchy step c_k^{c} was defined along the direction $-\bar{D}_k^2 \bar{g}_k$. To simplify this discussion, suppose that there are no bounds on u. In this case the trust-region constraint is of the form $\|W_k s_u\| \le \delta_k$. The presence of W_k gives the trust region an ellipsoidal shape. The steepest-descent direction for the quadratic (25) in the norm $\|W_k \cdot\|$ at $s_u = 0$ is given by $-(W_k^T W_k)^{-1} \bar{g}_k$. Our analysis still holds for this case since $\{\|(W_k^T W_k)^{-1}\|\}$ is a bounded sequence. The reason why we avoid the term $(W_k^T W_k)^{-1}$ is that in many applications there is no reasonable way to solve systems with $W_k^T W_k$. We will show in section 10.2 how this affects the use of conjugate gradients (see Remark 10.2). Finally, we point out that this problem does not arise if the decoupled approach is used.

5.2.4. Optimal decrease for the tangential component. The conditions in the previous subsection are sufficient to guarantee global convergence to a point satisfying first-order necessary KKT conditions, but they are too weak to guarantee global convergence to a point satisfying second-order necessary KKT conditions. To accomplish this, just as in the unconstrained case [46], [59], in the box-constrained case [13] and the equality-constrained case [19], we need to make sure that s_u satisfies an appropriate fraction of optimal decrease condition.

First we consider the decoupled approach and let o_k^{d} be an optimal solution of the trust-region subproblem (26)–(27). It follows from the KKT conditions for this trust-region subproblem that there exists $\gamma_k \ge 0$ such that

$$W_k^T H_k W_k + E_k \bar{D}_k^{-2} + \gamma_k \bar{D}_k^{-2} \quad \text{is positive semidefinite,} \tag{32}$$

$$\Big(W_k^T H_k W_k + E_k \bar{D}_k^{-2} + \gamma_k \bar{D}_k^{-2}\Big) o_k^{\mathsf{d}} = -\bar{g}_k, \text{ and} \tag{33}$$

$$\gamma_k(\delta_k - \|\bar{D}_k^{-1} o_k^{\mathsf{d}}\|) = 0.$$

(For practical algorithms to compute o_k^{d} see references [46], [53], [55], and [60]. These conditions are also sufficient for o_k^{d} to be an optimal solution [22], [59].) Since $u_k + o_k^{\mathsf{d}}$ might not be strictly feasible, we consider $\tau_k o_k^{\mathsf{d}}$, where τ_k is given by

$$\tau_k = \sigma_k \min_{i=1,\ldots,n-m} \left\{ 1,\ \max \left\{ \frac{b_i - (u_k)_i}{(o_k^{\mathsf{d}})_i},\ \frac{a_i - (u_k)_i}{(o_k^{\mathsf{d}})_i} \right\} \right\}. \tag{34}$$

The tangential component $(s_k)_u$ is then required to satisfy the following fraction of optimal decrease condition

$$\begin{aligned} &\Psi_k(0) - \Psi_k((s_k)_u) \geq \beta_2^{\mathsf{d}} \left(\Psi_k(0) - \Psi_k(\tau_k o_k^{\mathsf{d}}) \right) \quad \text{and} \\ &\|\bar{D}_k^{-1}(s_k)_u\| \leq \beta_3^{\mathsf{d}} \delta_k, \end{aligned} \tag{35}$$

where $\beta_2^{\mathsf{d}}, \beta_3^{\mathsf{d}}$ are positive parameters.

From conditions (32), (33), and (35), and $\tau_k < 1$, we can write

$$\begin{aligned} \Psi_k(0) - \Psi_k((s_k)_u) &\geq \beta_2^{\mathsf{d}} \left(-\tau_k \bar{g}_k^T o_k^{\mathsf{d}} - \frac{1}{2}\tau_k^2 (o_k^{\mathsf{d}})^T \left(W_k^T H_k W_k + E_k \bar{D}_k^{-2} \right) (o_k^{\mathsf{d}}) \right) \\ &\geq \beta_2^{\mathsf{d}} \tau_k \left(-\bar{g}_k^T o_k^{\mathsf{d}} - \frac{1}{2} (o_k^{\mathsf{d}})^T \left(W_k^T H_k W_k + E_k \bar{D}_k^{-2} + \gamma_k \bar{D}_k^{-2} \right) (o_k^{\mathsf{d}}) \right) \\ &\qquad + \frac{1}{2} \beta_2^{\mathsf{d}} \tau_k^2 \gamma_k (o_k^{\mathsf{d}})^T \bar{D}_k^{-2} (o_k^{\mathsf{d}}) \\ &\geq \frac{1}{2} \beta_2^{\mathsf{d}} \tau_k \|R_k o_k^{\mathsf{d}}\|^2 + \frac{1}{2} \beta_2^{\mathsf{d}} \tau_k^2 \gamma_k \delta_k^2 \\ &\geq \frac{1}{2} \beta_2^{\mathsf{d}} \tau_k^2 \gamma_k \delta_k^2, \end{aligned} \tag{36}$$

where $W_k^T H_k W_k + E_k \bar{D}_k^{-2} + \gamma_k \bar{D}_k^{-2} = R_k^T R_k$.

Now let us focus on the coupled approach and let o_k^{c} be the optimal solution of the trust-region subproblem (28)–(29). It follows from the KKT conditions for this trust-region subproblem, and the equality

$$\left(C_y(x_k)^{-1} C_u(x_k) \right)^T C_y(x_k)^{-1} C_u(x_k) \ = \ W_k^T W_k - I_{n-m},$$

that there exists $\gamma_k \geq 0$ such that

$$W_k^T H_k W_k + E_k \bar{D}_k^{-2} + \gamma_k \left(\bar{D}_k^{-2} + W_k^T W_k - I_{n-m} \right) \text{ is positive semidefinite,} \tag{37}$$

$$\begin{aligned} &\left(W_k^T H_k W_k + E_k \bar{D}_k^{-2} + \gamma_k \left(\bar{D}_k^{-2} + W_k^T W_k - I_{n-m} \right) \right) o_k^{\mathsf{c}} = -\bar{g}_k, \text{ and} \\ &\gamma_k \left(\delta_k - \left\| \begin{pmatrix} -C_y(x_k)^{-1} C_u(x_k) o_k^{\mathsf{c}} \\ \bar{D}_k^{-1} o_k^{\mathsf{c}} \end{pmatrix} \right\| \right) = 0. \end{aligned} \tag{38}$$

Now we damp o_k^{c} with τ_k given as in (34) but with o_k^{d} replaced by o_k^{c}. Thus, the resulting step $u_k + \tau_k o_k^{\mathsf{c}}$ is strictly feasible. We impose the following fraction of optimal decrease condition on the tangential component $(s_k)_u$:

$$\begin{aligned} &\Psi_k(0) - \Psi_k((s_k)_u) \geq \beta_2^{\mathsf{c}} \Big(\Psi_k(0) - \Psi_k(\tau_k o_k^{\mathsf{c}}) \Big) \quad \text{and} \\ &\left\| \begin{pmatrix} -C_y(x_k)^{-1} C_u(x_k)(s_k)_u \\ \bar{D}_k^{-1}(s_k)_u \end{pmatrix} \right\| \leq \beta_3^{\mathsf{c}} \delta_k. \end{aligned} \tag{39}$$

In this case it can be shown in a way similar to (36) that

$$\Psi_k(0) - \Psi((s_k)_u) \geq \frac{1}{2}\beta_2^{\mathfrak{c}}\tau_k^2\gamma_k\delta_k^2. \tag{40}$$

5.3. Reduced and full Hessians. In the previous section we considered an approximation H_k to the full Hessian. The algorithms and theory presented in this paper are also valid if we use an approximation $\widehat{H}_k$ to the reduced Hessian $W_k^T\nabla_{xx}^2\ell_k W_k$. In this case we set

$$H_k = \begin{pmatrix} 0 & 0 \\ 0 & \widehat{H}_k \end{pmatrix}. \tag{41}$$

Due to the form of W_k, we have

$$W_k^T H_k W_k = \widehat{H}_k.$$

This allows us to obtain the expansion (24) in the context of a reduced Hessian approximation.

For the algorithms with reduced Hessian approximation, the following observations are useful:

$$\begin{aligned} H_k d &= \begin{pmatrix} 0 \\ \widehat{H}_k d_u \end{pmatrix}, \\ d^T H_k d &= d_u^T \widehat{H}_k d_u, \\ W_k^T H_k d &= \widehat{H}_k d_u. \end{aligned} \tag{42}$$

5.4. Outline of the algorithms. We need to introduce a merit function and the corresponding actual and predicted reductions. The merit function used is the augmented Lagrangian

$$L(x, \lambda; \rho) = f(x) + \lambda^T C(x) + \rho C(x)^T C(x).$$

We follow [15] and define the actual decrease at iteration k as

$$ared(s_k; \rho_k) = L(x_k, \lambda_k; \rho_k) - L(x_k + s_k, \lambda_{k+1}; \rho_k),$$

and the predicted decrease as

$$pred(s_k; \rho_k) = L(x_k, \lambda_k; \rho_k) - \left(q_k(s_k) + \Delta\lambda_k^T(J_k s_k + C_k) + \rho_k\|J_k s_k + C_k\|^2\right),$$

with $\Delta\lambda_k = \lambda_{k+1} - \lambda_k$.

Remark 5.1. A possible redefining of the actual and predicted decreases is obtained by subtracting the term $\frac{1}{2}(s_k)_u^T\left(E_k\bar{D}_k^{-2}\right)(s_k)_u$ from both $ared(s_k; \rho_k)$ and $pred(s_k; \rho_k)$. This type of modification has been suggested in [13] for minimization with simple bounds, and it does not affect the global and local results given in this paper.

To decide whether to accept or reject a trial step s_k, we evaluate the ratio

$$\frac{ared(s_k; \rho_k)}{pred(s_k; \rho_k)}.$$

To update the penalty parameter ρ_k we use the scheme proposed by El-Alem [20]. Other schemes to update the penalty parameter have been suggested in [21] and [40].

We can now outline the main procedures of the trust-region interior-point SQP algorithms and leave the practical computation of $s_k^{\mathfrak{n}}$, $(s_k)_u$, and λ_k to section 10.

ALGORITHMS 5.1 (trust-region interior-point SQP algorithms).

1. Choose x_0 such that $a < u_0 < b$, pick $\delta_0 > 0$, and calculate λ_0. Choose α_1, η_1, σ, $\delta_{\min}$, $\delta_{\max}$, $\bar{\rho}$, and ρ_{-1} such that $0 < \alpha_1, \eta_1, \sigma < 1$, $0 < \delta_{\min} \le \delta_{\max}$, $\bar{\rho} > 0$, and $\rho_{-1} \ge 1$.
2. For $k = 0, 1, 2, \ldots$ do
   2.1. Compute s_k^{n} such that $\|s_k^{\mathsf{n}}\| \le \delta_k$.
   Compute $(s_k)_u$ based on the subproblem (26)–(27) (or (28)–(29) for the coupled approach) satisfying
   $$\sigma_k(a - u_k) \le (s_k)_u \le \sigma_k(b - u_k),$$
   with $\sigma_k \in [\sigma, 1)$. Set $s_k = s_k^{\mathsf{n}} + s_k^{\mathsf{t}} = s_k^{\mathsf{n}} + W_k(s_k)_u$.
   2.2. Compute λ_{k+1} and set $\Delta\lambda_k = \lambda_{k+1} - \lambda_k$.
   2.3. Compute $pred(s_k; \rho_{k-1})$:
   $$\begin{aligned} pred(s_k; \rho_{k-1}) = {} & q_k(0) - q_k(s_k) - \Delta\lambda_k^T(J_k s_k + C_k) \\ & + \rho_{k-1}\left(\|C_k\|^2 - \|J_k s_k + C_k\|^2\right). \end{aligned}$$
   If $pred(s_k; \rho_{k-1}) \ge \frac{\rho_{k-1}}{2}\Big(\|C_k\|^2 - \|J_k s_k + C_k\|^2\Big)$, then set $\rho_k = \rho_{k-1}$. Otherwise set
   $$\rho_k = \frac{2\left(q_k(s_k) - q_k(0) + \Delta\lambda_k^T(J_k s_k + C_k)\right)}{\|C_k\|^2 - \|J_k s_k + C_k\|^2} + \bar{\rho}.$$
   2.4. If $\frac{ared(s_k;\rho_k)}{pred(s_k;\rho_k)} < \eta_1$, set
   $$\delta_{k+1} = \alpha_1 \max\left\{\|s_k^{\mathsf{n}}\|, \|\bar{D}_k^{-1}(s_k)_u\|\right\} \text{ in the decoupled case or}$$
   $$\delta_{k+1} = \alpha_1 \max\left\{\|s_k^{\mathsf{n}}\|, \left\|\begin{pmatrix} -C_y(x_k)^{-1}C_u(x_k)(s_k)_u \\ \bar{D}_k^{-1}(s_k)_u \end{pmatrix}\right\|\right\} \text{ in the}$$
   coupled case, and reject s_k.
   Otherwise accept s_k and choose δ_{k+1} such that
   $$\max\{\delta_{\min}, \delta_k\} \le \delta_{k+1} \le \delta_{\max}.$$
   2.5. If s_k was rejected set $x_{k+1} = x_k$ and $\lambda_{k+1} = \lambda_k$. Otherwise set $x_{k+1} = x_k + s_k$ and $\lambda_{k+1} = \lambda_k + \Delta\lambda_k$.

Of course the rules to update the trust radius in the previous algorithm can be much more involved, but the above suffices to prove convergence results and to understand the trust-region mechanism.

5.5. Assumptions. In order to establish local and global convergence results we need some general assumptions. We list these assumptions below. Let Ω be an open subset of $\mathbb{R}^n$ such that for all iterations k, x_k and $x_k + s_k$ are in Ω.

A.1. The functions $f(x)$, $c_i(x)$, $i = 1, \ldots, m$, are twice continuously differentiable in Ω.
A.2. The partial Jacobian $C_y(x)$ is nonsingular for all $x \in \Omega$.
A.3. The functions $f(x)$, $\nabla f(x)$, $\nabla^2 f(x)$, $C(x)$, $J(x)$, $\nabla^2 c_i(x)$, $i = 1, \ldots, m$ are bounded in Ω.
A.4. The sequences $\{W_k\}$, $\{H_k\}$, and $\{\lambda_k\}$ are bounded.
A.5. The matrix $C_y^{-1}(x)$ is uniformly bounded in Ω.
A.6. The sequence $\{u_k\}$ is bounded.

It is equivalent to Assumptions A.3–A.6 that there exist positive constants $\nu_0, \ldots, \nu_9$ independent of k such that

$$|f(x)| \le \nu_0, \quad \|\nabla f(x)\| \le \nu_1, \quad \|\nabla^2 f(x)\| \le \nu_2, \quad \|C(x)\| \le \nu_3, \quad \|J(x)\| \le \nu_4,$$
$$\|\nabla^2 c_i(x)\| \le \nu_5, \quad i = 1, \ldots, m, \quad \text{and} \quad \|C_y(x)^{-1}\| \le \nu_6$$

for all $x \in \Omega$, and

$$\|W_k\| \le \nu_6, \quad \|H_k\| \le \nu_7, \quad \|\lambda_k\| \le \nu_8, \quad \text{and} \quad \|\bar{D}_k\| \le \nu_9$$

for all k.

For the rest of this paper we suppose that Assumptions A.1–A.6 are always satisfied.

As we have pointed out earlier, our approach is related to the Newton method presented in section 4. The u component $(s_k^{\mathsf{N}})_u$ of the Newton step $s_k^{\mathsf{N}} = s_k^{\mathsf{n}} + W_k(s_k^{\mathsf{N}})_u$, whenever it is defined, is given by

$$\begin{aligned}(s_k^{\mathsf{N}})_u &= -\left(\bar{D}_k^2 W_k^T H_k W_k + E_k\right)^{-1} \bar{D}_k^2 \bar{g}_k \\ &= -\bar{D}_k \left(\bar{D}_k W_k^T H_k W_k \bar{D}_k + E_k\right)^{-1} \bar{D}_k \bar{g}_k,\end{aligned} \tag{43}$$

where

$$s_k^{\mathsf{n}} = \begin{pmatrix} -C_y(x_k)^{-1} C_k \\ 0 \end{pmatrix}, \tag{44}$$

and $\bar{g}_k = W_k^T(H_k s_k^{\mathsf{n}} + \nabla f_k)$. From (43) we see that the Newton step is well defined in a neighborhood of a nondegenerate point that satisfies the second-order sufficient KKT conditions and for which $W_k^T H_k s_k^{\mathsf{n}}$ is sufficiently small. To guarantee strict feasibility of this step we consider a damped Newton step given by

$$s_k^{\mathsf{n}} + W_k \tau_k^{\mathsf{N}} (s_k^{\mathsf{N}})_u, \tag{45}$$

where $(s_k^{\mathsf{N}})_u$ and s_k^{n} are given by (43) and (44), and

$$\tau_k^{\mathsf{N}} = \sigma_k \min_{i=1,\ldots,n-m} \left\{ 1, \max\left\{ \frac{b_i - (u_k)_i}{((s_k^{\mathsf{N}})_u)_i}, \frac{a_i - (u_k)_i}{((s_k^{\mathsf{N}})_u)_i} \right\} \right\}. \tag{46}$$

If Algorithms 5.1 are particularized to satisfy the following conditions on the steps, on the quadratic model, and on the Lagrange multipliers, then we can prove global and local convergence.

C.1. The quasi-normal component s_k^{n} satisfies conditions (18), (21), and (22).
The tangential component $(s_k)_u$ satisfies the fraction of Cauchy decrease condition (30) ((31) for the coupled approach).
The parameter σ_k is chosen in $[\sigma, 1)$, where $\sigma \in (0, 1)$ is fixed for all k.

C.2. The tangential component $(s_k)_u$ satisfies the fraction of optimal decrease condition (35) ((39) for the coupled approach).

C.3. The second derivatives of f and c_i, $i = 1, \ldots, m$ are Lipschitz continuous in Ω.
The approximation to the Hessian matrix is exact, i.e., $H_k = \nabla^2_{xx}\ell(x_k, \lambda_k)$ with Lagrange multiplier $\lambda_k = -C_y(x_k)^{-T}\nabla_y f(x_k)$.

C.4. The step s_k is given by (45) provided $(s_k^{\mathsf{N}})_u$ exists, $(s_k^{\mathsf{n}})_y$ lies inside the trust region (20), and $\tau_k^{\mathsf{N}}(s_k^{\mathsf{N}})_u$ lies inside the trust region (27) ((29) for the coupled approach).

The parameter σ_k is chosen such that $\sigma_k \geq \sigma$ and $|\sigma_k - 1|$ is $\mathcal{O}(\|\bar{D}_k \bar{g}_k\|)$.

Condition C.1 assures global convergence to a first-order KKT point. Global convergence to a point that satisfies the second-order necessary KKT conditions requires Conditions C.1–C.3. To prove local q-quadratic convergence, we need Conditions C.1, C.3, and C.4. It should be pointed out that the satisfaction of C.2 or C.4 does not necessarily imply the satisfaction of C.1.

6. Intermediate results. We start by pointing out that (22), together with the fact that the tangential component lies in the null space of J_k, imply

$$\|C_k\|^2 - \|J_k s_k + C_k\|^2 \geq \kappa_2 \|C_k\| \min\{\kappa_3 \|C_k\|, \delta_k\}. \tag{47}$$

We calculated the first derivatives of $\lambda(x) = -C_y(x)^{-T} \nabla_y f(x)$ in section 4. It is clear that under Assumptions A.3 and A.5 these derivatives are bounded in Ω. Thus, if λ_k is computed as stated in Condition C.3, then there exists a positive constant ν_{10} independent of k such that

$$\|\Delta \lambda_k\| \leq \nu_{10} \|s_k\|. \tag{48}$$

From $\|s_k^{\mathsf{q}}\| \leq \delta_{\max}$ and Assumptions A.3–A.4 we also have

$$\|\bar{g}_k\| = \left\| W_k^T \left(H_k s_k^{\mathsf{q}} + \nabla f_k \right) \right\| \leq \nu_{11}, \tag{49}$$

where $\nu_{11} = \nu_6(\nu_7 \delta_{\max} + \nu_1)$.

The following lemma is required for the convergence theory.

LEMMA 6.1. *Every trial step satisfies*

$$\|s_k\| \leq \kappa_4 \delta_k \tag{50}$$

and, if s_k is rejected in step 2.4 *of Algorithms* 5.1, *then*

$$\delta_{k+1} \geq \kappa_5 \|s_k\|, \tag{51}$$

where κ_4 and κ_5 are positive constants independent of k.

Proof. In the coupled trust-region approach we bound s_k^{t} as follows:

$$\begin{aligned} \left\| \begin{pmatrix} -C_y(x_k)^{-1} C_u(x_k) s_u \\ s_u \end{pmatrix} \right\| &\leq \left\| \begin{pmatrix} I_m & 0 \\ 0 & \bar{D}_k \end{pmatrix} \right\| \left\| \begin{pmatrix} -C_y(x_k)^{-1} C_u(x_k) s_u \\ \bar{D}_k^{-1} s_u \end{pmatrix} \right\| \\ &\leq (1 + \nu_9)\, \delta_k\,, \end{aligned}$$

where ν_9 is a uniform bound for $\|\bar{D}_k\|$, see Assumption A.6. Since $\|s_k^{\mathsf{n}}\| \leq \delta_k$, we obtain $\|s_k\| \leq (2 + \nu_9)\, \delta_k$. It is not difficult to see now that in step 2.4 we have $\delta_{k+1} \geq \frac{\alpha_1}{2} \min\{1, \frac{1}{1+\nu_9}\} \|s_k\|$.

In the decoupled approach, $\|s_k\| = \|s_k^{\mathsf{n}} + W_k (s_k)_u\| \leq (1 + \nu_6 \nu_9) \delta_k$ and similarly $\delta_{k+1} \geq \frac{\alpha_1}{2} \min\{1, \frac{1}{\nu_6 \nu_9}\} \|s_k\|$, where ν_6 is a uniform bound for $\|W_k\|$; see Assumption A.4.

We can combine these bounds to obtain

$$\begin{aligned} \|s_k\| &\leq \max\{2 + \nu_9, 1 + \nu_6 \nu_9\}\, \delta_k, \\ \delta_{k+1} &\geq \tfrac{\alpha_1}{2} \min\left\{1, \tfrac{1}{1+\nu_9}, \tfrac{1}{\nu_6 \nu_9}\right\} \|s_k\|. \end{aligned}$$

In the case where fraction of optimal decrease (35) or (39) is imposed on $(s_k)_u$, the constants κ_4 and κ_5 depend also on β_3^{d} and β_3^{c}. ∎

In the following lemma we rewrite the fraction of Cauchy decrease conditions (30) and (31) in a more useful form for the analysis.

LEMMA 6.2. *If $(s_k)_u$ satisfies Condition C.1, then*

$$q_k(s_k^{\mathsf{n}}) - q_k(s_k^{\mathsf{n}} + W_k(s_k)_u) \geq \kappa_6 \|\bar{D}_k \bar{g}_k\| \min\left\{\kappa_7 \|\bar{D}_k \bar{g}_k\|, \kappa_8 \delta_k\right\}, \tag{52}$$

where κ_6, κ_7, and κ_8 are positive constants independent of the iteration k.

Proof. From the definition (25) of Ψ_k we find

$$\begin{aligned} q_k(s_k^{\mathsf{n}}) - q_k(s_k^{\mathsf{n}} + W_k(s_k)_u) &\geq q_k(s_k^{\mathsf{n}}) - q_k(s_k^{\mathsf{n}} + W_k(s_k)_u) - \frac{1}{2}(s_k)_u^T \left(E_k \bar{D}_k^{-2}\right)(s_k)_u \\ &= \Psi_k(0) - \Psi_k((s_k)_u). \end{aligned} \tag{53}$$

Let $\tilde{\delta}_k$ be the maximum $\|\bar{D}_k^{-1} \ \cdot\|$ norm of a step, say $(\tilde{s}_k)_u$, along $-\bar{D}_k \frac{\tilde{g}_k}{\|\tilde{g}_k\|}$ allowed inside the trust region. Here $\tilde{g}_k = \bar{D}_k \bar{g}_k$.

If the trust region is given by (27), then

$$\delta_k = \tilde{\delta}_k. \tag{54}$$

If the trust region is given by (29), then we can use Assumptions A.4–A.6 to deduce the inequality

$$\begin{aligned} \delta_k^2 &= \left\| \begin{pmatrix} -C_y(x_k)^{-1} C_u(x_k)(\tilde{s}_k)_u \\ \bar{D}_k^{-1}(\tilde{s}_k)_u \end{pmatrix} \right\|^2 \\ &= \| - C_y(x_k)^{-1} C_u(x_k) \bar{D}_k \bar{D}_k^{-1}(\tilde{s}_k)_u\|^2 + \|\bar{D}_k^{-1}(\tilde{s}_k)_u\|^2 \\ &\leq (\nu_6^2 \nu_9^2 + 1) \|\bar{D}_k^{-1}(\tilde{s}_k)_u\|^2 \\ &= (\nu_6^2 \nu_9^2 + 1)\, \tilde{\delta}_k^2 \end{aligned}$$

or, equivalently,

$$\tilde{\delta}_k \geq \frac{1}{\sqrt{\nu_6^2 \nu_9^2 + 1}}\, \delta_k. \tag{55}$$

Define $\psi : \mathbb{R}^+ \longrightarrow \mathbb{R}$ as $\psi(t) = \Psi_k(-t \bar{D}_k \frac{\tilde{g}_k}{\|\tilde{g}_k\|}) - \Psi_k(0)$. Then $\psi(t) = -\|\tilde{g}_k\| t + \frac{r_k}{2} t^2$, where $r_k = \frac{\tilde{g}_k^T \tilde{H}_k \tilde{g}_k}{\|\tilde{g}_k\|^2}$ and $\tilde{H}_k = \bar{D}_k (W_k^T H_k W_k + E_k \bar{D}_k^{-2}) \bar{D}_k$. Now we need to minimize ψ in $[0, T_k]$, where T_k is given by

$$T_k = \min\left\{\tilde{\delta}_k,\ \sigma_k \min\left\{\frac{\|\bar{D}_k \bar{g}_k\|}{(\bar{g}_k)_i} : (\bar{g}_k)_i > 0\right\},\ \sigma_k \min\left\{-\frac{\|\bar{D}_k \bar{g}_k\|}{(\bar{g}_k)_i} : (\bar{g}_k)_i < 0\right\}\right\}.$$

Let t_k^* be the minimizer of ψ in $[0, T_k]$. If $t_k^* \in (0, T_k)$, then

$$\psi(t_k^*) = -\frac{1}{2} \frac{\|\tilde{g}_k\|^2}{r_k} \leq -\frac{1}{2} \frac{\|\tilde{g}_k\|^2}{\|\tilde{H}_k\|}. \tag{56}$$

If $t_k^* = T_k$, then either $r_k > 0$, in which case $\frac{\|\tilde{g}_k\|}{r_k} \geq T_k$, or $r_k \leq 0$, in which case $r_k T_k \leq \|\tilde{g}_k\|$. In either event,

$$\psi(t_k^*) = \psi(T_k) = -T_k\|\tilde{g}_k\| + \frac{r_k}{2}T_k^2 \leq -\frac{T_k}{2}\|\tilde{g}_k\|. \tag{57}$$

We can combine (53), (56), and (57) with

$$\Psi_k(0) - \Psi_k((s_k)_u) \geq \beta_1^{\mathsf{d}}\left(\Psi_k(0) - \Psi_k(c_k^{\mathsf{d}})\right) = -\beta_1^{\mathsf{d}}\psi(t_k^*)$$

to get

$$q_k(s_k^{\mathsf{n}}) - q_k(s_k^{\mathsf{n}} + W_k(s_k)_u) \geq \frac{1}{2}\beta_1^{\mathsf{d}}\|\tilde{g}_k\| \min\left\{\frac{\|\tilde{g}_k\|}{\|\tilde{H}_k\|}, T_k\right\}.$$

The facts that $\sigma_k \geq \sigma$ and $\|\bar{g}_k\| \leq \nu_{11}$ (see (49)) imply that

$$\begin{aligned}&\Psi_k(0) - \Psi_k((s_k)_u)\\ &\geq \frac{1}{2}\beta_1^{\mathsf{d}}\|\bar{D}_k\bar{g}_k\| \min\left\{\frac{\|\bar{D}_k\bar{g}_k\|}{\|\bar{D}_k^T\left(W_k^T H_k W_k + E_k\bar{D}_k^{-2}\right)\bar{D}_k\|}, \min\left\{\tilde{\delta}_k, \frac{\sigma}{\nu_{11}}\|\bar{D}_k\bar{g}_k\|\right\}\right\}.\end{aligned}$$

To complete the proof, we use (54), (55), the Assumptions A.1–A.6, and the fact that $\delta_k \leq \delta_{\max}$ to establish (52) with $\kappa_6 = \frac{1}{2}\min\{\beta_1^{\mathsf{d}}, \beta_1^{\mathsf{c}}\}$, $\kappa_7 = \min\{\frac{1}{\nu_7\nu_6^2\nu_9^2+\nu_1\nu_6}, \frac{\sigma}{\nu_{11}}\}$, and $\kappa_8 = \min\{1, \frac{1}{\sqrt{\nu_6^2\nu_9^2+1}}\}$. ∎

Now we state the convenient form of the fraction of optimal decrease conditions (35) and (39).

LEMMA 6.3. *If $(s_k)_u$ satisfies Condition C.2, then*

$$q_k(s_k^{\mathsf{n}}) - q_k(s_k^{\mathsf{n}} + W_k(s_k)_u) \geq \kappa_9\tau_k^2\gamma_k\delta_k^2, \tag{58}$$

where κ_9 is a positive constant independent of the iteration k.

Proof. The proof follows immediately from observation (53) and conditions (36) and (40). ∎

We also need the following two inequalities.

LEMMA 6.4. *Under Condition C.1 there exists a positive constant κ_{10} such that*

$$q_k(0) - q_k(s_k^{\mathsf{n}}) - \Delta\lambda_k^T(J_k s_k + C_k) \geq -\kappa_{10}\|C_k\|. \tag{59}$$

Moreover, if we assume Condition C.3, then

$$q_k(0) - q_k(s_k^{\mathsf{n}}) - \Delta\lambda_k^T(J_k s_k + C_k) \geq -\kappa_{11}\|C_k\|\left(\|s_k^{\mathsf{n}}\| + \|s_k\|\right). \tag{60}$$

Proof. The term $q_k(0) - q_k(s_k^{\mathsf{n}})$ can be bounded using (21) and $\|s_k^{\mathsf{n}}\| \leq \delta_k$ in the following way:

$$\begin{aligned}q_k(0) - q_k(s_k^{\mathsf{n}}) &= -\nabla_x\ell_k^T s_k^{\mathsf{n}} - \tfrac{1}{2}(s_k^{\mathsf{n}})^T H_k(s_k^{\mathsf{n}})\\ &\geq -\kappa_1\left(\|\nabla_x\ell_k\| + \tfrac{1}{2}\delta_k\|H_k\|\right)\|C_k\|.\end{aligned}$$

On the other hand, it follows from $\|J_k s_k + C_k\| \leq \|C_k\|$ that

$$-\Delta\lambda_k^T(J_k s_k + C_k) \geq -\|\Delta\lambda_k\|\,\|C_k\|. \tag{61}$$

Combining these two bounds with Assumptions A.3 and A.4 we get (59).

To prove (60) we first observe that, due to the definition of λ_k in Condition C.3 and to the form (18) of the quasi-normal component s_k^{n},

$$\nabla_x \ell_k^T s_k^{\mathsf{n}} = \begin{pmatrix} 0 \\ \nabla_u f_k + C_u(x_k)^T \lambda_k \end{pmatrix}^T \begin{pmatrix} (s_k^{\mathsf{n}})_y \\ 0 \end{pmatrix} = 0. \tag{62}$$

Thus,

$$q_k(0) - q_k(s_k^{\mathsf{n}}) \geq -\frac{1}{2}\kappa_1 \|H_k\| \, \|C_k\| \, \|s_k^{\mathsf{n}}\| \geq -\frac{1}{2}\kappa_1 \nu_7 \, \|C_k\| \, \|s_k^{\mathsf{n}}\|. \tag{63}$$

Also, by appealing to (48) and (61),

$$-\Delta\lambda_k^T (J_k s_k + C_k) \geq -\nu_{10} \|s_k\| \, \|C_k\|. \tag{64}$$

The proof of (60) is complete by combining (63) and (64). ∎

The convergence theory for trust regions traditionally requires consistency of actual and predicted decreases. This is given in the following lemma.

LEMMA 6.5. *Under Condition C.1 there exists a positive constant* κ_{12} *such that*

$$|ared(s_k;\rho_k) - pred(s_k;\rho_k)| \leq \kappa_{12} \left(\|s_k\|^2 + \rho_k \left(\|s_k\|^3 + \|C_k\| \, \|s_k\|^2 \right) \right). \tag{65}$$

Moreover, if Condition C.3 is also valid, then

$$|ared(s_k;\rho_k) - pred(s_k;\rho_k)| \leq \kappa_{13} \rho_k \left(\|s_k\|^3 + \|C_k\| \, \|s_k\|^2 \right). \tag{66}$$

Proof. Adding and subtracting $\ell(x_{k+1}, \lambda_k)$ to $ared(s_k;\rho_k) - pred(s_k;\rho_k)$, and using Taylor expansion, we obtain

$$\begin{aligned} ared(s_k;\rho_k) - pred(s_k;\rho_k) \quad = \quad & \tfrac{1}{2} s_k^T \left(H_k - \nabla^2_{xx} \ell(x_k + t_k^1 s_k, \lambda_k) \right) s_k \\ & - \tfrac{1}{2} \textstyle\sum_{i=1}^m (\Delta\lambda_k)_i s_k^T \nabla^2 c_i(x_k + t_k^2 s_k) s_k \\ & - \rho_k \left(\textstyle\sum_{i=1}^m c_i(x_k + t_k^3 s_k)(s_k)^T \nabla^2 c_i(x_k + t_k^3 s_k)(s_k) \right. \\ & \qquad + (s_k)^T J(x_k + t_k^3 s_k)^T J(x_k + t_k^3 s_k)(s_k) \\ & \qquad \left. - (s_k)^T J(x_k)^T J(x_k)(s_k) \right), \end{aligned}$$

where t_k^1, t_k^2, and t_k^3 are in $(0,1)$. By expanding $c_i(x_k + t_k^3 s_k)$ around $c_i(x_k)$ and using Assumptions A.3 and A.4 we get (65).

The estimate (66) follows from (48), $\rho_k \geq 1$, and the Lipschitz continuity of the second derivatives. ∎

The last result in this section is a direct consequence of the scheme that updates ρ_k in step 2.3 of Algorithms 5.1.

LEMMA 6.6. *The sequence* $\{\rho_k\}$ *satisfies*

$$\begin{aligned} & \rho_k \geq \rho_{k-1} \geq 1 \quad \textit{and} \\ & pred(s_k;\rho_k) \geq \frac{\rho_k}{2} \Big(\|C_k\|^2 - \|J_k s_k + C_k\|^2 \Big). \end{aligned} \tag{67}$$

7. Global convergence to a first-order KKT point. The proof of the global convergence to a first-order KKT point (Theorem 7.5) established in this section follows the structure of the convergence theory presented in [15] for the equality-constrained optimization problem. This proof is by contradiction and is based on Condition C.1. We show that the supposition

$$\|\bar{D}_k \bar{g}_k\| + \|C_k\| > \epsilon_{tol},$$

for all k, leads to a contradiction.

The following three lemmas are necessary to bound the predicted decrease.

LEMMA 7.1. *Under Condition C.1, the predicted decrease in the merit function satisfies*

$$\begin{aligned} pred(s_k;\rho) \quad \geq \quad & \kappa_6 \|\bar{D}_k \bar{g}_k\| \min\left\{\kappa_7 \|\bar{D}_k \bar{g}_k\|, \kappa_8 \delta_k\right\} \\ & -\kappa_{10}\|C_k\| + \rho\Big(\|C_k\|^2 - \|J_k s_k + C_k\|^2\Big) \end{aligned} \tag{68}$$

for every $\rho > 0$.

Proof. The inequality (68) follows from a direct application of (59) and from the lower bound (52). ∎

LEMMA 7.2. *Assume that Condition C.1 and $\|\bar{D}_k \bar{g}_k\| + \|C_k\| > \epsilon_{tol}$ are satisfied. If $\|C_k\| \leq \alpha \delta_k$, where α is a positive constant satisfying*

$$\alpha \leq \min\left\{\frac{\epsilon_{tol}}{3\delta_{\max}}, \frac{\kappa_6 \epsilon_{tol}}{3\kappa_{10}} \min\left\{\frac{2\kappa_7 \epsilon_{tol}}{3\delta_{\max}}, \kappa_8\right\}\right\}, \tag{69}$$

then

$$pred(s_k;\rho) \geq \frac{\kappa_6}{2}\|\bar{D}_k \bar{g}_k\| \min\left\{\kappa_7 \|\bar{D}_k \bar{g}_k\|, \kappa_8 \delta_k\right\} + \rho\left(\|C_k\|^2 - \|J_k s_k + C_k\|^2\right), \tag{70}$$

for every $\rho > 0$.

Proof. From $\|\bar{D}_k \bar{g}_k\| + \|C_k\| > \epsilon_{tol}$ and the first bound on α given by (69), we get

$$\|\bar{D}_k \bar{g}_k\| > \frac{2}{3}\epsilon_{tol}. \tag{71}$$

If we use this, (68), and the second bound on α given by (69), we obtain

$$\begin{aligned} pred(s_k;\rho) \geq & \ \tfrac{\kappa_6}{2}\|\bar{D}_k \bar{g}_k\| \min\left\{\kappa_7 \|\bar{D}_k \bar{g}_k\|, \kappa_8 \delta_k\right\} + \tfrac{\kappa_6 \epsilon_{tol}}{3} \min\left\{\tfrac{2\kappa_7 \epsilon_{tol}}{3}, \kappa_8 \delta_k\right\} \\ & - \kappa_{10}\|C_k\| + \rho\Big(\|C_k\|^2 - \|J_k s_k + C_k\|^2\Big) \\ \geq & \ \tfrac{\kappa_6}{2}\|\bar{D}_k \bar{g}_k\| \min\left\{\kappa_7 \|\bar{D}_k \bar{g}_k\|, \kappa_8 \delta_k\right\} + \rho\Big(\|C_k\|^2 - \|J_k s_k + C_k\|^2\Big). \end{aligned}$$ ∎

We can use Lemma 7.2 with $\rho = \rho_{k-1}$ and conclude that if $\|\bar{D}_k \bar{g}_k\| + \|C_k\| > \epsilon_{tol}$ and $\|C_k\| \leq \alpha \delta_k$, then the penalty parameter at the current iteration does not need to be increased. See step 2.3 of Algorithms 5.1. This is equivalent to Lemma 7.7 in [15]. The next lemma states the same result as Lemma 7.8 in [15] but with a different choice of α.

LEMMA 7.3. *Assume Condition C.1 and* $\|\bar{D}_k\bar{g}_k\| + \|C_k\| > \epsilon_{tol}$. *If* $\|C_k\| \leq \alpha\delta_k$, *where* α *satisfies* (69), *then there exists a positive constant* $\kappa_{14} > 0$ *such that*

$$pred(s_k;\rho_k) \geq \kappa_{14}\delta_k. \tag{72}$$

Proof. From (70), with $\rho = \rho_k$ and $\|\bar{D}_k\bar{g}_k\| \geq \frac{2}{3}\epsilon_{tol}$, cf. (71), we obtain

$$\begin{aligned} pred(s_k;\rho_k) &\geq \tfrac{\kappa_6\epsilon_{tol}}{3}\min\{\tfrac{2\kappa_7\epsilon_{tol}}{3}, \kappa_8\delta_k\} \\ &\geq \tfrac{\kappa_6\epsilon_{tol}}{3}\min\{\tfrac{2\kappa_7\epsilon_{tol}}{3\delta_{\max}}, \kappa_8\}\delta_k. \end{aligned}$$

Hence (72) holds with

$$\kappa_{14} = \frac{\kappa_6\epsilon_{tol}}{3}\min\left\{\frac{2\kappa_7\epsilon_{tol}}{3\delta_{\max}}, \kappa_8\right\}. \qquad \square$$

The following lemma is also required.

LEMMA 7.4. *Under Condition C.1, if* $\|\bar{D}_k\bar{g}_k\| + \|C_k\| > \epsilon_{tol}$ *for all* k, *then the sequences* $\{\rho_k\}$ *and* $\{L_k\}$ *are bounded and* δ_k *is uniformly bounded away from zero.*

Proof. See Lemmas 7.9–7.13 and 8.2 in [15]. $\square$

Our first global convergence result follows.

THEOREM 7.5. *Under Condition C.1, the sequences of iterates generated by the trust-region interior-point SQP Algorithms* 5.1 *satisfy*

$$\liminf_k \left(\|D_k W_k^T \nabla f_k\| + \|C_k\|\right) = 0. \tag{73}$$

Proof. The proof is by contradiction. Suppose that for all k,

$$\|\bar{D}_k\bar{g}_k\| + \|C_k\| > \epsilon_{tol}. \tag{74}$$

At each iteration k, either $\|C_k\| \leq \alpha\delta_k$ or $\|C_k\| > \alpha\delta_k$, where α satisfies (69). In the first case we appeal to Lemmas 7.3 and 7.4 and obtain

$$pred(s_k;\rho_k) \geq \kappa_{14}\delta_*,$$

where δ_* is the lower bound on δ_k given by Lemma 7.4. If $\|C_k\| > \alpha\delta_k$, we have from $\rho_k \geq 1$, (47), (67), and Lemma 7.4, that

$$pred(s_k;\rho_k) \geq \frac{\kappa_2}{2}\alpha\min\{\kappa_3\alpha, 1\}\delta_*.$$

Hence $pred(s_k;\rho_k) \geq \kappa_{15}$ for all k, where the positive constant κ_{15} does not depend on k. From this and (65) we establish

$$\left|\frac{ared(s_k;\rho_k) - pred(s_k;\rho_k)}{pred(s_k;\rho_k)}\right| \leq \frac{\kappa_{12}}{\kappa_{15}}\left(\|s_k\|^2 + \rho_*\left(\|s_k\|^3 + \|C_k\|\,\|s_k\|^2\right)\right) \leq \kappa_{16}\delta_k^2,$$

where ρ_* is the upper bound on ρ_k guaranteed by Lemma 7.4. From the rules that update δ_k in step 2.4 of Algorithms 5.1, this inequality tells us that an acceptable step is always found after a finite number of unsuccessful iterations. Using this fact, we can ignore the rejected steps and work only with successful iterates. So, without loss of generality, we have

$$L_k - L_{k+1} = ared(s_k;\rho_k) \geq \eta_1 pred(s_k;\rho_k) \geq \eta_1\kappa_{15}.$$

Now, if we let k go to infinity, this contradicts the boundedness of $\{L_k\}$ guaranteed by Lemma 7.4. Hence the supposition (74) is false, and we must have that

$$\liminf_k \left(\|\bar{D}_k \bar{g}_k\| + \|C_k\| \right) = 0. \tag{75}$$

Let $\{k_j\}$ be a subsequence with $\lim_j (\|\bar{D}_{k_j} \bar{g}_{k_j}\| + \|C_{k_j}\|) = 0$. Together with (21) and the boundedness of $\{H_k\}$, this implies $\lim_j (\|\bar{D}_{k_j} W_{k_j}^T \nabla f_{k_j}\| + \|C_{k_j}\|) = 0$. To establish (73), it remains to show that $\bar{D}_{k_j}$, which is the scaling matrix defined with the reduced gradient $W_{k_j}^T (H_{k_j} s_{k_j}^{\mathsf{n}} + \nabla f_{k_j})$, can be replaced by D_{k_j}. This can be shown by standard arguments. Let $i \in \{1, \ldots, n-m\}$ be arbitrary. Assume there exists $\epsilon_1 > 0$ and a subsequence of $\{k_j\}$, for simplicity again denoted by $\{k_j\}$, such that

$$\left| \left((\bar{D}_{k_j} - D_{k_j}) W_{k_j}^T \nabla f_{k_j} \right)_i \right| > \epsilon_1. \tag{76}$$

If $(W_{k_j}^T \nabla f_{k_j})_i \to 0$, then the boundedness of $\bar{D}_{k_j}$ and D_{k_j} yields a contradiction to (76). Thus, there must exist $\epsilon_2 > 0$ and a subsequence of $\{k_j\}$, again denoted by $\{k_j\}$, such that $|(W_{k_j}^T \nabla f_{k_j})_i| > \epsilon_2$. Since $\lim_j H_{k_j} s_{k_j}^{\mathsf{n}} = 0$, the definitions of $\bar{D}$ and D imply that $|(\bar{D}_{k_j} - D_{k_j})_i| \to 0$, which again leads to a contradiction of (76). Consequently, the previous assumption cannot be satisfied and (73) is proven. ∎

Using the continuity of $C(x)$, $D(x)W(x)^T \nabla f(x)$, and Theorem 7.5, we can deduce the following result.

COROLLARY 7.6. *Let the conditions of Theorem* 7.5 *be valid. If $\{x_k\}$ is a bounded sequence, then $\{x_k\}$ has a limit point satisfying the first-order KKT conditions.*

8. Global convergence to a second-order KKT point. In this section we establish global convergence to a point that satisfies the second-order necessary KKT conditions.

THEOREM 8.1. *Under Conditions C.1–C.3, the sequences of iterates generated by the trust-region interior-point SQP Algorithms* 5.1 *satisfy*

$$\liminf_k \left(\|\bar{D}_k \bar{g}_k\| + \|C_k\| + \tau_k^2 \gamma_k \right) = 0, \tag{77}$$

where γ_k is the Lagrange multiplier corresponding to the trust-region constraint; see (32), (37)*, and τ_k is the damping parameter defined in* (34).

Proof. The proof is again by contradiction. Suppose that for all k,

$$\|\bar{D}_k \bar{g}_k\| + \|C_k\| + \tau_k^2 \gamma_k > \frac{5}{3} \epsilon_{tol}. \tag{78}$$

(i) Suppose that $\|C_k\| \le \alpha' \delta_k$, where

$$\alpha' = \min \left\{ \alpha, \frac{\kappa_9 \epsilon_{tol}}{3\kappa_{11}(1+\kappa_4)} \right\} \tag{79}$$

and α satisfies (69). From the first bound on α in (69), we get

$$\|\bar{D}_k \bar{g}_k\| + \tau_k^2 \gamma_k > \frac{4}{3} \epsilon_{tol}.$$

Thus, either $\|\bar{D}_k \bar{g}_k\| > \frac{2}{3}\epsilon_{tol}$ or $\tau_k^2 \gamma_k > \frac{2}{3}\epsilon_{tol}$. In the first case we proceed exactly as in Lemmas 7.2, 7.3 and obtain

$$\begin{aligned} pred(s_k; \rho) &\ge \frac{\kappa_6}{2} \|\bar{D}_k \bar{g}_k\| \min \left\{ \kappa_7 \|\bar{D}_k \bar{g}_k\|, \kappa_8 \delta_k \right\} + \rho \left(\|C_k\|^2 - \|J_k s_k + C_k\|^2 \right) \\ &\ge \frac{\kappa_{14}}{\delta_{\max}} \delta_k^2 \end{aligned} \tag{80}$$

for every $\rho > 0$. If $\tau_k^2 \gamma_k > \frac{2}{3}\epsilon_{tol}$, then from (50), (58), (60), $\|s_k^{\mathsf{n}}\| \leq \delta_k$, and the second bound on α' given in (79), we can write

$$\begin{aligned} pred(s_k;\rho) &= q_k(s_k^{\mathsf{n}}) - q_k(s_k^{\mathsf{n}} + W_k(s_k)_u) + q_k(0) - q_k(s_k^{\mathsf{n}}) - \Delta\lambda_k^T(J_k s_k + C_k) \\ &\quad + \rho\left(\|C_k\|^2 - \|J_k s_k + C_k\|^2\right) \\ &\geq \frac{1}{2}\kappa_9 \tau_k^2 \gamma_k \delta_k^2 + \left(\frac{1}{3}\kappa_9 \epsilon_{tol}\delta_k - \kappa_{11}\|C_k\|(1+\kappa_4)\right)\delta_k \\ &\quad + \rho\left(\|C_k\|^2 - \|J_k s_k + C_k\|^2\right) \qquad (81) \\ &\geq \frac{1}{2}\kappa_9 \tau_k^2 \gamma_k \delta_k^2 + \rho\left(\|C_k\|^2 - \|J_k s_k + C_k\|^2\right) \\ &\geq \frac{\kappa_9 \epsilon_{tol}}{3}\delta_k^2 \end{aligned}$$

for every $\rho > 0$. From the two bounds (80), (81), we conclude that if $\|C_k\| \leq \alpha'\delta_k$ then the penalty parameter does not increase. See step 2.3 of Algorithms 5.1. Moreover, these two bounds on $pred(s_k;\rho_k)$ show the existence of a positive constant κ_{17} independent of k such that

$$pred(s_k;\rho_k) \geq \kappa_{17}\delta_k^2, \qquad (82)$$

provided $\|C_k\| \leq \alpha'\delta_k$.

(ii) Now we prove that $\{\rho_k\}$ is bounded. If ρ_k is increased at iteration k, then it is updated according to the rule

$$\rho_k = 2\left(\frac{q_k(s_k) - q_k(0) + \Delta\lambda_k^T(J_k s_k + C_k)}{\|C_k\|^2 - \|J_k s_k + C_k\|^2}\right) + \bar{\rho}.$$

We can write

$$\begin{aligned} \tfrac{\rho_k}{2}\left(\|C_k\|^2 - \|J_k s_k + C_k\|^2\right) &= q_k(s_k) - q_k(s_k^{\mathsf{n}}) \\ &\quad -\left(q_k(0) - q_k(s_k^{\mathsf{n}})\right) + \Delta\lambda_k^T(J_k s_k + C_k) \\ &\quad +\tfrac{\bar{\rho}}{2}\left(\|C_k\|^2 - \|J_k s_k + C_k\|^2\right). \end{aligned}$$

By applying (47) to the left-hand side and applying (50), (58), (60), and $\|s_k^{\mathsf{n}}\| \leq \delta_k$ to the right-hand side, we obtain

$$\begin{aligned} \frac{\rho_k}{2}\kappa_2\|C_k\|\min\{\kappa_3\|C_k\|,\delta_k\} &\leq \kappa_{11}(1+\kappa_4)\delta_k\|C_k\| + \frac{\bar{\rho}}{2}\left(-2(J_k^T C_k)^T s_k - \|J_k s_k\|^2\right) \\ &\leq (\kappa_{11}(1+\kappa_4) + \bar{\rho}\nu_4\kappa_4)\delta_k\|C_k\|. \qquad (83) \end{aligned}$$

If ρ_k is increased at iteration k, then, because of part (i), $\|C_k\| > \alpha'\delta_k$. Now we use this fact to establish that

$$\left(\frac{\kappa_2}{2}\min\{\kappa_3\alpha', 1\}\right)\rho_k \leq \kappa_{11}(1+\kappa_4) + \bar{\rho}\nu_4\kappa_4.$$

This proves that $\{\rho_k\}$ and $\{L_k\}$ are bounded sequences.

(iii) The next step is to prove that δ_k is bounded away from zero.

If s_{k-1} was an acceptable step, then $\delta_k \geq \delta_{\min}$; see step 2.4 in Algorithms 5.1.

If s_{k-1} was rejected, then $\delta_k \geq \kappa_5 \|s_{k-1}\|$; see (51). We consider two cases. In both cases we will use the fact that

$$1 - \eta_1 \leq \left| \frac{ared(s_{k-1}; \rho_{k-1})}{pred(s_{k-1}; \rho_{k-1})} - 1 \right|.$$

In the first case we will assume that $\|C_{k-1}\| \leq \alpha' \delta_{k-1}$. From (82) we have $pred(s_{k-1}; \rho_{k-1}) \geq \kappa_{17} \delta_{k-1}^2$. Thus, we can use $\|s_{k-1}\| \leq \kappa_4 \delta_{k-1}$ (see (50)) and (66) with k replaced by $k-1$ to obtain

$$\left| \frac{ared(s_{k-1}; \rho_{k-1})}{pred(s_{k-1}; \rho_{k-1})} - 1 \right| \leq \frac{\kappa_{13} \rho_* \left(\kappa_4^2 \delta_{k-1}^2 + \kappa_4 \alpha' \delta_{k-1}^2 \right)}{\kappa_{17} \delta_{k-1}^2} \|s_{k-1}\|.$$

This gives $\delta_k \geq \kappa_5 \|s_{k-1}\| \geq \frac{\kappa_5 (1-\eta_1) \kappa_{17}}{\kappa_{13} \rho_* (\kappa_4^2 + \alpha' \kappa_4)} \equiv \kappa_{18}$.

The other case is $\|C_{k-1}\| > \alpha' \delta_{k-1}$. In this case we get, from (47) and (67) with k replaced by $k-1$, that

$$\begin{aligned} pred(s_{k-1}; \rho_{k-1}) &\geq \tfrac{\rho_{k-1}}{2} \kappa_2 \|C_{k-1}\| \min\{\kappa_3 \|C_{k-1}\|, \delta_{k-1}\} \\ &\geq \rho_{k-1} \kappa_{19} \delta_{k-1} \|C_{k-1}\| \\ &\geq \rho_{k-1} \alpha' \kappa_{19} \delta_{k-1}^2, \end{aligned}$$

where $\kappa_{19} = \frac{\kappa_2}{2} \min\{\kappa_3 \alpha', 1\}$. Again we use $\rho_{k-1} \geq 1$ and (66) with k replaced by $k-1$, this time with the last two lower bounds on $pred(s_{k-1}; \rho_{k-1})$, and we write

$$\begin{aligned} \left| \frac{ared(s_{k-1}; \rho_{k-1})}{pred(s_{k-1}; \rho_{k-1})} - 1 \right| &\leq \frac{\kappa_{13} \rho_{k-1} \|s_{k-1}\|^3}{|pred(s_{k-1}; \rho_{k-1})|} + \frac{\kappa_{13} \rho_{k-1} \|C_{k-1}\| \, \|s_{k-1}\|^2}{|pred(s_{k-1}; \rho_{k-1})|} \\ &\leq \left(\frac{\kappa_{13} \rho_{k-1} \kappa_4^2 \delta_{k-1}^2}{\rho_{k-1} \alpha' \kappa_{19} \delta_{k-1}^2} + \frac{\kappa_{13} \rho_{k-1} \kappa_4 \delta_{k-1} \|C_{k-1}\|}{\rho_{k-1} \kappa_{19} \delta_{k-1} \|C_{k-1}\|} \right) \|s_{k-1}\|. \end{aligned}$$

Hence $\delta_k \geq \kappa_5 \|s_{k-1}\| \geq \frac{\kappa_5 (1-\eta_1) \alpha' \kappa_{19}}{\kappa_{13} (\kappa_4^2 + \alpha' \kappa_4)} \equiv \kappa_{20}$.

Combining the two cases yields

$$\delta_k \geq \delta_* = \min\{\delta_{\min}, \kappa_{18}, \kappa_{20}\}$$

for all k.

(iv) The rest of the proof consists of proving that an acceptable trial step is always found after a finite number of iterations and then concluding from this that the supposition (78) is false. The proof of these facts is exactly the proof of Theorem 7.5, where α is now α' and $\kappa_{14} \delta_*$ is replaced by $\kappa_{17} \delta_*^2$. ∎

The following result finally establishes global convergence to a point satisfying the second-order necessary KKT conditions. The proof uses ideas applied in [13, Lem. 3.8]. However, we show that convergence to a limit point satisfies the second-order necessary conditions even in the degenerate case.

THEOREM 8.2. *Let $\{x_k\}$ be a bounded sequence of iterates generated by the trust-region interior-point SQP Algorithms* 5.1 *under Conditions C.1–C.3. Then $\{x_k\}$ has a limit point x_* satisfying the first-order KKT conditions. Furthermore, x_* satisfies the second-order necessary KKT conditions.*

Proof. Consider the subsequence of $\{x_k\}$ for which the limit in (77) is zero. Since this subsequence is bounded, we can use the same arguments as in the proof of Theorem 7.5 to show that it has a convergent subsequence indexed by $\{k_j\}$ such that

$$\lim_j \left(\|\bar{D}_{k_j} \bar{g}_{k_j}\| + \|C_{k_j}\| \right) = \lim_j \left(\|D_{k_j} W_{k_j}^T \nabla f_{k_j}\| + \|C_{k_j}\| \right) = 0. \tag{84}$$

Moreover,

$$\lim_j \tau_{k_j}^2 \gamma_{k_j} = 0, \tag{85}$$

where τ_{k_j} is given by (34). Let x_* denote the limit of $\{x_{k_j}\}$. It follows from (84) and the continuity of $C(x)$ and $D(x)W(x)^T\nabla f(x)$ that x_* satisfies the first-order KKT conditions.

Next, we will prove that $\lim_j \gamma_{k_j} = 0$. First we consider the decoupled approach. Define the vector-valued function h as follows:

$$h(x)_i = \begin{cases} 1 & \text{if } \left(W(x)^T\nabla f(x)\right)_i = 0 \text{ and } \left(D(x)_{ii}\right) = 0, \\ \left(W(x)^T\nabla f(x)\right)_i & \text{otherwise,} \end{cases}$$

for all $i = 1, \ldots, n-m$. The function h is used to identify the active indices. By definition of h and since x_* satisfies the first-order KKT conditions, the implications

$$D(x_*)_{ii} = 0 \iff h(x_*)_i \neq 0, \quad i = 1, \ldots, n-m \tag{86}$$

are valid. (If x_* is nondegenerate, then $h(x_*) = W(x_*)^T\nabla f(x_*)$.) Moreover,

$$\lim_{x \to x_*} D(x)h(x) = 0. \tag{87}$$

Since $\lim_j x_{k_j} = x_*$, (86) implies the existence of $\epsilon_0 \in (0,1)$ such that

$$\min\left\{ (u_{k_j})_i - a_i,\, b_i - (u_{k_j})_i \right\} + \left|\left(h_{k_j}\right)_i\right| > 2\epsilon_0, \qquad i = 1, \ldots, n-m \tag{88}$$

for large enough j, and

$$2\epsilon_0 < \min\{b_i - a_i,\ i = 1, \ldots, n-m\}.$$

Without loss of generality, we will only consider the cases where $\tau_{k_j} \leq \sigma_{k_j} < 1$. In the following the index i will be the index defining τ_{k_j} in (34). (The index i is really i_j but we drop the j from i_j to alleviate the notation.) We also assume that j is large enough such that

$$\left|\left(\bar{D}_{k_j}^2 h_{k_j}\right)_i\right| < \epsilon_0^2; \tag{89}$$

cf. (87).

Multiplying both sides of (33) by $\bar{D}_{k_j}^2$ gives

$$\left(E_{k_j} + \gamma_{k_j} I_{n-m}\right) o_{k_j}^{\mathsf{d}} = \bar{D}_{k_j}^2 \left(-\bar{g}_{k_j} - W_{k_j}^T H_{k_j} W_{k_j} o_{k_j}^{\mathsf{d}} \right),$$

which in turn implies

$$\gamma_{k_j} |(o_{k_j}^{\mathsf{d}})_i| \leq (\bar{D}_{k_j}^2)_{ii} \left| \left(-\bar{g}_{k_j} - W_{k_j}^T H_{k_j} W_{k_j} o_{k_j}^{\mathsf{d}} \right)_i \right|. \tag{90}$$

Also, Assumption A.6 implies $\|o_{k_j}^{\mathsf{d}}\| \leq \nu_9 \delta_{k_j} \leq \nu_9 \delta_{\max}$. From this, (49), and Assumptions A.3–A.4, we can write

$$\frac{1}{(o_{k_j}^{\mathsf{d}})_i} \geq \frac{\gamma_{k_j}}{\kappa_{21}(\bar{D}_{k_j})_{ii}^2} \tag{91}$$

for some κ_{21} independent of k. Now we distinguish between two cases.

In the first case we consider $|(h_{k_j})_i| \leq \epsilon_0$ and appeal to (88) to get $\min\{(u_{k_j})_i - a_i,\, b_i - (u_{k_j})_i\} > \epsilon_0$. Thus, from (91) and the definition (34) of τ_{k_j} we obtain

$$\tau_{k_j} \geq \frac{\sigma_{k_j}\gamma_{k_j}\epsilon_0}{\kappa_{21}(\bar{D}_{k_j})_{ii}^2}. \tag{92}$$

Now we analyze the case $|(h_{k_j})_i| > \epsilon_0$. Two possibilities can occur.

(i) The first possibility is that the value of the numerator defining τ_{k_j} is equal to $(\bar{D}_{k_j})_{ii}^2$. In this situation, (91) immediately implies

$$\tau_{k_j} \geq \frac{\sigma_{k_j}\gamma_{k_j}}{\kappa_{21}}. \tag{93}$$

(ii) The other possibility is that the value of the numerator defining τ_{k_j} is not equal to $(\bar{D}_{k_j})_{ii}^2$. In this case we have from (89) that $(\bar{D}_{k_j})_{ii}^2 < \epsilon_0$ and, since $b_i - a_i > 2\epsilon_0$, the numerator in the definition (34) of τ_{k_j} is bigger than ϵ_0. Thus,

$$\tau_{k_j} \geq \frac{\sigma_{k_j}\gamma_{k_j}\epsilon_0}{\kappa_{21}(\bar{D}_{k_j})_{ii}^2}. \tag{94}$$

Using (85), (92), (93), (94), $\sigma_{k_j} \geq \sigma$, and the boundedness of $\bar{D}_{k_j}$ this proves that

$$\lim_j \gamma_{k_j} = 0.$$

By (32) we know that

$$\bar{D}_{k_j} W_{k_j}^T H_{k_j} W_{k_j} \bar{D}_{k_j} + E_{k_j} + \gamma_{k_j} I_{n-m}$$

is positive semidefinite. Hence condition (84), the continuity of $W(x)^T \nabla_{xx}^2 \ell(x,\lambda) W(x)$, and the limits $\lim_j \|W_{k_j}^T H_{k_j} s_{k_j}^{\mathsf{n}}\| = 0$ and $\lim_j \gamma_{k_j} = 0$ imply that the limit of the principal submatrix of $W_{k_j}^T H_{k_j} W_{k_j}$ corresponding to indices l such that $a_l < (u_*)_l < b_l$ is positive semidefinite. Hence the second-order necessary KKT conditions are satisfied at x_*. This completes the proof for the decoupled approach.

The proof for the coupled trust-region approach differs only from the proof for the decoupled approach in the use of equations (37) and (38) and in the use of $\|W_{k_j} o_{k_j}^{\mathsf{c}}\| \leq (1+\nu_9)\delta_{\max}$ to bound the right-hand side of inequality (90). ∎

Remark 8.1. The global convergence results of sections 7 and 8 hold true if the quadratic $\Psi_k(s_u)$ is redefined as $\Psi_k(s_u) = q_k(s_k^{\mathsf{n}} + W_k s_u)$ (see (24) and (25)) without the Newton augmentation term $\frac{1}{2} s_u^T (E_k \bar{D}_k^{-2}) s_u$. They are valid also if the matrices D_k and $\bar{D}_k$ are redefined, respectively, as D_k^p and $\bar{D}_k^p$ with $p \geq 1$.

9. Local rate of convergence. We will now analyze the local behavior of Algorithms 5.1 under Conditions C.1, C.3, and C.4. We start by looking at the behavior of the trust radius close to a nondegenerate point that satisfies the second-order sufficient KKT conditions. For this purpose we require the following lemma.

LEMMA 9.1. *Under Condition C.1, the quasi-normal component satisfies*

$$\|s_k^{\mathsf{n}}\| \leq \kappa_{22}\|s_k\|, \tag{95}$$

where κ_{22} is positive and independent of the iteration counter k.

Proof. From $s_k = s_k^{\mathsf{n}} + W_k(s_k)_u$, we obtain

$$\|s_k^{\mathsf{n}}\| \leq \|s_k\| + \|W_k\|\,\|(s_k)_u\|.$$

But since $\|s_k\|^2 = \|(s_k)_y\|^2 + \|(s_k)_u\|^2$, we use Assumption A.4 to obtain

$$\|s_k^{\mathsf{n}}\| \leq (1+\nu_6)\,\|s_k\|,$$

and (95) holds with $\kappa_{22} = 1+\nu_6$. ∎

THEOREM 9.2. *Let $\{x_k\}$ be a sequence of iterates generated by the trust-region interior-point SQP Algorithms 5.1 under Conditions C.1 and C.3. If x_k converges to a nondegenerate point x_* satisfying the second-order sufficient KKT conditions, then δ_k is uniformly bounded away from zero and eventually all the iterations will be successful.*

Proof. It follows from $\lim_k x_k = x_*$ and $C(x_*) = 0$ that $\lim_k \|C_k\| = 0$. This fact, condition (21), and Assumptions A.3–A.4, together imply

$$\lim_k \|W_k^T H_k s_k^{\mathsf{n}}\| = 0.$$

Since x_k converges to a nondegenerate point that satisfies the second-order sufficient KKT conditions and $\lim_k \|W_k^T H_k s_k^{\mathsf{n}}\| = 0$, there exists a $\bar\gamma > 0$ such that the smallest eigenvalue of $\bar D_k W_k^T H_k W_k \bar D_k + E_k$ is greater than $\bar\gamma$ for k sufficiently large.

First we will proof that $\{\rho_k\}$ is a bounded sequence. Since $\Psi_k(0) - \Psi_k((s_k)_u) \geq 0$, we obtain

$$\begin{aligned}\tfrac{1}{2}(\bar D_k^{-1}(s_k)_u)^T\left(\bar D_k W_k^T H_k W_k \bar D_k + E_k\right)(\bar D_k^{-1}(s_k)_u) &\leq -(\bar D_k^{-1}(s_k)_u)^T(\bar D_k \bar g_k)\\ &\leq \|\bar D_k^{-1}(s_k)_u\|\,\|\bar D_k \bar g_k\|,\end{aligned}$$

which, by using the upper bounds on W_k and $\bar D_k$ given by Assumptions A.4 and A.6, implies

$$\|s_k^{\mathsf{t}}\| = \|W_k(s_k)_u\| \leq \frac{2\nu_6\nu_9}{\bar\gamma}\|\bar D_k \bar g_k\|. \tag{96}$$

Using (52) and (96), we find that

$$\begin{aligned} q_k(s_k^{\mathsf{n}}) - q_k(s_k^{\mathsf{n}} + W_k(s_k)_u) &\geq \kappa_6\|\bar D_k\bar g_k\|\min\{\kappa_7\|\bar D_k\bar g_k\|, \kappa_8\delta_k\}\\ &\geq \kappa_{23}\|s_k^{\mathsf{t}}\|^2, \end{aligned} \tag{97}$$

where $\kappa_{23} = \frac{\kappa_6\bar\gamma}{2\nu_6\nu_9}\min\{\frac{\kappa_7\bar\gamma}{2\nu_6\nu_9}, \frac{\kappa_8}{\nu_6\nu_9}, \frac{\kappa_8}{1+\nu_9}\}$ accounts for the decoupled and coupled cases.

Next, we prove that if $\|C_k\| \leq \alpha''\|s_k\|$, where α'' will be defined later, then the penalty parameter does not need to be increased. From (21) and $\|C_k\| \leq \alpha''\|s_k\|$, we get

$$\begin{aligned}\|s_k\|^2 \leq \left(\|s_k^{\mathsf{n}}\| + \|s_k^{\mathsf{t}}\|\right)^2 &\leq 2\|s_k^{\mathsf{n}}\|^2 + 2\|s_k^{\mathsf{t}}\|^2\\ &\leq 2\alpha''\kappa_1^2\|C_k\|\,\|s_k\| + 2\|s_k^{\mathsf{t}}\|^2.\end{aligned}$$

This estimate, (21), (60), (97), and $\|C_k\| \le \alpha''\|s_k\|$ yield

$$\begin{aligned} pred(s_k;\rho) &= q_k(s_k^{\mathsf{n}}) - q_k(s_k^{\mathsf{n}} + W_k(s_k)_u) + q_k(0) - q_k(s_k^{\mathsf{n}}) - \Delta\lambda_k^T(J_k s_k + C_k) \\ &\quad + \rho\left(\|C_k\|^2 - \|J_k s_k + C_k\|^2\right) \\ &\ge \frac{1}{4}\kappa_{23}\|s_k\|^2 + \left(\frac{1}{4}\kappa_{23}\|s_k\| - (\alpha''\kappa_1^2\kappa_{23} + \kappa_{11}(\alpha''\kappa_1 + 1))\|C_k\|\right)\|s_k\| \\ &\quad + \rho\left(\|C_k\|^2 - \|J_k s_k + C_k\|^2\right), \end{aligned} \tag{98}$$

for every $\rho > 0$. If $\|C_k\| \le \alpha''\|s_k\|$, where α'' satisfies

$$(4\kappa_{11})\ \alpha'' + \left(4\kappa_1^2\kappa_{23} + 4\kappa_1\kappa_{11}\right)\ (\alpha'')^2 \le \kappa_{23}\,, \tag{99}$$

then we set $\rho = \rho_{k-1}$ in (98) and deduce that the penalty parameter does not need to be increased. See step 2.3 of Algorithms 5.1. Hence if ρ_k is increased, then the inequality $\|C_k\| > \alpha''\|s_k\|$ must hold, and we can proceed as in Theorem 8.1, equation (83), and write

$$\frac{\rho_k}{2}\kappa_2\|C_k\| \min\left\{\kappa_3\|C_k\|, \frac{1}{\kappa_4}\|s_k\|\right\} \le (\kappa_{11}(\kappa_{22}+1) + \bar\rho\nu_4)\|s_k\|\,\|C_k\|$$

(here we used inequality (95)), which in turn implies

$$\left(\frac{\kappa_2}{2}\min\left\{\kappa_3\alpha'', \frac{1}{\kappa_4}\right\}\right)\rho_k \le \kappa_{11}(\kappa_{22}+1) + \bar\rho\nu_4.$$

This gives the uniform boundedness of the penalty parameter

$$\rho_k \le \rho_*$$

for all k.

Given the boundedness of $\{\rho_k\}$ we can complete the proof of the theorem. If $\|C_k\| > \alpha''\|s_k\|$, where α'' satisfies (99), then from (47) and (67) we find that

$$pred(s_k;\rho_k) \ge \rho_k\frac{\kappa_2}{2}\|C_k\|\min\{\kappa_3\|C_k\|, \delta_k\} \ge \rho_k\kappa_{24}\|s_k\|^2, \tag{100}$$

where $\kappa_{24} = \frac{\kappa_2\alpha''}{2}\min\{\kappa_3\alpha'', \frac{1}{\kappa_4}\}$. In this case it follows from (66) and (100) that

$$\left|\frac{ared(s_k;\rho_k)}{pred(s_k;\rho_k)} - 1\right| \le \frac{\kappa_{13}}{\kappa_{24}}\left(\|s_k\| + \|C_k\|\right). \tag{101}$$

Now, suppose that $\|C_k\| \le \alpha''\|s_k\|$. From (98) with $\rho = \rho_k$ we obtain $pred(s_k;\rho_k) \ge \frac{\kappa_{23}}{4}\|s_k\|^2$. Now we use (66) and $\rho_k \le \rho_*$ to get

$$\left|\frac{ared(s_k;\rho_k)}{pred(s_k;\rho_k)} - 1\right| \le \frac{4\kappa_{13}\rho_*}{\kappa_{23}}\left(\|s_k\| + \|C_k\|\right). \tag{102}$$

Finally from (101), (102), $\lim_k x_k = x_*$, and $\lim_k \|C_k\| = 0$, we get

$$\lim_k \frac{ared(s_k;\rho_k)}{pred(s_k;\rho_k)} = 1,$$

which by the rules for updating the trust radius given in step 2.4 of Algorithms 5.1, shows that δ_k is uniformly bounded away from zero. ∎

We use the following straightforward globalization of the quasi-normal component s_k^{n} of the Newton step given in (44). The new quasi-normal component is given by

$$s_k^{\mathsf{n}} = \begin{pmatrix} -\xi_k C_y(x_k)^{-1} C_k \\ 0 \end{pmatrix}, \tag{103}$$

where

$$\xi_k = \begin{cases} 1 & \text{if } \|C_y(x_k)^{-1} C_k\| \le \delta_k, \\ \frac{\delta_k}{\|C_y(x_k)^{-1} C_k\|} & \text{otherwise.} \end{cases} \tag{104}$$

Before we state the q-quadratic rate of convergence, we prove the following important result.

LEMMA 9.3. *The quasi–normal component* (103) *satisfies conditions* (18), (21), *and* (22) *for some positive* κ_1, κ_2, *and* κ_3 *independent of* k.

Proof. It is obvious that (18) holds. Condition (21) is a direct consequence of the condition (22). In fact, using $\|C_y(x_k)(s_k^{\mathsf{n}})_y + C_k\| \le \|C_k\|$ and the boundedness of $\{C_y(x_k)^{-1}\}$, we find that

$$\begin{aligned} \|s_k^{\mathsf{n}}\| &= \|s_k^{\mathsf{n}} + C_y(x_k)^{-1} C_k - C_y(x_k)^{-1} C_k\| \\ &\le \|C_y(x_k)^{-1}\| \Big(\|C_y(x_k)(s_k^{\mathsf{n}})_y + C_k\| + \|C_k\| \Big) \quad \le \quad 2\nu_6 \|C_k\|. \end{aligned} \tag{105}$$

So, let us prove (22). A simple manipulation shows that

$$\begin{aligned} \|C_k\|^2 - \|C_y(x_k)(s_k^{\mathsf{n}})_y + C_k\|^2 &= \|C_k\|^2 - \| - \xi_k C_y(x_k) C_y(x_k)^{-1} C_k + C_k\|^2 \\ &= \|C_k\|^2 - \Big((1-\xi_k)\|C_k\| \Big)^2 \\ &= \xi_k (2 - \xi_k) \|C_k\|^2 \ \ge \ \xi_k \|C_k\|^2. \end{aligned}$$

We need to consider two cases. If $\xi_k = 1$, then

$$\|C_k\|^2 - \|C_y(x_k)(s_k^{\mathsf{n}})_y + C_k\|^2 \ge \|C_k\| \min\{\|C_k\|, \delta_k\}.$$

Otherwise $\xi_k = \frac{\delta_k}{\|C_y(x_k)^{-1} C_k\|}$. In this case we get

$$\|C_k\|^2 - \|C_y(x_k)(s_k^{\mathsf{n}})_y + C_k\|^2 \ge \frac{1}{\nu_6} \|C_k\| \, \delta_k \ge \frac{1}{\nu_6} \|C_k\| \min\{\|C_k\|, \delta_k\}.$$

Thus, the result holds with $\kappa_2 = \min\{1, \frac{1}{\nu_6}\}$ and $\kappa_3 = 1$. ∎

COROLLARY 9.4. *Let* $\{x_k\}$ *be a sequence of iterates generated by the trust-region interior-point SQP Algorithms* 5.1 *under Conditions C.1, C.3, and C.4. If* x_k *converges to a nondegenerate point* x_* *satisfying the second-order sufficient KKT conditions, then* x_k *converges q-quadratically.*

Proof. We start by showing that $|\tau_k^{\mathsf{N}} - 1|$ is $\mathcal{O}(\|x_k - x_*\|)$, where τ_k^{N} is given by (46). Since $\lim_k \|W_k^T H_k s_k^{\mathsf{n}}\| = 0$, we have that $|\frac{\tau_k^{\mathsf{N}}}{\sigma_k} - 1|$ is $\mathcal{O}(\|(s_k^{\mathsf{N}})_u\|)$ (see [12, Eq. (6.4) and Lem. 12]). Also since by Condition C.4 $|\sigma_k - 1|$ is $\mathcal{O}(\|\bar{D}_k \bar{g}_k\|)$ and $\bar{D}_k \bar{g}_k$ is $\mathcal{O}(\|(s_k^{\mathsf{N}})_u\|)$ (see (43)), we can see that $|\sigma_k - 1|$ is also $\mathcal{O}(\|(s_k^{\mathsf{N}})_u\|)$. Furthermore,

$$|\tau_k^{\mathsf{N}} - 1| \le \sigma_k \left| \frac{\tau_k^{\mathsf{N}}}{\sigma_k} - 1 \right| + |\sigma_k - 1|.$$

Hence $|\tau_k^{\mathsf{N}}-1|$ is $\mathcal{O}(\|(s_k^{\mathsf{N}})_u\|)$. But $(s_k^{\mathsf{N}})_u$ is $\mathcal{O}(\|x_k+s_k^{\mathsf{n}}-x_*\|)$ and s_k^{n} is $\mathcal{O}(\|x_k-x_*\|)$ and this shows that $|\tau_k^{\mathsf{N}}-1|$ is $\mathcal{O}(\|x_k-x_*\|)$.

We need to prove that Condition C.4 does not conflict with Condition C.1 so that Theorem 9.2 can be applied. In other words, we need to show that the decrease conditions given in Condition C.1 hold for the Newton damped step (45) whenever it is taken. In Lemma 9.3 we showed that the quasi-normal component s_k^{n} given in (103) satisfies (18), (21), and (22). From Condition C.4, s_k^{n} given by (44) is used when it coincides with the s_k^{n} given by (103). Thus s_k^{n} given by (44) satisfies also (18), (21), and (22). It remains to prove that $\tau_k^{\mathsf{N}}(s_k^{\mathsf{N}})_u$ satisfies the Cauchy decrease condition (30) ((31) for the coupled approach). This is indeed the case, since

$$\begin{aligned}
&\Psi_k(0)-\Psi_k(\tau_k^{\mathsf{N}}(s_k^{\mathsf{N}})_u)\\
&\quad\geq -\tau_k^{\mathsf{N}}\bar{g}_k^T(s_k^{\mathsf{N}})_u-\frac{1}{2}(\tau_k^{\mathsf{N}})^2((s_k^{\mathsf{N}})_u)^T\left(W_k^TH_kW_k+E_k\bar{D}_k^{-2}\right)((s_k^{\mathsf{N}})_u)\\
&\quad\geq \tau_k^{\mathsf{N}}\left(-\bar{g}_k^T(s_k^{\mathsf{N}})_u-\frac{1}{2}((s_k^{\mathsf{N}})_u)^T\left(W_k^TH_kW_k+E_k\bar{D}_k^{-2}\right)((s_k^{\mathsf{N}})_u)\right)\\
&\quad\geq \tau_k^{\mathsf{N}}\left(\Psi_k(0)-\Psi_k(c_k^{\mathsf{d}})\right),
\end{aligned}$$

and $|\tau_k^{\mathsf{N}}-1|$ is $\mathcal{O}(\|x_k-x_*\|)$.

Now we need to show that eventually s_k is given by (45). Since $\{x_k\}$ converges to a nondegenerate point satisfying the second-order sufficient KKT conditions, $(s_k^{\mathsf{N}})_u$ exists for k sufficiently large. Furthermore, $(s_k^{\mathsf{n}})_y=-C_y(x_k)^{-1}C_k$ for k large enough because $\lim_k\|C_y(x_k)^{-1}C_k\|=0$, and from Theorem 9.2, δ_k is eventually bounded away from zero. Using a similar argument we see that $\tau_k^{\mathsf{N}}(s_k^{\mathsf{N}})_u$ is inside the trust region (27) for the decoupled approach or (29) for the coupled approach. So, from Condition C.4 we conclude that there exists a positive integer $\bar{k}$ such that s_k is given by (45) for $k\geq\bar{k}$.

Using the fact that $(s_k^{\mathsf{N}})_u$ is $\mathcal{O}(\|x_k-x_*\|)$, we conclude that $\tau_k^{\mathsf{N}}(s_k^{\mathsf{N}})_u-(s_k^{\mathsf{N}})_u$ is $\mathcal{O}(\|x_k-x_*\|^2)$. Thus,

$$s_k-s_k^{\mathsf{N}}=\begin{pmatrix} s_k^{\mathsf{n}}-C_y(x_k)^{-1}C_u(x_k)\tau_k^{\mathsf{N}}(s_k^{\mathsf{N}})_u\\ \tau_k^{\mathsf{N}}(s_k^{\mathsf{N}})_u\end{pmatrix}-\begin{pmatrix} s_k^{\mathsf{n}}-C_y(x_k)^{-1}C_u(x_k)(s_k^{\mathsf{N}})_u\\ (s_k^{\mathsf{N}})_u\end{pmatrix}$$

is $\mathcal{O}(\|x_k-x_*\|^2)$. This completes the proof since s_k^{N} can be seen as a Newton step on a given vector function of the type (17). This function vanishes at x_* and is continuously differentiable with Lipschitz continuous derivatives and a nonsingular Jacobian matrix in an open neighborhood of x_*. See the discussion at the end of section 4. Thus, the q-quadratic rate of convergence follows from [17, Thm. 5.2.1] and from the fact that $s_k-s_k^{\mathsf{N}}$ is $\mathcal{O}(\|x_k-x_*\|^2)$. ∎

10. Trial steps and multiplier estimates. When we described the trust-region interior-point SQP algorithms, we deferred the practical computation of the quasi-normal and tangential components and of the multiplier estimates. In the following sections we address these issues.

10.1. Computation of the quasi-normal component. The quasi-normal component s_k^{n} is an approximate solution of the trust-region subproblem

$$\begin{aligned}&\text{minimize } \frac{1}{2}\|C_y(x_k)(s^{\mathsf{n}})_y+C_k\|^2\\ &\text{subject to } \|(s^{\mathsf{n}})_y\|\leq\delta_k,\end{aligned}\tag{106}$$

and it is required for global convergence to a point that satisfies the necessary KKT conditions to satisfy conditions (18), (21), and (22). As we saw in equation (105) of the proof of Lemma 9.3, property (21) is a consequence of (22). Whether property (22) holds depends on the way in which the quasi-normal component is computed. We will show below that (22) is satisfied by many reasonable ways to compute s_k^{n}.

There are various ways to compute the quasi-normal component s_k^{n} for large scale problems. For example, one can use the conjugate-gradient method as suggested in [61] and [63], or one can use the Lanczos bidiagonalization as described in [26]. Both methods compute an approximate minimizer to the least squares functional in (106) from a subspace which contains its negative gradient $-C_y(x_k)^T C_k$. Thus, the components s_k^{n} generated by these methods satisfy $||s_k^{\mathsf{n}}|| \leq \delta_k$ and

$$\frac{1}{2}\|C_y(x_k)(s_k^{\mathsf{n}})_y + C_k\|^2 \leq \min\left\{\frac{1}{2}\|C_y(x_k)s + C_k\|^2 : s \in span\{-C_y(x_k)^T C_k\}, \|s\| \leq \delta_k\right\}.$$

We can appeal to a classical result due to Powell (see [52, Thm. 4], [45, Lem. 4.8]) to show that

$$\|C_k\|^2 - \|C_y(x_k)(s_k^{\mathsf{n}})_y + C_k\|^2 \geq \frac{1}{2}\|C_y(x_k)^T C_k\| \min\left\{\frac{\|C_y(x_k)^T C_k\|}{\|C_y(x_k)^T C_y(x_k)\|}, \delta_k\right\}.$$

Now one can use the fact that $\{C_y(x_k)\}$ and $\{C_y(x_k)^{-T}\}$ are bounded and can write

$$\|C_k\|^2 - \|C_y(x_k)(s_k^{\mathsf{n}})_y + C_k\|^2 \geq \kappa_2\|C_k\| \min\{\kappa_3\|C_k\|, \delta_k\},$$

where κ_2 and κ_3 are positive and do not depend on k.

An alternative to the previous procedures is to compute the solution of $C_y(x_k)s = -C(x_k)$ and to scale this solution back into the trust region (see (103)). In Lemma 9.3, we proved that (103) satisfies conditions (18), (21), and (22).

10.2. Computation of the tangential component. In this section we show how to derive conjugate-gradient algorithms to compute $(s_k)_u$. Other practical algorithms to compute trial steps for box-constrained minimization trust-region subproblems are introduced in [7] using three-dimensional subspace approximations and conjugate gradients.

Let us consider first the decoupled trust-region approach given in section 5.2.1. If we ignore the bound constraints for the moment, we can apply the conjugate-gradient algorithm proposed by Steihaug [61] and Toint [63] to solve the problem

$$\begin{aligned} &\text{minimize } \Psi_k(s_u) \\ &\text{subject to } \|\bar{D}_k^{-1} s_u\| \leq \delta_k. \end{aligned}$$

However, we also need to incorporate the constraints

$$\sigma_k(a - u_k) \leq s_u \leq \sigma_k(b - u_k).$$

This leads to the following algorithm:

ALGORITHM 10.1 (computation of $s_k = s_k^{\mathsf{n}} + W_k(s_k)_u$ (decoupled approach)).

1. Set $s_u^0 = 0$, $r_0 = -\bar{g}_k = -W_k^T \nabla q_k(s_k^{\mathsf{n}})$, $q_0 = \bar{D}_k^2 r_0$, $d_0 = q_0$, and $\epsilon > 0$.
2. For $i = 0, 1, 2, \ldots$ do
   2.1. Compute $\gamma_i = \frac{r_i^T q_i}{d_i^T (W_k^T H_k W_k + E_k \bar{D}_k^{-2}) d_i}$.

2.2. Compute
$$\tau_i = \max\{\tau > 0 \quad : \quad \|\bar{D}_k^{-1}(s_u^i + \tau d_i)\| \le \delta_k, \\ \sigma_k(a - u_k) \le s_u^i + \tau d_i \le \sigma_k(b - u_k)\}.$$
2.3. If $\gamma_i \le 0$, or if $\gamma_i > \tau_i$, then set $(s_k)_u = s_u^i + \tau_i d_i$, where τ_i is given as in 2.2 and go to 3; otherwise set $s_u^{i+1} = s_u^i + \gamma_i d_i$.
2.4. Update the residuals: $r_{i+1} = r_i - \gamma_i(W_k^T H_k W_k + E_k \bar{D}_k^{-2})d_i$ and $q_{i+1} = \bar{D}_k^2 r_{i+1}$.
2.5. Check truncation criteria: if $\sqrt{\frac{r_{i+1}^T q_{i+1}}{r_0^T q_0}} \le \epsilon$, set $(s_k)_u = s_u^{i+1}$ and go to 3.
2.6. Compute $\alpha_i = \frac{r_{i+1}^T q_{i+1}}{r_i^T q_i}$ and set $d_{i+1} = q_{i+1} + \alpha_i d_i$.
3. Compute $s_k = s_k^{\mathsf{n}} + W_k(s_k)_u$ and stop.

Step 2 iterates entirely in the vector space of the u variables. After the u component of the step s_k has been computed, step 3 finds its y component. The decoupled approach allows an efficient use of an approximation $\widehat{H}_k$ to the reduced Hessian $W_k^T \nabla_{xx}^2 \ell_k W_k$. In this case, only two linear systems are required, one with $C_y(x_k)^T$ in step 1 to compute $\bar{g}_k$ and the other with $C_y(x_k)$ in step 3 to compute $W_k(s_k)_u$. If the Hessian $\nabla_{xx}^2 \ell_k$ is being approximated, then the total number of linear systems is $2I(k) + 2$, where $I(k)$ is the number of conjugate-gradient iterations.

One can transform this algorithm to work in the whole space rather than in the reduced space by considering the coupled trust-region approach given in section 5.2.2. This alternative is presented below.

ALGORITHM 10.2 (computation of $s_k = s_k^{\mathsf{n}} + W_k(s_k)_u$ (coupled approach)).
1. Set $s^0 = 0$, $r_0 = -\bar{g}_k = -W_k^T \nabla q_k(s_k^{\mathsf{n}})$, $q_0 = \bar{D}_k^2 r_0$, $d_0 = W_k q_0$, and $\epsilon > 0$.
2. For $i = 0, 1, 2, \ldots$ do
2.1. Compute $\gamma_i = \frac{r_i^T q_i}{d_i^T H_k d_i + (d_i)_u^T E_k \bar{D}_k^{-2}(d_i)_u}$.
2.2. Compute
$$\tau_i = \max\left\{\tau > 0 \quad : \quad \left\| \begin{pmatrix} -C_y(x_k)^{-1} C_u(x_k)\tau(d_i)_u \\ \bar{D}_k^{-1}\tau(d_i)_u \end{pmatrix} \right\| \le \delta_k, \right. \\ \left. \sigma_k(a - u_k) \le s_u^i + \tau(d_i)_u \le \sigma_k(b - u_k) \right\}.$$
2.3. If $\gamma_i \le 0$, or if $\gamma_i > \tau_i$, then $s_k^{\mathsf{t}} = s^i + \tau_i d_i$, where τ_i is given as in 2.2, and go to 3; otherwise set $s^{i+1} = s^i + \gamma_i d_i$.
2.4. Update the residuals: $r_{i+1} = r_i - \gamma_i \left(W_k^T H_k d_i + E_k \bar{D}_k^{-2}(d_i)_u\right)$ and $q_{i+1} = \bar{D}_k^2 r_{i+1}$.
2.5. Check truncation criteria: if $\sqrt{\frac{r_{i+1}^T q_{i+1}}{r_0^T q_0}} \le \epsilon$, set $s_k^{\mathsf{t}} = s^{i+1}$ and go to 3.
2.6. Compute $\alpha_i = \frac{r_{i+1}^T q_{i+1}}{r_i^T q_i}$ and set $d_{i+1} = W_k(q_{i+1} + \alpha_i d_i)$.
3. Compute $s_k = s_k^{\mathsf{n}} + s_k^{\mathsf{t}}$ and stop.

Note that in step 2 both the y and the u components of the tangential component are being computed. The coupled approach is suitable particularly when an approximation H_k to the full Hessian $\nabla_{xx}^2 \ell_k$ is used. The coupled approach can be used also with an approximation $\widehat{H}_k$ to the reduced Hessian $W_k^T \nabla_{xx}^2 \ell_k W_k$. In this case, we consider H_k that is given by (41) and use the equalities (42) to compute the terms involving H_k in Algorithm 10.2. If the Hessian $\nabla_{xx}^2 \ell_k$ is approximated, the total number of linear systems is $2I(k) + 2$, where $I(k)$ is the number of conjugate-

gradient iterations. If the reduced Hessian $W_k^T \nabla_{xx}^2 \ell_k W_k$ is approximated, this number is $I(k) + 2$.

Two final important remarks are in order.

Remark 10.1. If $W_k^T W_k$ was included as a preconditioner in Algorithm 10.2, then the conjugate-gradient iterates would monotonically increase in the norm $\|W_k \cdot\|$. Dropping this preconditioner means that the conjugate-gradient iterates do not necessarily increase in this norm (see [61]). As a result, if the quasi-Newton step is inside the trust region, Algorithm 10.2 can terminate prematurely by stopping at the boundary of the trust region.

Remark 10.2. Since the conjugate-gradient Algorithms 10.1 and 10.2 start by minimizing the quadratic function $\Psi_k(s_u)$ along the direction $-\bar{D}_k^2 \bar{g}_k$, it is quite clear that they produce reduced tangential components $(s_k)_u$ that satisfy (30) and (31), respectively, with $\beta_1^{\mathsf{d}} = \beta_1^{\mathsf{c}} = 1$.

10.3. Multiplier estimates. A convenient estimate for the Lagrange multipliers is the adjoint update

$$\lambda_k = -C_y(x_k)^{-T} \nabla_y f_k, \tag{107}$$

which we use after each successful step. However, we also consider the following update:

$$\lambda_{k+1} = -C_y(x_k)^{-T} \nabla_y q_k(s_k^{\mathsf{n}}) = -C_y(x_k)^{-T} \left((H_k s_k^{\mathsf{n}})_y + \nabla_y f_k \right). \tag{108}$$

Here the use of (108) instead of

$$\lambda_{k+1} = -C_y(x_k + s_k)^{-T} \nabla_y f(x_k + s_k), \tag{109}$$

might be justified, since we obtain (108) without any further cost from the first iteration of any of the conjugate-gradient algorithms described above. The updates (107), (108), and (109) satisfy the requirement given by A.4 needed to prove global convergence to a first-order KKT point.

11. Numerical example. A typical application that has the structure described in this paper is the control of a heating process. In this section we introduce a simplified model discussed in [8] for the heating of a probe in a kiln. The temperature $y(x,t)$ inside the probe is governed by a nonlinear partial differential equation. The spatial domain is given by $(0,1)$. The boundary $x = 1$ is the inside of the probe and $x = 0$ is the boundary of the probe.

The goal is to control the heating process in such a way that the temperature inside the probe follows a certain desired temperature profile $y_d(t)$. The control $u(t)$ acts on the boundary $x = 0$. The problem can be formulated as follows.

$$\text{minimize } \frac{1}{2} \int_0^T [(y(1,t) - y_d(t))^2 + \gamma u^2(t)] dt \tag{110}$$

subject to

$$\begin{aligned}
\tau(y(x,t)) \tfrac{\partial y}{\partial t}(x,t) - \partial_x(\kappa(y(x,t)) \partial_x y(x,t)) &= q(x,t), \quad (x,t) \in (0,1) \times (0,T), \\
\kappa(y(0,t)) \partial_x y(0,t) &= g[y(0,t) - u(t)], \quad t \in (0,T), \\
\kappa(y(1,t)) \partial_x y(1,t) &= 0, \quad t \in (0,T), \\
y(x,0) &= y_0(x), \quad x \in (0,1), \\
u_{low} \leq \quad u \quad &\leq u_{upp},
\end{aligned}$$

where $y \in L^2(0,T;H^1(0,1))$ and $u \in L^2(0,T)$. The functions $\tau : \mathbb{R} \to \mathbb{R}$ and $\kappa : \mathbb{R} \to \mathbb{R}$ denote the specific heat capacity and the heat conduction, respectively, y_0 is the initial temperature distribution, q is the source term, g is a given scalar, and γ is a regularization parameter. Here $u_{low}, u_{upp} \in L^\infty(0,T)$ are given functions.

If the partial differential equation and the integral are discretized, we obtain an optimization problem of the form (1). The discretization uses finite elements and was introduced in [8] (see also [29] and [39]). The spatial domain $(0,1)$ is divided into N_x subintervals of equidistant length, and the spatial discretization is done using piecewise linear finite elements. The time discretization is performed by partitioning the interval $[0,T]$ into N_t equidistant subintervals. Then the backward Euler method is used to approximate the state space in time, and piecewise constant functions are used to approximate the control space. This leads to a discretized problem with dimension $n = N_t(N_x + 1) + N_t$ and $m = N_t(N_x + 1)$. Under the assumptions on the coefficient functions κ and τ stated in [8] and [39] which guarantee the well posedness of the infinite-dimensional problem, it is shown in [39] that the constraints $C(y,u)$ of the discretized problem satisfy the assumptions A.3 and A.5, provided the discretization parameters N_x and N_t are chosen appropriately. For more details we refer to the comprehensive treatments in [8] and [39].

The algorithms studied in this paper have been implemented in `FORTRAN 77`. The resulting software package TRICE (trust-region interior-point SQP algorithms for optimal control and engineering design problems) is available via the internet [16].

We use the formula (103) to compute the quasi-normal component, and use Algorithms 10.1 and 10.2 to calculate the tangential component. The numerical test computations were done on a Sun Sparcstation 10 in double precision. These results demonstrate the effectiveness of the algorithms.

With this discretization scheme, $C_y(x)$ is a block bidiagonal matrix with tridiagonal blocks. Hence linear systems with $C_y(x)$ and $C_y(x)^T$ can be solved efficiently by block forward substitution or block backward substitution, respectively. In each substitution step, only a small system with tridiagonal system has to be solved. In the implementation we use the LINPACK subroutine DGTSL to solve the tridiagonal systems. Notice that direct factorizations are only applied to the small $(N_x+1)\times(N_x+1)$ tridiagonal subblocks of $C_y(x)$ but not to the entire Jacobian matrix $(C_y(x)\ C_u(x))$. See also [39].

As we pointed out in section 1, the inner products and norms used in the trust-region interior-point SQP algorithms are not necessarily the Euclidean ones. In our implementation [16], we call subroutines to calculate the inner products $\langle y^1, y^2\rangle$ and $\langle u^1, u^2\rangle$ with $y^1, y^2 \in \mathbb{R}^m$ and $u^1, u^2 \in \mathbb{R}^{n-m}$. The user may supply these subroutines to incorporate a specific scaling. If the inner product $\langle x^1, x^2\rangle$ is required, then it is calculated as $\langle y^1, y^2\rangle + \langle u^1, u^2\rangle$. In this example, we used discretizations of the $L^2(0,T)$ and $L^2(0,T;H^1(0,1))$ norms for the control and the state spaces, respectively. This is important for the correct computation of the adjoint and the appropriate scaling of the problem.

In our numerical example we use the functions

$$\tau(y) = q_1 + q_2 y, \quad y \in \mathbb{R}, \quad \kappa(y) = r_1 + r_2 y, \quad y \in \mathbb{R},$$

with parameters $r_1 = q_1 = 4$, $r_2 = -1$, and $q_2 = 1$. The desired and initial tempera-

tures, and the right-hand side, are given by

$$\begin{aligned} y_d(t) &= 2 - e^{\eta t}, \\ y_0(x) &= 2 + \cos \pi x, \quad \text{and} \\ q(x,t) &= [\eta(q_1 + 2q_2) + \pi^2(r_1 + 2r_2)]e^{\eta t} \cos \pi x \\ &\quad -r_2\pi^2 e^{2\eta t} + (2r_2\pi^2 + \eta q_2)e^{2\eta t} \cos^2 \pi x, \end{aligned}$$

with $\eta = -1$. The final temperature is chosen to be $T = 0.5$ and the scalar $g = 1$ is used in the boundary condition. The functions in this example are those used in [39, Ex. 4.1]. The size of the problem tested is $n = 2200$, $m = 2100$ corresponding to the values $N_t = 100$, $N_x = 20$.

The scheme used to update the trust radius is the following fairly standard one:

- If $\text{ratio}(s_k; \rho_k) < 10^{-4}$, reject s_k and set $\delta_{k+1} = 0.5 \text{ norm}(s_k)$;
- if $10^{-4} \leq \text{ratio}(s_k; \rho_k) < 0.1$, reject s_k and set $\delta_{k+1} = 0.5 \text{ norm}(s_k)$;
- if $0.1 \leq \text{ratio}(s_k; \rho_k) < 0.75$, accept s_k and set $\delta_{k+1} = \delta_k$;
- if $\text{ratio}(s_k; \rho_k) \geq 0.75$, accept s_k and set $\delta_{k+1} = \min\{2\delta_k, 10^{10}\}$;

where $\text{ratio}(s_k; \rho_k) = \frac{ared(s_k;\rho_k)}{pred(s_k;\rho_k)}$,

$$\text{norm}(s_k) = \max\{\|s_k^{\mathfrak{n}}\|, \|\bar{D}_k^{-1}(s_k)_u\|\}$$

in the decoupled approach, and

$$\text{norm}(s_k) = \max\left\{\|s_k^{\mathfrak{n}}\|, \left\|\begin{pmatrix} -C_y(x_k)^{-1}C_u(x_k)(s_k)_u \\ \bar{D}_k^{-1}(s_k)_u \end{pmatrix}\right\|\right\}$$

in the coupled approach. The algorithms are stopped if the trust radius gets below 10^{-8}.

We have used $\sigma_k = \sigma = 0.99995$ for all k; $\delta_0 = 1$ as initial trust radius; $\rho_{-1} = 1$ and $\bar{\rho} = 10^{-2}$ in the penalty scheme. The tolerance used in the conjugate-gradient iteration was $\epsilon = 10^{-4}$. The upper and lower bounds were $b_i = 10^{-2}$, $a_i = -1000$, $i = 1, \ldots, n - m$. The starting vector was $x_0 = 0$.

For both the decoupled and the coupled approaches, we did tests using approximations to reduced and to full Hessians. We approximate these matrices with the limited memory BFGS representations given in [10] with a memory size of five pairs of vectors. For the reduced Hessian we use a null-space secant update (see [49], [67]). The initial approximation chosen was γI_{n-m} for the reduced Hessian and γI_n for the full Hessian, where γ is the user specified regularization parameter in the objective function (110).

In our implementation we use the following form of the diagonal matrix $\bar{D}_k$:

$$\left(\bar{D}_k\right)_{ii} = \begin{cases} \min\{1, (b - u_k)_i\} & \text{if} \quad (\bar{g}_k)_i < 0, \\ \min\{1, (u_k - a)_i\} & \text{if} \quad (\bar{g}_k)_i \geq 0, \end{cases} \tag{111}$$

for $i = 1, \ldots, n - m$. This form of $\bar{D}_k$ gives a better transition between the infinite and finite bound and is less sensitive to the introduction of meaningless bounds. See also Remark 3.1.

The algorithms were stopped when

$$\|D_k W_k^T \nabla f_k\| + \|C_k\| < 10^{-8}.$$

TABLE 1
Numerical results for $\gamma = 10^{-2}$.

	Decoupled		Coupled	
	Reduced $\widehat{H}_k$	Full H_k	Reduced $\widehat{H}_k$	Full H_k
number of iterations k^*	14	20	17	18
$\|C_{k^*}\|$	$.5082E-11$	$.1370E-10$	$.7122E-12$	$.8804E-11$
$\|D_{k^*} W_{k^*}^T \nabla f_{k^*}\|$	$.4033E-08$	$.1389E-08$	$.6365E-10$	$.2641E-08$
$\|s_{k^*-1}\|$	$.1230E-04$	$.1461E-04$	$.3546E-05$	$.1445E-04$
δ_{k^*-1}	$.1638E+05$	$.1049E+07$	$.1311E+06$	$.2621E+06$
ρ_{k^*-1}	$.1000E+01$	$.1000E+01$	$.1000E+01$	$.1000E+01$

TABLE 2
Numerical results for $\gamma = 10^{-3}$.

	Decoupled		Coupled	
	Reduced $\widehat{H}_k$	Full H_k	Reduced $\widehat{H}_k$	Full H_k
number of iterations k^*	16	18	17	19
$\|C_{k^*}\|$	$.6233E-11$	$.1115E-10$	$.6487E-11$	$.1246E-09$
$\|D_{k^*} W_{k^*}^T \nabla f_{k^*}\|$	$.5161E-08$	$.2539E-08$	$.7282E-09$	$.4696E-08$
$\|s_{k^*-1}\|$	$.1626E-04$	$.1703E-04$	$.1530E-04$	$.4659E-04$
δ_{k^*-1}	$.6554E+05$	$.2621E+06$	$.1311E+06$	$.5243E+06$
ρ_{k^*-1}	$.1000E+01$	$.1000E+01$	$.1000E+01$	$.1000E+01$

The results are shown in Tables 1 and 2 corresponding to the values $\gamma = 10^{-2}$ and $\gamma = 10^{-3}$, respectively. There were no rejected steps. The different alternatives tested performed quite similarly. The decoupled approach with reduced Hessian approximation seems to be the best for this example. Note that in this case the computation of each trial step costs only three linear system solvers with $C_y(x_k)$ and $C_y(x_k)^T$—one to compute the quasi-normal component and two for the computation of the tangential component.

We performed an experiment to compare the use of the Coleman–Li affine scaling with the Dikin–Karmarkar affine scaling. When applied to our class of problems, the Coleman–Li affine scaling is given by the matrices D_k and $\bar{D}_k$. A study of the Dikin–Karmarkar affine scaling for steepest descent is given in [54]. For our class of problems, this scaling is given by

$$\Big(K_k\Big)_{ii} = \min\{1,\, (u_k - a)_i,\, (b - u_k)_i\},\ i = 1, \ldots, n-m, \tag{112}$$

and has no dual information built in. We ran the trust-region interior-point SQP algorithm with the decoupled and reduced Hessian approximation and with (111) replaced by (112). The algorithm took only 11 iterations to reduce $\|K_k W_k^T \nabla f_k\| + \|C_k\|$ to 10^{-8}. However, as we can see from the plots of the controls in Figures 1 and 2, the algorithm did not find the correct solution when it used the Dikin–Karmarkar affine scaling (112). Some of the variables are at the wrong bound corresponding to negative multipliers.

12. Conclusions. In this paper we have introduced and analyzed some trust-region interior-point SQP algorithms for an important class of nonlinear programming problems that appear in many engineering applications. These algorithms use the structure of the problem, and they combine trust-region techniques for equality-constrained optimization with an affine scaling interior-point approach for simple bounds. We have proved global and local convergence results for these algorithms

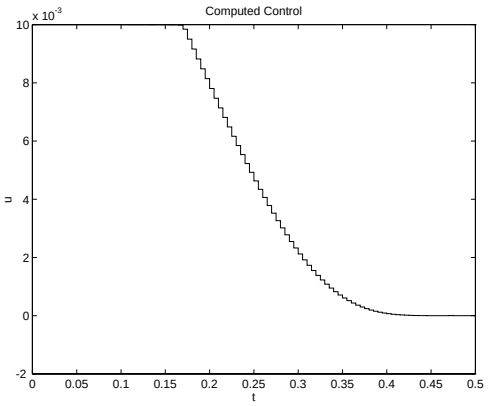


FIG. 1. *Coleman–Li affine scaling.*

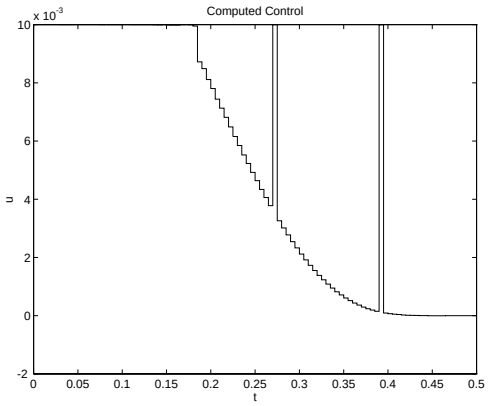


FIG. 2. *Dikin–Karmarkar affine scaling.*

that includes as special cases both the results established for equality constraints [15], [19] and those for simple bounds [13].

We have implemented the trust-region interior-point SQP algorithms covering several trial step computations and second-order approximations. In this paper we have reported numerical results for the solution of a specific optimal control problem governed by a nonlinear heat equation. In [11], [30], and [31], these algorithms have been applied to other optimal control problems. The numerical results have been quite satisfactory.

We are investigating extensions of these algorithms to handle bounds on the state variables y. See [66]. We are also developing an inexact analysis to deal with trial step computations that allow for inexact linear system solvers and inexact directional derivatives [31]. The formulation and analysis of these methods in an infinite-dimensional framework is also part of our current studies.

Acknowledgments. This research collaboration on the tailoring of current optimization algorithm design to exploit the structure of problems in computational design and control was begun as the result of conversations held during the April 1994 Workshop on Optimal Design and Control held at Virginia Polytechnic Institute and State University, Blacksburg. This workshop was organized by the Interdisciplinary Center for Applied Mathematics at Virginia Tech and sponsored by the AFOSR.

We would like to thank two referees for their patient and careful reading and their helpful comments on an earlier draft of this paper.

REFERENCES

[1] A. BARCLAY, P. E. GILL, AND J. B. ROSEN, *SQP Methods and Their Application to Numerical Optimal Control*, Numerical Analysis Rep. 97–3, Department of Mathematics, University of California, San Diego, La Jolla, CA, 1997.

[2] J. T. BETTS AND P. D. FRANK, *A sparse nonlinear optimization algorithm*, J. Optim. Theory Appl., 82 (1994), pp. 519–541.

[3] L. T. BIEGLER, J. NOCEDAL, AND C. SCHMID, *A reduced Hessian method for large-scale constrained optimization*, SIAM J. Optim., 5 (1995), pp. 314–347.

[4] L. T. BIEGLER, C. SCHMID, AND D. TERNET, *A multiplier-free, reduced Hessian method for process optimization*, in Large-Scale Optimization with Applications, Part II, IMA Vol. Math. Appl. 93, Springer-Verlag, New York, 1997, pp. 101–127.

[5] P. T. Boggs, *Sequential quadratic programming*, in Acta Numerica 1995, A. Iserles, ed., Cambridge University Press, Cambridge, 1995, pp. 1–51.
[6] J. Bonnans and C. Pola, *A trust region interior point algorithm for linearly constrained optimization*, SIAM J. Optim., 7 (1997), pp. 717–731.
[7] M. A. Branch, T. F. Coleman, and Y. Li, *A Subspace, Interior, and Conjugate Gradient Method for Large-Scale Bound-constrained Minimization Problems*, Tech. Rep. CTC95TR217, Advancing Computing Research Institute, Cornell University, Ithaca, NY, 1995.
[8] J. Burger and M. Pogu, *Functional and numerical solution of a control problem originating from heat transfer*, J. Optim. Theory Appl., 68 (1991), pp. 49–73.
[9] R. H. Byrd, M. E. Hribar, and J. Nocedal, *An Interior Point Algorithm for Large Scale Nonlinear Programming*, Tech. Rep. OTC 97/05, Optimization Technology Center, Northwestern University, 1997.
[10] R. H. Byrd, J. Nocedal, and R. B. Schnabel, *Representations of quasi–Newton matrices and their use in limited memory methods*, Math. Programming, 63 (1994), pp. 129–156.
[11] E. M. Cliff, M. Heinkenschloss, and A. Shenoy, *An optimal control problem for flows with discontinuities*, J. Optim. Theory Appl., 94 (1997), pp. 273–309.
[12] T. F. Coleman and Y. Li, *On the convergence of interior–reflective Newton methods for nonlinear minimization subject to bounds*, Math. Programming, 67 (1994), pp. 189–224.
[13] T. F. Coleman and Y. Li, *An interior trust region approach for nonlinear minimization subject to bounds*, SIAM J. Optim., 6 (1996), pp. 418–445.
[14] T. F. Coleman and J. Liu, *An Interior Newton Method for Quadratic Programming*, Tech. Rep. TR93–1388, Department of Computer Science, Cornell University, Ithaca, NY, 1993.
[15] J. E. Dennis, M. El-Alem, and M. C. Maciel, *A global convergence theory for general trust-region-based algorithms for equality constrained optimization*, SIAM J. Optim., 7 (1997), pp. 177–207.
[16] J. E. Dennis, M. Heinkenschloss, and L. N. Vicente, *TRICE: Trust-region interior-point SQP algorithms for optimal control and engineering design problems*, http://www.caam.rice.edu/∼trice.
[17] J. E. Dennis and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice–Hall, Englewood Cliffs, NJ, 1983.
[18] J. E. Dennis and L. N. Vicente, *Trust-region interior-point algorithms for minimization problems with simple bounds*, in Applied Mathematics and Parallel Computing, Festschrift for Klaus Ritter, H. Fisher, B. Riedmüller, and S. Schäffler, eds., Physica-Verlag, Springer-Verlag, New York, 1996, pp. 97–107.
[19] J. E. Dennis and L. N. Vicente, *On the convergence theory of general trust-region-based algorithms for equality-constrained optimization*, SIAM J. Optim., (1997), pp. 927–950.
[20] M. El-Alem, *A global convergence theory for the Celis–Dennis–Tapia trust-region algorithm for constrained optimization*, SIAM J. Numer. Anal., 28 (1991), pp. 266–290.
[21] M. El-Alem, *A robust trust-region algorithm with a nonmonotonic penalty parameter scheme for constrained optimization*, SIAM J. Optim., 5 (1995), pp. 348–378.
[22] D. M. Gay, *Computing optimal locally constrained steps*, SIAM J. Sci. Statist. Comput., 2 (1981), pp. 186–197.
[23] P. E. Gill, W. Murray, and M. A. Saunders, *SNOPT: An SQP Algorithm for Large-Scale Constrained Optimization*, Numerical Analysis Rep. 97–2, Department of Mathematics, University of California, San Diego, La Jolla, CA, 1997.
[24] P. E. Gill, W. Murray, M. A. Saunders, and M. H. Wright, *User's Guide for NPSOL (version 4.0): A FORTRAN Package for Nonlinear Programming*, Technical Rep. SOL 86–2, Systems Optimization Laboratory, Department of Operations Research, Stanford University, Stanford, CA, 1986.
[25] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 2nd ed., The John Hopkins University Press, Baltimore, MD, 1989.
[26] G. H. Golub and U. von Matt, *Quadratically constrained least squares and quadratic problems*, Numer. Math., 59 (1991), pp. 561–580.
[27] J. Goodman, *Newton's method for constrained optimization*, Math. Programming, 33 (1985), pp. 162–171.
[28] W. A. Gruver and E. W. Sachs, *Algorithmic Methods In Optimal Control*, Pitman, London, 1980.
[29] M. Heinkenschloss, *Projected sequential quadratic programming methods*, SIAM J. Optim., 6 (1996), pp. 373–417.
[30] M. Heinkenschloss, *SQP Methods for the Solution of Optimal Control Problems Governed by the Navier Stokes Equations*, Tech. Rep. in Interdisciplinary Center for Applied Math-

ematics, Virginia Polytechnic Institute and State University, Blacksburg, VA, 1996, in preparation.

[31] M. Heinkenschloss and L. N. Vicente, *Analysis of inexact trust-region interior-point SQP algorithms*, Tech. Rep. TR95–18, Department of Computational and Applied Mathematics, Rice University, 1995; revised April 1996. Appeared also as Tech. Rep. 95–06–01, Interdisciplinary Center for Applied Mathematics, Virginia Polytechnic Institute and State University, 1995.

[32] K. Ito and K. Kunisch, *The augmented Lagrangian method for parameter estimation in elliptic systems*, SIAM J. Control Optim., 28 (1990), pp. 113–136.

[33] C. T. Kelley and E. W. Sachs, *Solution of optimal control problems by a pointwise projected Newton method*, SIAM J. Control Optim., 33 (1995), pp. 1731–1757.

[34] C. T. Kelley and S. J. Wright, *Sequential quadratic programming for certain parameter identification problems*, Math. Programming, 51 (1991), pp. 281–305.

[35] K. Kunisch and G. Peichl, *Estimation of a temporally and spatially varying diffusion coefficient in a parabolic system by an augmented Lagrangian technique*, Numer. Math., 59 (1991), pp. 473–509.

[36] K. Kunisch and E. Sachs, *Reduced SQP methods for parameter identification problems*, SIAM J. Numer. Anal., 29 (1992), pp. 1793–1820.

[37] F.-S. Kupfer, *An infinite-dimensional convergence theory for reduced SQP methods in Hilbert space*, SIAM J. Optim., 6 (1996), pp. 126–163.

[38] F.-S. Kupfer and E. W. Sachs, *A prospective look at SQP methods for semilinear parabolic control problems*, in Optimal Control of Partial Differential Equations, Irsee 1990, Lecture Notes in Control and Inform. Sci. 149, K.-H. Hoffmann and W. Krabs, eds., Springer-Verlag, New York, 1991, pp. 143–157.

[39] F.-S. Kupfer and E. W. Sachs, *Numerical solution of a nonlinear parabolic control problem by a reduced SQP method*, Comput. Optim. Appl., 1 (1992), pp. 113–135.

[40] M. Lalee, J. Nocedal, and T. Plantenga, *On the implementation of an algorithm for large-scale equality constrained optimization*, SIAM J. Optim., 8 (1998), pp. 682–706.

[41] F. Leibfritz and E. W. Sachs, *Numerical solution of parabolic state constrained control problems using SQP- and interior-point-methods*, in Large Scale Optimization: State of the Art, W. W. Hager, D. Hearn, and P. Pardalos, eds., Kluwer Academic Publishers, Norwell, MA, 1994, pp. 251–264.

[42] D. B. Leineweber, H. G. Bock, J. P. Schlöder, J. V. Gallitzendörfer, A. Schäfer, and P. Jansohn, *A Boundary Value Problem Approach to the Optimization of Chemical Processes Described by DAE Models*, Tech. Rep. 97–14, Interdisziplinäres Zentrum für Wissenschaftliches Rechnen (IWR), Universtiät Heidelberg, 1997.

[43] Y. Li, *On Global Convergence of a Trust Region and Affine Scaling Method for Nonlinearly Constrained Minimization*, Tech. Rep. CTC94TR197, Advanced Computing Research Institute, Cornell University, Ithaca, New York, 1994.

[44] Y. Li, *A Trust Region and Affine Scaling Method for Nonlinearly Constrained Minimization*, Tech. Rep. CTC94TR198, Advanced Computing Research Institute, Cornell University, Ithaca, NY, 1994.

[45] J. J. Moré, *Recent developments in algorithms and software for trust regions methods*, in Mathematical Programming: The State of the Art, A. Bachem, M. Grotschel, and B. Korte, eds., Springer-Verlag, New York, 1983, pp. 258–287.

[46] J. J. Moré and D. C. Sorensen, *Computing a trust region step*, SIAM J. Sci. Statist. Comput., 4 (1983), pp. 553–572.

[47] W. Murray, *Sequential quadratic programming methods for large problems*, Comput. Optim. Appl., 7 (1997), pp. 127–142.

[48] W. Murray and F. J. Prieto, *A sequential quadratic programming algorithm using an incomplete solution of the subproblem*, SIAM J. Optim., 5 (1995), pp. 590–640.

[49] J. Nocedal and M. L. Overton, *Projected Hessian updating algorithms for nonlinearly constrained optimization*, SIAM J. Numer. Anal., 22 (1985), pp. 821–850.

[50] T. Plantenga, *Large-Scale Nonlinear Constrained Optimization using Trust Regions*, Ph.D. thesis, Northwestern University, Evanston, IL, 1994.

[51] E. Polak, *Computational Methods in Optimization. A Unified Approach*, Academic Press, New York, 1971.

[52] M. J. D. Powell, *A new algorithm for unconstrained optimization*, in Nonlinear Programming, J. B. Rosen, O. L. Mangasarian, and K. Ritter, eds., Academic Press, New York, 1970.

[53] F. Rendl and H. Wolkowicz, *A semidefinite framework for trust region subproblems with applications to large scale minimization. Semidefinite programming.* Math. Programming, Ser. B, 77 (1997), pp. 273–299.

[54] C. M. Samuelson, *The Dikin–Karmarkar Principle for Steepest Descent*, Ph.D. thesis, Department of Computational and Applied Mathematics, Rice University, Houston, TX, 1992; Tech. Rep. TR92–29.

[55] S. A. Santos and D. C. Sorensen, *A new matrix–free algorithm for the large-scale trust-region subproblem*, Tech. Rep. TR95–20, Department of Computational and Applied Mathematics, Rice University, Houston, TX, 1994.

[56] K. Schittkowski, *NLPQL: A* fortran *subroutine solving constrained nonlinear programming problems*, Ann. Oper. Res., 5 (1985), pp. 485–500.

[57] C. Schmid and L. T. Biegler, *A simultaneous approach for flowsheet optimization with existing modelling procedures*, Trans. I. Chem. Eng., Part A, 72 (1994), pp. 382–388.

[58] V. Schulz, *Reduced SQP methods for large–scale optimal control problems in DAE with application to path planning problems for satellite mounted robots*, Tech. Rep. 96–12, Interdisziplinäres Zentrum für Wissenschaftliches Rechnen (IWR), Universtiät Heidelberg, 1996.

[59] D. C. Sorensen, *Newton's method with a model trust region modification*, SIAM J. Numer. Anal., 19 (1982), pp. 409–426.

[60] D. C. Sorensen, *Minimization of a large scale quadratic function subject to an spherical constraint*, SIAM J. Optim., 7 (1997), pp. 141–161.

[61] T. Steihaug, *The conjugate gradient method and trust regions in large scale optimization*, SIAM J. Numer. Anal., 20 (1983), pp. 626–637.

[62] M. Steinbach, *Fast Recursive SQP Methods for Large–scale Optimal Control Problems*, Tech. Rep. 95–27, Interdisziplinäres Zentrum für Wissenschaftliches Rechnen (IWR), Universtät Heidelberg, 1995.

[63] P. L. Toint, *Towards an efficient sparsity exploiting Newton method for minimization*, in Sparse Matrices and Their Uses, I. S. Duff, ed., Academic Press, New York, 1981, pp. 57–87.

[64] M. Ulbrich, S. Ulbrich, and M. Heinkenschloss, *Global convergence of affine-scaling interior-point newton methods for infinite-dimensional nonlinear problems with pointwise bounds*, Tech. Rep. TR97–04, Department of Computational and Applied Mathematics, Rice University, Houston, TX, 1997. Available electronically at http://www.caam.rice.edu/∼heinken/Papers.html.

[65] L. N. Vicente, *Trust-Region Interior-Point Algorithms for a Class of Nonlinear Programming Problems*, Ph.D. thesis, Department of Computational and Applied Mathematics, Rice University, Houston, TX, 1996.

[66] L. N. Vicente, *On interior-point Newton algorithms for discretized optimal control problems with state constraints*, Optimization Methods and Software, 8 (1998), pp. 249–275.

[67] Y. Xie, *Reduced Hessian Algorithms for Solving Large–Scale Equality Constrained Optimization Problems*, Ph.D. thesis, Dept. of Computer Science, University of Colorado, Boulder, CO, 1991.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 5, pp. 1795–1814, September 1998

012

A PENALIZED NEUMANN CONTROL APPROACH FOR SOLVING AN OPTIMAL DIRICHLET CONTROL PROBLEM FOR THE NAVIER–STOKES EQUATIONS*

L. S. HOU† AND S. S. RAVINDRAN‡

Abstract. We introduce a penalized Neumann boundary control approach for solving an optimal Dirichlet boundary control problem associated with the two- or three-dimensional steady-state Navier–Stokes equations. We prove the convergence of the solutions of the penalized Neumann control problem, the suboptimality of the limit, and the optimality of the limit under further restrictions on the data. We describe the numerical algorithm for solving the penalized Neumann control problem and report some numerical results.

Key words. optimal control, Neumann control, Dirichlet control, Navier–Stokes equations, finite element method

AMS subject classifications. 35B40, 35B37, 35Q30, 65M60

PII. S0363012996304870

1. Introduction.

Optimal control for the Navier–Stokes equations has been the subject of extensive study in recent years and much progress has been made both mathematically and computationally; see, e.g., [AT], [FS1], [FS2], [FS3], [Fu1], [Fu2], [Fu3], [Gun], [GHS1], [GHS2], [GHS3], [HS], [HY1], [HY2], [HYR], [Li], [S1], and [S2]. In this work we confine ourselves to optimal Dirichlet control problems for the steady-state Navier–Stokes equations. Dirichlet controls, i.e., boundary velocity controls or boundary mass flux controls, are common in applications. For instance, one often attempts, through the suction and injection of fluid through orifices on the boundary to reduce the drag on a body moving through a fluid. Optimal Dirichlet control problems for time-dependent Navier–Stokes equations were studied in [FGH] for general Dirichlet controls and in [FS1], [FS2], [FS3] and [S1], [S2] for Dirichlet controls in a special case, namely, when the control is of the separation-of-variable type. Optimal Dirichlet control problems for steady-state Navier–Stokes equations were studied in [GHS2], [GHS3], and [HS]. In [GHS3], optimal Dirichlet controls of finite dimensions were analyzed and some numerical results presented. In [GHS2], the existence and regularity of optimal solutions for optimal Dirichlet control problems were proved; an optimality system of equations was derived; and finite element approximations were defined and optimal error estimates established. In [HS], optimal control problems with smooth Dirichlet controls were studied; in particular, an optimality system of equations was derived. The optimality systems in [GHS2] and [HS] involve a boundary Laplacian or a boundary biharmonic equation that complicates the numerical resolution of the optimality systems. In finite element approximations of (uncontrolled)

*Received by the editors June 7, 1996; accepted for publication (in revised form) June 16, 1997; published electronically June 25, 1998.
http://www.siam.org/journals/sicon/36-5/30487.html

†Department of Mathematics and Statistics, York University, North York, Ontario M3J 1P3, Canada. Current address: Department of Mathematics, Iowa State University, Ames, IA 50011 (hou@math.iastate.edu). The research of this author was supported in part by the Natural Science and Engineering Research Council of Canada under grant OGP-0169786.

‡Center for Research in Scientific Computation, Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205 (ravi@eos.ncsu.edu). Current address: Flow Modeling and Control Branch, NASA Langley Research Center, Mail Stop 170, Hampton, VA 23681 (ravi@fmd00.larc.nasa.gov).

boundary value problems for partial differential equations, Neumann boundary conditions are generally easier to handle than the Dirichlet ones; and the same is true of optimal boundary control problems. Inspired by the penalty method for solving Dirichlet problems for (uncontrolled) elliptic partial differential equations (see [Ba]), we propose in this article a penalty method for solving the optimal Dirichlet control problem. The proposed penalty approach avoids the boundary Laplacian or boundary biharmonic equations that appeared in [GHS2] and [HS]. The advantages (as well as disadvantages) of the penalty method in solving uncontrolled Dirichlet boundary value problems essentially hold true in solving optimal Dirichlet boundary control problems.

The optimal Dirichlet control problem we consider is to minimize the vorticity of viscous, incompressible flow by choosing an appropriate boundary velocity. Precisely, we will study the following optimal control problem: find a triplet $(\mathbf{u}, p, \mathbf{g})$ such that the functional

$$\mathcal{J}(\mathbf{u}, \mathbf{g}) = \frac{\alpha}{2}\int_{\Omega} |\operatorname{curl} \mathbf{u}|^2 \, d\mathbf{x} + \frac{\beta}{2}\int_{\Gamma} |\mathbf{g}|^2 \, ds \tag{1.1}$$

is minimized subject to the steady-state Navier–Stokes equations

$$-\nu \Delta \mathbf{u} + (\mathbf{u} \cdot \nabla)\mathbf{u} + \nabla p = \mathbf{f} \qquad \text{in } \Omega\,, \tag{1.2}$$

$$\operatorname{div} \mathbf{u} = 0 \qquad \text{in } \Omega, \tag{1.3}$$

and

$$\mathbf{u} = \mathbf{g} \qquad \text{on } \Gamma\,. \tag{1.4}$$

Here, Ω is a two- or three-dimensional bounded and simply connected flow domain (Ω is assumed to be of class $C^{1,1}$ or convex in $\mathbb{R}^2$ and of class $C^{1,1}$ in $\mathbb{R}^3$); Γ denotes the boundary of Ω; $\nu > 0$ denotes the constant viscosity; $\mathbf{u}$ and p denote the velocity field and the pressure field, respectively; $\mathbf{f}$ is a prescribed forcing term; and $\mathbf{g}$ is the boundary velocity—the control field. Because of the divergence-free condition on $\mathbf{u}$, $\mathbf{g}$ must necessarily satisfy $\int_{\Gamma} \mathbf{g} \cdot \mathbf{n}\, ds = 0$. The constants α and β appearing in the functional (1.1) are two positive parameters that adjust the relative weights of the two terms in the functional. Note that we use the same notation curl to denote the curl operators in two dimensions and three dimensions, although they are defined differently. The choice of the functional is motivated by the fact that irrotational flows have no local flow recirculations. We hope that minimizing the L^2-norm of the vorticity will lead to reduction in flow recirculations.

The plan of the paper is as follows. In section 2, we review mathematical background materials related to the steady-state Navier–Stokes equations and give a precise description of the optimal control problem we consider. In section 3, we introduce the penalized Neumann control approach and prove the existence of an optimal solution for the penalized Neumann control problem. In section 4, we demonstrate the convergence of the penalized optimal boundary control solutions and show that the limit is suboptimal. In section 5, we show that the limit found in section 4 is indeed an optimal solution for the optimal Dirichlet control problem. Finally in section 6, we describe the formal procedures for computing an approximate optimal solution and present some numerical results.

2. Preliminaries. Throughout, C or C_i (where i is any subscript) denotes a constant depending only on the domain Ω. We denote by $L^2(\Omega)$ the collection of Lebesgue square-integrable functions defined on Ω. Let $H^1(\Omega) = \{v \in L^2(\Omega) : \frac{\partial v}{\partial x_i} \in L^2(\Omega) \text{ for } i = 1, \ldots, d\}$, where $d = 2$ or 3; $H_0^1(\Omega) = \{v \in H^1(\Omega) : v|_\Gamma = 0\}$; $L_0^2(\Omega) = \{q \in L^2(\Omega) : \int_\Omega q \, d\mathbf{x} = 0\}$; and $H^m(\Omega) = \{v \in L^2(\Omega) : \frac{\partial^{|\alpha|} v}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}} \in L^2(\Omega) \text{ for all } \alpha = (\alpha_1, \ldots, \alpha_d) \text{ with } |\alpha| \le m\}$, where $d = 2$ or 3. Here $m > 0$ is an integer. For the definition of fractional ordered Sobolev spaces $H^s(\Omega)$ (s noninteger), see [Ad]. Negative ordered Sobolev spaces $H^{-s}(\Omega)$ ($s > 0$) are defined as the dual space, i.e., $H^{-s}(\Omega) = \{H_0^s(\Omega)\}^*$. Vector-valued counterparts of these spaces are denoted by boldface symbols, e.g., $\mathbf{H}^1(\Omega) = [H^1(\Omega)]^d$, where $d = 2$ or 3. The trace spaces $H^r(\Gamma)$ are the restriction to the boundary of $H^{r+1/2}(\Omega)$. We denote the norms and inner products for $H^s(\Omega)$ or $\mathbf{H}^s(\Omega)$ by $\|\cdot\|_s$ and $(\cdot,\cdot)_s$, respectively. The $L^2(\Omega)$ or $\mathbf{L}^2(\Omega)$ inner product is denoted by $(\cdot,\cdot)$. We denote the norms and inner products for $H^r(\Gamma)$ or $\mathbf{H}^r(\Gamma)$ by $\|\cdot\|_{r,\Gamma}$ and $(\cdot,\cdot)_{r,\Gamma}$, respectively. The $L^2(\Gamma)$ or $\mathbf{L}^2(\Gamma)$ inner product is denoted by $(\cdot,\cdot)_\Gamma$. The duality pairing between a Sobolev space $H^s(\Omega)$ ($s > 0$) and its dual space is denoted by $\langle\cdot,\cdot\rangle$. The duality pairing between a trace space $H^r(\Gamma)$ ($r > 0$) and its dual space is denoted by $\langle\cdot,\cdot\rangle_\Gamma$.

We define the following standard bilinear, trilinear forms associated with the Navier–Stokes equations

$$a(\mathbf{u}, \mathbf{v}) = \int_\Omega (\nabla \mathbf{u}) : (\nabla \mathbf{v}) \, d\mathbf{x} \qquad \forall\, \mathbf{u}, \mathbf{v} \in \mathbf{H}^1(\Omega)\,,$$

$$b(\mathbf{u}, q) = -\int_\Omega q \operatorname{div} \mathbf{u} \, d\mathbf{x} \qquad \forall\, \mathbf{u} \in \mathbf{H}^1(\Omega), \forall\, q \in L^2(\Omega),$$

and

$$c(\mathbf{u}, \mathbf{v}, \mathbf{w}) = \int_\Omega (\mathbf{u} \cdot \nabla)\mathbf{v} \cdot \mathbf{w} \, d\mathbf{x} \qquad \forall\, \mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathbf{H}^1(\Omega)\,.$$

We now summarize some properties of these linear forms. We have the coercivity relations associated with $a(\cdot,\cdot)$:

$$a(\mathbf{u}, \mathbf{u}) = \|\nabla \mathbf{u}\|_0^2 \ge C_0 \|\mathbf{u}\|_1^2 \qquad \forall\, \mathbf{u} \in \mathbf{H}_0^1(\Omega) \tag{2.1}$$

(which is a direct consequence of Poincaré inequality) and

$$\int_\Gamma |\mathbf{v}|^2 \, ds + \int_\Omega \left|\nabla \mathbf{v}\right|^2 d\mathbf{x} \ge C_1 \|\mathbf{v}\|_1^2 \qquad \forall\, \mathbf{v} \in \mathbf{H}^1(\Omega) \tag{2.2}$$

(whose proof can be found in [Ne]). The forms $a(\cdot,\cdot)$, $b(\cdot,\cdot)$, and $c(\cdot,\cdot,\cdot)$ are all continuous; in particular, we have

$$\left|c(\mathbf{u}, \mathbf{v}, \mathbf{w})\right| \le C_2 \|\mathbf{u}\|_1 \|\mathbf{v}\|_1 \|\mathbf{w}\|_1\,. \tag{2.3}$$

The bilinear form $b(\cdot,\cdot)$ satisfies the following inf-sup conditions:

$$\inf_{q \in L^2(\Omega)} \sup_{\mathbf{v} \in \mathbf{H}^1(\Omega)} \frac{\int_\Omega q \operatorname{div} \mathbf{v} \, d\mathbf{x}}{\|q\|_0 \|\mathbf{v}\|_1} \ge C_3 \tag{2.4}$$

and

$$\inf_{q\in L_0^2(\Omega)} \sup_{\mathbf{v}\in \mathbf{H}_0^1(\Omega)} \frac{\int_\Omega q \operatorname{div} \mathbf{v}\, d\mathbf{x}}{\|q\|_0 \, \|\mathbf{v}\|_1} \geq C_3 \,. \tag{2.5}$$

The proof of (2.5) can be found in [GR], and that of (2.4) in [Ma]. Using integration-by-parts techniques we may deduce

$$\int_\Omega (\mathbf{v}\cdot\nabla)\mathbf{v}\cdot\mathbf{v}\, d\mathbf{x} = \frac{1}{2}\int_\Gamma (\mathbf{v}\cdot\mathbf{n})|\mathbf{v}|^2\, ds \qquad \forall\, \mathbf{v}\in \mathbf{H}^1(\Omega) \text{ with } \operatorname{div}\mathbf{v} = 0\,. \tag{2.6}$$

We now give the definition of a solution for the Navier–Stokes equations with a Dirichlet boundary condition. Throughout, we assume $\mathbf{f}\in \mathbf{L}^2(\Omega)$.

DEFINITION 2.1. *Let* $\mathbf{g}\in \mathbf{H}^{1/2}(\Gamma)$. *A pair* $(\mathbf{u},p)\in \mathbf{H}^1(\Omega)\times L_0^2(\Omega)$ *is said to be a solution of the Navier–Stokes equations* (1.2)–(1.4) *iff*

$$\nu\, a(\mathbf{u},\mathbf{v}) + c(\mathbf{u},\mathbf{u},\mathbf{v}) + b(\mathbf{v},p) = \int_\Omega \mathbf{f}\cdot\mathbf{v}\, d\mathbf{x} \qquad \forall\, \mathbf{v}\in \mathbf{H}_0^1(\Omega)\,, \tag{2.7}$$

$$b(\mathbf{u},q) = 0 \qquad \forall\, q\in L_0^2(\Omega), \tag{2.8}$$

and

$$\mathbf{u}\big|_\Gamma = \mathbf{g}\,. \tag{2.9}$$

A proof of the existence of a solution in the sense of Definition 2.1 can be found in [GR] and [Te].

The optimal Dirichlet control problem we consider can be stated as:

(P) seek a $(\mathbf{u},p,\mathbf{g})\in \mathbf{H}^1(\Omega)\times L_0^2(\Omega)\times \mathbf{L}^2(\Gamma)$ such that (1.1) is minimized subject to (2.7)–(2.9).

We define the admissible set $\mathcal{U}_{ad}$ for (P) by

$$\mathcal{U}_{ad} = \big\{(\mathbf{u},p,\mathbf{g})\in \mathbf{H}^1(\Omega)\times L_0^2(\Omega)\times \mathbf{L}^2(\Gamma) : (\mathbf{u},p,\mathbf{g}) \text{ satisfies (2.7)–(2.9)}\big\}\,.$$

3. Penalized optimal Neumann control problems. For each $\epsilon\in(0,1/\nu)$, we consider the following Neumann control problem: find a $(\mathbf{u}_\epsilon,p_\epsilon,\mathbf{g}_\epsilon)\in \mathbf{H}^1(\Omega)\times L^2(\Omega)\times \mathbf{L}^2(\Gamma)$ such that the functional

$$\mathcal{J}(\mathbf{u},\mathbf{g}) = \frac{\alpha}{2}\int_\Omega |\operatorname{curl}\mathbf{u}|^2\, d\mathbf{x} + \frac{\beta}{2}\int_\Gamma |\mathbf{g}|^2\, ds \tag{3.1}$$

is minimized subject to the steady-state Navier–Stokes equations (1.2)–(1.3) with the nonlinear Neumann (or Robin)-type boundary condition

$$-p\mathbf{n} + \nu\frac{\partial\mathbf{u}}{\partial\mathbf{n}} - \frac{1}{2}(\mathbf{u}\cdot\mathbf{n})\mathbf{u} + \frac{1}{\epsilon}\,\mathbf{u} = \frac{1}{\epsilon}\,\mathbf{g} \qquad \text{on } \Gamma\,. \tag{3.2}$$

Formally, we see that as $\epsilon\to 0$, the Neumann boundary condition (3.2) reduces to the Dirichlet boundary condition (1.4), and therefore we expect that optimal solutions for the Neumann boundary control problems approach an optimal solution for the

Dirichlet boundary control problem. Here, ϵ acts as a penalty constant. By formally multiplying (1.2) by a test function $\mathbf{v}$ and integrating by parts, we obtain

$$\nu\int_\Omega(\nabla\mathbf{u}):(\nabla\mathbf{v})\,d\mathbf{x}-\nu\int_\Gamma\frac{\partial\mathbf{u}}{\partial\mathbf{n}}\cdot\mathbf{v}\,ds+\int_\Omega(\mathbf{u}\cdot\nabla)\mathbf{u}\cdot\mathbf{v}\,d\mathbf{x}$$
$$-\int_\Omega p\,\mathrm{div}\,\mathbf{v}\,d\mathbf{x}+\int_\Gamma p\mathbf{n}\cdot\mathbf{v}\,ds=\int_\Omega\mathbf{f}\cdot\mathbf{v}\,d\mathbf{x}\qquad\forall\,\mathbf{v}\in\mathbf{H}^1(\Omega)\,.$$

Eliminating $-p\mathbf{n}+\nu\frac{\partial\mathbf{u}}{\partial\mathbf{n}}$ in the boundary integrals using (3.2), we are led to the following definition of a (weak) solution for the Navier–Stokes equations with the Neumann boundary condition (3.2).

DEFINITION 3.1. *Let $\mathbf{g}\in\mathbf{L}^2(\Gamma)$. A pair $(\mathbf{u},p)\in\mathbf{H}^1(\Omega)\times L^2(\Omega)$ is said to be a solution of* (1.2)–(1.3) *with the Neumann condition* (3.2) *iff $(\mathbf{u},p)$ satisfies*

$$\begin{aligned}&\nu\int_\Omega(\nabla\mathbf{u}):(\nabla\mathbf{v})\,d\mathbf{x}+\frac{1}{\epsilon}\int_\Gamma\mathbf{u}\cdot\mathbf{v}\,ds-\frac{1}{2}\int_\Gamma(\mathbf{u}\cdot\mathbf{n})\mathbf{u}\cdot\mathbf{v}\,ds+\int_\Omega(\mathbf{u}\cdot\nabla)\mathbf{u}\cdot\mathbf{v}\,d\mathbf{x}\\&\quad-\int_\Omega p\,\mathrm{div}\,\mathbf{v}\,d\mathbf{x}=\int_\Omega\mathbf{f}\cdot\mathbf{v}\,d\mathbf{x}+\frac{1}{\epsilon}\int_\Gamma\mathbf{g}\cdot\mathbf{v}\,ds\qquad\forall\,\mathbf{v}\in\mathbf{H}^1(\Omega)\end{aligned}\tag{3.3}$$

and

$$-\int_\Omega q\,\mathrm{div}\,\mathbf{u}\,d\mathbf{x}=0\qquad\forall\,q\in L^2(\Omega)\,.\tag{3.4}$$

For each $\epsilon>0$, the penalized optimal Neumann control problems we consider can be stated as follows:

$$\begin{aligned}&\text{seek a }(\mathbf{u}_\epsilon,p_\epsilon,\mathbf{g}_\epsilon)\in\mathbf{H}^1(\Omega)\times L^2(\Omega)\times\mathbf{L}^2(\Gamma)\text{ such}\\&\text{that (3.1) is minimized subject to (3.3)–(3.4)}\,.\end{aligned}\tag{P}_\epsilon$$

In this section we will derive an estimate for solutions of the constraint equations (3.3)–(3.4) and then prove the existence of a solution for the optimal control problem $(\mathrm{P})_\epsilon$.

LEMMA 3.2. *Assume $\epsilon\in(0,1/\nu)$ and $\mathbf{g}\in\mathbf{L}^2(\Gamma)$. Then there exists a $(\mathbf{u},p)\in\mathbf{H}^1(\Omega)\times L^2(\Omega)$ satisfying* (3.3)–(3.4)*; furthermore,*

$$\frac{\nu}{2}\int_\Omega|\nabla\mathbf{u}|^2\,d\mathbf{x}+\frac{1}{4\epsilon}\int_\Gamma|\mathbf{u}|^2\,ds\le\frac{1}{2\nu C_1}\int_\Omega|\mathbf{f}|^2\,d\mathbf{x}+\frac{1}{\epsilon}\int_\Gamma|\mathbf{g}|^2\,ds\tag{3.5}$$

and

$$\|\overline{p}\|_0\le\frac{1}{C_3}\big(\nu\,\|\mathbf{u}\|_1+C_2\|\mathbf{u}\|_1^2+\|\mathbf{f}\|_0\big)\,,\tag{3.6}$$

where $\overline{p}=p-\big(1/|\Omega|\big)\int_\Omega p\,d\mathbf{x}$.

Proof. Since $\epsilon\in(0,1/\nu)$, we may use (2.2) to obtain

$$\nu\int_\Omega|\nabla\mathbf{v}|\,d\mathbf{x}+\frac{1}{\epsilon}\int_\Gamma|\mathbf{v}|^2\,ds\ge\nu\Big(\int_\Omega|\nabla\mathbf{v}|\,d\mathbf{x}+\int_\Gamma|\mathbf{v}|^2\,ds\Big)\ge\nu C_1\|\mathbf{v}\|_1^2\tag{3.7}$$

for every $\mathbf{v}\in\mathbf{H}^1(\Omega)$. This coercivity relation together with the inf-sup condition (2.4) allow us to prove the existence of a solution for (3.3)–(3.4) by using standard techniques for proving the existence of a solution for the Navier–Stokes equations with

homogeneous Dirichlet conditions (see [Te] or [GR]). Here we also used the fact that $\mathbf{H}^1(\Omega)\big|_\Gamma = \mathbf{H}^{1/2}(\Gamma)$ and $\mathbf{H}^{1/2}(\Gamma)$ is continuously embedded into $\mathbf{L}^3(\Gamma)$ so that we have the continuity of the trilinear term $\int_\Gamma (\mathbf{u}\cdot\mathbf{n})\mathbf{w}\cdot\mathbf{v}\,ds$ on $\mathbf{H}^1(\Omega)\times\mathbf{H}^1(\Omega)\times\mathbf{H}^1(\Omega)$. It remains to show that estimates (3.5)–(3.6) hold. Setting $\mathbf{v}=\mathbf{u}$ in (3.3) and using (3.4) we obtain

$$\nu\int_\Omega |\nabla\mathbf{u}|^2\,d\mathbf{x} + \frac{1}{\epsilon}\int_\Gamma |\mathbf{u}|^2\,ds \le \frac{1}{2\nu C_1}\int_\Omega |\mathbf{f}|^2\,d\mathbf{x} + \frac{\nu C_1}{2}\int_\Omega |\mathbf{u}|^2\,d\mathbf{x} + \frac{1}{\epsilon}\int_\Gamma |\mathbf{g}|^2\,ds + \frac{1}{4\epsilon}\int_\Gamma |\mathbf{u}|^2\,ds,$$

so that using (3.7) we are led to

$$\frac{\nu}{2}\int_\Omega |\nabla\mathbf{u}|^2\,d\mathbf{x} + \frac{1}{4\epsilon}\int_\Gamma |\mathbf{u}|^2\,ds \le \frac{1}{2\nu C_1}\int_\Omega |\mathbf{f}|^2\,d\mathbf{x} + \frac{1}{\epsilon}\int_\Gamma |\mathbf{g}|^2\,ds\,;$$

i.e., (3.5) is proved. For test functions $\mathbf{v}\in\mathbf{H}_0^1(\Omega)$, equation (3.3) reduces to

$$\nu\, a(\mathbf{u},\mathbf{v}) + c(\mathbf{u},\mathbf{u},\mathbf{v}) + b(\mathbf{v},p) = (\mathbf{f},\mathbf{v}) \qquad \forall\,\mathbf{v}\in\mathbf{H}_0^1(\Omega)\,.$$

Note that $\overline{p}\in L_0^2(\Omega)$, where $\overline{p} = p - \big(1/|\Omega|\big)\int_\Omega p\,d\mathbf{x}$, and

$$\begin{aligned} b(\mathbf{v},\overline{p}) &= -\int_\Omega \Big(p - \frac{1}{|\Omega|}\int_\Omega p\,d\mathbf{x}\Big)\operatorname{div}\mathbf{v}\,d\mathbf{x} = -\int_\Omega p\operatorname{div}\mathbf{v}\,d\mathbf{x} + \frac{1}{|\Omega|}\int_\Omega p\,d\mathbf{x}\int_\Omega \operatorname{div}\mathbf{v}\,d\mathbf{x} \\ &= b(\mathbf{v},p) + \frac{1}{|\Omega|}\int_\Omega p\,d\mathbf{x}\int_\Gamma \mathbf{v}\cdot\mathbf{n}\,ds = b(\mathbf{v},p) \qquad \forall\,\mathbf{v}\in\mathbf{H}_0^1(\Omega)\,. \end{aligned}$$

Using the last two relations and the second inf-sup condition (2.5) we easily obtain the estimate for $\overline{p}$:

$$\|\overline{p}\|_0 \le \frac{1}{C_3}\big(\nu\,\|\mathbf{u}\|_1 + C_2\|\mathbf{u}\|_1^2 + \|\mathbf{f}\|_0\big)\,. \qquad \square$$

We will make use of the following two lemmas to prove the existence of a solution for $(\mathrm{P})_\epsilon$.

LEMMA 3.3. *There exists a positive constant C_4 such that*

$$\|\mathbf{w}\|_1^2 \le C_4\Big(\int_\Omega |\operatorname{div}\mathbf{w}|^2\,d\mathbf{x} + \int_\Omega |\operatorname{curl}\mathbf{w}|^2\,d\mathbf{x} + \int_\Gamma |\mathbf{w}|^2\,ds\Big) \qquad \forall\,\mathbf{w}\in\mathbf{H}^1(\Omega)\,.$$

Proof. The proof follows standard techniques dealing with norm equivalence on Sobolev spaces (see, e.g., [Ne]). It proceeds as follows. Assume Lemma 3.3 is false. Then we may choose a sequence $\{\mathbf{w}^{(n)}\}_{n=1}^\infty \subset \mathbf{H}^1(\Omega)$ such that $\|\mathbf{w}^{(n)}\|_1 = 1$ for all n and

$$\int_\Omega |\operatorname{curl}\mathbf{w}^{(n)}|^2\,d\mathbf{x} + \int_\Omega |\operatorname{div}\mathbf{w}^{(n)}|^2\,d\mathbf{x} + \int_\Gamma |\mathbf{w}^{(n)}|^2\,ds < \frac{1}{n}\,. \tag{3.8}$$

The boundedness of $\{\mathbf{w}^{(n)}\}$ in $\mathbf{H}^1(\Omega)$ implies that there exists a $\mathbf{w}\in\mathbf{H}^1(\Omega)$ and a subsequence of $\{\mathbf{w}^{(n)}\}$, still denoted by $\{\mathbf{w}^{(n)}\}$, such that as $n\to\infty$,

$$\mathbf{w}^{(n)} \rightharpoonup \mathbf{w} \quad \text{in } \mathbf{H}^1(\Omega)\,, \qquad \mathbf{w}^{(n)} \to \mathbf{w} \quad \text{in } \mathbf{L}^2(\Omega)\,,$$

$$\operatorname{curl}\mathbf{w}^{(n)} \rightharpoonup \operatorname{curl}\mathbf{w} \quad \text{in } L^2(\Omega) \qquad \text{and} \qquad \operatorname{div}\mathbf{w}^{(n)} \rightharpoonup \operatorname{div}\mathbf{w}^{(n)} \quad \text{in } L^2(\Omega)\,.$$

Also, since the trace of $\mathbf{H}^1(\Omega)$ equals $\mathbf{H}^{1/2}(\Gamma)$ and the space $\mathbf{H}^{1/2}(\Gamma)$ is continuously imbedded into $\mathbf{L}^2(\Gamma)$, we have that

$$\mathbf{w}^{(n)} \rightharpoonup \mathbf{w} \quad \text{in } \mathbf{L}^2(\Gamma)\,.$$

From (3.8) we deduce that

$$\operatorname{curl}\mathbf{w}^{(n)} \to 0 \quad \text{in } L^2(\Omega)\,, \qquad \operatorname{div}\mathbf{w}^{(n)} \to 0 \quad \text{in } L^2(\Omega)\,,$$

and

$$\mathbf{w}^{(n)}|_\Gamma \to \mathbf{0} \quad \text{in } \mathbf{L}^2(\Gamma)\,.$$

By uniqueness of weak limits we have that

$$\operatorname{curl}\mathbf{w} = 0\,, \qquad \operatorname{div}\mathbf{w} = 0, \qquad \text{and} \qquad \mathbf{w}\big|_\Gamma = \mathbf{0}\,.$$

Since the boundary value problem $\operatorname{curl}\mathbf{w} = 0$, $\operatorname{div}\mathbf{w} = 0$, and $(\mathbf{w}\cdot\mathbf{n})|_\Gamma = 0$ admits a unique trivial solution (see [GR, Theorem I.3.6, p. 48]), we conclude $\mathbf{w} = \mathbf{0}$. This, of course, contradicts $\|\mathbf{w}\|_1 \geq \liminf_{n\to\infty}\|\mathbf{w}^{(n)}\|_1 = 1$. Thus the lemma is proved. □

LEMMA 3.4. *Assume* $\mathbf{u} \in \mathbf{V} \equiv \{\mathbf{v} \in \mathbf{H}^1(\Omega) : \operatorname{div}\mathbf{v} = 0\}$ *is a solution of*

$$\begin{aligned}\nu\int_\Omega (\nabla\mathbf{u}) : (\nabla\mathbf{v})\,d\mathbf{x} + \frac{1}{\epsilon}\int_\Gamma \mathbf{u}\cdot\mathbf{v}\,ds - \frac{1}{2}\int_\Gamma (\mathbf{u}\cdot\mathbf{n})\mathbf{u}\cdot\mathbf{v}\,ds + \int_\Omega (\mathbf{u}\cdot\nabla)\mathbf{u}\cdot\mathbf{v}\,d\mathbf{x}\\ = \int_\Omega \mathbf{f}\cdot\mathbf{v}\,d\mathbf{x} + \frac{1}{\epsilon}\int_\Gamma \mathbf{g}\cdot\mathbf{v}\,ds \qquad \forall\,\mathbf{v}\in\mathbf{V}.\end{aligned}$$

Then, there exists a $p \in L^2(\Omega)$ *such that* (3.3)–(3.4) *hold.*

Proof. The result follows directly from the first inf-sup condition (2.4) and [GR, Theorem IV.1.4, p. 283]. □

We are now in a position to prove the existence of a solution to $(\mathrm{P})_\epsilon$.

THEOREM 3.5. *Assume* $\epsilon \in (0, 1/\nu)$. *Then there exists a solution* $(\mathbf{u}_\epsilon, p_\epsilon, \mathbf{g}_\epsilon) \in \mathbf{H}^1(\Omega)\times L^2(\Omega)\times\mathbf{L}^2(\Gamma)$ *for the optimal control problem* $(\mathrm{P})_\epsilon$.

Proof. From Lemma 3.2 it is obvious that there exists a $(\mathbf{u}, p, \mathbf{g}) \in \mathbf{H}^1(\Omega)\times L^2(\Omega)\times\mathbf{L}^2(\Gamma)$ such that (3.3)–(3.4) holds. Hence we may choose a minimizing sequence $\{(\mathbf{u}_m, p_m, \mathbf{g}_m)\} \subset \mathbf{H}^1(\Omega)\times L^2(\Omega)\times\mathbf{L}^2(\Gamma)$ such that

$$\begin{aligned}\nu\int_\Omega \nabla\mathbf{u}_m : \nabla\mathbf{v}\,d\mathbf{x} + \frac{1}{\epsilon}\int_\Gamma \mathbf{u}_m\cdot\mathbf{v}\,ds - \frac{1}{2}\int_\Gamma (\mathbf{u}_m\cdot\mathbf{n})\mathbf{u}_m\cdot\mathbf{v}\,ds - \int_\Omega p_m \operatorname{div}\mathbf{v}\,d\mathbf{x}\\ + \int_\Omega (\mathbf{u}_m\cdot\nabla)\mathbf{u}_m\cdot\mathbf{v}\,d\mathbf{x} = \int_\Omega \mathbf{f}\cdot\mathbf{v}\,d\mathbf{x} + \frac{1}{\epsilon}\int_\Gamma \mathbf{g}_m\cdot\mathbf{v}\,ds \qquad \forall\,\mathbf{v}\in\mathbf{H}^1(\Omega)\,,\end{aligned} \tag{3.9}$$

$$-\int_\Omega q\operatorname{div}\mathbf{u}_m\,d\mathbf{x} = 0 \qquad \forall\,q\in L^2(\Omega)\,, \tag{3.10}$$

and

$$\lim_{m\to\infty}\mathcal{J}(\mathbf{u}_m,\mathbf{g}_m) = \inf\big\{\mathcal{J}(\mathbf{u},\mathbf{g}) : (\mathbf{u},p,\mathbf{g})\in\mathbf{H}^1(\Omega)\times L^2(\Omega)\times\mathbf{L}^2(\Gamma) \text{ and } (\mathbf{u},p,\mathbf{g}) \text{ satisfies (3.3)–(3.4)}\big\}\,.$$

The boundedness of $\{\mathcal{J}(\mathbf{u}_m, \mathbf{g}_m)\}$ implies the boundedness of $\{\|\mathbf{g}_m\|_{0,\Gamma}\}$. Then using (3.5) we see that the set $\{\|\mathbf{u}_m\|_1\}$ is also bounded independent of m (although the bound depends on ϵ, which is fixed). Hence we may extract subsequences (still denoted by $\mathbf{u}_m$ and $\mathbf{g}_m$, respectively) such that

$$\mathbf{u}_m \rightharpoonup \mathbf{u}_\epsilon \quad \text{in } \mathbf{H}^1(\Omega)\,, \qquad \text{and} \qquad \mathbf{g}_m \rightharpoonup \mathbf{g}_\epsilon \quad \text{in } \mathbf{L}^2(\Gamma)$$

for some $(\mathbf{u}_\epsilon, \mathbf{g}_\epsilon) \in \mathbf{H}^1(\Omega) \times \mathbf{L}^2(\Gamma)$, as $m \to \infty$. Compact imbedding results imply the strong convergence $\mathbf{u}_m \to \mathbf{u}_\epsilon$ in $\mathbf{L}^4(\Omega)$ as $m \to \infty$. Using standard techniques in proving the existence of a solution to the steady-state Navier–Stokes equations, we may pass to the limit in (3.9)–(3.10) as $m \to \infty$ to conclude that $\mathbf{u}_\epsilon \in \mathbf{V}$ and $(\mathbf{u}_\epsilon, \mathbf{g}_\epsilon)$ satisfies

$$\begin{aligned}\nu \int_\Omega (\nabla \mathbf{u}_\epsilon) : (\nabla \mathbf{v})\, d\mathbf{x} + \frac{1}{\epsilon}\int_\Gamma \mathbf{u}_\epsilon \cdot \mathbf{v}\, ds - \frac{1}{2}\int_\Gamma (\mathbf{u}_\epsilon \cdot \mathbf{n})\mathbf{u}_\epsilon \cdot \mathbf{v}\, ds \\ + \int_\Omega (\mathbf{u}_\epsilon \cdot \nabla)\mathbf{u}_\epsilon \cdot \mathbf{v}\, d\mathbf{x} = \int_\Omega \mathbf{f} \cdot \mathbf{v}\, d\mathbf{x} + \frac{1}{\epsilon}\int_\Gamma \mathbf{g}_\epsilon \cdot \mathbf{v}\, ds \qquad \forall\, \mathbf{v} \in \mathbf{V}\,.\end{aligned}$$

The last equations and Lemma 3.4 imply that there exists a $p_\epsilon \in L^2(\Omega)$ such that

$$\begin{aligned}\nu \int_\Omega (\nabla \mathbf{u}_\epsilon) : (\nabla \mathbf{v})\, d\mathbf{x} + \frac{1}{\epsilon}\int_\Gamma \mathbf{u}_\epsilon \cdot \mathbf{v}\, ds - \frac{1}{2}\int_\Gamma (\mathbf{u}_\epsilon \cdot \mathbf{n})\mathbf{u}_\epsilon \cdot \mathbf{v}\, ds - \int_\Omega p_\epsilon \operatorname{div} \mathbf{v}\, d\mathbf{x} \\ + \int_\Omega (\mathbf{u}_\epsilon \cdot \nabla)\mathbf{u}_\epsilon \cdot \mathbf{v}\, d\mathbf{x} = \int_\Omega \mathbf{f} \cdot \mathbf{v}\, d\mathbf{x} + \frac{1}{\epsilon}\int_\Gamma \mathbf{g}_\epsilon \cdot \mathbf{v}\, ds \qquad \forall\, \mathbf{v} \in \mathbf{H}^1(\Omega)\end{aligned} \tag{3.11}$$

and

$$\int_\Omega q \operatorname{div} \mathbf{u}_\epsilon\, d\mathbf{x} = 0 \qquad \forall\, q \in L^2(\Omega)\,; \tag{3.12}$$

i.e., $(\mathbf{u}_\epsilon, p_\epsilon, \mathbf{g}_\epsilon) \in \mathbf{H}^1(\Omega) \times L^2(\Omega) \times \mathbf{L}^2(\Gamma)$ satisfies the constraint equations (3.3)–(3.4). Finally, using the sequential weak lower semicontinuity of the functional $\mathcal{J}(\cdot,\cdot)$ we obtain

$$\begin{aligned}\mathcal{J}(\mathbf{u}_\epsilon, \mathbf{g}_\epsilon) &\le \liminf_{m\to\infty} \mathcal{J}(\mathbf{u}_m, \mathbf{g}_m) \\ &= \inf\big\{\mathcal{J}(\mathbf{u}, \mathbf{g}) : (\mathbf{u}, p, \mathbf{g}) \in \mathbf{H}^1(\Omega) \times L^2(\Omega) \times \mathbf{L}^2(\Gamma) \\ &\qquad \text{and } (\mathbf{u}, p, \mathbf{g}) \text{ satisfies (3.3)–(3.4)}\big\}\,.\end{aligned}$$

Hence, we have shown that $(\mathbf{u}_\epsilon, p_\epsilon, \mathbf{g}_\epsilon)$ is a solution for problem $(\mathrm{P})_\epsilon$. □

4. Convergence of solutions of Neumann control problems and suboptimality of the limit. Having shown the existence of a solution for $(\mathrm{P})_\epsilon$ for each ϵ, we now examine the convergence of $(\mathbf{u}_\epsilon, p_\epsilon, \mathbf{g}_\epsilon)$ as $\epsilon \to 0$.

THEOREM 4.1. *For each $\epsilon \in (0, 1/\nu)$, let $(\mathbf{u}_\epsilon, p_\epsilon, \mathbf{g}_\epsilon) \in \mathbf{H}^1(\Omega) \times L^2(\Omega) \times \mathbf{L}^2(\Gamma)$ be a solution of the optimal Neumann control problem $(\mathrm{P})_\epsilon$. Then there exists a $(\widehat{\mathbf{u}}, \widehat{p}, \widehat{\mathbf{g}}) \in \mathcal{U}_{ad}$ and a subsequence $\{\epsilon_k\}_{k=1}^\infty$ such that as $k \to \infty$,*

$$\mathbf{u}_{\epsilon_k} \rightharpoonup \widehat{\mathbf{u}} \quad \textit{in } \mathbf{H}^1(\Omega)\,, \qquad \overline{p_{\epsilon_k}} \rightharpoonup \widehat{p} \quad \textit{in } L^2_0(\Omega) \qquad \textit{and} \qquad \mathbf{g}_{\epsilon_k} \rightharpoonup \widehat{\mathbf{g}} \quad \textit{in } \mathbf{L}^2(\Gamma)\,,$$

where $\overline{p_{\epsilon_k}} = p_{\epsilon_k} - (1/|\Omega|)\int_\Omega p_{\epsilon_k}\, d\mathbf{x}$. Moreover,

$$\mathbf{u}_{\epsilon_k} \to \widehat{\mathbf{u}} \quad \textit{in } \mathbf{L}^2(\Omega)\,.$$

Proof. We first prove that the sets $\{\|\mathbf{u}_\epsilon\|_1\}$, $\{\|\overline{p_\epsilon}\|_0\}$, and $\{\|\mathbf{g}_\epsilon\|_{0,\Gamma}\}$ are all bounded independent of ϵ. Let $(\widetilde{\mathbf{u}}_\epsilon, \widetilde{p}_\epsilon)$ be the solution of (3.3)–(3.4) with $\mathbf{g} = \mathbf{0}$, i.e.,

$$\nu\int_\Omega (\nabla\widetilde{\mathbf{u}}_\epsilon) : (\nabla\mathbf{v})\,d\mathbf{x} + \frac{1}{\epsilon}\int_\Gamma \widetilde{\mathbf{u}}_\epsilon\cdot\mathbf{v}\,ds - \frac{1}{2}\int_\Gamma (\widetilde{\mathbf{u}}_\epsilon\cdot\mathbf{n})\widetilde{\mathbf{u}}_\epsilon\cdot\mathbf{v}\,ds + \int_\Omega (\widetilde{\mathbf{u}}_\epsilon\cdot\nabla)\widetilde{\mathbf{u}}_\epsilon\cdot\mathbf{v}\,d\mathbf{x} - \int_\Omega \widetilde{p}_\epsilon \operatorname{div}\mathbf{v}\,d\mathbf{x} = \int_\Omega \mathbf{f}\cdot\mathbf{v}\,d\mathbf{x} \qquad \forall\,\mathbf{v}\in\mathbf{H}^1(\Omega)$$

and

$$-\int_\Omega q\operatorname{div}\widetilde{\mathbf{u}}_\epsilon\,d\mathbf{x} = 0 \qquad \forall\, q\in L^2(\Omega)\,.$$

Lemma 3.2 gives us the estimate

$$\frac{\nu}{2}\int_\Omega |\nabla\widetilde{\mathbf{u}}_\epsilon|^2\,d\mathbf{x} + \frac{1}{4\epsilon}\int_\Gamma |\widetilde{\mathbf{u}}_\epsilon|^2\,ds \le \frac{1}{2\nu C_1}\int_\Omega |\mathbf{f}|^2\,d\mathbf{x}\,.$$

Since $(\widetilde{\mathbf{u}}_\epsilon, \widetilde{p}_\epsilon, \mathbf{0})$ is an admissible element for $(\mathrm{P})_\epsilon$, we have that

$$\mathcal{J}(\mathbf{u}_\epsilon, \mathbf{g}_\epsilon) \le \mathcal{J}(\widetilde{\mathbf{u}}_\epsilon, \mathbf{0})\,,$$

so that

$$\begin{aligned}\frac{\alpha}{2}\int_\Omega |\operatorname{curl}\mathbf{u}_\epsilon|^2\,d\mathbf{x} + \frac{\beta}{2}\int_\Gamma |\mathbf{g}_\epsilon|^2\,ds &\le \frac{\alpha}{2}\int_\Omega |\operatorname{curl}\widetilde{\mathbf{u}}_\epsilon|^2\,d\mathbf{x}\\ &\le \frac{\alpha}{2}\int_\Omega |\nabla\widetilde{\mathbf{u}}_\epsilon|^2\,d\mathbf{x} \le \frac{\alpha}{2\nu^2 C_1}\int_\Omega |\mathbf{f}|^2\,d\mathbf{x}\,,\end{aligned}$$

which implies

$$\int_\Omega |\operatorname{curl}\mathbf{u}_\epsilon|^2\,d\mathbf{x} \le \frac{1}{\nu^2 C_1}\int_\Omega |\mathbf{f}|^2\,d\mathbf{x}$$

and

$$\int_\Gamma |\mathbf{g}_\epsilon|^2\,ds \le \frac{\alpha}{\beta\nu^2 C_1}\int_\Omega |\mathbf{f}|^2\,d\mathbf{x}\,.$$

Since $(\mathbf{u}_\epsilon, p_\epsilon, \mathbf{g}_\epsilon)$ satisfies (3.3)–(3.4), we have the estimate (from Lemma 3.2)

$$\frac{\nu}{2}\int_\Omega |\nabla\mathbf{u}_\epsilon|^2\,d\mathbf{x} + \frac{1}{4\epsilon}\int_\Gamma |\mathbf{u}_\epsilon|^2\,ds \le \frac{1}{2\nu C_1}\int_\Omega |\mathbf{f}|^2\,d\mathbf{x} + \frac{1}{\epsilon}\int_\Gamma |\mathbf{g}_\epsilon|^2\,ds$$

so that

$$\int_\Gamma |\mathbf{u}_\epsilon|^2\,ds \le \frac{2\epsilon}{\nu C_1}\int_\Omega |\mathbf{f}|^2\,d\mathbf{x} + 4\int_\Gamma |\mathbf{g}_\epsilon|^2\,ds \le \frac{2+4\alpha/\beta}{\nu^2 C_1}\int_\Omega |\mathbf{f}|^2\,d\mathbf{x}\,.$$

Using Lemma 3.3 and the divergence-free condition of $\mathbf{u}_\epsilon$ we easily deduce that

$$\|\mathbf{u}_\epsilon\|_1^2 \le C_4\Big(\int_\Omega |\operatorname{curl}\mathbf{u}_\epsilon|^2\,d\mathbf{x} + \int_\Gamma |\mathbf{u}_\epsilon|^2\,ds\Big) \le \frac{C_4\big(3+4\alpha/\beta\big)}{\nu^2 C_1}\int_\Omega |\mathbf{f}|^2\,d\mathbf{x}\,.$$

Combining this last estimate with (3.6) we easily see that

$$\|\overline{p_\epsilon}\|_0 \le C_3\left(\sqrt{\frac{C_4(3+4\alpha/\beta)}{\nu C_1}}\|\mathbf{f}\|_0 + \frac{C_2C_4(3+4\alpha/\beta)}{\nu^2C_1}\|\mathbf{f}\|_0^2 + \|\mathbf{f}\|_0\right),$$

where $\overline{p_\epsilon} = p_\epsilon - (1/|\Omega|)\int_\Omega p_\epsilon\, d\mathbf{x}$. Thus we may extract a subsequence $\{\mathbf{u}_{\epsilon_k}\}$, $\{\overline{p_{\epsilon_k}}\}$, and $\{\mathbf{g}_{\epsilon_k}\}$ such that as $k \to \infty$,

$$\epsilon_k \to 0\,, \quad \mathbf{u}_{\epsilon_k} \rightharpoonup \widehat{\mathbf{u}} \quad \text{in } \mathbf{H}^1(\Omega)\,, \quad \overline{p_{\epsilon_k}} \rightharpoonup \widehat{p} \quad \text{in } L^2_0(\Omega)\,, \quad \text{and } \mathbf{g}_{\epsilon_k} \rightharpoonup \widehat{\mathbf{g}} \quad \text{in } \mathbf{L}^2(\Gamma)$$

for some $(\widehat{\mathbf{u}}, \widehat{p}, \widehat{\mathbf{g}}) \in \mathbf{H}^1(\Omega) \times L^2_0(\Omega) \times \mathbf{L}^2(\Gamma)$. Compact imbedding implies $\mathbf{u}_{\epsilon_k} \to \widehat{\mathbf{u}}$ in $\mathbf{L}^4(\Omega)$. We recall that $(\mathbf{u}_\epsilon, p_\epsilon, \mathbf{g}_\epsilon)$ satisfies equations (3.11)–(3.12). For each $\mathbf{v} \in \mathbf{H}^1_0(\Omega)$ and when $\epsilon = \epsilon_k$, equation (3.11) reduces to

$$\nu\int_\Omega \nabla\mathbf{u}_{\epsilon_k} : \nabla\mathbf{v}\, d\mathbf{x} + \int_\Omega (\mathbf{u}_{\epsilon_k}\cdot\nabla)\mathbf{u}_{\epsilon_k}\cdot\mathbf{v}\, d\mathbf{x} - \int_\Omega p_{\epsilon_k}\operatorname{div}\mathbf{v}\, d\mathbf{x} = \int_\Omega \mathbf{f}\cdot\mathbf{v}\, d\mathbf{x}\,.$$

Letting $k \to \infty$ yields

$$\nu\int_\Omega \nabla\widehat{\mathbf{u}} : \nabla\mathbf{v}\, d\mathbf{x} + \int_\Omega (\widehat{\mathbf{u}}\cdot\nabla)\widehat{\mathbf{u}}\cdot\mathbf{v}\, d\mathbf{x} - \int_\Omega \widehat{p}\operatorname{div}\mathbf{v}\, d\mathbf{x} = \int_\Omega \mathbf{f}\cdot\mathbf{v}\, d\mathbf{x} \qquad \forall\,\mathbf{v} \in \mathbf{H}^1_0(\Omega)\,.$$

Letting $k \to \infty$ in (3.12) yields

$$-\int_\Omega q\operatorname{div}\widehat{\mathbf{u}}\, d\mathbf{x} = 0 \qquad \forall\, q \in L^2_0(\Omega)\,.$$

Multiplying (3.11) (where we set $\epsilon = \epsilon_k$) by ϵ_k and letting $k \to \infty$ we obtain

$$\int_\Gamma \widehat{\mathbf{u}}\cdot\mathbf{v}\, ds = \int_\Gamma \widehat{\mathbf{g}}\cdot\mathbf{v}\, ds \qquad \forall\,\mathbf{v} \in \mathbf{H}^1(\Omega)\,,$$

which implies $\widehat{\mathbf{u}}\big|_\Gamma = \widehat{\mathbf{g}}$. This last relation and trace theorems imply $\widehat{\mathbf{g}} \in \mathbf{H}^{1/2}(\Gamma)$. Hence we have shown that $(\widehat{\mathbf{u}}, \widehat{p}, \widehat{\mathbf{g}})$ satisfies (2.7)–(2.9), i.e., that $(\widehat{\mathbf{u}}, \widehat{p}, \widehat{\mathbf{g}})$ is an admissible element for the optimal control problem (P). The strong convergence $\mathbf{u}_{\epsilon_k} \to \widehat{\mathbf{u}}$ in $\mathbf{L}^2(\Omega)$ follows from the compact imbedding $\mathbf{H}^1(\Omega) \hookrightarrow\hookrightarrow \mathbf{L}^2(\Omega)$. ∎

REMARK. *In the Neumann control problem* $(\mathrm{P})_\epsilon$ *we do not require* $\int_\Gamma \mathbf{g}_\epsilon\cdot\mathbf{n}\, ds = 0$. *However, the limit* $\widehat{\mathbf{g}}$ *automatically satisfies* $\int_\Gamma \widehat{\mathbf{g}}\cdot\mathbf{n}\, ds = 0$ *from the fact that* $\operatorname{div}\widehat{\mathbf{u}} = 0$ *and* $\widehat{\mathbf{u}}\big|_\Gamma = \widehat{\mathbf{g}}$. *The fact that* $\widehat{\mathbf{u}}\big|_\Gamma = \widehat{\mathbf{g}}$ *also implies* $\mathbf{g} \in \mathbf{H}^{1/2}(\Gamma)$, *although each* $\mathbf{g}_\epsilon$ *is merely in* $\mathbf{L}^2(\Gamma)$. ∎

We wish to show that the limit $(\widehat{\mathbf{u}}, \widehat{p}, \widehat{\mathbf{g}})$ is indeed a solution of the optimal Dirichlet control problem (P); namely, we will verify that

$$\mathcal{J}(\widehat{\mathbf{u}}, \widehat{\mathbf{g}}) \le \mathcal{J}(\mathbf{w}, \mathbf{z}) \qquad \forall\,(\mathbf{w}, r, \mathbf{z}) \in \mathcal{U}_{ad}. \tag{4.1}$$

In the remainder of this section we will prove that $(\widehat{\mathbf{u}}, \widehat{p}, \widehat{\mathbf{g}})$ is suboptimal in the sense that (4.1) is satisfied if $(\mathbf{w}, r)$ satisfies the additional condition

$$-r\mathbf{n} + \nu\frac{\partial\mathbf{w}}{\partial\mathbf{n}} \in \mathbf{L}^2(\Gamma)\,. \tag{4.2}$$

The optimality of $(\widehat{\mathbf{u}}, \widehat{p}, \widehat{\mathbf{g}})$ will be studied in the next section.

We will need the following lemma on integration by parts for functions in the space $H(\mathrm{div}\,,\Omega) \equiv \{\mathbf{v} \in \mathbf{L}^2(\Omega) : \mathrm{div}\,\mathbf{v} \in L^2(\Omega)\}$.

LEMMA 4.2. *Let* $\mathbf{w} \in H(\mathrm{div}\,,\Omega)$. *Then* $(\mathbf{w}\cdot\mathbf{n})\big|_\Gamma \in H^{-1/2}(\Gamma)$ *and*

$$\langle \mathbf{w}\cdot\mathbf{n}, v\rangle_\Gamma = \int_\Omega v\,\mathrm{div}\,\mathbf{w}\,d\mathbf{x} + \int_\Omega \mathbf{w}\cdot\nabla v\,d\mathbf{x} \qquad \forall\, v \in H^1(\Omega)\,,$$

where $\langle\cdot,\cdot\rangle_\Gamma$ *is the duality pairing between* $H^{-1/2}(\Gamma)$ *and* $H^{1/2}(\Gamma)$.

Proof. See [GR, equation (I.2.17), p. 28]. ∎

THEOREM 4.3. *Assume that* $(\widehat{\mathbf{u}},\widehat{p},\widehat{\mathbf{g}}) \in \mathcal{U}_{ad}$ *is the limit defined in Theorem* 4.1. *Then*

$$\mathcal{J}(\widehat{\mathbf{u}},\widehat{\mathbf{g}}) \le \mathcal{J}(\mathbf{w},\mathbf{z}) \qquad \forall\,(\mathbf{w},r,\mathbf{z}) \in \mathcal{U}_{ad} \text{ satisfying } (4.2).$$

Proof. Let $(\mathbf{w},r,\mathbf{z})$ be an arbitrary element in $\mathcal{U}_{ad}$ satisfying (4.2). By the definition of $\mathcal{U}_{ad}$, $(\mathbf{w},r,\mathbf{z})$ is a solution of

$$-\nu\Delta\mathbf{w} + (\mathbf{w}\cdot\nabla)\mathbf{w} + \nabla r = \mathbf{f} \qquad \text{in } \Omega\,, \tag{4.3}$$

$$\mathrm{div}\,\mathbf{w} = 0 \qquad \text{in } \Omega, \tag{4.4}$$

and

$$\mathbf{w}\big|_\Gamma = \mathbf{z}\,. \tag{4.5}$$

From (4.2) and the regularity results for the Navier–Stokes equations we obtain $(\mathbf{w},r) \in \mathbf{H}^{3/2}(\Omega)\times H^{1/2}(\Omega)$ and $-r\mathbf{n} + \nu\frac{\partial\mathbf{w}}{\partial\mathbf{n}} \in \mathbf{L}^2(\Gamma)$. Using (4.3) and the imbedding results for Sobolev spaces we obtain

$$\mathrm{div}\,\big(-rI + \nu\nabla\mathbf{w}\big) = \nu\Delta\mathbf{w} - \nabla r = -\mathbf{f} + (\mathbf{w}\cdot\nabla)\mathbf{w} \in \mathbf{L}^2(\Omega)\,.$$

By the integration-by-parts formula (Lemma 4.2) we have

$$\begin{aligned}\int_\Gamma [(-rI+\nu\nabla\mathbf{w})\cdot\mathbf{n}]\cdot\mathbf{v}\,ds &= \int_\Omega [-\nabla r + \nu\Delta\mathbf{w}]\cdot\mathbf{v}\,d\mathbf{x} + \int_\Omega [-rI+\nu\nabla\mathbf{w}] : \nabla\mathbf{v}\,d\mathbf{x}\\ &= \int_\Omega [-\mathbf{f} + (\mathbf{w}\cdot\nabla)\mathbf{w}]\cdot\mathbf{v}\,d\mathbf{x} - \int_\Omega r\,\mathrm{div}\,\mathbf{v}\,d\mathbf{x} + \nu\int_\Omega \nabla\mathbf{w} : \nabla\mathbf{v}\,d\mathbf{x},\end{aligned}$$

so that using (4.5) and adding/subtracting terms, we are led to

$$\begin{aligned}\nu\int_\Omega \nabla\mathbf{w} : \nabla\mathbf{v}\,d\mathbf{x} + \frac{1}{\epsilon}\int_\Gamma \mathbf{w}\cdot\mathbf{v}\,ds - \frac{1}{2}\int_\Gamma (\mathbf{w}\cdot\mathbf{n})\mathbf{w}\cdot\mathbf{v}\,ds + \int_\Omega (\mathbf{w}\cdot\nabla)\mathbf{w}\cdot\mathbf{v}\,d\mathbf{x}\\ - \int_\Omega r\,\mathrm{div}\,\mathbf{v}\,d\mathbf{x} = \int_\Omega \mathbf{f}\cdot\mathbf{v}\,d\mathbf{x} + \frac{1}{\epsilon}\int_\Gamma \mathbf{z}_\epsilon\cdot\mathbf{v}\,ds \qquad \forall\,\mathbf{v}\in\mathbf{H}^1(\Omega)\end{aligned}$$

where

$$\mathbf{z}_\epsilon \equiv \mathbf{z} + \epsilon\Big(-r\mathbf{n} + \nu\frac{\partial\mathbf{w}}{\partial\mathbf{n}}\Big) - \frac{\epsilon}{2}(\mathbf{w}\cdot\mathbf{n})\mathbf{w} \in \mathbf{L}^2(\Gamma)\,.$$

Thus, $(\mathbf{w},r,\mathbf{z}_\epsilon)$ is an admissible element for $(\mathrm{P})_\epsilon$, so that

$$\mathcal{J}(\mathbf{w},\mathbf{z}_\epsilon) \ge \mathcal{J}(\mathbf{u}_\epsilon,\mathbf{g}_\epsilon)\,.$$

Combining the last inequality with

$$\begin{aligned}\mathcal{J}(\mathbf{w},\mathbf{z}_\epsilon) &\equiv \frac{\alpha}{2}\int_\Omega |\operatorname{curl}\mathbf{w}|^2\,d\mathbf{x} + \frac{\beta}{2}\int_\Gamma \Big|\mathbf{z} - \epsilon r\mathbf{n} + \epsilon\nu\frac{\partial\mathbf{w}}{\partial\mathbf{n}} - \frac{\epsilon}{2}(\mathbf{w}\cdot\mathbf{n})\mathbf{w}\Big|^2\,ds \\ &= \mathcal{J}(\mathbf{w},\mathbf{z}) + \frac{\epsilon^2\beta}{2}\int_\Gamma \Big|-r\mathbf{n} + \nu\frac{\partial\mathbf{w}}{\partial\mathbf{n}} - \frac{1}{2}(\mathbf{w}\cdot\mathbf{n})\mathbf{w}\Big|^2\,ds \\ &\quad + \epsilon\beta\int_\Gamma \mathbf{z}\cdot\Big(-r\mathbf{n} + \nu\frac{\partial\mathbf{w}}{\partial\mathbf{n}} - \frac{1}{2}(\mathbf{w}\cdot\mathbf{n})\mathbf{w}\Big)\,ds,\end{aligned}$$

we obtain

$$\begin{aligned}\mathcal{J}(\mathbf{w},\mathbf{z}) \geq \mathcal{J}(\mathbf{u}_\epsilon,\mathbf{g}_\epsilon) &- \frac{\epsilon^2\beta}{2}\int_\Gamma \Big|-r\mathbf{n} + \nu\frac{\partial\mathbf{w}}{\partial\mathbf{n}} - \frac{1}{2}(\mathbf{w}\cdot\mathbf{n})\mathbf{w}\Big|^2\,ds \\ &- \epsilon\beta\int_\Gamma \mathbf{z}\cdot\Big(-r\mathbf{n} + \nu\frac{\partial\mathbf{v}}{\partial\mathbf{n}} - \frac{1}{2}(\mathbf{w}\cdot\mathbf{n})\mathbf{w}\Big)\,ds\,.\end{aligned}$$

Setting $\epsilon = \epsilon_k$ in the above relation (where ϵ_k is as defined in Theorem 4.1) and letting $k \to \infty$ we obtain

$$\mathcal{J}(\mathbf{w},\mathbf{z}) \geq \liminf_{k\to\infty} \mathcal{J}(\mathbf{u}_{\epsilon_k},\mathbf{g}_{\epsilon_k}) \geq \mathcal{J}(\widehat{\mathbf{u}},\widehat{\mathbf{g}})\,. \qquad \square$$

5. Optimality of the limit. In this section we will show that under certain restrictions on the data ν, $\mathbf{f}$, etc., the limit $(\widehat{\mathbf{u}},\widehat{p},\widehat{\mathbf{g}})$ defined in Theorem 4.1 is indeed a solution of the optimal Dirichlet control problem (P).

LEMMA 5.1. *Assume that* $(\widehat{\mathbf{u}},\widehat{p},\widehat{\mathbf{g}})$ *is the limit defined in Theorem* 4.1. *Then*

$$\mathcal{J}(\widehat{\mathbf{u}},\widehat{\mathbf{g}}) \leq \frac{\alpha}{2\nu^2 C_1}\int_\Omega |\mathbf{f}|^2\,d\mathbf{x}$$

Proof. Let $(\mathbf{u}_0,p_0) \in \mathbf{H}^1(\Omega)\times L^2(\Omega)$ be the solution of the Navier–Stokes equations (2.7)–(2.9) with the zero Dirichlet condition. Then $(\mathbf{u}_0,p_0,\mathbf{0}) \in \mathcal{U}_{ad}$. Lemma 3.2 gives us the estimate

$$\frac{\nu}{2}\int_\Omega |\nabla\mathbf{u}_0|^2\,d\mathbf{x} \leq \frac{1}{2\nu C_1}\int_\Omega |\mathbf{f}|^2\,d\mathbf{x}\,.$$

The regularity theory for the Navier–Stokes equations implies $(\mathbf{u}_0,p_0) \in \mathbf{H}^2(\Omega)\times H^1(\Omega)$ so that $-p_0\mathbf{n} + \nu\frac{\partial\mathbf{u}_0}{\partial\mathbf{n}} \in \mathbf{L}^2(\Gamma)$. Hence by Theorem 4.3,

$$\mathcal{J}(\widehat{\mathbf{u}},\widehat{\mathbf{g}}) \leq \mathcal{J}(\mathbf{u}_0,\mathbf{0}) = \frac{\alpha}{2}\int_\Omega |\operatorname{curl}\mathbf{u}_0|^2\,d\mathbf{x} \leq \frac{\alpha}{2}\int_\Omega |\nabla\mathbf{u}_0|^2\,d\mathbf{x} \leq \frac{\alpha}{2\nu^2 C_1}\int_\Omega |\mathbf{f}|^2\,d\mathbf{x}\,. \qquad \square$$

LEMMA 5.2. *Define* $\mathbf{H}_n^{1/2}(\Gamma) \equiv \big\{\mathbf{z}\in\mathbf{H}^{1/2}(\Gamma) : \int_\Gamma \mathbf{z}\cdot\mathbf{n}\,ds = 0\big\}$. *Then there exist a constant* $C_5 > 0$ *(depending on* Ω *only) and an extension operator* $E : \mathbf{H}_n^{1/2}(\Gamma) \to \mathbf{V}$ *such that* $\|E\mathbf{z}\|_1 \leq C_5\|\mathbf{z}\|_{1/2,\Gamma}$ *for every* $\mathbf{z}\in\mathbf{H}_n^{1/2}(\Gamma)$.

Proof. For each $\mathbf{z}\in\mathbf{H}_n^{1/2}(\Gamma)$ we define $\mathbf{w} = E\mathbf{z} \in \mathbf{V}$ as the unique solution of the Stokes problem

$$-\Delta\mathbf{w} + \nabla r = \mathbf{0} \qquad \text{in } \Omega\,,$$

$$\operatorname{div}\mathbf{w} = 0 \qquad \text{in } \Omega,$$

and

$$\mathbf{w}\big|_{\Gamma} = \mathbf{z}\,.$$

Clearly E maps $\mathbf{H}_n^{1/2}(\Gamma)$ into $\mathbf{V}$ linearly. The estimate $\|E\mathbf{z}\|_1 = \|\mathbf{w}\|_1 \le C_5\|\mathbf{z}\|_{1/2,\Gamma}$ follows from the estimates for Dirichlet boundary value problems for the steady-state Stokes equations (see [Te]). ∎

LEMMA 5.3. *Assume that* $(\mathbf{w}, r, \mathbf{z}) \in \mathbf{H}^1(\Omega) \times L_0^2(\Omega) \times \mathbf{H}_n^{1/2}(\Gamma)$ *and* $(\widetilde{\mathbf{w}}, \widetilde{r}, \widetilde{\mathbf{z}}) \in \mathbf{H}^1(\Omega) \times L_0^2(\Omega) \times \mathbf{H}_n^{1/2}(\Gamma)$ *satisfy, respectively,*

$$\begin{cases} -\nu\Delta\mathbf{w} + (\mathbf{w}\cdot\nabla)\mathbf{w} + \nabla r = \mathbf{f} & \text{in } \Omega\,, \\ \operatorname{div}\mathbf{w} = 0 & \text{in } \Omega, \\ \mathbf{w} = \mathbf{z} & \text{on } \Gamma, \end{cases}$$

and

$$\begin{cases} -\nu\Delta\widetilde{\mathbf{w}} + (\widetilde{\mathbf{w}}\cdot\nabla)\widetilde{\mathbf{w}} + \nabla\widetilde{r} = \mathbf{f} & \text{in } \Omega \\ \operatorname{div}\widetilde{\mathbf{w}} = 0 & \text{in } \Omega, \\ \widetilde{\mathbf{w}} = \widetilde{\mathbf{z}} & \text{on } \Gamma\,. \end{cases}$$

Assume further that $\|\mathbf{w}\|_1 < \frac{\nu C_0}{4C_2}$ *and* $\|\mathbf{z} - \widetilde{\mathbf{z}}\|_{1/2,\Gamma} \le \frac{\nu C_0}{4C_2C_5}$. *Then*

$$\|\widetilde{\mathbf{w}} - \mathbf{w}\|_1 \le \Big(\frac{4C_5}{C_0} + \sqrt{2}C_5 + C_5\Big)\|\widetilde{\mathbf{z}} - \mathbf{z}\|_{1/2,\Gamma} + \frac{4C_2C_5^2}{\nu C_0}\|\widetilde{\mathbf{z}} - \mathbf{z}\|_{1/2,\Gamma}^2\,.$$

Proof. Set $\boldsymbol{\xi} = \widetilde{\mathbf{w}} - \mathbf{w}$ and $\sigma = \widetilde{r} - r$. Then, by subtracting the relevant equations for $(\widetilde{\mathbf{w}}, \widetilde{r})$ and $(\mathbf{w}, r)$, we see that $(\boldsymbol{\xi}, \sigma) \in \mathbf{H}^1(\Omega) \times L_0^2(\Omega)$ satisfies

$$-\nu\Delta\boldsymbol{\xi} + (\boldsymbol{\xi}\cdot\nabla)\mathbf{w} + (\mathbf{w}\cdot\nabla)\boldsymbol{\xi} + (\boldsymbol{\xi}\cdot\nabla)\boldsymbol{\xi} + \nabla\sigma = \mathbf{0} \qquad \text{in } \Omega\,,$$

$$\operatorname{div}\boldsymbol{\xi} = 0 \qquad \text{in } \Omega,$$

and

$$\boldsymbol{\xi} = \widetilde{\mathbf{z}} - \mathbf{z} \qquad \text{on } \Gamma\,.$$

Put $\boldsymbol{\eta} = E(\widetilde{\mathbf{z}} - \mathbf{z})$, where E is the extension operator defined in Lemma 5.2. Then $\boldsymbol{\eta} \in \mathbf{H}^1(\Omega)$, $\boldsymbol{\eta}\big|_{\Gamma} = \widetilde{\mathbf{z}} - \mathbf{z}$, and $\|\boldsymbol{\eta}\|_1 \le C_5\|\widetilde{\mathbf{z}} - \mathbf{z}\|_{1/2,\Gamma}$. Setting $\boldsymbol{\zeta} = \boldsymbol{\xi} - \boldsymbol{\eta}$ we see that

$$\begin{aligned} -\nu\Delta\boldsymbol{\zeta} + (\boldsymbol{\zeta}\cdot\nabla)\mathbf{w} + (\boldsymbol{\eta}\cdot\nabla)\mathbf{w} + (\mathbf{w}\cdot\nabla)\boldsymbol{\zeta} + (\mathbf{w}\cdot\nabla)\boldsymbol{\eta} + (\boldsymbol{\eta}\cdot\nabla)\boldsymbol{\zeta} \\ + (\boldsymbol{\zeta}\cdot\nabla)\boldsymbol{\eta} + (\boldsymbol{\eta}\cdot\nabla)\boldsymbol{\eta} + (\boldsymbol{\zeta}\cdot\nabla)\boldsymbol{\zeta} + \nabla\sigma = \nu\Delta\boldsymbol{\eta} \qquad \text{in } \Omega\,, \end{aligned}$$

$$\operatorname{div}\boldsymbol{\zeta} = 0 \qquad \text{in } \Omega,$$

and

$$\boldsymbol{\zeta} = \mathbf{0} \qquad \text{on } \Gamma\,.$$

Using the weak form of these equations (see Definition 2.1) and the fact that $c(\mathbf{u}, \mathbf{v}, \mathbf{v}) = 0$ for all $\mathbf{v} \in \mathbf{H}_0^1(\Omega)$ and all $\mathbf{u} \in \mathbf{V}$, we obtain

$$\nu\|\nabla\boldsymbol{\zeta}\|_0^2 + c(\boldsymbol{\zeta}, \mathbf{w}, \boldsymbol{\zeta}) + c(\boldsymbol{\eta}, \mathbf{w}, \boldsymbol{\zeta}) + c(\mathbf{w}, \boldsymbol{\eta}, \boldsymbol{\zeta}) + c(\boldsymbol{\zeta}, \boldsymbol{\eta}, \boldsymbol{\zeta}) + c(\boldsymbol{\eta}, \boldsymbol{\eta}, \boldsymbol{\zeta}) = -\nu\int_{\Omega} (\nabla\boldsymbol{\eta}) : (\nabla\boldsymbol{\zeta})\, d\mathbf{x}$$

so that using inequalities (2.1), (2.3), and $rs \le \delta r^2 + \frac{1}{4\delta}s^2$, we are led to (for any $\delta > 0$)

$$\big(\nu C_0 - C_2\|\mathbf{w}\|_1 - C_2\|\boldsymbol{\eta}\|_1 - 3\delta\big)\|\boldsymbol{\zeta}\|_1^2 - \frac{C_2^2}{4\delta}\big(2\|\boldsymbol{\eta}\|_1^2\,\|\mathbf{w}\|_1^2 + \|\boldsymbol{\eta}\|_1^4\big) \le \frac{\nu^2}{4\delta}\|\nabla\boldsymbol{\eta}\|_0^2 + \delta\|\boldsymbol{\zeta}\|_1^2\,.$$

Choosing $\delta = \frac{\nu C_0}{16}$ and noting that $C_2\|\boldsymbol{\eta}\|_1 \le C_2C_5\|\widetilde{\mathbf{z}} - \mathbf{z}\|_{1/2,\Gamma} \le \frac{\nu C_0}{4}$, we obtain

$$\frac{\nu C_0}{4}\|\boldsymbol{\zeta}\|_1^2 \le \frac{\nu^2C_5^2}{4\delta}\|\widetilde{\mathbf{z}} - \mathbf{z}\|_{1/2,\Gamma}^2 + \frac{2C_2^2C_5^2}{4\delta}\Big(\frac{\nu C_0}{4C_2}\Big)^2\|\widetilde{\mathbf{z}} - \mathbf{z}\|_{1/2,\Gamma}^2 + \frac{C_2^2C_5^4}{4\delta}\|\widetilde{\mathbf{z}} - \mathbf{z}\|_{1/2,\Gamma}^4\,,$$

so that

$$\|\boldsymbol{\zeta}\|_1^2 \le \Big(\frac{16C_5^2}{C_0^2} + 2C_5^2\Big)\|\widetilde{\mathbf{z}} - \mathbf{z}\|_{1/2,\Gamma}^2 + \frac{16C_2^2C_5^4}{\nu^2C_0^2}\|\widetilde{\mathbf{z}} - \mathbf{z}\|_{1/2,\Gamma}^4\,.$$

Hence, using the inequality $(r^2 + s^2) \le (r + s)^2$ we are led to

$$\|\boldsymbol{\zeta}\|_1 \le \Big(\frac{4C_5}{C_0} + \sqrt{2}C_5\Big)\|\widetilde{\mathbf{z}} - \mathbf{z}\|_{1/2,\Gamma} + \frac{4C_2C_5^2}{\nu C_0}\|\widetilde{\mathbf{z}} - \mathbf{z}\|_{1/2,\Gamma}^2\,.$$

Finally, we use the triangle inequality to derive the estimate for $\boldsymbol{\xi}$:

$$\|\boldsymbol{\xi}\|_1 \le \|\boldsymbol{\eta}\|_1 + \|\boldsymbol{\zeta}\|_1 \le \Big(\frac{4C_5}{C_0} + \sqrt{2}C_5 + C_5\Big)\|\widetilde{\mathbf{z}} - \mathbf{z}\|_{1/2,\Gamma} + \frac{4C_2C_5^2}{\nu C_0}\|\widetilde{\mathbf{z}} - \mathbf{z}\|_{1/2,\Gamma}^2\,. \qquad \square$$

THEOREM 5.4. *Assume that*

$$\frac{\|\mathbf{f}\|_0}{\nu^2}\sqrt{\frac{\alpha}{\min\{\alpha,\beta\}}} < \frac{1}{4}\,\frac{C_0}{C_2}\sqrt{\frac{C_1}{C_4}} \tag{5.1}$$

and let $(\widehat{\mathbf{u}}, \widehat{p}, \widehat{\mathbf{g}})$ *be the limit defined in Theorem* 4.1. *Then* $(\widehat{\mathbf{u}}, \widehat{p}, \widehat{\mathbf{g}})$ *is a solution of the optimal Dirichlet control problem* (P).

Proof. Let $(\mathbf{w}, r, \mathbf{z}) \in \mathcal{U}_{ad}$ be given. We need to prove that $\mathcal{J}(\widehat{\mathbf{u}}, \widehat{\mathbf{g}}) \le \mathcal{J}(\mathbf{w}, \mathbf{z})$. Using Lemma 3.3 and the facts that $\operatorname{div}\mathbf{w} = 0$ and $\mathbf{w}\big|_\Gamma = \mathbf{z}$, we obtain

$$\mathcal{J}(\mathbf{w}, \mathbf{z}) \ge \frac{1}{2C_4}\min\{\alpha,\beta\}\|\mathbf{w}\|_1^2\,.$$

Hence, if $\frac{1}{2C_4}\min\{\alpha,\beta\}\|\mathbf{w}\|_1^2 > \mathcal{J}(\widehat{\mathbf{u}}, \widehat{\mathbf{g}})$, then $\mathcal{J}(\mathbf{w}, \mathbf{z}) > \mathcal{J}(\widehat{\mathbf{u}}, \widehat{\mathbf{g}})$. So we only need to consider the case where

$$\frac{1}{2C_4}\min\{\alpha,\beta\}\|\mathbf{w}\|_1^2 \le \mathcal{J}(\widehat{\mathbf{u}}, \widehat{\mathbf{g}})\,. \tag{5.2}$$

We assume (5.2) holds. Then using (5.1)–(5.2) and Lemma 5.1 we obtain

$$\begin{aligned}\|\mathbf{w}\|_1 \le \Big\{\frac{2C_4}{\min\{\alpha,\beta\}}\mathcal{J}(\widehat{\mathbf{u}}, \widehat{\mathbf{g}})\Big\}^{1/2} &\le \Big\{\frac{2C_4}{\min\{\alpha,\beta\}}\,\frac{\alpha}{2\nu^2C_1}\int_\Omega |\mathbf{f}|^2\,d\mathbf{x}\Big\}^{1/2}\\ &\le \Big\{\frac{2\alpha C_4}{\nu^2C_1\min\{\alpha,\beta\}}\Big\}^{1/2}\|\mathbf{f}\|_0 \le \frac{\nu C_0}{4C_2}\,.\end{aligned}$$

Using the denseness of $\mathbf{C}^\infty(\Gamma)\cap\mathbf{H}_n^{1/2}(\Gamma)$ in $\mathbf{H}_n^{1/2}(\Gamma)$ we may choose a sequence $\{\mathbf{z}_m\} \subset \mathbf{C}^\infty(\Gamma) \cap \mathbf{H}_n^{1/2}(\Gamma)$ such that $\|\mathbf{z}_m - \mathbf{z}\|_{1/2,\Gamma} \to 0$ as $m \to \infty$. For sufficiently large m, we have

$$\|\mathbf{z}_m - \mathbf{z}\|_{1/2,\Gamma} < \frac{\nu C_0}{4C_2C_5}\,.$$

Let $(\mathbf{w}_m, r_m) \in \mathbf{H}^1(\Omega) \times L_0^2(\Omega)$ be the solution of (2.7)–(2.8) with the Dirichlet condition $\mathbf{w}_m\big|_\Gamma = \mathbf{z}_m$. Using Lemma 5.3 we obtain $\|\mathbf{w}_m - \mathbf{w}\|_1 \to 0$ as $m \to \infty$. The regularity theories for the Navier–Stokes equations imply $(\mathbf{w}_m, r_m) \in \mathbf{H}^{3/2}(\Omega) \times H^{1/2}(\Omega)$ and $-r_m\mathbf{n} + \nu\frac{\partial \mathbf{w}_m}{\partial \mathbf{n}} \in \mathbf{L}^2(\Gamma)$. From Theorem 4.3 we obtain $\mathcal{J}(\mathbf{w}_m, \mathbf{z}_m) \geq \mathcal{J}(\widehat{\mathbf{u}}, \widehat{\mathbf{g}})$. Upon letting $m \to \infty$ we conclude that

$$\mathcal{J}(\mathbf{w}, \mathbf{z}) = \lim_{m\to\infty} \mathcal{J}(\mathbf{w}_m, \mathbf{z}_m) \geq \mathcal{J}(\widehat{\mathbf{u}}, \widehat{\mathbf{g}})\,. \qquad \square$$

REMARK. *Combining Theorems* 4.1 *and* 5.4 *we see that if the solution for* (P) *is unique, then we have the convergence as* $\epsilon \to 0$ *(instead of merely a subsequence convergence):*

$$\mathbf{u}_\epsilon \rightharpoonup \widehat{\mathbf{u}}\,, \quad \overline{p_\epsilon} \rightharpoonup \widehat{p}\,, \quad \textit{and } \mathbf{g}_\epsilon \rightharpoonup \widehat{\mathbf{g}}\,. \qquad \square$$

REMARK. *The small data requirement* (5.1) *is due to the small data requirements in Lemma* 5.3. *On the other hand, the small data requirements in Lemma* 5.3 *are those that are needed in order to show the continuous dependence of solutions on Dirichlet data for the Navier–Stokes equations. It is well known that the Navier–Stokes equations do not always have a unique solution; thus it seems hopeless to prove, without the small data requirement, the continuous dependence on Dirichlet data of solutions of the Navier–Stokes equations. This in turn suggests that it seems hopeless to prove Theorem* 5.4 *for arbitrary data. But for most practical purposes, one should be content with the suboptimal result of Theorem* 4.3 (*which does not require the smallness of data*). $\square$

6. Finite element approximations and numerical results. We have shown that the optimal solutions of Neumann control problems $(\mathrm{P})_\epsilon$ converge to an optimal solution of the Dirichlet control problem (P). Thus, we may choose a sufficiently small ϵ and solve $(\mathrm{P})_\epsilon$ to obtain an approximate solution for (P). In this section we briefly describe the solution procedures for $(\mathrm{P})_\epsilon$ with a fixed ϵ and present some numerical results. The purpose of this section is merely to confirm numerically the convergence of the optimal Neumann boundary control solutions which we have proven rigorously. Thus the presentation of this section is mostly formal.

The solution procedures for $(\mathrm{P})_\epsilon$ are as follows. First, by introducing the Lagrangian for $(\mathrm{P})_\epsilon$,

$$\begin{aligned}\mathcal{L}(\mathbf{u}, p, \mathbf{g}, \boldsymbol{\mu}, \rho) = \mathcal{J}(\mathbf{u}, \mathbf{g}) - \Big(\nu\, a(\mathbf{u}, \boldsymbol{\mu}) + \frac{1}{\epsilon}\int_\Gamma \mathbf{u}\cdot\boldsymbol{\mu}\, ds - \frac{1}{2}\int_\Gamma (\mathbf{u}\cdot\mathbf{n})\mathbf{u}\cdot\boldsymbol{\mu}\, ds \\ + c(\mathbf{u}, \mathbf{u}, \boldsymbol{\mu}) + b(\boldsymbol{\mu}, p) + b(\mathbf{u}, \rho) - \int_\Omega \mathbf{f}\cdot\boldsymbol{\mu}\, d\mathbf{x} - \frac{1}{\epsilon}\int_\Gamma \mathbf{g}\cdot\boldsymbol{\mu}\, ds\Big),\end{aligned}$$

and differentiating the Lagrangian with respect to each of its arguments we obtain the following optimality system of equations that the optimal solution for $(\mathrm{P})_\epsilon$ must satisfy:

$$\begin{aligned}\nu\int_\Omega (\nabla\mathbf{u}) : (\nabla\mathbf{v})\, d\mathbf{x} + \frac{1}{\epsilon}\int_\Gamma \mathbf{u}\cdot\mathbf{v}\, ds - \frac{1}{2}\int_\Gamma (\mathbf{u}\cdot\mathbf{n})\mathbf{u}\cdot\mathbf{v}\, ds + \int_\Omega (\mathbf{u}\cdot\nabla)\mathbf{u}\cdot\mathbf{v}\, d\mathbf{x} \\ -\int_\Omega p\,\mathrm{div}\,\mathbf{v}\, d\mathbf{x} = \int_\Omega \mathbf{f}\cdot\mathbf{v}\, d\mathbf{x} + \frac{1}{\epsilon}\int_\Gamma \mathbf{g}\cdot\mathbf{v}\, ds \qquad \forall\, \mathbf{v} \in \mathbf{H}^1(\Omega)\,,\end{aligned} \tag{6.1}$$

$$-\int_\Omega q\,\mathrm{div}\,\mathbf{u}\, d\mathbf{x} = 0 \qquad \forall\, q \in L^2(\Omega)\,, \tag{6.2}$$

$$\begin{aligned}
&\nu\int_\Omega(\nabla\boldsymbol{\mu}):(\nabla\mathbf{w})\,d\mathbf{x}+\frac{1}{\epsilon}\int_\Gamma\boldsymbol{\mu}\cdot\mathbf{w}\,ds-\frac{1}{2}\int_\Gamma(\mathbf{w}\cdot\mathbf{n})\mathbf{u}\cdot\boldsymbol{\mu}\,ds\\
&\quad-\frac{1}{2}\int_\Gamma(\mathbf{u}\cdot\mathbf{n})\mathbf{w}\cdot\boldsymbol{\mu}\,ds+\int_\Omega(\mathbf{u}\cdot\nabla)\mathbf{w}\cdot\boldsymbol{\mu}\,d\mathbf{x}+\int_\Omega(\mathbf{w}\cdot\nabla)\mathbf{u}\cdot\boldsymbol{\mu}\,d\mathbf{x}\\
&\quad-\int_\Omega\rho\,\mathrm{div}\,\mathbf{w}\,d\mathbf{x}=\int_\Omega(\mathrm{curl}\,\mathbf{u})\cdot(\mathrm{curl}\,\mathbf{w})\,d\mathbf{x}\qquad\forall\,\mathbf{w}\in\mathbf{H}^1(\Omega)\,,
\end{aligned}\tag{6.3}$$

$$-\int_\Omega r\,\mathrm{div}\,\boldsymbol{\mu}\,d\mathbf{x}=0\qquad\forall\,r\in L^2(\Omega),\tag{6.4}$$

and

$$\int_\Gamma\Big(\beta\mathbf{g}+\frac{1}{\epsilon}\boldsymbol{\mu}\Big)\cdot\mathbf{z}\,ds=0\qquad\forall\,\mathbf{z}\in\mathbf{L}^2(\Gamma)\,.$$

Note that we may use the last relation to eliminate $\mathbf{g}$ in (6.1) to obtain

$$\begin{aligned}
&\nu\int_\Omega\nabla\mathbf{u}:\nabla\mathbf{v}\,d\mathbf{x}+\frac{1}{\epsilon}\int_\Gamma\mathbf{u}\cdot\mathbf{v}\,ds-\frac{1}{2}\int_\Gamma(\mathbf{u}\cdot\mathbf{n})\mathbf{u}\cdot\mathbf{v}\,ds+\int_\Omega(\mathbf{u}\cdot\nabla)\mathbf{u}\cdot\mathbf{v}\,d\mathbf{x}\\
&\quad-\int_\Omega p\,\mathrm{div}\,\mathbf{v}\,d\mathbf{x}=\int_\Omega\mathbf{f}\cdot\mathbf{v}\,d\mathbf{x}-\frac{1}{\epsilon^2\beta}\int_\Gamma\boldsymbol{\mu}\cdot\mathbf{v}\,ds\qquad\forall\,\mathbf{v}\in\mathbf{H}^1(\Omega).
\end{aligned}\tag{6.5}$$

The system formed by (6.2)–(6.5) will be called *an optimality system of equations.*

Next we choose finite element subspaces and define finite element approximations of the optimality system. The finite element spaces $\mathbf{X}_h\subset\mathbf{H}^1(\Omega)$ and $S_h\subset L^2(\Omega)$ are chosen such that

$$\inf_{\mathbf{v}_h\in\mathbf{X}_h}\|\mathbf{v}_h-\mathbf{v}\|_1\le Ch^m\|\mathbf{v}\|_{m+1}\qquad\forall\,\mathbf{v}\in\mathbf{H}^{m+1}(\Omega)\,,$$

$$\inf_{q_h\in S_h}\|q_h-q\|_1\le Ch^m\|q\|_m\qquad\forall\,q\in H^m(\Omega)\,,$$

and

$$\inf_{q_h\in S_h}\sup_{\mathbf{v}_h\in\mathbf{X}_h}\frac{b(\mathbf{v}_h,q_h)}{\|\mathbf{v}_h\|_1\|q_h\|_0}\ge C_6\,.$$

The last discrete inf-sup condition is needed in finite element approximations of the Navier–Stokes equations (see, e.g., [GR]) and naturally is also needed in the approximations of the optimality system of equations. We define finite element approximations of the optimality system (6.2)–(6.5) as follows:

$$\begin{aligned}
&\nu\int_\Omega\nabla\mathbf{u}_h:\nabla\mathbf{v}_h\,d\mathbf{x}+\frac{1}{\epsilon}\int_\Gamma\mathbf{u}_h\cdot\mathbf{v}_h\,ds-\frac{1}{2}\int_\Gamma(\mathbf{u}_h\cdot\mathbf{n})\mathbf{u}_h\cdot\mathbf{v}_h\,ds-\int_\Omega p_h\,\mathrm{div}\,\mathbf{v}_h\,d\mathbf{x}\\
&\quad+\int_\Omega(\mathbf{u}_h\cdot\nabla)\mathbf{u}_h\cdot\mathbf{v}_h\,d\mathbf{x}=\int_\Omega\mathbf{f}\cdot\mathbf{v}_h\,d\mathbf{x}-\frac{1}{\epsilon^2\beta}\int_\Gamma\boldsymbol{\mu}_h\cdot\mathbf{v}_h\,ds\qquad\forall\,\mathbf{v}_h\in\mathbf{X}_h\,,
\end{aligned}\tag{6.6}$$

$$-\int_\Omega q_h\,\mathrm{div}\,\mathbf{u}_h\,d\mathbf{x}=0\qquad\forall\,q_h\in S_h\,,\tag{6.7}$$

$$\begin{aligned}
&\nu\int_\Omega(\nabla\boldsymbol{\mu}_h):(\nabla\mathbf{w}_h)\,d\mathbf{x}+\frac{1}{\epsilon}\int_\Gamma\boldsymbol{\mu}_h\cdot\mathbf{w}_h\,ds-\frac{1}{2}\int_\Gamma(\mathbf{w}_h\cdot\mathbf{n})\mathbf{u}_h\cdot\boldsymbol{\mu}_h\,ds\\
&\quad-\frac{1}{2}\int_\Gamma(\mathbf{u}_h\cdot\mathbf{n})\mathbf{w}_h\cdot\boldsymbol{\mu}_h\,ds+\int_\Omega(\mathbf{u}_h\cdot\nabla)\mathbf{w}_h\cdot\boldsymbol{\mu}_h\,d\mathbf{x}-\int_\Omega\rho_h\,\mathrm{div}\,\mathbf{w}_h\,d\mathbf{x}\\
&\quad+\int_\Omega(\mathbf{w}_h\cdot\nabla)\mathbf{u}_h\cdot\boldsymbol{\mu}_h\,d\mathbf{x}=\int_\Omega(\mathrm{curl}\,\mathbf{u}_h)\cdot(\mathrm{curl}\,\mathbf{w}_h)\,d\mathbf{x}\qquad\forall\,\mathbf{w}_h\in\mathbf{X}_h,
\end{aligned}\tag{6.8}$$

TABLE 1
$\mathbf{L}^2(\Omega)$ *errors of each two consecutive optimal solutions.*

i	1	2	3	4	5	6	7	8
ϵ_i	10	1	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
$\left\|\mathbf{u}_i - \mathbf{u}_{i+1}\right\|_2^2$	.3553	.04725	.01763	.002506	.0002618	.00002627	.000009	

and

$$-\int_\Omega r_h \operatorname{div} \boldsymbol{\mu}_h \, d\mathbf{x} = 0 \qquad \forall\, r_h \in S_h\,. \tag{6.9}$$

We solve the discrete, nonlinear system of equations (6.6)–(6.9) by Newton's method with the initial guess obtained from solving the corresponding linear system of equations (simply dropping all nonlinear terms in the optimality system).

It is possible to use the techniques of [GHS1] to mathematically justify these solution procedures, e.g., to prove the existence of a solution $(\mathbf{u}, p, \boldsymbol{\mu}, \rho)$ for (6.2)–(6.5) such that $(\mathbf{u}, p, -\boldsymbol{\mu}/(\beta\epsilon))$ gives a solution of $(\mathrm{P})_\epsilon$; to prove that for each solution of (6.2)–(6.5) and for each sufficiently small h, there exists a solution $(\mathbf{u}_h, p_h, \boldsymbol{\mu}_h, \rho_h)$ such that as $h \to 0$,

$$\mathbf{u}_h \to \mathbf{u}, \quad p_h \to p, \quad \boldsymbol{\mu}_h \to \boldsymbol{\mu}, \quad \text{and} \quad \rho_h \to \rho\,;$$

and to prove that if $(\mathbf{u}_h, p_h, \boldsymbol{\mu}_h, \rho_h) \in \mathbf{H}^{m+1}(\Omega) \times \mathbf{H}^m(\Omega) \times \mathbf{H}^{m+1}(\Omega) \times \mathbf{H}^m(\Omega)$, then

$$\begin{aligned}\|\mathbf{u}_h - \mathbf{u}\|_1 + \|p_h - p\|_0 + \|\boldsymbol{\mu}_h - \boldsymbol{\mu}\|_1 + \|\rho_h - \rho\|_0 \\ \le Ch^m\big(\|\mathbf{u}\|_{m+1} + \|p\|_m + \|\boldsymbol{\mu}\|_{m+1} + \|\rho\|_m\big)\,.\end{aligned}$$

However, the detailed justification of these results are beyond the scope of this paper.

We conclude this paper by presenting some numerical results for two test problems. These results confirm numerically the convergence results we established, i.e., Theorem 4.1. Further computational studies of the proposed method will be reported elsewhere.

In the first example we consider a Dirichlet optimal control problem for the Navier–Stokes equations (1.2)–(1.4) with the following data: Ω is the unit square; $\nu = 0.1$; and the prescribed body force $\mathbf{f} = (f_1, f_2)^T$, where $f_1 = 0.8\pi^2 \sin(2\pi x)\cos(2\pi y) + 2\pi \sin(2\pi x)\cos(2\pi x)$ and $f_2 = -0.8\pi^2 \cos(2\pi x)\sin(2\pi y) + 2\pi \sin(2\pi y)\cos(2\pi y)$. The functional is given by (1.1), wherein we choose $\alpha = \beta = 1$. For each sufficiently small ϵ, we can compute an optimal solution for $(\mathrm{P})_\epsilon$ by solving the discrete optimality system (6.6)–(6.9). We used a uniform mesh in Ω with 162 triangles and chose the finite element spaces to be continuous piecewise quadratics for the velocity/adjoint velocity and continuous piecewise linear functions for the pressure/adjoint pressure. We computed the optimal solutions for a sequence of ϵ values: $\epsilon_1 = 10$, $\epsilon_2 = 1$, $\epsilon_3 = 10^{-1}$, $\epsilon_4 = 10^{-2}$, $\epsilon_5 = 10^{-3}$, $\epsilon_6 = 10^{-4}$, $\epsilon_7 = 10^{-5}$, and $\epsilon_8 = 10^{-6}$. We also computed the $\mathbf{L}^2(\Omega)$ norms of $\widehat{\mathbf{u}}_{\epsilon_{i+1}} - \widehat{\mathbf{u}}_{\epsilon_i}$, and these are summarized in Table 1.

The computational results in Table 1 are consistent with the convergence results of Theorem 4.1. The solution $(\mathbf{u}_0, p_0)$ of the equations with $\mathbf{g} = \mathbf{0}$ is given by $\mathbf{u}_0 = \big(\sin(2\pi x)\cos(2\pi y), -\cos(2\pi x)\sin(2\pi y)\big)^T$ and $p_0 = 0$, and $\int_\Omega |\operatorname{curl} \mathbf{u}_0|^2 \, d\mathbf{x} = 38.8605$. The values of $\int_\Omega |\operatorname{curl} \mathbf{u}_\epsilon|^2 \, d\mathbf{x}$ for $\epsilon \le 10^{-5}$ are around 18.

In the second example we consider a Dirichlet optimal control problem for the Navier–Stokes equations (1.2)–(1.4) with the following data: Ω is the unit square;

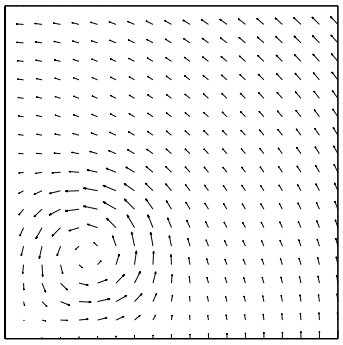


FIG. 1. *Uncontrolled velocity field.*

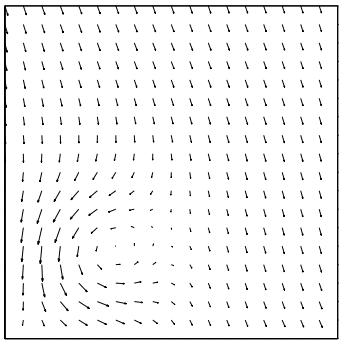


FIG. 2. *Optimal velocity field* ($\epsilon = 1$).

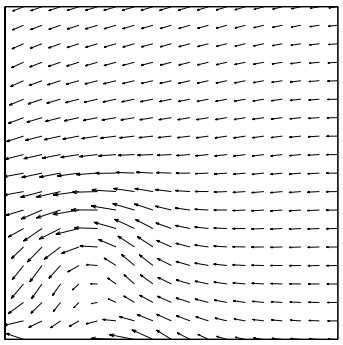


FIG. 3. *Optimal velocity field* ($\epsilon = 10^{-1}$).

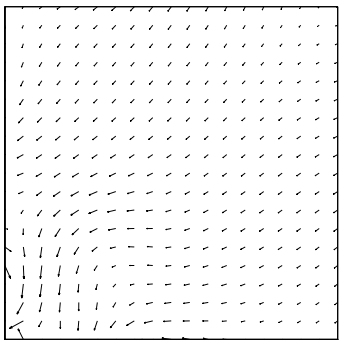


FIG. 4. *Optimal velocity field* ($\epsilon = 10^{-3}$).

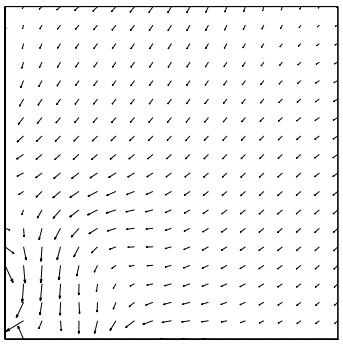


FIG. 5. *Optimal velocity field* ($\epsilon = 10^{-5}$).

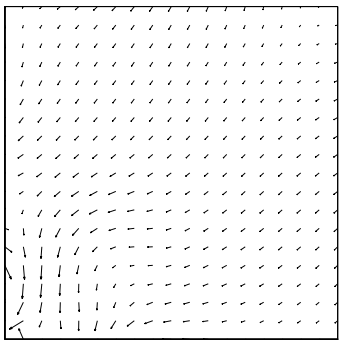


FIG. 6. *Optimal velocity field* ($\epsilon = 10^{-7}$).

TABLE 2
The $\mathbf{L}^2(\Omega)$-norm of the vorticity of optimal solutions.

ϵ_i	1	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}
$\left\|\left\|\operatorname{curl}\mathbf{u}_\epsilon\right\|\right\|_2^2$	12.20	11.81	11.64	11.63	11.62	11.62	11.62	11.62

$\nu = 0.1$; and the prescribed body force $\mathbf{f} = (f_1, f_2)^T$, where

$$\mathbf{f} = \begin{pmatrix} -50\nu\pi\,\cos((x-0.25)\pi/0.4)\,\sin((y-0.25)\pi/0.4) - \frac{20}{\pi}\,\sin((x-0.25)\pi/0.2) \\ 50\nu\pi\,\sin((x-0.25)\pi/0.4)\,\cos((y-0.25)\pi/0.4) - \frac{20}{\pi}\,\sin((y-0.25)\pi/0.2) \end{pmatrix}$$

in the region $\{(x,y) : |x-0.25| \le 0.2, |y-0.25| \le 0.2\}$ and $\mathbf{f} = (-0.25x, -0.25y)^T$ elsewhere on the unit square. The functional is given by (1.1), wherein we choose $\alpha = 100$ and $\beta = 1$. We used the same mesh as in the first example. We computed the optimal solutions for a sequence of ϵ values by solving the discrete optimality system (6.6)–(6.9): $\epsilon_1 = 1$, $\epsilon_2 = 10^{-1}$, $\epsilon_3 = 10^{-2}$, $\epsilon_4 = 10^{-3}$, $\epsilon_5 = 10^{-4}$, $\epsilon_6 = 10^{-5}$, $\epsilon_7 = 10^{-6}$, and $\epsilon_8 = 10^{-7}$. We also computed the $\mathbf{L}^2(\Omega)$-norms of $\operatorname{curl}\widehat{\mathbf{u}}_{\epsilon_i}$ as shown in Table 2. The $\mathbf{L}^2(\Omega)$-norm of the the vorticity of the uncontrolled velocity field is 39.77.

Figure 1 shows the uncontrolled flow field. Figures 2–6 depict the optimal velocity fields for various ϵ values we tested. We could easily visualize from these figures the convergence of the optimal solutions as $\epsilon \to 0$. Also, by comparing the uncontrolled flow field with the optimal flow fields (for small ϵ), we clearly see the reduction in recirculation in the optimal control solutions.

REFERENCES

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

[AT] F. ABERGEL AND R. TEMAM, *On some control problems in fluid mechanics*, Theoret. Comput. Fluid Dynamics, 1 (1990), pp. 303–325.

[Ba] I. BABUŠKA, *The finite element method with penalty*, Math. Comp., 27 (1973), pp. 221–228.

[FS1] H. FATTORINI AND S. SRITHARAN, *Necessary and sufficient conditions for optimal controls in viscous flow problems*, Proc. Roy. Soc. Edinburgh Ser. A, 124A (1994), pp. 211–251.

[FS2] H. FATTORINI AND S. SRITHARAN, *Optimal chattering control for viscous flows*, Nonlinear Anal., 25 (1995), pp. 763–797.

[FS3] H. FATTORINI AND S. SRITHARAN, *Existence of optimal controls for viscous flow problems*, Proc. Roy. Soc. London Ser. A, 439 (1992), pp. 81–102.

[Fu1] A. FURSIKOV, *On some control problems and results concerning the unique solvability of a mixed boundary value problem for the three-dimensional Navier-Stokes and Euler systems*, Soviet Math. Dokl., 3 (1980), pp. 889–893.

[Fu2] A. FURSIKOV, *Control problems and theorems concerning the unique solvability of a mixed boundary value problem for the three-dimensional Navier-Stokes and Euler equations*, Math USSR Sb., 43 (1982), pp. 281–307.

[Fu3] A. FURSIKOV, *Properties of solutions of some extremal problems connected with the Navier-Stokes system*, Math USSR Sb., 46 (1983), pp. 323–351.

[FGH] A. FURSIKOV, M. GUNZBURGER, AND L. HOU, *Boundary value problems and optimal boundary control for the Navier–Stokes system: The two-dimensional case*, SIAM J. Control and Optim., 36 (1998), pp. 852–894.

[Gun] M. GUNZBURGER, ED., *Flow Control*, IMA Vol. Math. Appl. 68, Springer-Verlag, New York, 1995.

[GHS1] M. GUNZBURGER, L. HOU, AND T. SVOBODNY, *Analysis and finite element approximation of optimal control problems for the stationary Navier-Stokes equations with distributed and Neumann controls*, Math. Comp., 57 (1991), pp. 123–151.

[GHS2] M. GUNZBURGER, L. HOU, AND T. SVOBODNY, *Analysis and finite element approximation of optimal control problems for the stationary Navier-Stokes equations with Dirichlet controls*, Modél. Math. Anal. Numér., 25 (1991), pp. 711–748.

[GHS3] M. Gunzburger, L. Hou, and T. Svobodny, *Boundary velocity control of incompressible flow with an application to viscous drag reduction*, SIAM J. Control Optim., 30 (1992), pp. 167–181.

[GR] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier-Stokes Equations*, Springer-Verlag, Berlin, 1986.

[HS] L. Hou and T. Svobodny, *Optimization problems for the Navier-Stokes equations with regular boundary controls*, J. Math. Anal. Appl., 177 (1993), pp. 342–367.

[HY1] L. Hou and Y. Yan, *Dynamics for controlled Navier–Stokes systems with distributed controls*, SIAM J. Control Optim., 35 (1997), pp. 654–677.

[HY2] L. Hou and Y. Yan, *Dynamics and approximations of a velocity tracking problem for incompressible flows with piecewise distributed controls*, SIAM J. Control Optim., 35 (1997), pp. 1847–1885.

[HRY] L. Hou, S. Ravindran, and Y. Yan, *Numerical solutions of optimal distributed control problems for incompressible flows*, International Journal of Computational Fluid Dynamics, 8 (1997), pp. 99–114.

[Li] J.-L. Lions, *Control of Distributed Singular Systems*, Bordas, Paris, 1985.

[Ma] H. Manouzi, *The stokes problem and the mixed boundary conditions*, C. R. Math. Rep. Acad. Sci. Canada, 12 (1990), pp. 155–160.

[Ne] J. Nečas, *Les méthodes directes en théorie des équations elliptiques*, Masson, Paris, 1967.

[S1] S. Sritharan, *Dynamic programming of the Navier-Stokes equations*, System Control Lett., 16 (1991), pp. 299–307.

[S2] S. Sritharan, *An optimal control problem in exterior hydrodynamics*, Roy. Soc. Edinburgh Proc. A, 121 (1992), pp. 5–32.

[Te] R. Temam, *Navier-Stokes Equations, Theory and Numerical Methods*, North-Holland, Amsterdam, 1979.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 5, pp. 1815–1832, September 1998

013

APPROXIMATE JACOBIAN MATRICES FOR NONSMOOTH CONTINUOUS MAPS AND C^1-OPTIMIZATION*

V. JEYAKUMAR† AND D. T. LUC‡

Abstract. The notion of approximate Jacobian matrices is introduced for a continuous vector-valued map. It is shown, for instance, that the Clarke generalized Jacobian is an approximate Jacobian for a locally Lipschitz map. The approach is based on the idea of convexificators of real-valued functions. Mean value conditions for continuous vector-valued maps and Taylor's expansions for continuously Gâteaux differentiable functions (i.e., C^1-functions) are presented in terms of approximate Jacobians and approximate Hessians, respectively. Second-order necessary and sufficient conditions for optimality and convexity of C^1-functions are also given.

Key words. generalized Jacobians, nonsmooth analysis, mean value conditions, optimality conditions

AMS subject classifications. 49A52, 90C30, 26A24

PII. S0363012996311745

1. Introduction.

Over the past two decades, a great deal of research has focused on the study of first- and second-order analysis of real-valued nonsmooth functions [2, 3, 4, 5, 11, 12, 14, 15, 21, 23, 24, 20, 25, 27, 28, 29, 30, 34, 35]. The results of nonsmooth analysis of real-valued functions now provide basic tools of modern analysis in many branches of mathematics, such as mathematical programming, control, and mechanics. Indeed, the range of applications of nonsmooth calculus demonstrates its basic nature of nonsmooth phenomena in the mathematical and engineering sciences.

On the other hand, research in the area of nonsmooth analysis of vector-valued maps has been of substantial interest in recent years [2, 6, 7, 8, 9, 10, 18, 21, 22, 23, 24, 29, 31]. In particular, it is known that the development and analysis of generalized Jacobian matrices for nonsmooth vector-valued maps are crucial from the viewpoint of control problems and numerical methods of optimization. For instance, the Clarke generalized Jacobian matrices [2] of a locally Lipschitz map play an important role in the Newton-based numerical methods for solving nonsmooth equations and optimization problems (see [26] and other references therein, and see also [17, 18, 19] for other applications). Warga [32, 33] examined derivative (unbounded derivative) containers in the context of local and global inverse function theorems as set-valued derivatives for locally Lipschitz (continuous) vector-valued maps. Mordukhovich [21, 22] developed generalized differential calculus for general nonsmooth vector-valued maps using the set-valued derivatives, called coderivatives [9, 21].

Our aim in this paper is to introduce a new concept of approximate Jacobian matrices for continuous vector-valued maps that are not necessarily locally Lipschitz, develop certain calculus rules for approximate Jacobians, and apply the concept to optimization problems involving continuously Gâteaux differentiable functions. This

*Received by the editors November 8, 1996; accepted for publication (in revised form) October 2, 1997; published electronically July 9, 1998. This research was partially supported by a grant from the Australian Research Council.

http://www.siam.org/journals/sicon/36-5/31174.html

†Department of Applied Mathematics, University of New South Wales, Sydney 2052, Australia (jeya@maths.unsw.edu.au). Some of the work of this author was carried out while visiting the Centre for Experimental and Constructive Mathematics at the Simon Fraser University, Canada.

‡Institute for Mathematics, Hanoi, Vietnam (dtluc@thevinh.ac.vn). Some of the work of this author was done while visiting the University of New South Wales.

concept is a generalization of the idea of convexificators of real-valued functions, studied recently in [4, 5, 13], to vector-valued maps. Convexificators provide two-sided convex approximations [30] for real-valued functions. Unlike the set-valued generalized derivatives [9, 21, 22, 32, 33], mentioned above for vector-valued maps, the approximate Jacobian is defined as a closed subset of the space of $(n \times m)$ matrices for a vector-valued map from $\mathbb{R}^n$ into $\mathbb{R}^m$.

Approximate Jacobians not only extend the nonsmooth analysis of locally Lipschitz maps to continuous maps but also unify and strengthen various results of nonsmooth analysis. They also enjoy useful calculus, such as the generalized mean value property and chain rules. Moreover, approximate Jacobians allow us to present second-order optimality conditions in easily verifiable forms in terms of approximate Hessian matrices for C^1-optimization problems, extending the corresponding results for $C^{1,1}$-problems [7].

The outline of the paper is as follows. In section 2, approximate Jacobian matrices are introduced, and it is shown that for a locally Lipschitz map the Clarke generalized Jacobian is an approximate Jacobian. Various examples of approximate Jacobians are also given. Section 3 establishes mean value conditions for continuous vector-valued maps and provides necessary and sufficient conditions in terms of approximate Jacobians for a continuous map to be locally Lipschitz. Various calculus rules for approximate Jacobians are given in section 4. Approximate Hessian matrices are introduced in section 5, and their connections to $C^{1,1}$-functions are discussed. Section 6 presents generalizations of Taylor's expansions for C^1-functions. In section 7, second-order necessary and sufficient conditions for optimality and convexity of C^1-functions are given.

2. Approximate Jacobians for continuous maps. This section contains notation, definitions, and preliminaries that will be used throughout the paper. Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be a continuous function which has components $(f_1, \ldots, f_m)$. For each $v \in \mathbb{R}^m$, the composite function, $(vF) : \mathbb{R}^n \to \mathbb{R}$, is defined by

$$(vF)(x) = \langle v,\ F(x)\rangle = \sum_{i=1}^{m} v_i f_i(x).$$

The lower Dini directional derivative and the upper Dini directional derivative of vF at x in the direction $u \in \mathbb{R}^n$ are defined by

$$(vF)^-(x,u) := \liminf_{t\downarrow 0} \frac{(vF)(x+tu) - (vF)(x)}{t},$$

$$(vF)^+(x,u) := \limsup_{t\downarrow 0} \frac{(vF)(x+tu) - (vF)(x)}{t}.$$

We denote by $L(\mathbb{R}^n, \mathbb{R}^m)$ the space of all $(n \times m)$ matrices. The *convex hull* and the *closed convex hull* of a set A in a topological vector space are denoted by $co(A)$ and $\overline{co}(A)$, respectively.

DEFINITION 2.1. *The map $F : \mathbb{R}^n \to \mathbb{R}^m$ admits an* approximate Jacobian *$\partial^* F(x)$ at $x \in R^n$ if $\partial^* F(x) \subseteq L(\mathbb{R}^n, \mathbb{R}^m)$ is closed, and for each $v \in R^m$,*

$$(vF)^-(x,u) \le \sup_{M \in \partial^* F(x)} \langle Mv, u\rangle \ \ \forall u \in \mathbb{R}^n. \tag{2.1}$$

A matrix M of $\partial^* F(x)$ is called an approximate Jacobian matrix of F at x. Note that condition (2.1) is equivalent to the condition

$$(vF)^+(x,u) \geq \inf_{M\in\partial^* F(x)} \langle Mv, u\rangle \ \ \forall u \in \mathbb{R}^n. \tag{2.2}$$

It is worth noting that the inequality (2.1) means that the set $\partial^* F(x)v$ is an upper convexificator [13, 16] of the function vF at x. Similarly, the inequality (2.2) states that $\partial^* F(x)v$ is a lower convexificator of vF at x. In the case $m = 1$, the inequality (2.1) (or (2.2)) is equivalent to the condition

$$F^-(x,u) \leq \sup_{x^*\in\partial^* F(x)} \langle x^*, u\rangle \ \text{ and } \ F^+(x,u) \geq \inf_{x^*\in\partial^* F(x)} \langle x^*, u\rangle; \tag{2.3}$$

thus, the set $\partial^* F(x)$ is a convexificator of F at x. Also note that in the case $m = 1$, condition (2.3) is also equivalent to the condition that for each $\alpha \in \mathbb{R}$,

$$(\alpha F)^-(x,u) \leq \sup_{x^*\in\partial^* F(x)} \langle \alpha x^*, u\rangle \ \ \forall u \in \mathbb{R}^n. \tag{2.4}$$

Similarly, the condition (2.3) is also equivalent to the condition that for each $\alpha \in \mathbb{R}$,

$$(\alpha F)^+(x,u) \geq \inf_{x^*\in\partial^* F(x)} \langle \alpha x^*, u\rangle \ \ \forall u \in \mathbb{R}^n. \tag{2.5}$$

For applications of convexificators, see [5, 13, 16]. To clarify the definition, let us consider some examples.

Example 2.2. If $F : \mathbb{R}^n \to \mathbb{R}^m$ is continuously differentiable at x, then any closed subset $\Phi(x)$ of $L(\mathbb{R}^n, \mathbb{R}^m)$ containing the Jacobian $\nabla F(x)$ is an approximate Jacobian of F at x. In this case, for each $v \in \mathbb{R}^m$,

$$(vF)^-(x,u) = \langle \nabla F(x)v, u\rangle \leq \sup_{M\in\Phi(x)} \langle Mv, u\rangle \quad \forall u \in \mathbb{R}^n.$$

Observe from the definition of the approximate Jacobian that for any map $F : \mathbb{R}^n \to \mathbb{R}^m$, the whole space $L(\mathbb{R}^n, \mathbb{R}^m)$ serves as a trivial approximate Jacobian for F at any point in $\mathbb{R}^n$. Let us now examine approximate Jacobians for locally Lipschitz maps.

Example 2.3. Suppose that $F : \mathbb{R}^n \to \mathbb{R}^m$ is locally Lipschitz at x. Then the Clarke generalized Jacobian $\partial_C F(x)$ is an approximate Jacobian of F at x. Indeed, for each $v \in \mathbb{R}^m$,

$$\partial^\circ(vF)(x) = \partial_C F(x)v. \tag{2.6}$$

Consequently, for each $u \in \mathbb{R}^n$,

$$(vF)^\circ(x,u) = \max_{\xi\in\partial^\circ(vF)(x)} \langle \xi, u\rangle = \max_{M\in\partial_C F(x)} \langle Mv, u\rangle,$$

where

$$\partial_C F(x) = co\{\lim_{n\to\infty} \nabla F(x_n)^T : x_n \in \Omega,\ x_n \to x\},$$

Ω is the set of points in $\mathbb{R}^n$ where F is differentiable, and the Clarke directional derivative of vF is given by

$$(vF)^\circ(x,u) = \limsup_{\substack{x'\to x\\ t\downarrow 0}} \frac{\langle v, F(x'+tu) - F(x')\rangle}{t}.$$

Since for each $u \in \mathbb{R}^n$,

$$(vF)^-(x,u) \le (vF)^\circ(x,u) \ \ \forall u \in R^n,$$

the set $\partial_C F(x)$ is an approximate Jacobian of F at x.

For the locally Lipschitz map $F : \mathbb{R}^n \to \mathbb{R}^m$, the set

$$\partial_B F(x) := \{ \lim_{n\to\infty} \nabla F(x_n)^T : x_n \in \Omega,\ x_n \to x\}$$

is also an approximate Jacobian of F at x. The set $\partial_B F(x)$ is known as the B-subdifferential of F at x, which plays a significant role in the development of nonsmooth Newton methods (see [26]). In passing, note that for each $v \in \mathbb{R}^m$,

$$\partial^\circ(vF)(x) = co(\partial_M(vF)(x)) = co(D^*F(x)(v)),$$

where the set-valued mapping $D^*F(x)$ from $\mathbb{R}^m$ into $\mathbb{R}^n$ is the coderivative of F at x and $\partial_M(vF)(x)$ is the first-order subdifferential of vF at x in the sense of Mordukhovich [22]. However, for locally Lipschitz maps, the coderivative does not appear to have a representation of the form (2.6), which allowed us above to compare approximate Jacobians with the Clarke generalized Jacobian. The reader is referred to [9, 21, 22, 29] for a more general definition and associated properties of coderivatives. A second-order analogue of the coderivative for vector-valued maps is given recently in [10].

Let us look at a numerical example of a locally Lipschitz map where the Clarke generalized Jacobian strictly contains an approximate Jacobian.

Example 2.4. Consider the function $F : \mathbb{R}^2 \to \mathbb{R}^2$

$$F(x,y) = (|x|, |y|).$$

Then

$$\partial^* F(0) = \left\{ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} \right\}$$

is an approximate Jacobian of F at 0. On the other hand, the Clarke generalized Jacobian

$$\partial_C F(0) = \left\{ \begin{pmatrix} \alpha & 0 \\ 0 & \beta \end{pmatrix} \ :\ \alpha, \beta \in [-1,\ 1] \right\},$$

which is also an approximate Jacobian of F at 0 and contains $\partial^* F(0)$.

Observe in this example that $\partial_C F(0)$ is the convex hull of $\partial^* F(0)$. However, this is not always the case. The following example illustrates that even for the case where $m = 1$, the convex hull of an approximate Jacobian of a locally Lipschitz map may be strictly contained in the Clarke generalized Jacobian.

Example 2.5. Define $F : \mathbb{R}^2 \to \mathbb{R}$ by

$$F(x,y) = |x| - |y|.$$

Then it can easily be verified that

$$\partial_1^* F(0) = \{(1,1), (-1,-1)\} \ \text{ and } \ \partial_2^* F(0) = \{(1,-1), (-1,1)\}$$

are approximate Jacobians of F at 0, whereas

$$\partial_B F(0) = \{(1,1), (-1,1), (1,-1), (-1,-1)\}$$

and

$$\partial_C F(0) = co(\{(1,1), (-1,1), (1,-1), (-1,-1)\}).$$

It is also worth noting that

$$co(\partial_1^* F(0)) \subset co(\partial_M F(0)) = \partial_C F(0).$$

Clearly, this example shows that certain results, such as mean value conditions and necessary optimality conditions that are expressed in terms of $\partial^* F(x)$, may provide sharp conditions even for locally Lipschitz maps (see section 3).

Let us now present an example of a continuous map where the Clarke generalized Jacobian does not exist, whereas approximate Jacobians are quite easy to calculate.

Example 2.6. Define $F : \mathbb{R}^2 \to \mathbb{R}^2$ by

$$F(x,y) = (\sqrt{|x|}\,\operatorname{sgn}(x) + |y|, \quad \sqrt{|y|}\,\operatorname{sgn}(y) + |y|),$$

where $\operatorname{sgn}(x) = 1$ for $x > 0$, 0 for $x = 0$, and -1 for $x < 0$. Then F is not locally Lipschitz at $(0,0)$, and so the Clarke generalized Jacobian does not exist. However, for each $c \in \mathbb{R}$, the set

$$\partial^* F(0,0) = \left\{ \begin{pmatrix} \alpha & 1 \\ 0 & \beta \end{pmatrix}, \begin{pmatrix} \alpha & -1 \\ 0 & \beta \end{pmatrix} \,:\, \alpha, \beta \geq c \right\}$$

is an approximate Jacobian of F at $(0,0)$.

3. Generalized mean value theorems. In this section we derive mean value theorems for continuous maps in terms of approximate Jacobians and show how locally Lipschitz vector-valued maps can be characterized using approximate Jacobians.

THEOREM 3.1. *Let $a, b \in R^n$ and $F : \mathbb{R}^n \to \mathbb{R}^m$ be continuous. Assume that for each $x \in [a,b]$, $\partial^* F(x)$ is an approximate Jacobian of F at x. Then*

$$F(b) - F(a) \in \overline{co}(\partial^* F([a,b])(b-a)).$$

Proof. Let us first note that the right-hand side above is the closed convex hull of all points of the form $M(b-a)$, where $M \in \partial^* F(\zeta)$ for some $\zeta \in [a,b]$. Let $v \in \mathbb{R}^m$ be arbitrary and fixed. Consider the real-valued function $g : [0,1] \to \mathbb{R}$

$$g(t) = \langle v, F(a + t(b-a)) - F(a) + t(F(a) - F(b))\rangle.$$

Then g is continuous on $[0,1]$ with $g(0) = g(1)$. So g attains a minimum or a maximum at some $t_0 \in (0,1)$. Suppose that t_0 is a minimum point. Then, for each $\alpha \in \mathbb{R}$, $g^-(t_0, \alpha) \geq 0$. It now follows from direct calculations that

$$g^-(t_0, \alpha) = (vF)^-(a + t_0(b-a), \alpha(b-a)) + \alpha\langle v, F(a) - F(b)\rangle.$$

Hence, for each $\alpha \in \mathbb{R}$,

$$(vF)^-(a + t_0(b-a), \alpha(b-a)) \geq \alpha\langle v, F(b) - F(a)\rangle.$$

Now, by taking $\alpha = 1$ and $\alpha = -1$, we obtain that

$$-(vF)^-(a + t_0(b-a), a-b) \le \langle v, F(b) - F(a)\rangle \le (vF)^-(a + t_0(b-a), b-a)\rangle.$$

By (2.1), we get

$$\inf_{M\in\partial^* F(a+t_0(b-a))} \langle Mv, b-a\rangle \le \langle v, F(b) - F(a)\rangle \le \sup_{M\in\partial^* F(a+t_0(b-a))} \langle Mv, b-a\rangle.$$

Consequently,

$$\langle v, F(b) - F(a)\rangle \in \overline{co}(\partial^* F(a + t_0(b-a))v)(b-a),$$

and so

$$\langle v, F(b) - F(a)\rangle \in \overline{co}(\partial^* F([a,b])v)(b-a). \tag{3.1}$$

Since this inclusion holds for each $v \in \mathbb{R}^m$, we claim that

$$F(b) - F(a) \in \overline{co}(\partial^* F([a,b])(b-a)).$$

If this is not so, then it follows from the separation theorem

$$\langle p, F(b) - F(a)\rangle - \epsilon > \sup_{u\in\overline{co}(\partial^* F([a,b])(b-a))} \langle p, u\rangle$$

for some $p \in \mathbb{R}^m$ since $\overline{co}(\partial^* F([a,b])(b-a))$ is a closed convex subset of $\mathbb{R}^m$. This implies

$$\langle p, F(b) - F(a)\rangle > \sup\{\alpha : \alpha \in \overline{co}(\partial^* F([a,b])p)(b-a)\},$$

which contradicts (3.1).

Similarly, if t_0 is a maximum point, then $g^+(t_0, \alpha) \le 0$ for each $\alpha \in \mathbb{R}$. Using the same line of arguments as above, we arrive at the same conclusion, and so the proof is complete. □

COROLLARY 3.2. *Let $a, b \in \mathbb{R}^n$ and $F : \mathbb{R}^n \to \mathbb{R}^m$ be continuous. Assume that $\partial^* F(x)$ is a bounded approximate Jacobian of F at x for each $x \in [a,b]$. Then*

$$F(b) - F(a) \in co(\partial^* F([a,b])(b-a)). \tag{3.2}$$

Proof. Since for each $x \in [a,b]$, $\partial^* F(x)$ is compact, the set

$$co(\partial^* F([a,b])(b-a) = co\{\partial^* F([a,b])(b-a)\}$$

is closed, and so the conclusion follows from Theorem 3.1. □

In the following corollary we deduce the mean value theorem for locally Lipschitz maps (see [1, 6]) as a special case of Theorem 3.1.

COROLLARY 3.3. *Let $a, b \in R^n$ and $F : \mathbb{R}^n \to \mathbb{R}^m$ be locally Lipschitz on $\mathbb{R}^n$. Then*

$$F(b) - F(a) \in co(\partial_C F([a,b])(b-a)). \tag{3.3}$$

Proof. In this case the Clarke generalized Jacobian $\partial_C F(x)$ is a convex and compact approximate Jacobian of F at x. Hence, the conclusion follows from Corollary 3.2. □

Note that even for the case where F is locally Lipschitz, Corollary 3.2 provides a stronger mean value condition than condition (3.3) of Corollary 3.3. To see this, let $n = 2,\ m = 1,\ F(x, y) = |x| - |y|$, $a = (-1, -1)$, and $b = (1, 1)$. Then condition (3.2) of Corollary 3.2 is verified by

$$\partial^* F(0) = \{(1, -1), (-1, 1)\}.$$

However, condition (3.3) holds for $\partial_C F(0)$, where

$$\partial_C F(0) = co(\{(1, 1), (-1, -1), (1, -1), (-1, 1)\}) \supset \partial^* F(0).$$

As a special case of the above theorem, we see that if F is real-valued, then an asymptotic mean value equality is obtained. This was shown in [13].

COROLLARY 3.4. *Let $a, b \in X$ and $F : \mathbb{R}^n \to \mathbb{R}$ be continuous. Assume that, for each $x \in [a, b]$, $\partial^* F(x)$ is a convexificator of F. Then there exist $c \in (a, b)$ and a sequence $\{x_k^*\} \subset co(\partial^* F(c))$ such that*

$$F(b) - F(a) = \lim_{k\to\infty} \langle x_k^*,\ b - a\rangle.$$

Proof. The conclusion follows from the proof of Theorem 3.1 by noting that a convexificator $\partial^* F(x)$ is an approximate Jacobian of F at x. □

We now see how locally Lipschitz functions can be characterized using the above mean value theorem. We say that a set-valued mapping $G : \mathbb{R}^n \to L(\mathbb{R}^n, \mathbb{R}^m)$ is *locally bounded* at x if there exist a neighborhood U of x and a positive α such that $||A|| \le \alpha$ for each $A \in G(U)$. Recall that the map G is said to be *upper semicontinuous* at x if for each open set V containing $G(x)$ there is a neighborhood U of x such that $G(U) \subset V$. Clearly, if G is upper semicontinuous at x and if $G(x)$ is *bounded*, then G is locally bounded at x.

THEOREM 3.5. *Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be continuous. Then F has a locally bounded approximate Jacobian map $\partial^* F$ at x if and only if F is locally Lipschitz at x.*

Proof. Assume that $\partial^* F(y)$ is the approximate Jacobian of F for each y in a neighborhood U of x and that $\partial^* F$ is locally bounded on U. Without loss of generality, we may assume that U is convex. Then there exists $\alpha > 0$ such that $||A|| \le \alpha$ for each $A \in \partial^* F(U)$. Let $x, y \in U$. Then $[x, y] \subset U$, and by the mean value theorem,

$$F(x) - F(y) \in \overline{co}(\partial^* F([x, y])(x - y)) \subset \overline{co}(\partial^* F(U)(x - y)).$$

Hence,

$$\|F(x) - F(y)\| \le \|x - y\| \max\{\|A\| : A \in \partial^* F(U)\}.$$

This gives us that

$$\|F(x) - F(y)\| \le \alpha\|x - y\|,$$

and so F is locally Lipschitz at x.

Conversely, if F is locally Lipschitz at x, then the Clarke generalized Jacobian can be chosen as an approximate Jacobian for F, which is locally bounded at x. □

4. Calculus rules for approximate Jacobians. In this section, we present some basic calculus rules for approximate Jacobians. We begin by introducing the notion of regular approximate Jacobians which are useful in some applications.

DEFINITION 4.1. *The map $F : \mathbb{R}^n \to \mathbb{R}^m$ admits a regular approximate Jacobian, $\partial^* F(x)$ at $x \in R^n$ if $\partial^* F(x) \subseteq L(\mathbb{R}^n, \mathbb{R}^m)$ is closed, and for each $v \in R^m$,*

$$(vF)^+(x,u) = \sup_{M \in \partial^* F(x)} \langle Mv, u\rangle \ \ \forall u \in \mathbb{R}^n, \tag{4.1}$$

or equivalently,

$$(vF)^-(x,u) = \inf_{M \in \partial^* F(x)} \langle Mv, u\rangle \ \ \forall u \in \mathbb{R}^n. \tag{4.2}$$

Note that in the case $m = 1$, this definition collapses to the notion of the regular convexificator studied in [13]. Thus, a closed set $\partial^* h(x) \subset \mathbb{R}^n$ is a regular convexificator of the real-valued function h at x if for each $u \in \mathbb{R}^n$,

$$h^-(x,u) = \inf_{\xi \in \partial^* h(x)} \langle \xi, u\rangle \ \text{ and } \ h^+(x,u) = \sup_{\xi \in \partial^* h(x)} \langle \xi, u\rangle.$$

It is evident that these equalities follow from (4.1) by taking $F = h$ and $v = -1$ and $v = 1$, respectively.

It is immediate from the definition that if F is differentiable at x, then $\{\nabla f(x)\}$ is a regular approximate Jacobian of F at x. However, if F is locally Lipschitz at x, then the Clarke generalized Jacobian $\partial_C F(x)$ is not necessarily a regular approximate Jacobian of F at x. It is also worth noting that if $\partial_1^* F(x)$ and $\partial_2^* F(x)$ are two regular approximate Jacobians of F at x, then $\overline{co}(\partial_1^* F(x)) = \overline{co}(\partial_2^* F(x))$.

In passing, we note that if F is locally Lipschitz on a neighborhood U of x, then there exists a dense set $K \subset U$ such that F admits a regular approximate Jacobian at each point of K. By Rademacher's theorem, the dense subset can be chosen as the set where F is differentiable.

THEOREM 4.2 (Rule 1). *Let F and H be continuous maps from R^n to R^m. Assume that $\partial^* F(x)$ is an approximate Jacobian of F at x and $\partial^* H(x)$ is a regular approximate Jacobian of H at x. Then the set $\partial^* F(x) + \partial^* H(x)$ is an approximate Jacobian of $F + H$ at x.*

Proof. Let $v \in R^m,\ u \in R^n$ be arbitrary. By definition,

$$\langle v, F + H\rangle^-(x,u) = \liminf_{t \downarrow 0} \frac{\langle v, F(x+tu) - F(x) + H(x+tu) - H(x)\rangle}{t}.$$

Let $\{t_n\}$ be a sequence of positive numbers converging to 0 such that

$$\langle v, F + H\rangle^-(x,u) = \lim_{n \to \infty} \frac{\langle v, F(x+t_n u) - F(x) + H(x+t_n u) - H(x)\rangle}{t_n}.$$

Further, let $\{s_n\}$ be another sequence of positive numbers converging to 0 such that

$$\langle v, F\rangle^-(x,u) = \liminf_{t \downarrow 0} \frac{\langle v, F(x+tu) - F(x)\rangle}{t} = \lim_{n \to \infty} \frac{\langle v, F(x+s_n u) - F(x)\rangle}{s_n}.$$

Then we have

$$\lim_{n \to \infty} \frac{\langle v, F(x+s_n u) - F(x)\rangle}{s_n} \le \sup_{M \in \partial^* F(x)} \langle Mv,\ u\rangle$$

and

$$\limsup_{n\to\infty} \frac{\langle v, H(x+s_nu)-H(x)\rangle}{s_n} \le \langle v, H\rangle^+(x,u) = \sup_{M\in\partial^* H(x)} \langle Mv, u\rangle.$$

Consequently,

$$\begin{aligned}\langle v, F+H\rangle^-(x,u) &\le \lim_{n\to\infty} \frac{\langle v, F(x+s_nu)-F(x)\rangle}{s_n} + \frac{\langle v, H(x+s_nu)-H(x)\rangle}{s_n}\\ &\le \sup_{M\in\partial^* F(x)} \langle Mv, u\rangle + \sup_{N\in\partial^* H(x)} \langle Nv, u\rangle\\ &= \sup_{P\in\partial^* F(x)+\partial^* H(x)} \langle Pv, u\rangle.\end{aligned}$$

Since u and v are arbitrary, we conclude that $\partial^* F(x) + \partial^* H(x)$ is an approximate Jacobian of $F+H$ at x. □

Note that as in the case of convexificators of real-valued functions [18], the set $\partial^* F(x) + \partial^* H(x)$ is not necessarily regular at x.

THEOREM 4.3 (Rule 2). *Let $F : \mathbb{R}^n \to \mathbb{R}^m$ and $H : \mathbb{R}^m \to \mathbb{R}^l$ be continuous maps. Assume that $\partial^* F(x)$ is a bounded approximate Jacobian of F at x and $\partial^* H(x)$ is a bounded approximate Jacobian of H at $F(x)$. If the maps $\partial^* F$ and $\partial^* H$ are upper semicontinuous at x and $F(x)$, respectively, then $\partial^* H(F(x))\partial^* F(x)$ is an approximate Jacobian of $H \circ F$ at x.*

Proof. Let $w \in \mathbb{R}^l$ and $u \in \mathbb{R}^m$ be arbitrary. Consider the lower Dini directional derivative of $\langle w,\ H \circ F\rangle$ at x:

$$\langle w,\ H\circ F\rangle^-(x,u) = \liminf_{t\downarrow 0} \frac{\langle w,\ H(F(x+tu)) - H(F(x))\rangle}{t}.$$

By applying the mean value theorem (see Theorem 3.1) to H and F, we obtain

$$\begin{aligned} F(x+tu) - F(x) &\in t\overline{co}(\partial^* F([x, x+tu])u),\\ H(F(x+tu)) - H(F(x)) &\in \overline{co}(\partial^* H([F(x,),\ F(x+tu)])(F(x+tu)-F(x)))\end{aligned}$$

It now follows from the upper semicontinuity of $\partial^* F$ and $\partial^* H$ that for an arbitrary small positive ϵ we can find $t_0 > 0$ such that for $t \in (0, t_0)$ we have

$$\begin{aligned}\partial^* F([x, x+tu]) &\subseteq \partial^* F(x) + \epsilon B_1,\\ \partial^* H([F(x),\ F(x+tu)]) &\subseteq \partial^* H(F(x)) + \epsilon B_2,\end{aligned}$$

where B_1 and B_2 are the unit balls in $L(\mathbb{R}^n, \mathbb{R}^m)$ and $L(\mathbb{R}^m, \mathbb{R}^l)$, respectively. Using these inclusions, we obtain

$$\frac{\langle w,\ H(F(x+tu)) - H(F(x))\rangle}{t} \in \langle w, A\rangle,$$

where

$$A := \overline{co}((\partial^* H(F(x))\partial^* F(x) + \epsilon(\partial^* H(F(x))B_1 + B_2\partial^* F(x)) + \epsilon^2 B_2 B_1)u).$$

Since $\partial^* H(F(x))$ and $\partial^* F(x)$ are bounded, we can find $\alpha > 0$ such that $\|M\| \le \alpha$ for all $M \in \partial^* H(F(x))$ or $M \in \partial^* F(x)$. Consequently,

$$\langle w, H\circ F\rangle^-(x,u) \le \sup_{M\in\partial^* H(F(x))\partial^* F(x)} \langle Mw, u\rangle + 2\epsilon\|u\| + \epsilon^2\|u|.$$

As ϵ is arbitrary, we conclude that $\partial^* H(F(x))\partial^* F(x)$ is an approximate Jacobian of $H \circ F$ at x. □

5. Approximate Hessian matrices. In this section, unless stated otherwise, we assume that $f : \mathbb{R}^n \to \mathbb{R}$ is a C^1- function, that is, a continuously Gâteaux differentiable function, and introduce the notion of approximate Hessian for such functions. Note that the derivative of f, which is denoted by ∇f, is a map from $\mathbb{R}^n$ to $\mathbb{R}^n$.

DEFINITION 5.1. *The function f admits an* approximate Hessian $\partial^2_* f(x)$ *at x if this set is an approximate Jacobian to ∇f at x.*

Note that $\partial^2_* f(x) = \partial^* \nabla f(x)$ and the matrix $M \in \partial^2_* f(x)$ is an approximate Hessian matrix of F at x. Clearly, if f is twice differentiable at x, then $\nabla^2 f(x)$ is a symmetric approximate Hessian matrix of f at x.

Let us now examine the relationships between the approximate Hessians and the generalized Hessians, studied for $C^{1,1}$-functions, that is, Gâteaux differentiable functions with locally Lipschitz derivatives. Recall that if $f : \mathbb{R}^n \to \mathbb{R}$ is $C^{1,1}$, then the generalized Hessian in the sense of Hiriart-Urruty, Strodiot, and Hien Nguyen [7] is given by

$$\partial^2_H f(x) = co\{M \ : \ M = \lim_{n\to\infty} \nabla^2 f(x_n),\ x_n \in \Delta,\ x_n \to x\},$$

where Δ is the set of points in $\mathbb{R}^n$ where f is twice differentiable. Clearly, $\partial^2_H f(x)$ is a nonempty convex compact set of symmetric matrices. The second-order directional derivative of f at x in the directions $(u, v) \in \mathbb{R}^n \times \mathbb{R}^n$ is defined by

$$f^{\circ\circ}(x; u, v) = \limsup_{\substack{y\to x \\ s\to 0}} \frac{\langle \nabla f(y + su), v\rangle - \langle \nabla f(y), v\rangle}{s}.$$

Since $(v\nabla f)^-(x, u) \le f^{\circ\circ}(x; u, v)$, for each $(u, v) \in \mathbb{R}^n$ and

$$f^{\circ\circ}(x; u, v) = \max_{M\in\partial^2_H f(x)} \langle Mu, v\rangle = \max_{M\in\partial^2_H f(x)} \langle Mv, u\rangle,$$

$\partial^2_H f(x)$ is an approximate Hessian of f at x.

The generalized Hessian of f at x as a set-valued map, $\partial^{\circ\circ} f(x) : \mathbb{R}^n \to \mathbb{R}^n$, which was given in Cominetti and Correa [3], is defined by

$$\partial^{\circ\circ} f(x)(u) = \{x^* \in \mathbb{R}^n : f^{\circ\circ}(x; u, v) \ge \langle x^*, v\rangle \forall v \in \mathbb{R}^n\}.$$

It is known that the mapping $(u, v) \longrightarrow f^{\circ\circ}(x; u, v)$ is finite and sublinear and that $\partial^{\circ\circ} f(x)(u)$ is a nonempty, convex, and compact subset of $\mathbb{R}^n$, and for each $x, u, v \in \mathbb{R}^n$,

$$f^{\circ\circ}(x; u, v) = \max\{\langle x^*, v\rangle : x^* \in \partial^{\circ\circ} f(x)(u)\}.$$

Moreover, for each $u \in \mathbb{R}^n$,

$$\partial^{\circ\circ} f(x)(u) = \partial^2_H f(x)u.$$

If f is twice continuously differentiable at x, then the generalized Hessian $\partial^{\circ\circ} f(x)(u)$ is a singleton for every $u \in \mathbb{R}^n$.

In [34, 35], another generalized second-order directional derivative and a generalized Hessian set-valued map for a $C^{1,1}$ function f at x were given as follows:

$$f^{\diamond\diamond}(x; u, v) = \sup_{z\in\mathbb{R}^n} \limsup_{s\downarrow 0} \frac{\langle \nabla f(x + sz + su), v\rangle - \langle \nabla f(x + sz), v\rangle}{s},$$

$$\partial^{\diamond\diamond} f(x)(u) = \{x^* \in X^* : f^{\diamond\diamond}(x; u, v) \geq \langle x^*, v\rangle \, \forall v \in X\}.$$

It was shown that the mapping $(u, v) \longrightarrow f^{\diamond\diamond}(x; u, v)$ is finite and sublinear; $\partial^{\diamond\diamond} f(x)(u)$ is a nonempty, convex, and compact subset of $\mathbb{R}^n$; and $\partial^{\diamond\diamond} f(x)(u)$ is singled-valued for each $u \in \mathbb{R}^n$ if and only if f is twice Gâteaux differentiable at x. Further, for each $u \in \mathbb{R}^n$, $\partial^{\diamond\diamond} f(x)(u) \subset \partial^{\circ\circ} f(x)(u) = \partial^2_H f(x)u$. If for each $(u, v) \in \mathbb{R}^n$ the function $y \longrightarrow f^{\diamond\diamond}(y; u, v)$ is upper semicontinuous at x, then

$$\partial^{\diamond\diamond} f(x)(u) = \partial^2_H f(x)u.$$

The following proposition gives us necessary and sufficient conditions in terms of approximate Hessians for a C^1-function to be $C^{1,1}$.

PROPOSITION 5.2. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a C^1-function. Then f has a locally bounded approximate Hessian map $\partial^2_* f$ at x if and only if f is $C^{1,1}$ at x.*

Proof. This follows from Theorem 3.5 by taking F as ∇f. □

We complete this section with an example showing that for a $C^{1,1}$ function the approximate Hessian may be a singleton which is contained in the generalized Hessian of Hiriart-Urruty, Strodiot, and Hien Nguyen [7].

Example 5.3. Let g be an odd, linear piecewise continuous function on $\mathbb{R}$ as follows. $g(x) = x$ for $x \geq 1$ and $g(0) = 0$; $g(x) = 2x - 1$ for $x \in [\frac{1}{2}, 1]$; $g(x) = -\frac{1}{2}x + \frac{1}{4}$ for $x \in [\frac{1}{6}, \frac{1}{2}]$; $g(x) = 2x - \frac{1}{6}$ for $x \in [\frac{1}{12}, \frac{1}{6}]$; $g(x) = -\frac{1}{4}x + \frac{1}{48}$ for $x \in [\frac{1}{60}, \frac{1}{12}]$, etc. Let

$$G(x) = \int_0^{|x|} g(t)dt, \quad x \in \mathbb{R}.$$

Define

$$f(x, y) = G(x) + \frac{y^2}{2}.$$

Then the function f is a $C^{1,1}$ function, and the generalized Hessian of f at $(0, 0)$ is

$$\partial^2_H f(0) = \left\{ \begin{pmatrix} \alpha & 0 \\ 0 & 1 \end{pmatrix} \, : \, \alpha \in [0,\ 2] \right\}.$$

However, the approximate Hessian of f at $(0, 0)$ is the singleton

$$\partial^2_* f(0) = \left\{ \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \right\}.$$

6. Generalized Taylor's expansions for C^1-functions. In this section, we see how Taylor's expansions can be obtained for C^1- functions using approximate Hessians.

THEOREM 6.1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously Gâteaux differentiable on $\mathbb{R}^n$; let $x, y \in \mathbb{R}^n$. Suppose that for each $z \in [x, y]$, $\partial^2_* f(z)$ is an approximate Hessian of f at z. Then there exists $\zeta \in (x, y)$ such that*

$$f(y) \in f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2}\overline{co}\langle \partial^2_* f(\zeta)(y - x), (y - x)\rangle.$$

Proof. Let $h(t) = f(y + t(x - y)) + t\langle \nabla f(y + t(x - y)), y - x\rangle + \frac{1}{2}at^2 - f(y)$, where $a = -2(f(x) - f(y) + \langle \nabla f(x), y - x\rangle)$. Then $h(0) = 0, h(1) = f(x) - f(y) +$

$\langle \nabla f(x), y-x\rangle + \frac{1}{2}a = 0$, and h is continuous. So h attains its extremum at some $\gamma \in (0,1)$. Suppose that γ is a minimum point of h. Now, by necessary conditions, we have for all $v \in \mathbb{R}$

$$h^-(\gamma; v) \geq 0.$$

Then

$$\begin{aligned}
0 &\leq h^-(\gamma; v)\\
&= \liminf_{\lambda\to 0^+} \frac{h(\gamma+\lambda v) - h(\gamma)}{\lambda}\\
&= \lim_{\lambda\to 0^+} \frac{f(y+(\gamma+\lambda v)(x-y)) - f(y+\gamma(x-y))}{\lambda}\\
&\quad + \frac{1}{2}\lim_{\lambda\to 0^+} \frac{a(\gamma+\lambda v)^2 - a\gamma^2}{\lambda}\\
&\quad + \liminf_{\lambda\to 0^+} \frac{(\gamma+\lambda v)\langle \nabla f(y+(\gamma+\lambda v)(x-y)), y-x\rangle - \gamma\langle \nabla f(y+\gamma(x-y)), y-x\rangle}{\lambda}\\
&= v\langle \nabla f(y+\gamma(x-y)), x-y\rangle + a\gamma v + v\langle \nabla f(y+\gamma(x-y)), y-x\rangle\\
&\quad + \gamma \liminf_{\lambda\to 0^+} \frac{\langle \nabla f(y+(\gamma+\lambda v)(x-y)), y-x\rangle - \langle \nabla f(y+\gamma(x-y)), y-x\rangle}{\lambda}\\
&= a\gamma v + \gamma \liminf_{\lambda\to 0^+} \frac{\langle \nabla f(y+(\gamma+\lambda v)(x-y)), y-x\rangle - \langle \nabla f(y+\gamma(x-y)), y-x\rangle}{\lambda}.
\end{aligned}$$

Let $\zeta = y+\gamma(x-y)$. Then $\zeta \in (x,y)$, and for $v=1$ we get

$$\begin{aligned}
0 &\leq a\gamma + \gamma \liminf_{\lambda\to 0^+} \frac{\langle \nabla f(y+\gamma(x-y)+\lambda(x-y)), y-x\rangle - \langle \nabla f(y+\gamma(x-y)), y-x\rangle}{\lambda}\\
&\leq a + \sup_{M\in\partial_*^2 f(\zeta)} \langle M(y-x), x-y\rangle.
\end{aligned}$$

This gives us that

$$a \geq \inf_{M\in\partial_*^2 f(\zeta)} \langle M(y-x), y-x\rangle.$$

Similarly, for $v=-1$, we obtain

$$\begin{aligned}
0 &\leq -a\gamma + \gamma \liminf_{\lambda\to 0^+} \frac{\langle \nabla f(y+\gamma(x-y)+\lambda(y-x)), y-x\rangle - \langle \nabla f(y+\gamma(x-y)), y-x\rangle}{\lambda}\\
&\leq -a + \sup_{M\in\partial_*^2 f(\zeta)} \langle M(y-x), y-x\rangle;
\end{aligned}$$

thus,

$$a \leq \sup_{M\in\partial_*^2 f(\zeta)} \langle M(y-x), y-x\rangle.$$

Hence, it follows that

$$\inf_{M\in\partial_*^2 f(\zeta)} \langle M(y-x), y-x\rangle \leq a \leq \sup_{M\in\partial_*^2 f(\zeta)} \langle M(y-x), y-x\rangle,$$

and so

$$a \in \overline{co}\langle \partial_*^2 f(\zeta)(y-x), (y-x)\rangle;$$

thus,

$$f(y) - f(x) - \langle \nabla f(x), y - x\rangle = \frac{a}{2} \in \frac{1}{2}\overline{co}\langle \partial_*^2 f(\zeta)(y-x), (y-x)\rangle. \tag{6.1}$$

The case where γ is a maximum point of h also yields the same condition (6.1). The details are left to the reader. □

COROLLARY 6.2. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously Gâteaux differentiable on $\mathbb{R}^n$ and $x, y \in \mathbb{R}^n$. Suppose that for each $z \in [x, y]$, $\partial_*^2 f(z)$ is a convex and compact approximate Hessian of f at z. Then there exist $\zeta \in (x, y)$ and $M_\zeta \in \partial_*^2 f(\zeta)$ such that*

$$f(y) = f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2}\langle M_\zeta(y-x), y-x\rangle.$$

Proof. It follows from the hypothesis that for each $z \in [x, y]$, $\partial_*^2 f(z)$ is convex and compact, and so the $\overline{co}$ in the conclusion of the previous theorem is superfluous. Thus, the inequalities

$$\inf_{M \in \partial_*^2 f(\zeta)} \langle M(y-x), y-x\rangle \le a \le \sup_{M \in \partial_*^2 f(\zeta)} \langle M(y-x), y-x\rangle$$

give us that

$$a \in \langle \partial_*^2 f(\zeta)(y-x), (y-x)\rangle. \qquad \square$$

COROLLARY 6.3 (see [7]). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be $C^{1,1}$ and $x, y \in \mathbb{R}^n$. Then there exist $\zeta \in (x, y)$ and $M_\zeta \in \partial_H^2 f(\zeta)$ such that*

$$f(y) = f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2}\langle M_\zeta(y-x), y-x\rangle.$$

Proof. In this case, the conclusion follows from the above corollary by choosing the generalized Hessian $\partial_H^2 f(x)$ as an approximate Hessian of f for each x. □

7. Second-order conditions for optimality and convexity of C^1-functions. In this section, we present second-order necessary and sufficient conditions for optimality and convexity of C^1-functions using approximate Hessian matrices. Consider the optimization problem

$$\begin{aligned} (\mathrm{P}) \quad & \text{minimize } f(x) \\ & \text{subject to } x \in \mathbb{R}^n, \end{aligned}$$

where $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is a continuously Gâteaux differentiable function on $\mathbb{R}^n$. We say that a map $F : \mathbb{R}^n \to \mathbb{R}^m$ admits a *semiregular approximate Jacobian* $\partial^* F(x)$ at $x \in R^n$ if $\partial^* F(x) \subseteq L(\mathbb{R}^n, \mathbb{R}^m)$ is closed, and for each $v \in R^m$,

$$(vF)^+(x, u) \le \sup_{M \in \partial^* F(x)} \langle Mv, u\rangle \ \ \forall u \in \mathbb{R}^n.$$

Similarly, the C^1-function $f : \mathbb{R}^n \to \mathbb{R}$ admits a *semiregular approximate Hessian* $\partial_*^2 f(x)$ at x if this set is a semiregular approximate Jacobian to ∇f at x.

Of course, every semiregular approximate Hessian to f at x is an approximate Hessian at x. For a $C^{1,1}$ function $f : \mathbb{R}^n \to \mathbb{R}$, the generalized Hessian, $\partial^2_H f(x)$, of f at x is a bounded semiregular approximate Hessian of f at x since

$$(v\nabla f)^+(x,u) \le f^{\circ\circ}(x;u,v) = \max_{M\in\partial^2_H f(x)} \langle Mu, v\rangle = \max_{M\in\partial^2_H f(x)} \langle Mv, u\rangle.$$

THEOREM 7.1. *For the problem* (P)*, let* $\bar{x} \in \mathbb{R}^n$*. Assume that* $\partial^2_* f(\bar{x})$ *is a semiregular approximate Hessian of* f *at* $\bar{x}$*.*

(i) *If* $\bar{x}$ *is a local minimum of* (P)*, then* $\nabla f(\bar{x}) = 0$*, and for each* $u \in \mathbb{R}^n$*,*

$$\sup_{M\in\partial^2_* f(\bar{x})} \langle Mu, u\rangle \ge 0.$$

(ii) *If* $\bar{x}$ *is a local maximum of* (P)*, then* $\nabla f(\bar{x}) = 0$*, and for each* $u \in \mathbb{R}^n$*,*

$$\inf_{M\in\partial^2_* f(\bar{x})} \langle Mu, u\rangle \le 0.$$

Proof. Let $u \in \mathbb{R}^n$. Since $\bar{x}$ is a local minimum of (P), there exists $\delta > 0$ such that for each $s \in [0, \delta]$,

$$f(\bar{x} + su) \ge f(\bar{x}).$$

Then, by the mean value theorem, for each $s \in (0, \delta]$, there exists $0 < t < s$ such that

$$\langle \nabla f(\bar{x} + tu), u\rangle \ge 0.$$

So, there exists a positive sequence $\{t_n\} \downarrow 0$ such that $\langle \nabla f(\bar{x} + t_n u), u\rangle \ge 0$. Now, as $\nabla f(\bar{x}) = 0$, it follows that

$$\begin{aligned}(u\nabla f)^+(\bar{x};u) &= \limsup_{s\downarrow 0} \frac{\langle \nabla f(\bar{x}+su), u\rangle - \langle \nabla f(\bar{x}), u\rangle}{s}\\ &\ge 0.\end{aligned}$$

Since $\partial^2_* f(x)$ is a semiregular approximate Hessian of f at x, we have

$$(u\nabla f)^+(\bar{x};u) \le \sup_{M\in\partial^2_* f(\bar{x})} \langle Mu, u\rangle,$$

and hence,

$$\sup_{M\in\partial^2_* f(\bar{x})} \langle Mu, u\rangle \ge 0.$$

On the other hand, if f attains a local maximum at $\bar{x}$, then it follows by the similar arguments as above that for each $u \in \mathbb{R}^n$,

$$\inf_{M\in\partial^2_* f(\bar{x})} \langle Mu, u\rangle \le 0.$$

Note in this case that it is convenient to use the inequality

$$(u\nabla f)^-(\bar{x},u) \ge \inf_{M\in\partial^2_* f(\bar{x})} \langle Mu, u\rangle. \qquad \square$$

Let us look at a numerical example to illustrate the significance of the optimality conditions obtained in the previous theorem.

Example 7.2. Define $f : \mathbb{R}^2 \to \mathbb{R}$ by

$$f(x,y) = \frac{2}{3}|x|^{\frac{3}{2}} + \frac{1}{2}y^2.$$

Then f is C^1 but is not $C^{1,1}$ since the gradient

$$\nabla f(x,y) = \left(\sqrt{|x|}\,\operatorname{sgn}(x), y\right)$$

is not locally Lipschitz at $(0,0)$. Evidently, $(0,0)$ is a minimum point of f, $\nabla f(0,0) = (0,0)$, and

$$\partial_*^2 f(0) = \left\{ \begin{pmatrix} \alpha & 0 \\ 0 & 1 \end{pmatrix} \;:\; \alpha \geq 0 \right\}$$

is a semiregular approximate Hessian of f at $(0,0)$. And for each $u = (u_1, u_2) \in \mathbb{R}^2$,

$$\sup_{M \in \partial_*^2 f(0)} \langle Mu, u\rangle = \sup\{\alpha u_1^2 + u_2^2 \;:\; \alpha \geq 0\} \geq 0.$$

Hence, the statement (i) of Theorem 7.1 is verified. However, the generalized Hessians [7] do not apply to this function.

COROLLARY 7.3. *For the problem* (P)*, let* $\bar{x} \in \mathbb{R}^n$*. Suppose that* $\partial_*^2 f(\bar{x})$ *is a bounded semiregular approximate Hessian of* f *at* $\bar{x}$*.*

(i) *If* $\bar{x}$ *is a local minimum of* (P)*, then* $\nabla f(\bar{x}) = 0$*, and for each* $u \in \mathbb{R}^n$ *there exists a matrix* $M \in \partial_*^2 f(\bar{x})$ *such that* $\langle Mu, u\rangle \geq 0$*.*

(ii) *If* $\bar{x}$ *is a local maximum of* (P)*, then* $\nabla f(\bar{x}) = 0$*, and for each* $u \in \mathbb{R}^n$ *there exists a matrix* $M \in \partial_*^2 f(\bar{x})$ *such that* $\langle Mu, u\rangle \leq 0$*.*

Proof. Since $\partial_*^2 f(\bar{x})$ is closed and bounded, it follows from Theorem 7.1 that $\nabla f(\bar{x}) = 0$, and for each $u \in \mathbb{R}^n$,

$$\max_{M \in \partial_*^2 f(\bar{x})} \langle Mu, u\rangle \geq 0,$$

and so the first conclusion holds. The second conclusion similarly follows from Theorem 7.1. □

We now see how optimality conditions for the problem (P) where f is $C^{1,1}$ follows from Corollary 7.3 (cf. [7]).

COROLLARY 7.4. *For the problem* (P)*, assume that the function* f *is* $C^{1,1}$ *and* $\bar{x} \in \mathbb{R}^n$*.*

(i) *If* $\bar{x}$ *is a local minimum of* (P)*, then* $\nabla f(\bar{x}) = 0$*, and for each* $u \in \mathbb{R}^n$ *there exists a matrix* $M \in \partial_H^2 f(\bar{x})$ *such that* $\langle Mu, u\rangle \geq 0$*.*

(ii) *If* $\bar{x}$ *is a local maximum of* (P)*, then* $\nabla f(\bar{x}) = 0$*, and for each* $u \in \mathbb{R}^n$ *there exists a matrix* $M \in \partial_H^2 f(\bar{x})$ *such that* $\langle Mu, u\rangle \leq 0$*.*

Proof. The conclusion follows from Corollary 7.3 by choosing $\partial_H^2 f(\bar{x})$ as the semiregular bounded approximate Hessian $\partial_*^2 f(\bar{x})$ of f at $\bar{x}$. □

Clearly, the conditions of Theorem 7.1 are not sufficient for a local minimum, even for a C^2-function f. The generalized Taylor's expansion is now applied to obtain a version of second-order sufficient condition for a local minimum. For related results, see [34, 16].

THEOREM 7.5. *For the problem* (P), *let* $\bar{x} \in \mathbb{R}^n$. *Assume that for each* x *in a neighborhood of* $\bar{x}$, $\partial_*^2 f(x)$ *is a bounded approximate Hessian of* f *at* x. *If* $\nabla f(\bar{x}) = 0$ *and for* $0 < \alpha < 1$, *each* $u \in \mathbb{R}^n$ *satisfies* $u \neq 0$; *then the following holds:*

$$(\forall M \in \overline{co}(\partial_*^2 f(\bar{x} + \alpha u))), \quad \langle Mu, u\rangle \geq 0. \tag{7.1}$$

Then $\bar{x}$ *is a local minimum of* (P).

Proof. Suppose that $\bar{x}$ is not a local minimum of (P). Then there exists a sequence $\{x_n\}$ such that $x_n \neq \bar{x}$, $x_n \longrightarrow \bar{x}$ as $n \longrightarrow +\infty$, and $f(x_n) < f(\bar{x})$ for each n. Let $x_n = \bar{x} + u_n$, where $u_n \neq 0$. From the generalized Taylor expansion, Theorem 6.1, there exists $0 < \alpha_n < 1$ such that

$$f(x_n) \in f(\bar{x}) + \langle \nabla f(\bar{x}), x_n - \bar{x}\rangle + \frac{1}{2}\overline{co}\langle \partial_*^2 f(\bar{x} + \alpha_n u_n)(u_n), u_n\rangle.$$

Thus, there exists $M_n \in \overline{co}(\partial_*^2 f(\bar{x} + \alpha_n u_n))$ such that $f(x_n) = f(\bar{x}) + \langle M_n u_n, u_n\rangle$, and so $\langle M_n u_n, u_n\rangle < 0$. This contradicts (7.1). Hence, $\bar{x}$ is a local minimum of (P). □

The following theorem gives us second-order sufficient optimality conditions for a strict local minimum.

THEOREM 7.6. *For the problem* (P), *let* $\bar{x} \in \mathbb{R}^n$. *Assume that, for each* x *in a neighborhood of* $\bar{x}$, $\partial_*^2 f(x)$ *is a bounded approximate Hessian of* f *at* x. *If* $\nabla f(\bar{x}) = 0$ *and for* $0 < \alpha < 1$, *each* $u \in \mathbb{R}^n$ *satisfies* $u \neq 0$, *then the following holds:*

$$(\forall M \in \overline{co}(\partial_*^2 f(\bar{x} + \alpha u))), \quad \langle Mu, u\rangle > 0. \tag{7.2}$$

Then $\bar{x}$ *is a strict local minimum of* (P).

Proof. The method of proof is similar to the one given above for Theorem 7.5 and so it is omitted. □

We now see how the mean value theorem of section 3 and approximate Hessians can be used to characterize convexity of C^1- functions.

THEOREM 7.7. *Let* $f : \mathbb{R}^n \to R$ *be a continuously Gâteaux differentiable function. Assume that* $\partial_*^2 f(x)$ *is an approximate Hessian of* f *for each point* $x \in \mathbb{R}^n$. *If the matrices* $M \in \partial_*^2 f(x)$ *are positive semidefinite for each* $x \in \mathbb{R}^n$, *then* f *is convex.*

Proof. Let $x, u \in \mathbb{R}^n$. Then, by the mean value theorem,

$$\nabla f(x + u) - \nabla f(x) \in \overline{co}(\partial_*^2 f([x, x + u])u),$$

and so,

$$\langle \nabla f(x + u) - \nabla f(x),\ u\rangle \in \langle \overline{co}(\partial_*^2 f([x, x + u])u), u\rangle.$$

Thus, there exist $z \in [x, x + u]$ and $M \in \overline{co}(\partial_*^2 f(z))$ such that

$$\langle \nabla f(x + u) - \nabla f(x),\ u\rangle = \langle Mu, u\rangle.$$

It follows by the assumption that

$$\langle \nabla f(x + u) - \nabla f(x),\ u\rangle \geq 0.$$

Since $x, u \in \mathbb{R}^n$ are arbitrary, we get that ∇f is monotone in the sense that for each $x, u \in \mathbb{R}^n$,

$$\langle \nabla f(x + u) - \nabla f(x),\ u\rangle \geq 0.$$

The conclusion now follows from the standard result of convex analysis that f is convex if and only if ∇f is monotone. □

COROLLARY 7.8. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be $C^{1,1}$. Then f is convex if and only if for each $x \in \mathbb{R}^n$, the matrices $M \in \partial_H^2 f(x)$ are positive semidefinite.*

Proof. Since f is $C^{1,1}$ for each $x \in \mathbb{R}^n$, $\partial_H^2 f(x)$ is an approximate Hessian of f at x. Hence, it follows from Theorem 7.7 that f is convex.

Conversely, assume that f is convex. Let Δ be a set of points in $\mathbb{R}^n$ on which f is twice differentiable. Then, each matrix M of

$$\{ \lim_{n\to\infty} \nabla^2 f(x_n) \ : \ \{x_n\} \subset \Delta,\ x_n \to x\}$$

is positive semidefinite as it is a limit of a sequence of positive semidefinite matrices. Hence, each matrix M of

$$\partial_H^2 f(x) = co\{ \lim_{n\to\infty} \nabla^2 f(x_n) \ : \ \{x_n\} \subset \Delta,\ x_n \to x\}$$

is also positive semidefinite. □

Acknowledgments. The authors are grateful to the referees for their detailed comments and valuable suggestions which have contributed to the final preparation of the paper. The first author is grateful to Professor Jonathan Borwein for his helpful comments on the earlier version of the paper and for certain useful references. The second author wishes to thank the first author for his kind invitation and hospitality.

REFERENCES

[1] F. H. CLARKE, *Necessary Conditions for Pproblems in Optimal Control and Calculus of Variations*, Ph.D. Thesis, University of Washington, Seattle, 1973.
[2] F. H. CLARKE, *Optimization and nonsmooth analysis*, Wiley-Interscience, New York, 1983.
[3] R. COMINETTI AND R. CORREA, *A generalized second-order derivative in nonsmooth optimization*, SIAM J. Control and Optim., 28 (1990), pp. 789–809.
[4] V. F. DEMYANOV AND V. JEYAKUMAR, *Hunting for a smaller convex subdifferential*, J. Global Optim., 10 (1997), pp. 305–326.
[5] V. F. DEMYANOV AND A. M. RUBINOV, *Constructive Nonsmooth Analysis*, Verlag Peter Lang, Frankfurt am Main, 1995.
[6] J. -B. HIRIART-URRUTY, *Mean value theorems for vector valued mappings in nonsmooth optimization*, Numer. Funct. Anal. Optim., 2 (1980), pp. 1–30.
[7] J. B. HIRIART-URRUTY, J. J. STRODIOT, AND V. HIEN NGUYEN, *Generalized Hessian matrix and second-order optimality conditions for problems with $C^{1,1}$ data*, Appl. Math. Optim., 11 (1984), pp. 43–56.
[8] A. D. IOFFE, *Nonsmooth Analysis: differential calculus of nondifferential mappings*, Trans. Amer. Math. Soc., 266 (1981), pp. 1–56.
[9] A. D. IOFFE, *Approximate subdifferentials and applications* I: *The finite dimensional theory*, Trans. Amer. Math. Soc., 281 (1984), pp. 389–416.
[10] A. D. IOFFE AND J. -P. PENOT, *Limiting subhessians and limiting subjects and their calculus*, Trans. Amer. Math. Soc., 349 (1997), pp. 789–808.
[11] V. JEYAKUMAR, *On optimality conditions in nonsmooth inequality constrained minimization*, Numer. Funct. Anal. Optim., 9 (1987), pp. 535–546.
[12] V. JEYAKUMAR, *Composite nonsmooth programming with Gâteaux differentiability*, SIAM J. Optim., 1 (1991), pp. 30–41.
[13] V. JEYAKUMAR AND D. T. LUC, *Nonsmooth Calculus, Minimality and Monotonicity of Convexificators*, Applied Mathematics Research Report AMR96/29, University of New South Wales, Australia, 1996, submitted.
[14] V. JEYAKUMAR AND X. Q. YANG, *Convex composite multi-objective nonsmooth programming*, Math. Progr., 59 (1993), pp. 325–343.
[15] V. JEYAKUMAR AND X. Q. YANG, *Convex composite minimization with $C^{1,1}$ functions*, J. Optim. Theory Appl., 86 (1995), pp. 631–648.

[16] V. Jeyakumar and X. Q. Yang, *Approximate Generalized Hessians and Taylor's Expansions for Continuously Gâteaux Differentiable Functions*, Applied Mathematics Research Report AMR96/20, University of New South Wales, Australia, *Nonlinear Anal.*, 1998, to appear.
[17] D. T. Luc, *Taylor's formula for $C^{k,1}$ functions*, SIAM J. Optim., 5 (1995), pp. 659–669.
[18] D. T. Luc and S. Schaible, *On generalized monotone nonsmooth maps*, J. Convex Anal., 3 (1996), pp. 195–205.
[19] D. T. Luc and S. Swaminathan, *A characterization of convex functions*, Nonlinear Anal., 20 (1993), pp. 697–701.
[20] P. Michel and J.-P. Penot, *A generalized derivative for calm and stable functions*, Differential Integral Equations, 5 (1992), pp. 433–454.
[21] B. S. Mordukhovich, *Metric approximations and necessary optimality conditions for general classes of nonsmooth extremal problems*, Soviet. Math. Dokl., 22 (1980), pp. 526–530.
[22] B. S. Mordukhovich, *Generalized differential calculus for nonsmooth and set-valued mappings*, J. Math. Anal. Appl., 183 (1994), pp. 250–288.
[23] B. S. Mordukhovich and Y. Shao, *On nonconvex subdifferential calculus in Banach spaces*, J. Convex Anal., 2 (1995), pp. 211–228.
[24] B. S. Mordukhovich and Y. Shao, *Nonsmooth sequential analysis in Asplund spaces*, Trans. Amer. Math. Soc., 348 (1996), pp. 1235–1280.
[25] Z. Pales and V. Zeidan, *Generalized Hessian for $C^{1,1}$ functions in infinite dimensional normed spaces*, Math. Programming, 74 (1996), pp. 59–78.
[26] J. S. Pang and L. Qi, *Nonsmooth equations: Motivation and algorithms*, SIAM J. Optim., 3 (1993), pp. 443–465.
[27] R. T. Rockafellar, *Generalized directional derivatives and subgradient of nonconvex functions*, Canad. J. Math., 32 (1980), pp. 257–280.
[28] R. T. Rockafellar, *Second-order optimality conditions in nonlinear programming obtained by way of epi-derivatives*, Math. Oper. Res., 14 (1989), pp. 462–484.
[29] R. T. Rockafellar and J. B. Wets, *Variational Analysis*, Springer-Verlag, Berlin, New York, 1998, to appear.
[30] M. Studniarski and V. Jeyakumar, *A generalized mean-value theorem and optimality conditions in composite nonsmooth minimization*, Nonlinear Anal., 24 (1995), pp. 883–894.
[31] L. Thibault, *On generalized differentials and subdifferentials of Lipschitz vector valued functions*, Nonlinear Anal., 6 (1982), pp. 1037–1053.
[32] J. Warga, *Derivative containers, inverse functions and controllability*, in Calculus of Variations and Control Theory, D.L. Russell, ed., Academic Press, New York, 1976.
[33] J. Warga, *Fat homeomorphisms and unbounded derivative containers*, J. Math. Anal. Appl., 81 (1981), pp. 545–560.
[34] X. Q. Yang, *Generalized Second-Order Directional Derivatives and Optimality Conditions*, Ph.D. thesis, University of New South Wales, Australia, 1994.
[35] X. Q. Yang and V. Jeyakumar, *Generalized second-order directional derivatives and optimization with $C^{1,1}$ functions*, Optimization, 26 (1992), pp. 165–185.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 5, pp. 1833–1851, September 1998

014

$\mathbb{L}^2$ SUFFICIENT CONDITIONS FOR END-CONSTRAINED OPTIMAL CONTROL PROBLEMS WITH INPUTS IN A POLYHEDRON*

J. C. DUNN†

Abstract. An $\mathbb{L}^2$-local optimality sufficiency theorem is proved for a class of structured infinite-dimensional nonconvex programs with constraints of the form $u \in \Omega$ and $h(u) = 0$, where Ω is a set of Lebesgue measurable essentially bounded vector-valued functions $u(\cdot) : [0,1] \to \mathbb{R}^m$ with range in a polyhedron U, and h is a smooth map of the space of essentially bounded functions $u(\cdot)$ into R^k. The sufficiency theorem is based on formal counterparts of the finite-dimensional Karush–Kuhn–Tucker sufficient conditions in a Cartesian product of polyhedra, a strengthened variant of Pontryagin's necessary condition, and structure and continuity conditions on the first and second differentials of the objective function and equality constraint functions. The new sufficient conditions are directly applicable to nonconvex continuous-time Bolza optimal control problems with control-quadratic Hamiltonians, unqualified affine inequality constraints on vector-valued control inputs, and equality constraints on the terminal state vector or equivalent isoperimetric constraints on integrals of functions depending on the state and control variables.

Key words. infinite-dimensional programs, affine inequality constraints, nonconvex equality constraints, nonconvex objectives, $\mathbb{L}^2$-local optimality, second-order sufficient conditions, optimal control, constrained inputs, terminal state constraints

AMS subject classifications. 49M07, 49M10, 49K15, 65K10, 90C06

PII. S0363012995288513

1. Introduction.

In finite-dimensional spaces, all norms are equivalent and local optimality is a norm-invariant property. On the other hand, in infinite-dimensional spaces, norm-equivalence is lost and local optimality in one norm need not imply local optimality in another. This fact has computational and theoretical implications for refined finite-dimensional approximations to constrained minimization problems in infinite-dimensional function spaces [13], [15].

The distinction between strong and weak minimizing curves in the calculus of variations provides a classic illustration of norm-dependent local optimality in function spaces. For variational problems, strict versions of the necessary conditions of Legendre and Jacobi are sufficient for weak local optimality but not strong local optimality; however, strong local optimality can be deduced when a strict version of the Weierstrass necessary condition is added to the weak local optimality sufficient conditions. This classical development has a natural extension to Bolza optimal control problems for ordinary differential equations [3], [4], [21], [22], [23], [24], [28], [29], [30]. In the optimal control setting, some strict form of the Pontryagin necessary condition replaces the Weierstrass condition since pointwise constraints on control or state variables are generally present. Variants of the Pontryagin minimum principle are also invoked in the global optimality sufficient conditions of [5], [6], [7], and [8] for time-optimal control of state-constrained ordinary differential inclusions.

Alternative local optimality sufficient conditions for optimal control problems have been established with modifications of a basic proof strategy for finite-dimensional

*Received by the editors July 5, 1995; accepted for publication (in revised form) May 6, 1997; published electronically July 9, 1998. This research was supported by NSF grant DMS-9500908.
http://www.siam.org/journals/sicon/36-5/28851.html
†Mathematics Department, Box 8205, North Carolina State University, Raleigh, NC 27695-8205 (dunn@eos.ncsu.edu).

nonlinear programs. As one might expect, these sufficient conditions are closely related to the strict complementarity and coercivity hypotheses in the Karush–Kuhn–Tucker (KKT) theory. Initially, the KKT approach produced sufficient conditions for weak local optimality in the control context, i.e., local optimality in the $\mathbb{L}^\infty$ norm on control functions [9], [20], [21]; however, recent investigations have deduced both weak ($\mathbb{L}^\infty$) and strong ($\mathbb{L}^2$) local optimality from strict complementarity and coercivity hypotheses, and an additional condition of the Pontryagin type for specially structured infinite-dimensional nonlinear programs and related optimal control problems [11], [14], [15], [27].

The strong $\mathbb{L}^2$-local optimality sufficient conditions established in [14], [15], [27] apply to infinite-dimensional nonlinear programs,

$$\min J(u), \tag{1.1a}$$

subject to

$$u \in \Omega = \{u \in \mathbb{L}^\infty_m[0,1] : u(t) \overset{a.e.}{\in} U\}, \tag{1.1b}$$

where $\mathbb{L}^\infty_m[0,1]$ is the vector space of Lebesgue measurable essentially bounded functions $u(\cdot) : [0,1] \to \mathbb{R}^m$, U is a polyhedral convex set in $\mathbb{R}^m$, and the first and second Gâteaux differentials of J satisfy certain structure conditions and $\mathbb{L}^2$ continuity conditions described in section 2. The latter conditions have been shown to hold for Bolza optimal control problems with control-quadratic Hamiltonians and, more specifically, for nonconvex nonquadratic regulator optimal control problems [27]. In the present article, the $\mathbb{L}^2$-local optimality sufficiency proof strategy in [15] is extended to a larger class of smooth structured nonconvex constrained minimization problems,

$$\min J(u), \tag{1.2a}$$

subject to

$$u \in \Omega_h = \{u \in \Omega : h(u) = 0\}, \tag{1.2b}$$

where Ω is defined in (1.1b), h maps $\mathbb{L}^\infty_m[0,1]$ to $\mathbb{R}^k$, and J and $h_1, \ldots, h_k$ satisfy the structure and continuity conditions in section 2. The results obtained for (1.2) are immediately applicable to an important class of Bolza optimal control problems with nonconvex objective functions and end-constraint functions defined by

$$J(u) = P\left(x(u)(1)\right) + \int_0^1 f^0\left(t, x(u)(t), u(t)\right) dt, \tag{1.3a}$$

and

$$h_i(u) = \pi_i\left(x(u)(1)\right) + \int_0^1 \phi^0_i\left(t, x(u)(t), u(t)\right) dt, \quad i = 1, \ldots, k, \tag{1.3b}$$

where $x(u) : [0,1] \to \mathbb{R}^n$ is the unique (absolutely continuous) solution of an initial value problem

$$\frac{dx}{dt}(t) \overset{a.e.}{=} f\left(t, x(t), u(t)\right), \tag{1.3c}$$

$$x(0) = x^0. \tag{1.3d}$$

As in [27], the structure and continuity conditions of section 2 will hold for these problems when the Hamiltonians,

$$H(t,\psi,x,u) = f^0(t,x,u) + \langle \psi, f(t,x,u)\rangle, \tag{1.4}$$

and

$$H(t,\psi,x,u) = \phi_i^0(t,x,u) + \langle \psi, f(t,x,u)\rangle, \tag{1.5}$$

are quadratic in the control input vector $u \in \mathbb{R}^m$ and when P, π_i, f^0, ϕ_i^0, and f satisfy suitable smoothness and growth conditions. Note that (1.3b) admits both terminal state constraints ($\phi_i^0 = 0$) and isoperimetric constraints ($\pi_i = 0$).

In finite-dimensional nonconvex programming, second-order sufficient conditions are natural starting points in the development of sensitivity analyses and local convergence theories for gradient-related methods, multiplier methods, sequential quadratic programming methods, and other iterative computational schemes. Analogous theories rest on function space local optimality sufficient conditions and related growth estimates for J in the feasible set near an optimal u_* [3], [9], [15], [17], [18], [19], [26], [27]. In particular, a strengthened corollary of the $\mathbb{L}^2$-local optimality sufficient conditions in [11], [14], and [15] supports well-developed local convergence and active constraint identification theorems for gradient projection methods in the setting of problem (1.1) [15], [26], [27]. It seems likely that an analogous corollary of the $\mathbb{L}^2$-local optimality sufficient conditions established in section 4 will find similar uses in local convergence theories for augmented gradient projection methods and sequential quadratic programming methods applicable to problem (1.2).

2. Structure and continuity conditions. As in [15], problem (1.2) is set in the pre-Hilbert space $\{\mathbb{L}_m^\infty[0,1], \langle\cdot,\cdot\rangle_2\}$, with the standard $\mathbb{L}^2$ inner product,

$$\langle u, v\rangle_2 = \int_0^1 \langle u(t), v(t)\rangle dt,$$

associated norm,

$$\|u\|_2 = \sqrt{\langle u, u\rangle_2} = \left(\int_0^1 \|u(t)\|^2 dt\right)^{\frac{1}{2}},$$

and open balls,

$$B_2(u,\delta) = \{v \in \mathbb{L}_m^\infty[0,1] : \|v-u\|_2 < \delta\},$$

where $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ are the Euclidean inner product and norm on $\mathbb{R}^m$. The analysis in section 4 requires that the structure conditions and $\mathbb{L}^2$ continuity conditions imposed on J in [15] hold here for J *and* the constraint components h_i, $i = 1,\dots,k$. For a generic real functional F, these conditions specify that for all u in $\mathbb{L}_m^\infty[0,1]$, there exist $\nabla F(u) \in \mathbb{L}_m^\infty[0,1]$, $S_F(u) \in \mathbb{L}_{m\times m}^\infty[0,1]$, and $K_F(u) \in \mathbb{L}_{m\times m}^2([0,1]\times[0,1])$ such that

$$d^1F(u\,;\,v) = \langle \nabla F(u), v\rangle_2 = \int_0^1 \langle \nabla F(u)(t), v(t)\rangle dt, \tag{2.6}$$

$$d^2F(u\,;\,v,w) = \langle v, \nabla^2 F(u)w\rangle_2 = \int_0^1 \langle v(t), (\nabla^2 F(u)w)(t)\rangle dt, \tag{2.7a}$$

$$(\nabla^2 F(u)w)(t) = S_F(u)(t)w(t) + \int_0^1 K_F(u)(t,s)w(s)ds \tag{2.7b}$$

for all v and w in $\mathbb{L}_m^\infty[0,1]$ and almost all t in $[0,1]$ and (s,t) in $[0,1]\times[0,1]$, with

$$\lim_{\|v-u\|_2\to 0} \|S_F(v) - S_F(u)\|_\infty = 0, \tag{2.8a}$$

$$\lim_{\|v-u\|_2\to 0} \|K_F(v) - K_F(u)\|_2 = 0, \tag{2.8b}$$

where

$$\|S_F(v) - S_F(u)\|_\infty \stackrel{def}{=} ess \sup_{t\in[0,1]} \|S_F(v)(t) - S_F(u)(t)\|$$

and

$$\|K_F(v) - K_F(u)\|_2 \stackrel{def}{=} \left(\int_0^1 \int_0^1 \|K_F(v)(t,s) - K_F(u)(t,s)\|^2 dtds\right)^{\frac{1}{2}}.$$

It is also assumed that the $m\times m$ matrices $S_F(u)(t)$ and $K_F(u)(t,s)$ are symmetric, with $K_F(u)(t,s) = K_F(u)(s,t)$, and that the vector and matrix norms on $\mathbb{R}^m$ and $\mathbb{R}^{m\times m}$ in (2.6)–(2.8) are induced by the standard Euclidean inner product on $\mathbb{R}^m$. These conditions imply that F is twice continuously Fréchet differentiable on the pre-Hilbert space $\{\mathbb{L}_m^\infty[0,1], \|\cdot\|_2\}$, and therefore establish the Taylor formula,

$$\begin{aligned} F(u) - F(u_*) &= \langle \nabla F(u_*), u-u_*\rangle_2 \\ &\quad + \frac{1}{2}\langle u-u_*, \nabla^2 F(u_*)(u-u_*)\rangle_2 + r_F(u_*\,;\, u-u_*), \end{aligned} \tag{2.9a}$$

with

$$\lim_{\|u-u_*\|_2\to 0} \frac{r_F(u_*\,;\, u-u_*)}{\|u-u_*\|_2^2} = 0. \tag{2.9b}$$

Note 2.1. In the Bolza optimal control formulation (1.3), the vectors $\nabla J(u)(t)$ and matrices $S_J(u)(t)$ are formally derived from the u-gradient and u-Hessian of the Hamiltonian (1.4) evaluated on the state and co-state trajectories $x(u)(\cdot)$ and $\psi(u)(\cdot)$ corresponding to $u(\cdot)$. More precisely,

$$\nabla J(u)(t) = \nabla_u H(t, \psi(u)(t), x(u)(t), u(t))$$

and

$$S_J(u)(t) = \nabla_{uu}^2 H(t, \psi(u)(t), x(u)(t), u(t)),$$

where $x(u)(\cdot)$ solves the initial value problem, (1.3c) and (1.3d), and $\psi(u)(\cdot)$ solves the *adjoint* backward initial value problem,

$$\frac{d\psi}{dt}(t) \stackrel{a.e.}{=} -\nabla_x H(t, \psi, x(u)(t), u(t)),$$

$$\psi(1) = \nabla P(x(u)(1)).$$

In the same way, $\nabla h_i(u)(t)$ and $S_{h_i}(u)(t)$ are formally obtained by an analogous construction, with H defined by (1.5) instead of (1.4) and P replaced by π_i. If conditions (2.6)–(2.8) are to hold for J and h_i, then the corresponding Hamiltonians must be u-quadratic, and the functions P, π_i, f^0, ϕ_i^0, and f must satisfy additional differentiability and growth hypotheses to ensure existence, uniqueness, and smooth dependence on u in the $\mathbb{L}^2$ norm for the state and costate trajectories $x(u)(\cdot)$ and $\psi(u)(\cdot)$. The global hypotheses in [15] and [27] admit nontrivial nonconvex optimal control problems, including certain nonquadratic regulator problems with u-quadratic integrands f^0 and ϕ_i^0, and nonlinear state equations with u-linear right sides f.

3. Preliminary results. By definition [25], the cone of exterior normals at a point u in the convex set Ω is

$$\mathcal{N}_\Omega(u) = \{w \in \mathbb{L}_m^\infty[0,1] : \forall v \in \Omega \quad \langle w, v-u\rangle_2 \le 0\}. \tag{3.10}$$

This cone, the associated complementary orthogonal closed subspaces,

$$\mathbb{N}_\Omega(u) = \text{cl } span\mathcal{N}_\Omega(u), \tag{3.11a}$$

$$\mathbb{T}_\Omega(u) = \mathbb{N}_\Omega(u)^\perp, \tag{3.11b}$$

the constraint derivative null space,

$$\mathbb{T}_h(u_*) = \ker h'(u_*) = \{w \in \mathbb{L}_m^\infty[0,1] : \langle \nabla h_i(u_*), w\rangle_2 = 0, \quad i = 1,\dots,k\}, \tag{3.12}$$

and the subspace,

$$\mathbb{T}(u) = \mathbb{T}_\Omega(u) \cap \mathbb{T}_h(u), \tag{3.13}$$

are fundamental objects in the sufficiency theorem and proof of section 4. Basic properties of the orthogonal projection maps $P_{\mathbb{N}_\Omega(u)}$ and $P_{\mathbb{T}_\Omega(u)}$ established in [14] and [15] are reviewed in this section, and an elementary right inverse lemma is proved for the restriction of the linear map $h'(u)$ to the subspace $\mathbb{T}_\Omega(u)$. Note that the existence portion of the Hilbert space projection theorem and the standard Banach space right inverse lemma [1, pp. 79–80], [16, p. 155] can't be invoked here since the inner product space $\{\mathbb{L}_m^\infty[0,1], \langle\cdot,\cdot\rangle_2\}$ is incomplete. Our arguments rely instead on the orthogonality characterization of projector maps and the finite codimensionality of $\mathbb{T}_h(u)$. This approach is straightforward enough and works equally well for the counterpart of problem (1.2) in the Hilbert space $\{\mathbb{L}_m^2[0,1], \langle\cdot,\cdot\rangle_2\}$ (However, see Note 4.5).

References [14] and [15] supply the following key representations for the cone and subspaces in (3.10) and (3.11) and related projection decomposition formulas:

$$\mathcal{N}_\Omega(u) = \{w \in \mathbb{L}_m^\infty[0,1] : w(t) \overset{a.e}{\in} \mathcal{N}_U(u(t))\} \tag{3.14}$$

and

$$\mathbb{N}_\Omega(u) = \{w \in \mathbb{L}_m^\infty[0,1] : w(t) \overset{a.e}{\in} \mathbb{N}_U(u(t))\}, \tag{3.15a}$$

$$\mathbb{T}_\Omega(u) = \{w \in \mathbb{L}_m^\infty[0,1] : w(t) \overset{a.e}{\in} \mathbb{T}_U(u(t))\}, \tag{3.15b}$$

where

$$\mathcal{N}_U(\xi) = \{\zeta \in \mathbb{R}^m : \forall \eta \in U \quad \langle \zeta, \eta - \xi \rangle \leq 0\},$$

$$\mathbb{N}_U(\xi) = span\, \mathcal{N}_U(\xi),$$

and

$$\mathbb{T}_U(\xi) = \mathbb{N}_U(\xi)^{\perp}$$

for ξ in U. Note that the polyhedron U is the union of the relative interiors of its polyhedral faces $\{\mathcal{F}_1, \ldots, \mathcal{F}_d\}$, with $ri\ \mathcal{F}_i \cap ri\ \mathcal{F}_j = \emptyset$ for $i \neq j$, and $\mathcal{N}_U(\cdot)$ *constant* on each set $ri\ \mathcal{F}_i$. Hence, the set-valued functions $\mathcal{N}_U(u(\cdot))$, $\mathbb{N}_U(u(\cdot))$, and $\mathbb{T}_U(u(\cdot))$ are constant on the sets

$$\alpha_i(u) = u^{-1}\,[ri\ \mathcal{F}_i], \quad i = 1, \ldots, d, \tag{3.16a}$$

with

$$\alpha_i(u) \cap \alpha_j(u) = \emptyset \text{ for } i \neq j \tag{3.16b}$$

and

$$\mu\left([0,1] \setminus \cup_{i=1}^{d}\, \alpha_i(u)\right) = 0, \tag{3.16c}$$

where μ denotes Lebesgue measure. From this it follows easily that for each measurable essentially bounded z, the pointwise projection decomposition formulas,

$$\left(P_{\mathbb{N}_\Omega(u)} z\right)\ (t) = P_{\mathbb{N}_U(u(t))}\ z(t), \tag{3.17a}$$

$$\left(P_{\mathbb{T}_\Omega(u)} z\right)\ (t) = P_{\mathbb{T}_U(u(t))}\ z(t), \tag{3.17b}$$

produce measurable essentially bounded functions $P_{\mathbb{N}_\Omega(u)} z$ in $\mathbb{N}_\Omega(u)$ and $P_{\mathbb{T}_\Omega(u)} z$ in $\mathbb{T}_\Omega(u)$ such that $z - P_{\mathbb{N}_\Omega(u)} z$ and $z - P_{\mathbb{T}_\Omega(u)} z$ are orthogonal to $\mathbb{N}_\Omega(u)$ and $\mathbb{T}_\Omega(u)$, respectively. Thus, (3.17) defines *orthogonal projection maps*, $P_{\mathbb{N}_\Omega(u)}$ and $P_{\mathbb{T}_\Omega(u)}$, from $\mathbb{L}^\infty_m[0,1]$ into the complementary orthogonal closed subspaces $\mathbb{N}_\Omega(u)$ and $\mathbb{T}_\Omega(u)$. In fact, $\mathbb{N}_\Omega(u)$ and $\mathbb{T}_\Omega(u)$ are actually *pointwise orthogonal* in the sense that

$$\forall v \in \mathbb{N}_\Omega(u),\ \forall w \in \mathbb{T}_\Omega(u), \quad \langle v(t), w(t) \rangle \overset{a.e.}{=} 0. \tag{3.18}$$

A pre-Hilbert space variant of the Banach space right inverse lemma in [1, pp. 79–80] and [16, p. 155] is also needed in the sufficiency proof of section 4. The following simple lemmas lead directly to the required result in Corollary 3.3.

LEMMA 3.1. *Let* $\{a_1, \ldots, a_l\}$ *be a linearly independent set in a real inner product space* $\{\mathbb{U}, \langle \cdot, \cdot \rangle_{\mathbb{U}}\}$, *let* $\mathbb{V} = \{a_1, \ldots, a_l\}^{\perp}$, *and let* $\mathbb{U}/\mathbb{V}$ *denote the corresponding quotient space of cosets* $[u] = u + \mathbb{V}$ *with* u *in* $\mathbb{U}$. *Then* $\{[a_1], \ldots, [a_l]\}$ *is a basis for* $\mathbb{U}/\mathbb{V}$, *and consequently* $\dim \mathbb{U}/\mathbb{V} = l$.

Proof. Fix u in $\mathbb{U}$. Since $\{a_1, \ldots, a_l\}$ is linearly independent, the corresponding Gramian matrix is invertible and there is a unique $\alpha \in \mathbb{R}^l$ such that

$$\left\langle a_i, u - \sum_{j=1}^{l} \alpha_j a_j \right\rangle_{\mathbb{U}} = 0, \quad i = 1, \ldots, l,$$

or equivalently

$$u - \sum_{j=1}^{l} \alpha_j a_j \in \mathbb{V}.$$

It follows that for each $[u]$ in $\mathbb{U}/\mathbb{V}$, there is a unique $\alpha \in \mathbb{R}^l$ such that

$$[u] = \left[\sum_{j=1}^{l} \alpha_j a_j\right] = \sum_{j=1}^{l} \alpha_j [a_j]$$

with $\alpha = 0$ iff $[u] = 0$. □

LEMMA 3.2. *Let $\{a_1, \ldots, a_k\}$ be a finite set in a real inner product space $\{\mathbb{U}, \langle\cdot,\cdot\rangle_{\mathbb{U}}\}$, and define $A : \mathbb{U} \to \mathbb{R}^k$ by the rule*

$$\forall w \in \mathbb{U},\ \forall i = 1, \ldots, k, \quad (Aw)_i = \langle a_i, w\rangle_{\mathbb{U}}.$$

Then there is a real number $b > 0$ and a map $B : A(\mathbb{U}) \to \mathbb{U}$ such that

$$\forall \xi \in A(\mathbb{U}), \quad AB(\xi) = \xi, \text{ and } \|B(\xi)\|_{\mathbb{U}} \le b\|\xi\|. \tag{3.19}$$

Proof. If $a_i = 0$ for $i = 1, \ldots, k$, then $A(\mathbb{U}) = \{0\}$ and condition (3.19) holds trivially with $B(0) = 0$ and any $b \ge 0$. Suppose that $a_i \ne 0$ for some i, and relabel the a_i's if necessary so that $\{a_1, \ldots, a_l\}$ is a basis for $span\{a_1, \ldots, a_k\}$. By construction, the null space of A is $\mathbb{V} = \{a_1, \ldots, a_l\}^{\perp}$. By Lemma 3.1, $\dim\mathbb{U}/\mathbb{V} = l$. Hence the rule

$$\forall [u] \in \mathbb{U}/\mathbb{V}, \quad \hat{A}[u] = Au$$

defines a one-to-one linear map from the finite-dimensional space $\mathbb{U}/\mathbb{V}$ onto the finite-dimensional space $A(\mathbb{U}) \subset \mathbb{R}^k$, and the corresponding inverse map $\hat{A}^{-1} : A(\mathbb{U}) \to \mathbb{U}/\mathbb{V}$ is therefore automatically *bounded*, i.e.,

$$\exists \hat{b}\ \forall \xi \in A(\mathbb{U}), \quad \|\hat{A}^{-1}\xi\|_{\mathbb{U}/\mathbb{V}} \le \hat{b}\|\xi\|,$$

where

$$\|\hat{A}^{-1}\xi\|_{\mathbb{U}/\mathbb{V}} \stackrel{def}{=} \inf_{u \in \hat{A}^{-1}\xi} \|u\|_{\mathbb{U}}.$$

Put $b = \hat{b} + 1$. Then for each $\xi \in A(\mathbb{U})$, there is a $B(\xi) \in \hat{A}^{-1}\xi \subset \mathbb{U}$ such that

$$AB(\xi) = \hat{A}[B(\xi)] = \hat{A}\hat{A}^{-1}\xi = \xi$$

and

$$\|B(\xi)\|_{\mathbb{U}} \le \inf_{u \in \hat{A}^{-1}\xi} \|u\|_{\mathbb{U}} + \|\xi\|. \qquad \square$$

COROLLARY 3.3. *Let h be Gâteaux differentiable at a point u_* in Ω, with*

$$(h'(u_*)w)_i = d^1 h_i(u_*\ ;\ w) = \langle \nabla h_i(u_*), w\rangle_2, \quad i = 1, \ldots, k,$$

for some $\nabla h_i(u_*) \in \mathbb{L}^\infty_m[0,1]$ *and all* $w \in \mathbb{L}^\infty_m[0,1]$. *Then there is a real number* $b_* \geq 0$ *and a map* $B_* : h'(u_*)(\mathbb{T}_\Omega(u_*)) \to \mathbb{T}_\Omega(u_*)$ *such that for all* $w \in \mathbb{T}_\Omega(u_*)$,

$$h'(u_*)B_*(h'(u_*)w) = h'(u_*)w$$

and

$$\|B_*(h'(u_*)w)\|_2 \leq b_*\|h'(u_*)w\|.$$

Proof. Apply the lemma with $\mathbb{U} = \mathbb{T}_\Omega(u_*)$, $\langle\cdot,\cdot\rangle_\mathbb{U} = \langle\cdot,\cdot\rangle_2$, and $a_i = P_{\mathbb{T}_\Omega(u_*)}\nabla h_i(u_*)$ for $i = 1,\ldots,k$. ∎

4. $\mathbb{L}^2$-local optimality sufficient conditions. $\mathbb{L}^2$-local optimality sufficiency theorems are proved in [11], [14], and [15] for problem (1.1), where U is a polyhedral convex set in $\mathbb{R}^m$ and J satisfies the structure and continuity conditions of section 2. The developments in [14] and [15] are guided by an analogy between (1.1) and its finite-dimensional counterpart in Cartesian products $\Omega^k = U \times \cdots \times U \subset \mathbb{R}^{k\times m}$. This analogy suggests that sufficiency theorems for (1.1) may rest on the following pointwise strict complementarity condition and $\mathbb{L}^2$ coercivity condition at a point $u_* \in \Omega$:

$$-\nabla J(u_*)(t) \overset{a.e.}{\in} ri\, \mathcal{N}_U(u_*(t)), \tag{4.20a}$$

$$\forall w \in \mathbb{T}_\Omega(u_*), \qquad \langle w, \nabla^2 J(u_*)w\rangle_2 \geq c_T\, \|w\|_2^2, \tag{4.20b}$$

where $\mathcal{N}_U(\xi)$ is the cone of outer normals to the polyhedron U at $\xi \in U$, $T_\Omega(u_*)$ is the orthogonal complement of the normal cone $\mathcal{N}_\Omega(u_*)$ in the inner product space $\{\mathbb{L}^\infty_m[0,1], \langle\cdot,\cdot\rangle_2\}$ (section 3), and c_T is a positive real number. These conditions are indeed central hypotheses in the sufficiency theorems of [11], [14], and [15]; however, (4.20a) is so much weaker than its finite-dimensional componentwise counterpart in the sets Ω^k that (4.20a) and (4.20b) alone are *not* sufficient for local optimality, even in the $\mathbb{L}^\infty$ norm. More precisely, since (4.20a) does not imply that the gradient values $-\nabla J(u_*)(t)$ are essentially bounded away from the *relative boundary* $rb\, \mathcal{N}_U(u_*(t))$ for $t \in [0,1]$, and since the essential sup norm $\|u - u_*\|_\infty$ can increase without bound as $\|u - u_*\|_2$ approaches zero, it follows that (4.20a) does not ensure adequate growth of the first-order term $\langle\nabla J(u_*), u - u_*\rangle_2$ for $u \in \Omega$, $\|u - u_*\|_2$ small, and $u - u_*$ bounded away from $T_\Omega(u_*)$ *in direction* [14]. Additional hypotheses and modified proof techniques are therefore required to compensate for the deficiency in (4.20a) in the infinite-dimensional setting of (1.1). For the $\mathbb{L}^\infty$-local optimality sufficiency theorems in [11], [14], and [15], the hypotheses take the form of restrictions on the behavior of the operators $S_J(u_*)(t)$ in section 2 and the subspaces *span* $\mathcal{N}_U(u_*(t))$ near frontier points of the sets $\alpha_i \subset [0,1]$ in (3.16a), where $u_*(t)$ passes from the relative interior of one polyhedral face in U to another. On the other hand, in the $\mathbb{L}^2$-local optimality theorems of [11], [14], and [15], conditions (4.20a) and (4.20b) are supplemented by a strengthened variant of an $\mathbb{L}^2$-local optimality necessary condition akin to Pontryagin's minimum principle, namely,

$$\forall \xi \in U, \quad \mathcal{H}_J(u_*\,;\,\xi,t) - \mathcal{H}_J(u_*\,;\,u_*(t),t) \geq \frac{1}{2}\, c_P\|\xi - u_*(t)\|^2 \tag{4.20c}$$

a.e. in $[0,1]$, where c_P is a positive number and

$$\mathcal{H}_J(u_* \,;\, \xi, t) = \langle \nabla J(u_*)(t), \xi - u_*(t)\rangle + \frac{1}{2}\langle \xi - u_*(t), S_J(u_*)(t)\,(\xi - u_*(t))\rangle.$$

Conditions (4.20) appear once again in the present analysis for problem (1.2) with J replaced everywhere by a Lagrangian function,

$$L(\lambda_*, \cdot) = J(\cdot) + \langle \lambda_*, h(\cdot)\rangle,$$

and $\mathbb{T}_\Omega(u_*)$ replaced by the subspace $\mathbb{T}(u_*) = \mathbb{T}_\Omega(u_*) \cap \mathbb{T}_h(u_*)$ in (3.13). An $\mathbb{L}^2$-local optimality sufficiency theorem based on these conditions will now be proved with suitable modifications of the proof strategy developed in [15], and the following technical lemma.

LEMMA 4.1. *Suppose that h satisfies the hypotheses of Corollary* 3.3 *at some point u_* in Ω and that b_* is the nonnegative real number in Corollary* 3.3. *Assume that β is a Lebesgue measurable set in $[0,1]$ and $u \in \mathbb{L}^\infty_m[0,1]$, and put*

$$w = (1 - \chi_\beta)(u - u_*)$$

and

$$w_{\mathbb{T}_\Omega} = P_{\mathbb{T}_\Omega(u_*)} w,$$

where χ_β is the characteristic function of β. Then for some $\hat{w}$ in the subspace $\mathbb{T}(u_*)$,

$$\|w - \hat{w}\|_2 \le (1 + b_*\|h'(u_*)\|)\|w - w_{\mathbb{T}_\Omega}\|_2 \tag{4.21}$$

$$+ b_* \left(\int_\beta \sum_{i=1}^k \|\nabla h_i(u_*)(t)\|^2 dt \right)^{\frac{1}{2}} \|u - u_*\|_2 \tag{4.22}$$

$$+ b_* \|h'(u_*)(u - u_*)\|.$$

Proof. Let $B_* : h'(u_*)(\mathbb{T}_\Omega(u_*)) \to \mathbb{T}_\Omega(u_*)$ be the right inverse map in Corollary 3.3. For u in $\mathbb{L}^\infty_m[0,1]$, put

$$\hat{w} = w_{\mathbb{T}_\Omega} - B_*(h'(u_*) w_{\mathbb{T}_\Omega}).$$

By construction, $\hat{w} \in \mathbb{T}_\Omega(u_*)$ and $h'(u_*)\hat{w} = 0$, and therefore $\hat{w} \in \mathbb{T}(u_*)$. By the triangle inequality and Corollary 3.3,

$$\begin{aligned}
\|w - \hat{w}\|_2 &\le \|w - w_{\mathbb{T}_\Omega}\|_2 + \|w_{\mathbb{T}_\Omega} - \hat{w}\|_2 \\
&\le \|w - w_{\mathbb{T}_\Omega}\|_2 + b_*\|h'(u_*) w_{\mathbb{T}_\Omega}\|_2 \\
&\le (1 + b_*\|h'(u_*)\|)\|w - w_{\mathbb{T}_\Omega}\|_2 \\
&\quad + b_*(\|h'(u_*)(u - u_* - w)\| + \|h'(u_*)(u - u_*)\|)
\end{aligned}$$

with

$$\begin{aligned}
\|h'(u_*)(u - u_* - w)\| &= \left(\sum_{i=1}^k \langle \nabla h_i(u_*), \chi_\beta(u - u_*)\rangle_2^2 \right)^{\frac{1}{2}} \\
&\le \left(\int_\beta \sum_{i=1}^k \|\nabla h_i(u_*)(t)\|^2 dt \right)^{\frac{1}{2}} \|u - u_*\|_2. \qquad \square
\end{aligned}$$

THEOREM 4.2. *Suppose that the structure and continuity conditions* (2.6)–(2.8b) *are met by the objective function* J*, the constraint functions,* $h_1, \ldots, h_k$*, and hence the Lagrangians* $L(\lambda, \cdot)$ *for problem* (1.2)*. Let* $S(\lambda, u)$ *denote the corresponding matrix-valued function* $S_{L(\lambda,\cdot)}(u)$ *in conditions* (2.6)–(2.8) *for* $L(\lambda, \cdot)$*. Let* $u_* \in \Omega_h$ *and assume that for some* $\lambda_* \in \mathbb{R}^k$*,* $c_T > 0$*, and* $c_P > 0$*, the following conditions hold at* u_**:*

$$-\nabla L(\lambda_*, u_*)(t) \overset{a.e.}{\in} ri\, \mathcal{N}_U(u_*(t)), \tag{4.23a}$$

$$\forall w \in \mathbb{T}(u_*), \qquad \langle w, \nabla^2 L(\lambda_*, u_*) w\rangle_2 \geq c_T \, \|w\|_2^2, \tag{4.23b}$$

and

$$\forall \xi \in U, \quad \mathcal{H}(\lambda_*, u_* \,;\, \xi, t) - \mathcal{H}(\lambda_*, u_* \,;\, u_*(t), t) \geq \frac{1}{2}\, c_P \|\xi - u_*(t)\|^2 \tag{4.23c}$$

a.e. in $[0,1]$*, with* $\mathcal{H}(\lambda_*, u_* \,;\, \xi, t) = \mathcal{H}_{L(\lambda_*,\cdot)}(u_* \,; \xi, t)$*, i.e.,*

$$\mathcal{H}(\lambda_*, u_* \,;\, \xi, t) = \langle \nabla L(\lambda_*, u_*)(t), \xi - u_*(t)\rangle + \frac{1}{2}\langle \xi - u_*(t), S(\lambda_*, u_*)(t)\,(\xi - u_*(t))\rangle.$$

Then u_* *is an* $\mathbb{L}^2$*-local minimizer for problem* (1.2)*; more specifically, for each* c_2 *in the interval* $0 < c_2 < \min\{c_T, c_P\}$*, there is a corresponding* $\delta_2 > 0$ *such that*

$$J(u) - J(u_*) \geq \frac{1}{2} c_2 \|u - u_*\|_2^2 \tag{4.24}$$

for all $u \in \Omega_h \cap B_2(u_*, \delta_2)$*.*

Proof. Conditions (2.6)–(2.8) are satisfied by J and $h_1, \ldots, h_k$, and hence by the Lagrangian $L(\lambda_*, \cdot)$. Thus, Taylor's formula (2.9) is valid for $L(\lambda_*, \cdot)$ and the components of h, and it follows that for all u in Ω_h,

$$\begin{aligned} J(u) - J(u_*) &= \langle \nabla L(\lambda_*, u_*), u - u_*\rangle_2 \\ &\quad + \frac{1}{2}\langle u - u_*, \nabla^2 L(\lambda_*, u_*)(u - u_*)\rangle_2 + r_L(\lambda_*, u_* \,;\, u - u_*), \end{aligned} \tag{4.25a}$$

$$r_L(\lambda_*, u_* \,;\, u - u_*) = o(\|u - u_*\|_2^2), \tag{4.25b}$$

and

$$\|h'(u_*)(u - u_*)\| = \left(\sum_{i=1}^{k} \langle \nabla h_i(u_*), u - u_*\rangle_2^2\right)^{\frac{1}{2}} = o(\|u - u_*\|_2). \tag{4.26}$$

The desired estimate (4.24) will now be obtained from (4.23), (4.25), (4.26), and the decomposition,

$$u - u_* = (1 - \chi_{\varphi(u)})(u - u_*) + \chi_{\varphi(u)}(u - u_*), \tag{4.27}$$

where $\varphi(u)$ is a suitably constructed measurable set in $[0,1]$ corresponding to u.

Let $K(\lambda_*, u_*)$ denote the matrix-valued function $K_{L(\lambda_*,\cdot)}(u_*)$ in conditions (2.6)–(2.8) for $L(\lambda_*, \cdot)$. Then (2.6)–(2.8), (4.23c), (4.25), and (4.27) immediately produce

$$\begin{aligned} J(u) - J(u_*) \geq\ & \langle \nabla L(\lambda_*, u_*), w(u)\rangle_2 \\ & + \frac{1}{2}\langle w(u), \nabla^2 L(\lambda_*, u_*) w(u)\rangle_2 + \frac{1}{2} c_P \|\chi_{\varphi(u)}(u - u_*)\|_2^2 \\ & - \frac{1}{2}\left(\int\int_{(\varphi(u)^c \times \varphi(u)^c)^c} \|K(\lambda_*, u_*)(t,s)\|^2 \, dtds \right)^{\frac{1}{2}} \|u - u_*\|_2^2 \\ & + r_L(\lambda_*, u_* \,;\, u - u_*), \end{aligned}$$

with

$$w(u) = (1 - \chi_{\varphi(u)})(u - u_*),$$

where $\varphi^c = [0,1] \setminus \varphi$ and $(\varphi^c \times \varphi^c)^c = ([0,1] \times [0,1]) \setminus (\varphi^c \times \varphi^c)$. Hence (4.24) follows at once if the sets $\varphi(u)$ are constructed so that

$$\begin{aligned} -\frac{1}{2}\left(\int\int_{(\varphi(u)^c \times \varphi(u)^c)^c} \|K(\lambda_*, u_*)(t,s)\|^2 \, dtds \right)^{\frac{1}{2}} \|u - u_*\|_2^2 \\ + r_L(\lambda_*, u_* \,;\, u - u_*) \geq -\frac{1}{4}(\min\{c_T, c_P\} - c_2)\|u - u_*\|_2^2 \end{aligned} \tag{4.28}$$

and

$$\begin{aligned} \langle \nabla L(\lambda_*, u_*), w(u)\rangle_2 + \frac{1}{2}\langle w(u), \nabla^2 L(\lambda_*, u_*) w(u)\rangle_2 \\ \geq\ \frac{1}{2} c_T \|w(u)\|_2^2 - \frac{1}{4}(\min\{c_T, c_P\} - c_2)\|u - u_*\|_2^2 \end{aligned} \tag{4.29}$$

for all $u \in \Omega_h \cap B_2(u_*, \delta_2)$ with δ_2 sufficiently small. Condition (4.28) holds if δ_2 and the Lebesgue measure $\mu(\varphi(u))$ are merely sufficiently small; however, (4.29) also requires that $\sup_{t \in \varphi(u)^c} \|u(t) - u_*(t)\|$ is sufficiently small, and that $-\nabla L(\lambda_*, u_*)(t)$ is bounded away from the relative boundary of $\mathcal{N}_U(u_*(t))$ in $\mathbb{N}_U(u_*(t))$ for t in $\varphi(u)^c$. Suitable sets $\varphi(u)$ are described fully below, along with the estimates that establish (4.28) and (4.29). As in [14] and [15], the cone

$$C_\epsilon(u_*) = \{u \in \mathbb{L}_m^\infty[0,1] : \|w(u) - w(u)_{\mathbb{T}_\Omega}\|_2 \leq \epsilon \|w(u)\|_2\}$$

is at the center of this development. More specifically, (4.29) is proved by first applying (4.23a), (4.23b), and Lemma 4.1 for $w(u)$ in $C_\epsilon(u_*)$ with ϵ sufficiently small, and then invoking (4.23a) and the above-mentioned properties of $\varphi(u)$ for $w(u)$ in $C_\epsilon(u_*)^c$.

Fix c_2 in the interval $0 < c_2 < \min\{c_T, c_P\}$, and let b_* be the nonnegative real number in Lemma 4.1. Put

$$M = 1 + 3b_* \|h'(u_*)\|, \tag{4.30a}$$

and choose $\epsilon > 0$ so that

$$(2 + 3M\epsilon)(c_T + \|\nabla^2 L(\lambda_*, u_*)\|)M\epsilon \leq \frac{1}{2}(\min\{c_T, c_P\} - c_2). \tag{4.30b}$$

Note that $\mu((\varphi^c \times \varphi^c)^c) \leq 2\mu(\varphi)$, and that Lebesgue integrals are absolutely continuous functions of their domain sets. Hence, there is a $\nu \in (0,1]$ such that

$$\left(\int\int_{(\varphi^c \times \varphi^c)^c} \|K(\lambda_*, u_*)(t,s)\|^2 \, dtds \right)^{\frac{1}{2}} \leq \frac{1}{4}(\min\{c_T, c_P\} - c_2) \tag{4.31}$$

and

$$\left(\int_{\varphi} \sum_{i=1}^{k} \|\nabla h_i(u_*)(t)\|^2 dt \right)^{\frac{1}{2}} \leq \epsilon \|h'(u_*)\| \tag{4.32}$$

for all measurable sets $\varphi \subset [0,1]$ with $\mu(\varphi) \leq \nu$. Furthermore, by (4.25) and (4.26), there is a $\delta' > 0$ such that

$$|r_L(\lambda_*, u_* \,;\, u - u_*)| \leq \frac{1}{8}(\min\{c_T, c_P\} - c_2)\|u - u_*\|_2^2 \tag{4.33}$$

and

$$\|h'(u_*)(u - u_*)\| \leq \epsilon \|h'(u_*)\| \|u - u_*\|_2 \tag{4.34}$$

for all u in $\Omega_h \cap B_2(u_*, \delta')$. Recall that the set-valued map $\mathcal{N}_U(\cdot)$ is constant on each of the sets $\alpha_i(u_*)$ in (3.16a), and that for each nonempty set $\mathcal{S}$ in $\mathbb{R}^m$, the real function, $dist(\cdot, \mathcal{S}) : \mathbb{R}^m \to \mathbb{R}^1$, is continuous. Hence, if S is the relative boundary of $\mathcal{N}_U(u_*(t))$, then the formula,

$$\Delta(u_*)(t) = dist[-\nabla L(\lambda_*, u_*)(t), rb\mathcal{N}_U(u_*(t))],$$

defines a measurable extended real-valued function on $[0,1]$[1]; moreover, $\Delta(u_*)(t) > 0$ almost everywhere, in view of (4.23a). It follows that for some measurable set β in $[0,1]$,

$$\mu(\beta^c) \leq \frac{1}{2}\nu \tag{4.35a}$$

and

$$c_\beta \stackrel{def}{=} \inf_{t \in \beta} \Delta(u_*)(t) > 0. \tag{4.35b}$$

Now choose $\delta \in (0, \delta']$ so that

$$\frac{\epsilon^2 c_\beta}{\delta} - \frac{1}{2}\|\nabla^2 L(\lambda_*, u_*)\| \geq \frac{1}{2}c_T, \tag{4.36a}$$

and let

$$\delta_2 = \sqrt{\frac{\nu}{2}}\delta. \tag{4.36b}$$

Finally, for u in Ω_h, put

$$\varphi(u) = (\beta \cap \theta(u))^c = \beta^c \cup \theta(u)^c, \tag{4.37a}$$

[1] With $\Delta(u_*)(t) = +\infty$ when the relative boundary $rb\mathcal{N}_U(u_*(t))$ is empty.

with

$$\theta(u) = \{t \in [0,1] : \|u(t) - u_*(t)\| \leq \delta\}. \tag{4.37b}$$

By construction, $\delta_2 \leq \delta'$ and

$$\mu(\varphi(u)) \leq \mu(\beta^c) + \mu(\theta(u)^c) \leq \frac{\nu}{2} + \frac{\nu}{2}$$

for all u in $\Omega_h \cap B_2(u_*, \delta_2)$. Hence, (4.31) and (4.33) immediately yield (4.28) for all u in $\Omega_h \cap B_2(u_*, \delta_2)$. To see that (4.29) also holds, suppose that $u \in \Omega_h \cap B_2(u_*, \delta_2)$ and $w(u) \in C_\epsilon(u_*)$. Note that $w(u) = v(u) - u_*$, with $v(u) = (1 - \chi_{\varphi(u)})u + \chi_{\varphi(u)}u_* \in \Omega$. Hence, (4.23a) implies that

$$\langle \nabla L(\lambda_*, u_*), w(u)\rangle_2 \geq 0. \tag{4.38}$$

Furthermore, by Lemma 4.1 and the estimates (4.32) and (4.34), there is a $\hat{w}(u)$ in the subspace $\mathbb{T}(u_*)$ such that

$$\|w(u) - \hat{w}(u)\|_2 \leq M\epsilon\|u - u_*\|_2$$

and therefore

$$\|\hat{w}(u)\|_2 \leq (1 + M\epsilon)\|u - u_*\|_2$$

and

$$\begin{aligned}\|\hat{w}(u)\|_2^2 &= \|w(u)\|_2^2 - \langle w(u) + \hat{w}(u), w(u) - \hat{w}(u)\rangle_2 \\ &\geq \|w(u)\|_2^2 - (\|w(u)\|_2 + \|\hat{w}(u)\|_2)\|w(u) - \hat{w}(u)\|_2 \\ &\geq \|w(u)\|_2^2 - (2 + M\epsilon)M\epsilon\|u - u_*\|_2^2.\end{aligned}$$

Conditions (4.23b) and (4.30) now yield

$$\begin{aligned}\langle w(u), \nabla^2 L(\lambda_*, u_*)w(u)\rangle_2 &\geq c_T\|\hat{w}(u)\|_2^2 \\ &\quad - \|\nabla^2 L(\lambda_*, u_*)\|(2\|\hat{w}(u)\|_2\|w(u) - \hat{w}(u)\|_2 \\ &\qquad + \|w(u) - \hat{w}(u)\|_2^2) \\ &\geq c_T\|\hat{w}(u)\|_2^2 - \frac{1}{2}(\min\{c_T, c_P\} - c_2)\|u - u_*\|_2^2.\end{aligned}$$

This estimate and (4.38) establish (4.29) for $u \in \Omega_h \cap B_2(u_*, \delta_2)$ and $w(u) \in C_\epsilon(u_*)$. On the other hand, suppose that $u \in \Omega_h \cap B_2(u*, \delta_2)$ and $w(u) \in C_\epsilon(u_*)^c = \mathbb{L}_m^\infty[0,1] \setminus C_\epsilon(u_*)$. As in [15], put $z = c_\beta\delta^{-1}(w(u) - w(u)_{\mathbb{T}_\Omega})$ and note that $z(t) \overset{a.e.}{\in} \mathbb{N}_U(u_*(t))$ and $ess\sup_{t\in[0,1]} \|z(t)\| \leq c_\beta$, in view of (3.17) and (4.37). Conditions (4.35) and (4.37) then yield

$$-\nabla L(\lambda_*, u_*)(t) + z(t) \overset{a.e.}{\in} \mathcal{N}_U(u_*(t)),$$

in which case

$$\langle \nabla L(\lambda_*, u_*)(t), w(u)(t)\rangle \overset{a.e.}{\geq} \frac{c_\beta}{\delta}\|w(u)(t) - w(u)_{\mathbb{T}_\Omega}(t)\|^2,$$

and therefore

$$\langle \nabla L(\lambda_*, u_*), w(u)\rangle_2 \geq \frac{c_\beta}{\delta}\|w(u) - w(u)_{\mathbb{T}_\Omega}\|_2^2 \geq \frac{\epsilon^2 c_\beta}{\delta}\|w(u)\|_2^2.$$

This estimate and (4.36) establish (4.29) for $u \in \Omega_h \cap B_2(u*, \delta_2)$ and $w(u) \in C_\epsilon(u_*)^c$. □

COROLLARY 4.3. *Assume that the hypotheses of Theorem* 4.2 *hold, with the Pontryagin condition* (4.23c) *replaced by a stronger coercivity condition of the Legendre–Clebsch type, i.e., for some* $c_P > 0$,

$$\forall \xi \in \mathbb{R}^m, \quad \langle \xi, S(\lambda_*, u_*)(t)\xi \rangle \geq c_P \|\xi\|^2 \tag{4.39}$$

a.e. in $[0,1]$. *Then* u_* *is an* $\mathbb{L}^2$*-local minimizer of* J *in* Ω_h, *and the growth condition* (4.24) *holds in* $B_2(u_*, \delta_2)$ *for some* $\delta_2 > 0$.

The sufficient conditions in Theorem 4.2 imply the $\mathbb{L}^2$-quadratic growth property (4.24), which is clearly stronger than $\mathbb{L}^2$-local optimality, per se. Condition (4.24) and similar uniform local growth properties are needed in Liapunov-like local asymptotic stability analyses for iterative constrained minimization algorithms in infinite-dimensional spaces, where local minimizers (or even strict local minimizers) need not be stable local attractors for the standard iterative maps [10], [12], [13], [15], [26], [27], [9]. Examples 1 and 2 in [13] demonstrate the gap between (4.24) and $\mathbb{L}^2$-local optimality for (1.1); each of these problems has a convex objective function J and a unique global minimizer u_* that does not have property (4.24). The following simple examples accomplish the same purpose for (1.2).

Example 4.1. Let $J(u) = \int_0^1 u\ dt$, $U = [0, \infty)$, and $h(u) = \int_0^1 (1-2t)u\ dt$ in (1.2). Then $u_* = 0$ is the unique global (and hence $\mathbb{L}^2$-local) minimizer for the linear functional J in Ω_h. To see that (4.24) does not hold at u_*, let $\epsilon \in (0, 1/2)$, $u_\epsilon = \epsilon^{-\frac{1}{3}} \chi_{[\frac{1}{2}-\epsilon, \frac{1}{2}+\epsilon]}$, and note that $u_\epsilon \in \Omega_h$ and

$$J(u_\epsilon) - J(u_*) = \int_{\frac{1}{2}-\epsilon}^{\frac{1}{2}+\epsilon} \epsilon^{-\frac{1}{3}}\ dt = 2\epsilon^{\frac{2}{3}} = \frac{1}{2}\|u_\epsilon\|^4,$$

with

$$\lim_{\epsilon \to 0^+} \|u_\epsilon - u_*\|_2 = 0.$$

Now let $L(\lambda, u) = \int_0^1 u\ dt + \lambda \int_0^1 (1-2t)u\ dt$, and consider that $ri\ \mathcal{N}_U(u_*(t)) = (-\infty, 0)$, $\mathbb{T}_U(u_*(t)) = \{0\}$, and $\nabla L(\lambda, u)(t) = 1 + \lambda(1-2t)$ for $t \in [0,1]$. Moreover, (2.6)–(2.8) hold trivially with $S(\lambda, u) = 0$ and $K(\lambda, u) = 0$. Thus, (4.23a) holds at u_* for any $\lambda_* \in [-1, 1]$, and (4.23b) is satisfied trivially since $\mathbb{T}(u_*) = \{0\}$. In addition, the Pontryagin condition,

$$\forall \xi \in [0, \infty), \quad (1 + \lambda_*(1-2t))\xi \geq 0,$$

is satisfied almost everywhere in $[0,1]$ for any $\lambda_* \in [-1, 1]$. On the other hand, the strengthened Pontryagin condition (4.23c) requires that

$$\forall \xi \in [0, \infty), \quad (1 + \lambda_*(1-2t))\xi \geq \frac{1}{2} c_P \xi^2$$

almost everywhere in $[0,1]$, and this is impossible for $c_P > 0$ and $\lambda_* \in [-1, 1]$.

Example 4.2. Let $J(u) = \int_0^1 u^2\ dt - (\int_0^1 u\ dt)^2$, $U = [0,1]$, and $h(u) = \int_0^1 (1-2t)u\ dt$ in (1.2). By Cauchy's inequality, J has a global (and hence $\mathbb{L}^2$-local)

minimizer in Ω_h at $u_* = 0$. To see that (4.24) does not hold at u_*, let $\epsilon \in [0,1]$, $u_\epsilon(t) = \epsilon$ for $t \in [0,1]$, and note that $u_\epsilon \in \Omega_h$ and

$$J(u_\epsilon) - J(u_*) = \int_0^1 \epsilon^2 \, dt - \left(\int_0^1 \epsilon \, dt \right)^2 = 0,$$

with

$$\lim_{\epsilon \to 0^+} \|u_\epsilon - u_*\|_2 = 0.$$

Now let $L(\lambda, u) = \int_0^1 u^2 \, dt - (\int_0^1 u \, dt)^2 + \lambda \int_0^1 (1-2t)u \, dt$, and observe that $\mathcal{N}_U(u_*(t)) = (-\infty, 0]$, $ri\ \mathcal{N}_U(u_*(t)) = (-\infty, 0)$, $\mathbb{T}_U(u_*(t)) = \{0\}$, $\nabla L(\lambda, u)(t) = 2u(t) - 2\int_0^1 u \, dt + \lambda(1-2t)$, and $(\nabla^2 L(\lambda, u)w)(t) = (\nabla^2 J(u)w)(t) = 2w(t) - 2\int_0^1 w \, dt$. By Cauchy's inequality, $\nabla^2 J(u)$ is positive semidefinite and hence J and $L(\lambda, \cdot)$ are convex. Furthermore (2.6)–(2.8) hold for L with $S(\lambda, u)(t) = 2$ and $K(\lambda, u)(t, s) = -2$. Since $1 - 2t$ changes sign at $t = 1/2$, it is clear that (4.23a) can't hold at u_* for any λ_*. On the other hand, the weaker first-order necessary condition,

$$-\nabla L(\lambda_*, u_*)(t) \overset{a.e.}{\in} \mathcal{N}_U(u_*(t)), \tag{4.40}$$

is satisfied iff $\lambda_* = 0$, and condition (4.23b) holds trivially since $\mathbb{T}(u_*) = \{0\}$. Finally, if $\lambda_* = 0$, then $\nabla L(\lambda_*, u_*) = 0$ and the strengthened Pontryagin condition (4.23c) is satisfied, since $S(\lambda_*, u_*)(t) = 2 > 0$. □

The simple convex programs in Examples 4.1 and 4.2 show that the $\mathbb{L}^2$-quadratic growth property (4.24) may be lost if either of the conditions (4.23a) or (4.23c) is weakened appreciably. For nonconvex programs, similar relaxations in the hypotheses of Theorem 4.2 may admit functions u_* that not only fail to have property (4.24) but also are not locally optimal even in the weak $\mathbb{L}^\infty$ sense (cf. Example 1 in [11]). On the other hand, the sufficient conditions in Theorem 4.2 are certainly not *necessary* for the $\mathbb{L}^2$-quadratic growth property (4.24). In particular, suppose that h is affine and continuous in the $\mathbb{L}^2$ norm and that J is twice continuously Fréchet differentiable in the $\mathbb{L}^2$ norm and strongly convex with coercive Hessians, i.e.,

$$\forall u \exists c_u \forall w, \quad \langle w, \nabla^2 J(u) w \rangle_2 \geq c_u \|w\|_2^2.$$

Then for all $\lambda \in \mathbb{R}^k$, the Lagrangian $L(\lambda, \cdot)$ is also twice continuously Fréchet differentiable in the $\mathbb{L}^2$ norm and strongly convex, with $\nabla^2 L(\lambda, u) = \nabla^2 J(u)$ and therefore

$$J(u) - J(u_*) \geq \langle \nabla L(\lambda_*, u_*), u - u_* \rangle_2 + \frac{1}{2} c_{u_*} \|u - u_*\|_2^2 + o(\|u - u_*\|_2^2)$$

for all $u, u_* \in \Omega_h$ and $\lambda_* \in \mathbb{R}^k$. The growth property (4.24) now follows at once if the first-order necessary condition (4.40) holds for some $\lambda_* \in \mathbb{R}^k$. Thus, when h is affine, the strong coercivity condition on $\nabla^2 J$ yields (4.24) directly, *without* (2.6)–(2.8), (4.23a), and (4.23c). However, if (2.6)–(2.8) happen to hold, then the $\mathbb{L}^2$ coercivity condition on J implies that the operators $S(\lambda_*, u_*)(t)$ are essentially uniformly coercive (cf. the proof of Theorem 6.4 in [14]), and this immediately yields (4.23c).

Hypotheses (2.6)–(2.8), (4.23a), (4.23c) are also superfluous in Theorem 4.2 when $u_*(t) \overset{a.e.}{\in} \text{int } U$ (or more generally, when $u_*(t) \overset{a.e.}{\in} ri\ U$). In this exceptional case, it is possible to establish (4.24) with a variant of the classic sufficiency proof for

equality-constrained problems with feasible sets $h^{-1}(0)$ (cf. [1]). More specifically, if $u_*(t) \overset{a.e.}{\in} int\ U$, then $ri\ \mathcal{N}_U(u_*(t)) \overset{a.e}{=} \mathcal{N}_U(u_*(t)) \overset{a.e}{=} \{0\}$, $rb\ \mathcal{N}_U(u_*(t)) \overset{a.e}{=} \emptyset$, $dist\ (-\nabla L(\lambda_*, u_*)(t), rb\ \mathcal{N}_U(u_*(t))) \overset{a.e}{=} +\infty$, $\mathbb{T}_U(u_*(t) \overset{a.e.}{=} \mathbb{R}^m$, $\mathbb{T}_\Omega(u_*) = \mathbb{L}^\infty_m[0,1]$, and therefore $\mathbb{T}(u_*) = \mathbb{T}_h(u_*) = ker\ h'(u_*)$. If J and h are twice continuously Fréchet differentiable and the first-order necessary condition (4.40) holds at u_*, then Taylor's formula (2.9) reduces to

$$J(u) - J(u_*) = \frac{1}{2}\langle u - u_*, \nabla^2 L(\lambda_*, u_*)(u - u_*)\rangle_2 + o(\|u - u_*\|_2^2)$$

for $u \in \Omega_h = h^{-1}(0)$. An application of Lemma 4.1 with $\beta = \emptyset$ now produces the $\mathbb{L}^2$-quadratic growth estimate (4.24) directly from the coercivity condition (4.23b). Note that the hypothesis $u_*(t) \overset{a.e.}{\in} \text{int}\, U$ does *not* imply that $u_* \in \text{int}\, \Omega$ in either the $\mathbb{L}^2$ or the $\mathbb{L}^\infty$ sense; in fact, the $\mathbb{L}^2$ interior of Ω is *empty* whenever U is a proper subset of $\mathbb{R}^m$. Note also that if the structure and continuity conditions (2.6)–(2.8) are satisfied, then (4.40) and (4.23b) once again imply (4.23c).

Finally, if (2.6)–(2.8) hold with $K(\lambda_*, u_*) = 0$, then the proof of Theorem 4.2 can be drastically simplified. In this rare and essentially trivial case, Taylor's formula (2.9) reduces to

$$J(u) - J(u_*) = \int_0^1 [\langle \nabla L(\lambda_*, u_*)(t), u(t) - u_*(t)\rangle + \langle u(t) - u_*(t), S(\lambda_*, u_*)(t)(u(t) - u_*(t))\rangle]\, dt + o(\|u - u_*\|_2^2),$$

for $u \in \Omega_h$. The $\mathbb{L}^2$-quadratic growth condition (4.24) now follows at once from (4.23c), and the remaining hypotheses (4.23a) and (4.23b) in Theorem 4.2 are superfluous. However, it can be shown that (4.23c) implies (4.23b) in the present special circumstances.

Note 4.3. For the Bolza optimal control scheme (1.3), our earlier observations in Note 2.1 establish that $\nabla L(\lambda, u)(t)$ and $S(\lambda, u)(t)$ are formally obtained from

$$\nabla L(\lambda, u)(t) = \nabla_u H(\lambda, t, \psi(\lambda, u)(t), x(u)(t), u(t))$$

and

$$S(\lambda, u)(t) = \nabla^2_{uu} H(\lambda, t, \psi(\lambda, u)(t), x(u)(t), u(t)),$$

where $x(u)(\cdot)$ solves the initial value problem, (1.3c) and (1.3d), and $\psi(\lambda, u)(\cdot)$ solves the *adjoint* backward initial value problem,

$$\frac{d\psi}{dt}(t) \overset{a.e.}{=} -\nabla_x H(\lambda, t, \psi, x(u)(t), u(t)),$$

$$\psi(1) = \nabla P(x(u)(1)) + \sum_{i=1}^k \lambda_i \nabla \pi_i(x(u)(1)),$$

with

$$H(\lambda, t, \psi, x, u) = f^0(t, x, u) + \sum_{i=1}^k \lambda_i \phi_i^0(t, x, u) + \langle \psi, f(t, x, u)\rangle.$$

When $H(\lambda, t, \psi, x, u)$ is quadratic in u, it can now be seen that $\mathcal{H}(\lambda_*, u_* \,;\, \xi, t)$ coincides with the increment,

$$H(\lambda_*, t, \psi(\lambda_*, u_*)(t), x(u_*)(t), \xi) - H(\lambda_*, t, \psi(\lambda_*, u_*)(t), x(u_*)(t), u_*(t)).$$

Hence, condition (4.23c) amounts to a strengthening of the Pontryagin Minimum Principle for problem (1.3), i.e.,

$$H(\lambda_*, t, \psi(\lambda_*, u_*)(t), x(u_*)(t), u_*(t)) = \min_{\xi \in U} H(\lambda_*, t, \psi(\lambda_*, u_*)(t), x(u_*)(t), \xi) \tag{4.41}$$

a.e. in $[0, 1]$. Under minimal smoothness requirements on P, π_i, f^0, ϕ_i^0, and f in the Bolza formulation, condition (4.41) will hold at *normal* $\mathbb{L}^2$-local minimizers that satisfy additional end-constraint regularity and controllability conditions. Similarly, the pointwise strict complementarity condition (4.23a) is a strengthening of the Lagrange stationarity condition,

$$-\nabla L(\lambda_*, u_*)(t) \overset{a.e.}{\in} \mathcal{N}_U(u_*(t)),$$

which also follows as a corollary of the stronger Pontryagin condition (4.41) in the convex set U. Finally, the analysis in [15] for problem (1.1) suggests that the $\mathbb{L}^2$ coercivity condition (4.23b) may be viewed as a stronger version of a second-order necessary condition,

$$\forall w \in \mathbb{T}(u_*), \qquad \langle w, \nabla^2 L(\lambda_*, u_*) w \rangle_2 \geq 0; \tag{4.42}$$

however, while the necessity of (4.42) is certainly plausible, it remains to be proved in the context of problem (1.2), since the set Ω is not a polyhedron, and standard second-order necessary conditions in nonpolyhedral feasible sets require representation-dependent constraint qualifications for Ω that are not invoked in the present geometric development.

Note 4.4. For Bolza optimal control problems, $\mathbb{L}^2$ coercivity conditions like (4.23b) are implied by second-order necessary conditions, Legendre–Clebsch conditions, and disconjugacy conditions of the Jacobi type [21], [28], [29]. The Legendre–Clebsch condition (4.39) and Corollary 4.3 are also important in $\mathbb{L}^2$-local convergence theories for gradient projection methods [15], [27] and other familiar constrained minimization schemes whose iteration maps have fixed points at $\mathbb{L}^\infty$-local minimizers [13], [15]. Condition (4.39) is a natural requirement for nonconvex nonquadratic regulator Bolza problems with u-linear state equations, and integrands f^0 and ϕ_i^0 of the form $q(t, x) + \langle r(t, x), u \rangle + \langle u, s(t, x) u \rangle$, with uniformly positive-definite $m \times m$ matrices $s(t, x)$.

Note 4.5. As noted earlier, the proof techniques and results in this section are equally valid in the Hilbert space $\{\mathbb{L}^2_m[0, 1], \| \cdot \|_2\}$; however, in this complete inner product space, it is possible to treat equality constraint functions h with range in an infinite-dimensional Banach space $\mathbb{Y}$ by invoking the Banach space right inverse lemma in place of Lemma 3.2. In this setting, the Lagrangian is defined by $L(\lambda_*, \cdot) = J(u) + \lambda_*(h(\cdot))$ with λ_* a bounded linear functional on $\mathbb{Y}$, the structure and continuity hypotheses (2.6)–(2.8) are imposed directly on $L(\lambda, \cdot)$, h is assumed to be twice continuously Fréchet differentiable, the range of $h'(u_*)$ is assumed to be closed in $\mathbb{Y}$, and $h'(u_*)$ is required to satisfy the absolute continuity condition,

$$\lim_{\mu(\beta) \to 0} \sup_{\|w\|_2 = 1} \|h'(u_*) \chi_\beta w\| = 0.$$

Under these circumstances, counterparts of Lemma 4.1 and Theorem 4.2 can be proved exactly as before.

Acknowledgments. The author gratefully acknowledges several valuable exposition-related comments offered by one of the referees. These observations are incorporated in Examples 4.1 and 4.2 and the accompanying discussion between Corollary 4.3 and Note 4.3.

REFERENCES

[1] V. M. Alekseev, V. M. Tikhomirov, and S. V. Fomin, *Optimal Control*, Plenum, New York, 1987.
[2] W. Alt and K. Malanowski, *The Lagrange-Newton method for nonlinear optimal control problems*, Comput. Optim. Appl. 2 (1993), pp. 77–100.
[3] A. V. Arutyunov and N. T. Tynyanskii, *Second-order conditions in the time optimal problem*, Soviet Math. Dokl., 24 (1981), pp. 525–528.
[4] A. V. Arutyunov and N. T. Tynyanskii, *Conditions of the first and second order in a problem of time optimality*, Uspekhi Mat. Nauk., 36 (1981), pp. 199–200.
[5] V. I. Blagodatskikh, *Sufficient conditions of optimality for differential inclusions*, Math. USSR Izv., 8 (1974), pp. 621–630.
[6] V. I. Blagodatskikh, *On the theory of sufficient conditions of optimality*, Soviet Math. Dokl., 17 (1976), pp. 1680–1683.
[7] V. I. Blagodatskikh, *On the theory of sufficient conditions of optimality*, Trudy Mat. Inst., 142 (1976), pp. 78–87.
[8] V. I. Blagodatskikh, *Sufficient conditions for optimality in problems with state constraints*, Appl. Math. Optim., 7 (1981), pp. 149–157.
[9] A. S. Dontchev, W. W. Hager, A. B. Poore, and B. Yang, *Optimality, stability and convergence in nonlinear control*, Appl. Math. Optim., 31 (1995), pp. 297–326.
[10] J. C. Dunn, *On the convergence of projected gradient processes to singular critical points*, J. Optim. Theory Appl., 55 (1987), pp. 203–216.
[11] J. C. Dunn and T. Tian, *Variants of the Kuhn-Tucker sufficient conditions in cones of nonnegative functions*, SIAM J. Control Optim., 30 (1992), pp. 1361–1384.
[12] J. C. Dunn, *A subspace decomposition principle for scaled gradient projection methods: Local theory*, SIAM J. Control Optim., 31 (1993), pp. 219–246.
[13] J. C. Dunn, *Gradient-related constrained minimization algorithms in function spaces: Convergence properties and computational implications*, in Large Scale Optimization: State of the Art, Kluwer Academic Publishers, Dordrecht, 1994.
[14] J. C. Dunn, *Second order optimality conditions in sets of $\mathbb{L}^\infty$ functions with range in a polyhedron*, SIAM J. Control Optim., 33 (1995), pp. 1603–1635.
[15] J. C. Dunn, *On $\mathbb{L}^2$ sufficient conditions and the gradient projection method for optimal control problems*, SIAM J. Control Optim., 34 (1996), pp. 1270–1290.
[16] D. G. Luenberger, *Optimization by Vector Space Methods*, John Wiley, New York, 1969.
[17] K. Malanowski, *Sensitivity analysis of optimization problems in Hilbert space, with application to optimal control*, Appl. Math. Optim., 21 (1990), pp. 1–20.
[18] K. Malanowski, *Second order conditions and constraint qualifications in stability and sensitivity analysis of solutions to optimization problems in Hilbert space*, Appl. Math. Optim., 25 (1992), pp. 51–79.
[19] K. Malanowski, *Two-norm approach in stability and sensitivity analysis of optimization and optimal control problems*, Adv. Math. Sci. Appl., 2 (1993), pp. 397–443.
[20] H. Maurer and J. Zowe, *First and second-order necessary and sufficient optimality conditions for infinite-dimensional programming problems*, Math. Programming, 16 (1979), pp. 98–110.
[21] H. Maurer, *First and second order sufficient optimality conditions in mathematical programming and optimal control*, Math. Programming Stud., 14 (1981), pp. 163–177.
[22] H. Maurer, *The Two-Norm Approach for Second Order Sufficiency Conditions in Mathematical Programming and Optimal Control*, Tech. report 6/92-N, Inst. Angew. Math. Inform., Universität Münster, Germany, 1992.
[23] H. Maurer, *Solution differentiability for parametric nonlinear control problems with control-state constraints*, Control Cybernet., 23 (1994), pp. 201–227.

[24] D. Orrell and V. Zeidan, *Another Jacobi sufficiency criterion for optimal control with smooth constraints*, J. Optim. Theory Appl., 58 (1988), pp. 283–300.
[25] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[26] T. Tian, *Convergence Analysis of a Projected Gradient Method for a Class of Optimal Control Problems*, Ph.D. dissertation, North Carolina State University, Raleigh, NC, 1992.
[27] T. Tian and J. C. Dunn, *On the gradient projection method for optimal control problems with non-negative* $\mathbb{L}^2$ *inputs*, SIAM J. Control Optim., 32 (1994), pp. 516–537.
[28] V. Zeidan, *Sufficient conditions for the generalized problem of Bolza*, Trans. Amer. Math. Soc., 275 (1983), pp. 561–586.
[29] V. Zeidan, *Sufficiency criteria via focal points and via coupled points*, SIAM J. Control Optim., 30 (1992), pp. 82–98.
[30] V. Zeidan, *The Riccati equation for optimal control problems with mixed state-control constraints: Necessity and sufficiency*, SIAM J. Control Optim., 32 (1994), pp. 1297–1321.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 6, pp. 1853–1879, November 1998

001

PONTRYAGIN'S PRINCIPLE FOR STATE-CONSTRAINED CONTROL PROBLEMS GOVERNED BY PARABOLIC EQUATIONS WITH UNBOUNDED CONTROLS*

J. P. RAYMOND† AND H. ZIDANI†

Abstract. This paper deals with optimal control problems governed by semilinear parabolic equations with pointwise state constraints and unbounded controls. Under some strong stability assumption, we obtain necessary optimality conditions in the form of a Pontryagin's minimum principle in qualified form. A Pontryagin's principle in nonqualified form is also proved without any stability condition.

Key words. optimal control, nonlinear boundary controls, semilinear parabolic equations, state constraints, Pontryagin's minimum principle, unbounded controls

AMS subject classifications. 49K20, 93C20

PII. S0363012996302470

1. Introduction.

This article concerns optimal control problems for the following parabolic system:

$$\frac{\partial y}{\partial t} + Ay + f(x,t,y) = 0 \text{ in } Q, \quad \frac{\partial y}{\partial n_A} + g(s,t,y,v) = 0 \text{ on } \Sigma, \quad y(0) = w \text{ in } \Omega, \tag{1.1}$$

where $\Omega \subset \mathbb{R}^N$, $Q = \Omega \times]0,T[$, $\Sigma = \Gamma \times]0,T[$, Γ is the boundary of Ω, $T > 0$, v is a boundary control, w is a control of the initial condition, and A is a second order elliptic operator. Constraints of the form

$$v \in V_{ad} \subset L^{\sigma}(\Sigma), \qquad w \in W_{ad} \subset C(\overline{\Omega}),$$

$$\phi(y) \in \mathcal{C} \tag{1.2}$$

are imposed on the control variables v, w, and the state variable y (here ϕ is a continuous mapping from $C(\overline{Q})$ into $C(\overline{D})$, $\mathcal{C} \subset C(\overline{D})$ is a closed convex subset with nonempty interior in $C(\overline{D})$, and $\overline{D}$ is a nonempty compact subset of $\overline{Q}$). The control problem is

$$(P) \quad \inf \left\{ J(y,v,w) \mid (y,v,w) \in C(\overline{Q}) \times V_{ad} \times W_{ad},\ (y,v,w) \text{ satisfies (1.1), (1.2)} \right\},$$

where the cost functional is defined by

$$J(y,v,w) = \int_Q F(x,t,y)\,dx\,dt + \int_\Sigma G(s,t,y,v)\,ds\,dt + \int_\Omega L(x,y(T),w)dx. \tag{1.3}$$

We are mainly interested in optimality conditions for such problems, in the form of Pontryagin's principles. The existence of optimal solutions for (P) is a priori assumed.

In recent years there has been growing interest in optimality conditions for state-constrained control problems governed by partial differential equations (or variational

*Received by the editors April 22, 1996; accepted for publication (in revised form) July 7, 1997; published electronically July 17, 1998.
http://www.siam.org/journals/sicon/36-6/30247.html
†Universityé Paul Saltier, Laboratory MAP, URN CARS 9974, 31062 Toulouse cede 4, France (raymond@mip.ups-Tyler., zidani@mip.ups-Tyler).

inequalities). This is reflected by an important number of papers on this subject. For convex control problems we refer to [1], [4], [5], [7], and [33]. In the case of nonconvex control problems, the method of Lagrange multipliers provides optimality conditions for both bounded and unbounded controls [3], [11], [39], [34], [35]. When no qualification condition is assumed, optimality conditions are obtained in nonqualified form (optimality conditions of Fritz John type). To get optimality conditions in qualified form with a Lagrange multiplier theorem, a qualification condition such as the Zowe–Kurcyusz regularity condition is needed. (In many problems, this regularity condition corresponds to a Slater type qualification condition; see [39] and [34].)

Another method proceeds by penalizing the state constraints and then characterizing optimal solutions of the original problem as ε-solutions of the penalized problems. The characterization of ε-solutions is carried out thanks to the Ekeland variational principle. By this method optimality conditions are obtained in the form of Pontryagin principles, which are in general more precise than optimality conditions deduced from Lagrange multiplier theorems. Moreover, assumptions on the data of the problems (differentiability assumptions, convexity requirement, etc.) are less restrictive than those necessary for Lagrange multiplier theorems. To the best of our knowledge, except in [25], this method has so far been used only for problems with bounded controls (bounded in time [30], [31], [32] or in space and time [12], [23], [24], [26]).

There is a fundamental reason for this limitation. When we apply Ekeland's principle to obtain Pontryagin's principle, we need a complete metric space, let us say (V_{ad}, d_E) (the space of controls V_{ad}, endowed with the so-called Ekeland's metric d_E, in order to recover a Pontryagin principle), and a penalized functional $F_\varepsilon(v) = J_\varepsilon(y_v, v)$ (v is the control variable, y_v is the solution of the state equation corresponding to v) which must be lower semicontinuous on (V_{ad}, d_E). To prove this lower semicontinuity property we need some assumptions on J_ε and we have to prove that the mapping $\mathcal{T} : v \longmapsto y_v$ is continuous from (V_{ad}, d_E) into a Banach space Z (which depends on the considered problem). The continuity of $\mathcal{T}$ depends on regularity results for the state equation. In the problems studied in the articles mentioned above, V_{ad} is a subset of some Lebesgue space L^σ with $1 \leq \sigma \leq \infty$, and (thanks to regularity results for partial differential equations) it can be proved that the mapping $\mathcal{T}$ is continuous from L^σ into Z for every $\sigma > \bar{\sigma}$, where $\bar{\sigma}$ depends on the state equation.

If V_{ad} is bounded in L^∞ and if a sequence of controls converges for the Ekeland metric, it can easily be proved that this sequence still converges for the topology of L^σ for any $\sigma > \bar{\sigma}$. Therefore in this case, the mapping $\mathcal{T}$ is continuous and Ekeland's principle can be applied.

If V_{ad} is not bounded in L^∞, convergence in the Ekeland metric does not imply convergence in the Lebesgue space norm; moreover, (V_{ad}, d_E) is not necessarily complete. This is the reason why, up to now, in the presence of pointwise state constraints, Pontryagin's principles have only been proved for bounded controls (at least for nonconvex problems; indeed for convex problems the optimality conditions deduced from Lagrange multiplier theorems correspond to Pontryagin principles).

Let us stress that the growth conditions on the integrands and the nonlinear terms in the state equations, postulated in ([32, Chapter 4, Hypothesis 2, p. 130]) correspond to bounded controls. The same remark is valid for [12].

In [25], Fattorini and Sritharan prove a Pontryagin principle in nonqualified form for control problems of Navier–Stokes equations in which the controls are not necessarily bounded. Their idea is to work with bounded perturbations (see [25, p. 227]).

Here we consider a control set of the form

$$V_{ad} = \{v \in L^\sigma(\Sigma) \mid v(s,t) \in K_V(s,t) \text{ a.e. in, } \Sigma\},$$

where K_V is a measurable multimapping with nonempty and closed values in $\mathcal{P}(\mathbb{R})$ (see section 2). We do not think that the method developed in [25] can be applied to such a control set. Moreover, the method developed in [25] deals with Pontryagin principles in nonqualified form and requires some convexity condition on the cost functional (see Hypothesis 2.9 in [25]).

The purpose of this paper is to extend the method based on Ekeland's principle to problems with unbounded controls. In order to explain the main ideas of this extension let us recall the starting point of the method described above. If $\bar{v}$ is an ε^2-solution of the problem

$$(P_\varepsilon) \qquad \inf\{F_\varepsilon(v) \mid v \in V_{ad}\},$$

where F_ε and (V_{ad}, d_E) satisfy the assumptions of Ekeland's principle, then there exists another ε^2-solution v_ε such that

$$d_E(v_\varepsilon, \bar{v}) \le \varepsilon \quad \text{and} \quad F_\varepsilon(v_\varepsilon) - F_\varepsilon(v) \le \varepsilon d(v, v_\varepsilon) \quad \text{for every } v \in V_{ad}.$$

In order to exploit this optimality condition, v is replaced by some perturbation of v_ε. The methods developed in [8], [12], [21], [24], [26], [31], and [32] differ both in their choices of F_ε and in their choices of the perturbations. Pontryagin principles in qualified form are only obtained in [8] and [12] by choosing for F_ε a regularization of an exact penalized functional. A Pontryagin principle is then obtained under a strong stability condition. Pontryagin principles in nonqualified form are obtained in [8] under a weak stability condition by a method of spike perturbations. With another choice for the penalized functional and other kinds of perturbations, Pontryagin principles in nonqualified form are obtained in [24], [26], [31], and [32]. In [24] Fattorini and Murphy use a method of multispike perturbations. The type of perturbations used in [31], [40], [26], [12], [13], and [32] can be viewed as a generalization of multispike perturbations, which we call diffuse perturbations.

In contrast to spike or multispike perturbations, which are precisely localized, a diffuse perturbation is not localized around some points but is implicitly defined by some relations (see [12], [31], [40], [42], and [26]). The existence of diffuse perturbations satisfying relations a priori defined is proved in [29], [30], [26], and in [12] in a constructive manner. To our knowledge this kind of perturbation has been introduced for the first time by Yao [40] and Li [28]. We prove here that all the relations needed to define a diffuse perturbation can be obtained as a consequence of the Lyapunov convexity theorem. Connections with Lyapunov's convexity theorem or with Uhl's theorem are clarified in [42, p. 1315], and [30]. (See also [22] for another process.)

Preliminary results related to this topic were announced in [36]. The metric space used in [36] is different from the one defined in section 3.2. This is the reason why a convexity condition (assumption (A7)) is needed in [36] to ensure some semicontinuity property. Thus, the methods of the present paper improve upon those of [36].

The paper is organized as follows. In the next section we formulate the control problem governed by a semilinear parabolic equation and state the main results: the weak and strong Pontryagin's principles. In section 3 we give some regularity results for solutions of the state and adjoint equations. In section 4, we derive some technical results used in section 5 to prove the main result stated in section 2.

2. Assumptions and main results. Throughout the paper Ω is a bounded open subset of $\mathbb{R}^N$ ($N \geq 2$) of class $C^{2,\beta}$ for some $0 < \beta \leq 1$ (that is, the boundary Γ of Ω is an $(N-1)$-dimensional manifold of class $C^{2,\beta}$ such that Ω lies locally on one side of Γ. A function is of class $C^{2,\beta}$ if it is of class C^2 and if its second order derivatives are Hölder continuous of exponent β). We denote by q, σ positive numbers satisfying

$$q > N/2 + 1, \quad \sigma > N + 1 \quad \text{and} \quad q\sigma + q > qN + 2\sigma.$$

The differential operator A in equation (1.1) is defined by

$$Ay(x) = -\sum_{i,j=1}^{N} D_i(a_{ij}(x)D_j y(x)),$$

with coefficients a_{ij} belonging to $C^{1,\beta}(\overline{\Omega})$ and satisfying the conditions

$$(2.1)\quad a_{ij}(x) = a_{ji}(x) \qquad \text{for every } i,j \in \{1,...,N\}, \qquad m_0|\xi|^2 \leq \sum_{i,j=1}^{N} a_{ij}(x)\xi_j\xi_i$$

for all $x \in \overline{\Omega}$ and all $\xi \in \mathbb{R}^N$, with $0 < m_0$ (D_i denotes the partial derivative with respect to x_i). In (1.1), $\frac{\partial y}{\partial n_A}$ is the conormal derivative of y with respect to A, that is,

$$\frac{\partial y}{\partial n_A}(s,t) = \sum_{i,j} a_{ij}(s)D_j y(s,t)n_i(s),$$

where $n = (n_1, \ldots, n_N)$ is the unit normal to Γ outward Ω.

For all $1 \leq \tau \leq \infty$, the norms in the spaces $L^\tau(\Omega)$, $L^\tau(\Gamma)$, $L^\tau(Q)$, $L^\tau(\Sigma)$ will be denoted by $\|\cdot\|_{\tau,\Omega}$, $\|\cdot\|_{\tau,\Gamma}$, $\|\cdot\|_{\tau,Q}$, $\|\cdot\|_{\tau,\Sigma}$. The Hilbert space $W(0,T;H^1(\Omega),(H^1(\Omega))') = \{y \in L^2(0,T;H^1(\Omega)) \mid \frac{dy}{dt} \in L^2(0,T;(H^1(\Omega))')\}$, endowed with its usual norm, will be denoted by $W(0,T)$. Also set $\overline{\Omega}_0 = \overline{\Omega} \times \{0\}$ and $\overline{\Omega}_T = \overline{\Omega} \times \{T\}$.

2.1. Assumptions.

(A1) For every $y \in \mathbb{R}$, $f(\cdot,y)$ is measurable on Q. For almost every $(x,t) \in Q$, $f(x,t,\cdot)$ is of class C^1 on $\mathbb{R}$. The following estimates hold:

$$|f(x,t,0)| \leq M_1(x,t), \qquad C_0 \leq f'_y(x,t,y) \leq M_1(x,t)\eta(|y|),$$

where M_1 belongs to $L^q(Q)$, η is a nondecreasing function from $\mathbb{R}^+$ to $\mathbb{R}^+$, and $C_0 \in \mathbb{R}$. (We have denoted by f'_y the partial derivative of f with respect to y, and in the following we adopt the same kind of notation for other functions.)

(A2) For every $(y,v) \in \mathbb{R}^2$, $g(\cdot,y,v)$ is measurable on Σ. For almost every $(s,t) \in \Sigma$ and every $v \in \mathbb{R}$, $g(s,t,\cdot,v)$ is of class C^1 on $\mathbb{R}$. For almost every $(s,t) \in \Sigma$, $g(s,t,\cdot)$ and $g'_y(s,t,\cdot)$ are continuous on $\mathbb{R} \times \mathbb{R}$. The following estimates hold:

$$|g(s,t,0,v)| \leq M_2(s,t) + m_1|v|, \qquad C_0 \leq g'_y(s,t,y,v) \leq (M_2(s,t) + m_1|v|)\eta(|y|),$$

where M_2 belongs to $L^\sigma(\Sigma)$, $m_1 > 0$, and C_0 and η are as in (A1).

(A3) For every $(y,w) \in \mathbb{R}^2$, $L(\cdot,y,w)$ is measurable on Ω. For almost every $x \in \Omega$, $L(x,\cdot)$ is of class C^1 on $\mathbb{R} \times \mathbb{R}$. The following estimate holds:

$$|L(x,y,w)| + |L'_w(x,y,w)| + |L'_y(x,y,w)| \leq M_3(x)\eta(|w|)\eta(|y|),$$

where $M_3 \in L^1(\Omega)$, and η is as in (A1).

(A4) For every $y \in \mathbb{R}$, $F(\cdot, y)$ is measurable on Q. For almost every $(x,t) \in Q$, $F(x,t,\cdot)$ is of class C^1 on $\mathbb{R}$. The following estimate holds:

$$|F(x,t,y)| + |F'_y(x,t,y)| \le M_4(x,t)\eta(|y|),$$

where $M_4 \in L^1(Q)$, and η is as in (A1).

(A5) For every $(y,v) \in \mathbb{R}^2$, $G(\cdot,y,v)$ is measurable on Σ. For almost every $(s,t) \in \Sigma$ and every $v \in \mathbb{R}$, $G(s,t,\cdot,v)$ is of class C^1 on $\mathbb{R}$. For almost every $(s,t) \in \Sigma$, $G(s,t,\cdot)$ and $G'_y(s,t,\cdot)$ are continuous on $\mathbb{R} \times \mathbb{R}$. The following estimate holds:

$$|G(s,t,y,v)| + |G'_y(s,t,y,v)| \le (M_5(s,t) + m_1|v|^\sigma)\eta(|y|),$$

where $M_5 \in L^1(\Sigma)$, and m_1 and η are as in (A2).

(A6) The set of constraints on v is defined by

$$V_{ad} = \{v \in L^\sigma(\Sigma) \mid v(s,t) \in K_V(s,t) \text{ for a.e. } (s,t) \in \Sigma\},$$

where K_V is a measurable multimapping with nonempty and closed values in $\mathcal{P}(\mathbb{R})$ (that is, the set of all subsets of $\mathbb{R}$). The constraint on the initial condition is $w \in W_{ad}$, where W_{ad} is a closed convex subset of $C(\overline{\Omega})$.

(A7) In the state constraint (1.2), ϕ is a mapping of class C^1 from $C(\overline{Q})$ into $C(\overline{D})$, $\overline{D}$ is a nonempty compact subset of $\overline{Q}$, and $\mathcal{C} \subset C(\overline{D})$ is a closed convex subset with nonempty interior in $C(\overline{D})$.

The assumption "$\mathcal{C}$ has a nonempty interior in $C(\overline{D})$" is used, in nonqualified form of the Pontryagin's principle, to prove that the pair of multipliers is nonzero (see section 5.3). In section 3.1 we recall an existence and uniqueness result in $W(0,T) \cap C(\overline{Q})$ for (1.1), already proved in [37]. Therefore, the state constraint (1.2) makes sense because the weak solution of (1.1) is continuous on $\overline{Q}$. Let us give some examples of state constraints described by (1.2).

Example 2.1. If we choose $\overline{D} = \overline{\Omega} \times \{T\}$, we have a problem with a terminal state constraint. We may consider $\phi(y) = y|_{\overline{D}}$ ($y|_{\overline{D}}$ is the restriction of y to $\overline{D}$) and $\mathcal{C} = \{z \in C(\overline{D}) \mid ||z - y_T||_{C(\overline{D})} \le \epsilon\}$, where $\epsilon > 0$ and $y_T \in C(\overline{D})$ are given.

Example 2.2. We consider $\phi(y) = \psi(\cdot, y(\cdot))|_{\overline{D}}$, where $\psi \in C(\overline{Q} \times \mathbb{R})$ is such that ψ'_y (the partial derivative of ψ with respect to y) belongs to $C(\overline{Q} \times \mathbb{R})$, $\mathcal{C} = \{z \in C(\overline{D}) \mid z \le 0\}$, and $\overline{D}$ is any nonempty compact subset of $\overline{Q}$.

2.2. Strong stability assumption. For $\gamma \ge 0$, set

$$\mathcal{C}_\gamma = \left\{ \varphi \in C(\overline{D}) \mid \inf_{z \in \mathcal{C}} \|\varphi - z\|_{C(\overline{D})} \le \gamma \right\},$$

and consider the perturbed state constraint

$$\phi(y) \in \mathcal{C}_\gamma. \tag{2.2}$$

We denote by (P_γ) the problem

$$(P_\gamma) \quad \inf \left\{ J(y,v,w) \mid (y,v,w) \in C(\overline{Q}) \times V_{ad} \times W_{ad},\ (y,v,w) \text{ satisfies (1.1), (2.2)} \right\}.$$

Observe that (P) is identical to (P_0). Following [8], [12], and [13], we say that (P_γ) is strongly stable on the right if there exist $\tilde{\varepsilon} > 0$ and $\tilde{r} > 0$ such that, for every $\gamma' \in [\gamma, \gamma + \tilde{\varepsilon}]$, we have

$$\inf(P_\gamma) - \inf(P_{\gamma'}) \le \tilde{r}(\gamma - \gamma').$$

With the additional assumption (A8), a Pontryagin principle for (P) may be obtained in qualified form. Some remarks on (A8) are made after Theorem 2.1.

(A8) (P) is strongly stable on the right.

2.3. Statement of the main result. We define the boundary Hamiltonian function by

$$H_\Sigma(s,t,y,v,p,\nu) = \nu G(s,t,y,v) - pg(s,t,y,v)$$

for every $(s,t,y,v,p,\nu) \in \Gamma \times [0,T] \times \mathbb{R}^4$. The main result of this paper is the Pontryagin principle for (P), stated in the following theorem.

THEOREM 2.1. *If* (A1)–(A7) *are fulfilled and if* $(\bar y, \bar v, \bar w)$ *is a solution of* (P)*, then there exist* $\bar p \in L^1(0,T;W^{1,1}(\Omega))$*,* $\bar\nu \in \mathbb{R}$*,* $\bar\mu \in \mathcal{M}(\overline{D})$ *(the space of Radon measures on* $\overline{D}$*) and a measurable subset* $\widetilde{\Sigma} \subset \Sigma$ *such that*

$$(\bar\nu,\bar\mu) \neq 0, \quad \bar\nu \geq 0, \quad \langle \bar\mu, z - \phi(\bar y)\rangle_{\mathcal{M}(\overline{D})\times C(\overline{D})} \leq 0 \qquad \text{for all } z \in \mathcal{C}, \tag{2.3}$$

$$\begin{cases} -\dfrac{\partial \bar p}{\partial t} + A\bar p + f'_y(x,t,\bar y)\bar p = \bar\nu F'_y(x,t,\bar y) + [\phi'(\bar y)^*\bar\mu]|_Q & \text{in } Q, \\ \dfrac{\partial \bar p}{\partial n_A} + g'_y(s,t,\bar y,\bar v)\bar p = \bar\nu G'_y(s,t,\bar y,\bar v) + [\phi'(\bar y)^*\bar\mu]|_\Sigma & \text{on } \Sigma, \\ \bar p(T) = \bar\nu L'_y(x,\bar y(T),\bar w) + [\phi'(\bar y)^*\bar\mu]|_{\overline{\Omega}_T} & \text{in } \Omega, \end{cases} \tag{2.4}$$

$$H_\Sigma(s,t,\bar y(s,t),\bar v(s,t),\bar p(s,t),\bar\nu) = \min_{v\in K_V(s,t)} H_\Sigma(s,t,\bar y(s,t),v,\bar p(s,t),\bar\nu) \tag{2.5}$$

for all $(s,t) \in \widetilde{\Sigma}$*, with* $\mathcal{L}^N(\widetilde{\Sigma}) = \mathcal{L}^N(\Sigma)$,

$$\int_\Omega \bar\nu L'_w(x,\bar y(T),\bar w)(\bar w - w)\,dx + \langle \bar p(0) + [\phi'(\bar y)^*\bar\mu]|_{\overline{\Omega}_0}, \bar w - w\rangle_{\mathcal{M}(\overline{\Omega})\times C(\overline{\Omega})} \leq 0 \tag{2.6}$$

for all $w \in W_{ad}$*, where* $[\phi'(\bar y)^*\bar\mu]|_Q$ *is the restriction of* $[\phi'(\bar y)^*\bar\mu]$ *to* Q*,* $[\phi'(\bar y)^*\bar\mu]|_\Sigma$ *is the restriction of* $[\phi'(\bar y)^*\bar\mu]$ *to* Σ*,* $[\phi'(\bar y)^*\bar\mu]|_{\overline{\Omega}_T}$ *is the restriction of* $[\phi'(\bar y)^*\bar\mu]$ *to* $\overline{\Omega}_T$*, and* $[\phi'(\bar y)^*\bar\mu]|_{\overline{\Omega}_0}$ *is the restriction of* $[\phi'(\bar y)^*\bar\mu]$ *to* $\overline{\Omega}_0$*,* ($[\phi'(\bar y)^*\bar\mu]$ *is the Radon measure on* $\overline{Q}$ *defined by* $z \longmapsto \langle \bar\mu, \phi'(\bar y)z\rangle_{\mathcal{M}(\overline{D})\times C(\overline{D})}$ *for* $z \in C(\overline{Q})$ *and* $\mathcal{L}^N$ *denotes the* N*-dimensional Lebesgue measure*)*. Moreover, if* (A8) *is satisfied, we can take* $\bar\nu = 1$ *in* (2.4)*,* (2.5)*, and* (2.6)*.*

The meaning of weak solutions for (2.4), regularity results for $\bar p$, and the definition of $\bar p(0)$ are given in section 3.

The notion of stability considered in (A8) is closely related to the notion of calmness introduced by Clarke [16]. In the above setting this notion is due to Burke [9]. It has been used in control problems by Bonnans and Casas [8]. We do not know sufficient conditions ensuring that a state-constrained control problem is strongly stable on the right. However, even if (P) is not strongly stable on the right, (P_γ) will be strongly stable for all $\gamma > 0$, except on a subset of $\mathbb{R}^+$ of zero Lebesgue measure [8], [12]. In some situations, Pontryagin's principles in qualified form may be derived from a nonqualified form. Consider the example described below.

Example 2.3. Suppose that Ω is connected and consider the state equation

$$\frac{\partial y}{\partial t} - \Delta y = 0 \text{ in } Q, \quad \frac{\partial y}{\partial n} + y^4 = v \text{ on } \Sigma, \quad y(0) = y_0 \text{ on } \Omega, \tag{2.7}$$

where y_0 is a given function in $C(\overline{\Omega})$. We set $V_{ad} = \{v \in L^\sigma(\Sigma) \mid v(s,t) \geq 0 \text{ a.e. on } \Sigma\}$. The state constraints are defined by

$$0 \leq y(x,t) \leq \gamma_d \quad \text{on } \overline{Q},$$

for some given $\gamma_d > 0$. (We suppose that $0 \leq y_0(x) < \gamma_d$ on $\overline{\Omega}$.) Since for $v \in V_{ad}$, the solution y of (2.7) is nonnegative, then we can restrict the state constraints to $y(x,t) \leq \gamma_d$. We denote by J a cost functional defined as in (1.3), with $w \equiv y_0$, for which assumptions (A3)–(A5) are satisfied. We suppose that the control problem

$$(P_{ex}) \qquad \inf\{J(y,v) \mid y \in C(\overline{Q}),\ v \in V_{ad},\ (y,v) \text{ satisfies (2.7)},\ y(x,t) \leq \gamma_d \text{ on } \overline{Q}\}$$

admits solutions. We wish to prove that every solution of (P_{ex}) satisfies the Pontryagin principle in qualified form. For this we suppose that $(\bar{y}, \bar{v})$ is a solution which satisfies the Pontryagin principle in nonqualified form. For (P_{ex}), the adjoint equation (2.4) corresponding to $(\bar{y}, \bar{v})$ and $\bar{\nu} = 0$ is

$$-\frac{\partial \bar{p}}{\partial t} - \Delta \bar{p} = \bar{\mu}_Q \text{ in } Q, \quad \frac{\partial \bar{p}}{\partial n} + 4\bar{y}^3 \bar{p} = \bar{\mu}_\Sigma \text{ on } \Sigma, \quad \bar{p}(T) = \bar{\mu}_{\overline{\Omega}_T} \text{ on } \overline{\Omega},$$

where the measure $\bar{\mu} = \bar{\mu}_Q + \bar{\mu}_\Sigma + \bar{\mu}_{\overline{\Omega}_T}$ satisfies

$$\bar{\mu} \geq 0, \quad \bar{\mu} \neq 0, \quad \langle \bar{\mu}, \bar{y} - \gamma_d \rangle_{\mathcal{M}(\overline{Q}) \times C(\overline{Q})} = 0. \tag{2.8}$$

(Observe that, for (P_{ex}), (2.8) corresponds to (2.3) with $\bar{\nu} = 0$.) The Pontryagin principle in nonqualified form is expressed as

$$\bar{p}(s,t)(v - \bar{v}(s,t)) \geq 0 \quad \text{for all } v \geq 0 \quad \text{and for almost all } (s,t) \in \Sigma. \tag{2.9}$$

We set $\bar{t} = \inf\{t \in [0,T] \mid \bar{\mu}(\overline{\Omega} \times]t,T]) = 0\}$. Following [34, Theorems 4.2, 4.3, and Remark 4.7], we can define $\bar{p}_{\overline{\Omega}}(t^+)$ as the function in $L^1(\Omega)$ which satisfies the Green formula

$$\int_{\Omega \times]t,T[} \bar{p} \left(\frac{\partial y}{\partial t} - \Delta y \right) dx\, dt + \int_{\Gamma \times]t,T[} \bar{p} \left(\frac{\partial y}{\partial n} + 4\bar{y}^3 y \right) ds\, dt$$

$$= \langle \bar{\mu}_{\Omega \times]t,T[}, y \rangle_{\mathcal{M}_b(\Omega \times]t,T[) \times C_b(\Omega \times]t,T[)} + \langle \bar{\mu}_{\Gamma \times]t,T[}, y \rangle_{\mathcal{M}_b(\Gamma \times]t,T[) \times C_b(\Gamma \times]t,T[)}$$

$$+ \langle p(T), y(T) \rangle_{\mathcal{M}(\overline{\Omega}) \times C(\overline{\Omega})} - \langle p_{\overline{\Omega}}(t^+), y(t) \rangle_{\mathcal{M}(\overline{\Omega}) \times C(\overline{\Omega})},$$

for all $y \in C^2(\overline{Q})$ (for $A \subset \overline{Q}$, $\bar{\mu}_A$ denotes the restriction of $\bar{\mu}$ to A). With such a definition for $\bar{p}_{\overline{\Omega}}(t^+)$, the restriction of $\bar{p}$ to $\overline{\Omega} \times]0,t[$ is the unique solution of

$$-\frac{\partial p}{\partial t} - \Delta p = \bar{\mu}_{\Omega \times]0,t[} \text{ in } \Omega \times]0,t[, \quad \frac{\partial p}{\partial n} + 4\bar{y}^3 p = \bar{\mu}_{\Gamma \times]0,t[} \text{ on } \Gamma \times]0,t[,$$
$$p(t) = \bar{p}_{\overline{\Omega}}(t^+) + \bar{\mu}_{\overline{\Omega} \times \{t\}} \text{ on } \overline{\Omega}.$$

We can easily prove that if $\bar{p}_{\overline{\Omega}}(t^+) = 0$, then $\bar{\mu}_{\overline{\Omega}\times]t,T]} = 0$. From the definition of $\bar{t}$ we see that, for every $\varepsilon > 0$, $\bar{p}_{\overline{\Omega}}((\bar{t}-\varepsilon)^+) \geq 0$ and $\bar{p}_{\overline{\Omega}}((\bar{t}-\varepsilon)^+) \neq 0$. Let $\hat{p}$ be the solution of

$$-\frac{\partial p}{\partial t} - \Delta p = 0 \text{ in } \Omega\times]0,\bar{t}-\varepsilon[, \quad \frac{\partial p}{\partial n} + kp = 0 \text{ on } \Gamma\times]0,\bar{t}-\varepsilon[, \quad p(t) = \bar{p}_{\overline{\Omega}}((\bar{t}-\varepsilon)^+) \text{ on } \overline{\Omega},$$

where $k = \max\{4\bar{y}^3(x,t) \mid (x,t) \in \overline{Q}\}$. By a comparison principle, we can verify that $\bar{p} \geq \hat{p}$ on $\overline{\Omega}\times]0,\bar{t}-\varepsilon[$ and in particular on $\Gamma\times]0,\bar{t}-\varepsilon[$. Moreover, $\hat{p}$ belongs to $C^2(\overline{Q}\times[0,\bar{t}-\varepsilon[)$, and the function $\hat{p}(\bar{t}-2\varepsilon)$ belongs to $C^2(\overline{\Omega})$, is nonnegative, and not identically zero. Therefore, from the maximum principle for classical solutions of parabolic equations, we deduce that $\hat{p} > 0$ on $\overline{\Omega}\times[0,\bar{t}-2\varepsilon[$. Thus, $\bar{p}(s,t) > 0$ a.e. on $\Gamma\times]0,\bar{t}[$. With (2.9), this implies $\bar{v} \equiv 0$ on $\Gamma\times]0,\bar{t}[$. Thus, $0 \leq \bar{y}(x,t) \leq \max_{\overline{\Omega}} y_0 < \gamma_d$ on $\overline{\Omega}\times[0,\bar{t}]$. From (2.8), it follows $\bar{\mu}_{\overline{\Omega}\times[0,\bar{t}]} = 0$. Thus, $\bar{\mu} \equiv 0$ and we get a contradiction.

In this simple example we see that the Pontryagin principle in qualified form follows from the Pontryagin principle in nonqualified form.

3. State equation and adjoint equation.

3.1. Existence, uniqueness and regularity of the state variable.

THEOREM 3.1. *Under assumptions* (A1) *and* (A2), *if* $v \in L^\sigma(\Sigma)$ *and* $w \in C(\overline{\Omega})$, *then* (1.1) *admits a unique weak solution* y_{vw} *in* $W(0,T)\cap C(\overline{Q})$. *This solution satisfies*

$$\|y_{vw}\|_{\infty,Q} \leq C_1(\|v\|_{\sigma,\Sigma} + \|w\|_{\infty,\Omega} + 1),$$

where $C_1 = C_1(T,\Omega,N,q,\sigma,C_0)$. *Moreover, the mapping* $(v,w) \longmapsto y_{vw}$ *is continuous from* $L^\sigma(\Sigma)\times C(\overline{\Omega})$ *into* $C(\overline{Q})$.

Proof. The existence of a unique weak solution y_{vw} in $W(0,T)\cap C(\overline{Q})$ for equation (1.1), is proved in ([37, Theorem 3.1]). The last part of the theorem can be proved as in ([37], Proposition 4.3). □

COROLLARY 3.2. *For every* $k > 0$ *and every* $\varepsilon > 0$, *there exist* $C_2 = C_2(T,\Omega,N,q,\sigma,C_0,k)$, $C_3 = C_3(T,\Omega,N,q,\sigma,C_0,k,\varepsilon)$, *and* $\alpha > 0$ *such that, for every* $(v,w) \in V_{ad}\times W_{ad}$ *satisfying* $\|v\|_{L^\sigma(\Sigma)} + \|w\|_{\infty,\Omega} \leq k$, *the weak solution* y_{vw} *of* (1.1) *corresponding to* (v,w) *is Hölder continuous on* $[\varepsilon,T]\times\overline{\Omega}$ *and obeys*

$$\|y_{vw}\|_{C(\overline{Q})} \leq C_2, \qquad \|y_{vw}\|_{C^{\alpha,\frac{\alpha}{2}}(\overline{\Omega}\times[\varepsilon,T])} \leq C_3.$$

Moreover, if w *is Hölder continuous on* $\overline{\Omega}$, *then* y_{vw} *is Hölder continuous on* $\overline{Q}$.

Proof. Since y_{vw} belongs to $C(\overline{Q})$, thanks to (A1)–(A2), we see that y_{vw} is also the unique weak solution of

$$\frac{\partial y}{\partial t} + Ay = \tilde{f} \text{ in } Q, \quad \frac{\partial y}{\partial n_A} = \tilde{g} \text{ on } \Sigma, \quad y(0) = w \text{ in } \Omega,$$

where

$$\tilde{f}(\cdot) = -f(\cdot, y_{vw}(\cdot)) \in L^q(Q), \qquad \tilde{g}(\cdot) = -g(\cdot, y_{vw}(\cdot), v(\cdot)) \in L^\sigma(\Sigma).$$

We denote by γ_0 the trace operator from $L^{\sigma'}(0,T;W^{1,\nu}(\Omega))$ into $L^{\sigma'}(0,T;W^{1-\frac{1}{\nu},\nu}(\Gamma))$ (with $\nu = N/(\sigma N - N + 1)$), and by i the embedding from $L^{\sigma'}(0,T;W^{1-\frac{1}{\nu},\nu}(\Gamma))$ into $L^{\sigma'}(0,T;L^{\sigma'}(\Gamma))$, we can write

$$\langle \tilde{g}, (i\circ\gamma_0)z\rangle_{L^\sigma(\Sigma)\times L^{\sigma'}(\Sigma)} = \langle (\gamma_0^t \circ i^t)\tilde{g}, z\rangle_{L^\sigma(0,T;(W^{1,\nu}(\Omega))')\times L^{\sigma'}(0,T;W^{1,\nu}(\Omega))}$$

for every $z \in L^{\sigma'}(0,T;W^{1,\nu}(\Omega))$. We can identify $(\gamma_0^t \circ i^t)\tilde{g}$ with $(f_0, f_1, \ldots, f_N) \in (L^\sigma(0,T;L^{\nu'}(\Omega)))^{N+1}$ $(\nu' = \frac{N\sigma}{N-1})$ in the following manner:

$$\langle (\gamma_0^t \circ i^t)\tilde{g}, z\rangle_{L^\sigma(0,T;(W^{1,\nu}(\Omega))')\times L^{\sigma'}(0,T;W^{1,\nu}(\Omega))} = \int_Q (f_0 z + \Sigma_i f_i D_i z)\, dxdt$$

for every $z \in L^{\sigma'}(0,T;W^{1,\nu}(\Omega))$. Therefore, y_{vw} is the weak solution (in the sense of [17]) of the initial boundary value problem

$$\frac{\partial y}{\partial t} - \operatorname{div}(a(x,t,\nabla y)) = \tilde{f} + f_0 \text{ in } Q, \quad a(x,t,\nabla y).n = 0 \text{ on } \Sigma, \quad y(0) = w \text{ in } \Omega, \tag{3.1}$$

where

$$a(x,t,\nabla y) = \left(\sum_{j=1}^N a_{ij} D_j y - f_i\right)_{i=1,..,N}.$$

Now we can easily verify that assumptions of ([17, Chapter 3, Theorem 1.3]) are satisfied by system (3.1), and the Hölder continuity results of corollary 3.2 follow from this theorem. □

3.2. Metric space of controls. To apply the Ekeland variational principle, we have to define a metric space of controls in order that the mapping $(v,w) \longmapsto y_{vw}$ be continuous from this metric space to $C(\overline{Q})$. Thanks to Theorem 3.1, this continuity condition will be realized if convergence in the metric space of controls implies convergence in $L^\sigma(\Sigma) \times C(\overline{\Omega})$. In the case where boundary controls are bounded, convergence in (V_{ad}, d_E) (where d_E is the so-called Ekeland distance) implies convergence in $L^\sigma(\Sigma)$. This condition is no longer true for unbounded controls (see [25, p. 227]). To overcome this difficulty, we define a new metric space in the following way.

Let $\tilde{v}$ be in V_{ad} (in section 5, $\tilde{v}$ will be an optimal boundary control that we want to characterize). For $0 < k < \infty$, we define the set

$$V_{ad}(\tilde{v},k) = \{v \in V_{ad} \mid |v(s,t) - \tilde{v}(s,t)| \le k \text{ for a.e. } (s,t) \in \Sigma\}.$$

We endow the set $V_{ad}(\tilde{v},k) \times W_{ad}$ with the following metric:

$$d((v_1,w_1),(v_2,w_2)) = \mathcal{L}^N(\{(s,t) \mid v_1(s,t) \neq v_2(s,t)\}) + \|w_1 - w_2\|_{\infty,\Omega}.$$

Remark 3.1. From [19], we know that the mapping

$$d_E : (v_1,v_2) \longmapsto \mathcal{L}^N(\{(s,t) \mid v_1(s,t) \neq v_2(s,t)\})$$

is a distance on $V_{ad}(\tilde{v},k)$. Moreover, if $(v_n)_n \subset V_{ad}(\tilde{v},k)$, if $v \in V_{ad}(\tilde{v},k)$, and if $\lim_n d_E(v_n,v) = 0$, then $(v_n)_n$ converges to v in $L^\sigma(\Sigma)$. This is no longer true for any sequence $(v_n)_n$ included in V_{ad}.

LEMMA 3.3. *$(V_{ad}(\tilde{v},k) \times W_{ad}, d)$ is a complete metric space, and the mapping which associates $(y_{vw}, J(y_{vw},v,w))$ with (v,w) is continuous from $(V_{ad}(\tilde{v},k) \times W_{ad}, d)$ into $C(\overline{Q}) \times \mathbb{R}$.*

Proof. (i) To prove that $(V_{ad}(\tilde{v},k) \times W_{ad}, d)$ is a complete metric space, it remains to prove that $(V_{ad}(\tilde{v},k), d_E)$ is complete. Let $(v_n)_n$ be a Cauchy sequence in $(V_{ad}(\tilde{v},k), d_E)$. Following [19], we can prove that $(v_n)_n$ converges for d_E to some

measurable function v such that $v(s,t) \in K_V(s,t)$ and $|v(s,t) - \tilde{v}(s,t)| \leq k$ for almost all $(s,t) \in \Sigma$. Therefore, $v \in L^\sigma(\Sigma)$ and $v \in V_{ad}(\tilde{v},k)$.

(ii) Now, we consider $(v_n, w_n)_{n\geq 1} \subset V_{ad}(\tilde{v},k) \times W_{ad}$ and $(v,w) \in V_{ad}(\tilde{v},k) \times W_{ad}$ such that $(v_n,w_n)_n$ converges to (v,w) for the metric d. We denote by y and y_n $(n \geq 1)$ the solution of (1.1) corresponding, respectively, to (v,w) and to (v_n,w_n). To prove the continuity result, it remains to prove that the sequence $(y_n, J(y_n,v_n,w_n))_n$ converges to $(y, J(y,v,w))$ in $C(\overline{Q}) \times \mathbb{R}$.

For this, we observe that $(w_n)_n$ converges to w in $C(\overline{\Omega})$ and $(v_n)_n$ converges to v in $L^\sigma(\Sigma)$. We complete the proof thanks to the continuity assumptions on F, G, L and to the continuity results stated in Theorem 3.1. □

3.3. Adjoint equation. Let (a,b) be in $L^q(Q) \times L^\sigma(\Sigma)$ with $a \geq C_0$ and $b \geq C_0$. We consider the following terminal boundary value problem:

$$-\frac{\partial p}{\partial t} + Ap + ap = \mu_Q \text{ in } Q, \quad \frac{\partial p}{\partial n_A} + bp = \mu_\Sigma \text{ on } \Sigma, \quad p(T) = \mu_{\overline{\Omega}_T} \text{ on } \overline{\Omega}, \tag{3.2}$$

where $\mu = \mu_Q + \mu_\Sigma + \mu_{\overline{\Omega}_T}$ is a bounded Radon measure on $\overline{Q} \setminus \overline{\Omega}_0$, μ_Q is the restriction of μ to Q, μ_Σ is the restriction of μ to Σ, and $\mu_{\overline{\Omega}_T}$ is the restriction of μ to $\overline{\Omega}_T$.

DEFINITION 3.4. *We shall say that p is a weak solution of* (3.2) *in* $L^1(0,T;W^{1,1}(\Omega))$ *if and only if the two following conditions are fulfilled:*

(i) $ap \in L^1(Q)$ *and* $bp \in L^1(\Sigma)$,

(ii) *For every* $\varphi \in C^1(\overline{Q})$ *satisfying* $\varphi(x,0) = 0$ *on* $\overline{\Omega}$, *we have*

$$\int_Q \left\{ p\frac{\partial \varphi}{\partial t} + \Sigma_{i,j} a_{ij} D_j \varphi D_i p + a\varphi p \right\} dxdt + \int_\Sigma b\varphi p \, dsdt = \langle \varphi, \mu \rangle_{C_b(\overline{Q}\setminus\overline{\Omega}_0) \times \mathcal{M}_b(\overline{Q}\setminus\overline{\Omega}_0)}.$$

($C_b(\overline{Q}\setminus\overline{\Omega}_0)$ *denotes the space of bounded continuous functions on* $\overline{Q}\setminus\overline{\Omega}_0$, *while* $\mathcal{M}_b(\overline{Q}\setminus\overline{\Omega}_0)$ *denotes the space of bounded Radon measures on* $\overline{Q} \setminus \overline{\Omega}_0$, *that is, the topological dual of* $C_0(\overline{Q} \setminus \overline{\Omega}_0)$.)

In the following, we shall say that a pair $(\delta, d) \in \mathbb{R}^2$ fulfills the condition $(C_{q\sigma})$ if and only if

$$(C_{q\sigma}) \quad \begin{cases} \frac{N\sigma}{\sigma-2} < d \leq \frac{N\sigma}{N-1} \quad \text{and} \quad \frac{2d}{d-N} < \delta \leq \sigma & \text{if } \sigma \leq q, \\ \frac{Nq}{q-2} < d \leq \frac{N\sigma}{N-1} \quad \text{and} \quad \frac{2d}{d-N} < \delta \leq q & \text{if } N \leq q < \sigma, \\ \frac{Nq}{q-2} < d \leq \inf(\frac{N\sigma}{N-1}, \frac{Nq}{N-q}) \quad \text{and} \quad \frac{2d}{d-N} < \delta \leq q & \text{if } q < N. \end{cases}$$

Since $q > \frac{N}{2} + 1$, $\sigma > N+1$, and $q\sigma + q > qN + 2\sigma$, we notice that the set of pairs (δ, d) satisfying $(C_{q\sigma})$ is nonempty. These conditions appear in a natural manner when we study equation (3.2) (see Remark 3.3). We now recall an existence theorem for parabolic equations with measures as data stated in [34].

THEOREM 3.5. *Let* (a,b) *be in* $L^q(Q) \times L^\sigma(\Sigma)$ *satisfying* $a \geq C_0$, $b \geq C_0$ *and let* μ *be in* $\mathcal{M}_b(\overline{Q}\setminus\overline{\Omega}_0)$. *Equation* (3.2) *admits a unique weak solution* $p \in L^1(0,T;W^{1,1}(\Omega))$. *For every* (δ,d) *satisfying* $(C_{q\sigma})$, p *belongs to* $L^{\delta'}(0,T;W^{1,d'}(\Omega))$ *and we have*

$$\|p\|_{L^{\delta'}(0,T;W^{1,d'}(\Omega))} \leq C_4 \|\mu\|_{\mathcal{M}_b(\overline{Q}\setminus\overline{\Omega}_0)},$$

where $C_4 = C_4(T,\Omega,N,\delta,d,C_0)$ *is independent of* a *and* b. *Moreover, there exists a Radon measure on* $\overline{\Omega}$, *denoted by* $p(0)$ *such that*

$$\int_Q p\left\{\frac{\partial y}{\partial t}+Ay+ay\right\}dxdt+\int_\Sigma p\left\{\frac{\partial y}{\partial n_A}+by\right\}dsdt$$

$$=\langle y,\mu\rangle_{C_b(\overline{Q}\setminus\overline{\Omega}_0)\times\mathcal{M}_b(\overline{Q}\setminus\overline{\Omega}_0)}-\langle y(0),p(0)\rangle_{C(\overline{\Omega})\times\mathcal{M}(\overline{\Omega})}$$

for every $y\in Y=\{y\in W(0,T)\cap C(\overline{Q})\mid \frac{\partial y}{\partial t}+Ay\in L^q(Q),\ \frac{\partial y}{\partial n_A}\in L^\sigma(\Sigma)\}$.

Remark 3.2. If $p\in L^{\delta'}(0,T;W^{1,d'}(\Omega))$ (where (δ,d) satisfies $(C_{q\sigma})$), and if

$$\mathrm{div}_{xt}\left((\Sigma_j a_{ij}D_jp)_{1\le i\le N},p\right)=\frac{\partial p}{\partial t}-Ap\quad\text{belongs to }\mathcal{M}_b(Q),$$

then we can define the normal trace of the vector field $((\sum_j a_{ij}D_jp)_{1\le i\le N},p)$ in the space $W^{\frac{-1}{m},m}(\partial Q)$ (for some $1<m<\frac{N+1}{N}$). If we denote by $\gamma_n((\sum_j a_{ij}D_jp)_{1\le i\le N},p)$ this normal trace, we can prove (see Theorem 4.2 in [34]) that this normal trace belongs to $\mathcal{M}(\partial Q)$ and the restriction of $\gamma_n((\sum_j a_{ij}D_jp)_{1\le i\le N},p)$ to $\overline{\Omega}_T$ is equal to $\mu_{\overline{\Omega}_T}$, the restriction of $\gamma_n((\sum_j a_{ij}D_jp)_{1\le i\le N},p)$ to Σ is equal to $\mu_\Sigma-bp$, and if $p(0)$ is the measure on $\overline{\Omega}$ which satisfies the Green formula of Theorem 3.4, then $-p(0)$ is the restriction of $\gamma_n((\sum_j a_{ij}D_jp)_{1\le i\le N},p)$ to $\overline{\Omega}_0$. In fact it can be proved that $p(0)$ belongs to $L^1(\Omega)$ (see Theorem 4.3 in [34]).

Remark 3.3. Let us explain the origin of the condition $(C_{q\sigma})$. In [34], the existence of a weak solution in $L^1(0,T;W^{1,1}(\Omega))$ for equation (3.2) is proved by duality arguments and an approximation process. The condition $\delta>2d/(d-N)$ appears to get C^0-regularity results for some adjoint equation associated with (3.2) (see ([34, Theorem 4.1])). Condition $\delta>2d/(d-N)$, together with conditions $p\in L^{\delta'}(0,T,W^{1,d'}(\Omega))$, $a\in L^q(Q)$, $b\in L^\sigma(\Sigma)$, $ap\in L^1(Q)$, and $bp\in L^1(\Sigma)$ are equivalent to the condition "(δ,d) satisfies $(C_{q\sigma})$." In the case where $a\in L^\infty(Q)$ and $b\in L^\infty(\Sigma)$, condition $(C_{q\sigma})$ can be replaced by the only condition $\delta>2d/(d-N)$.

4. Existence of diffuse perturbations. In section 5, we consider control problems in which the state constraints are penalized. The penalization is chosen in such a way that the solution of (P) that we want to characterize will be an ε-solution of the penalized problem. In order to exploit optimality conditions deduced from Ekeland's variational principle, we need to construct admissible perturbations of approximate optimal solutions. For this we use a kind of perturbation that we call "diffuse perturbation" and which goes back to Yao [40] and Li [28]. A diffuse perturbation of a control $\bar v\in V_{ad}$ is a function v_ρ defined by

$$v_\rho(s,t)=\begin{cases}\bar v(s,t) & \text{on }\Sigma\setminus E_\rho,\\ v(s,t) & \text{on }E_\rho,\end{cases}$$

where $v\in V_{ad}$ and E_ρ is some measurable subset of Σ. It is clear that $v_\rho\in V_{ad}$. Contrary to spike or multispike perturbations, where E_ρ is precisely defined, here E_ρ must satisfy some relations such as (4.9), (4.10), and (4.11). As explained in Lemma 4.2, the existence of E_ρ follows from the Lyapunov convexity theorem. To get optimality conditions, we need some differential calculus rules for this type of perturbation, stated in the following theorem.

THEOREM 4.1. *Let* ρ *be positive constant such that* $0<\rho<1$. *For every* $v_1,v_2\in V_{ad}$ *and for every* $w_1,w_2\in C(\overline{\Omega})$, *there exists a measurable subset* $E_\rho\subset\Sigma$ *such that*

$$\mathcal{L}^N(E_\rho)=\rho\mathcal{L}^N(\Sigma),\tag{4.1}$$

$$\int_{E_\rho} (G(\cdot, y_1, v_2) - G(\cdot, y_1, v_1))\, dsdt = \rho \int_\Sigma (G(\cdot, y_1, v_2) - G(\cdot, y_1, v_1))\, dsdt, \tag{4.2}$$

$$y_\rho = y_1 + \rho z + r_\rho, \qquad \textit{with} \quad \lim_{\rho \to 0} \frac{1}{\rho} \|r_\rho\|_{C(\overline{Q})} = 0, \tag{4.3}$$

$$J(y_\rho, v_\rho, w_\rho) = J(y_1, v_1, w_1) + \rho \Delta J + o(\rho), \tag{4.4}$$

where v_ρ, w_ρ are the controls defined by

$$v_\rho(s,t) = \begin{cases} v_1(s,t) & \text{on } \Sigma \setminus E_\rho, \\ v_2(s,t) & \text{on } E_\rho, \end{cases} \tag{4.5}$$

$$w_\rho = w_1 + \rho w_2, \tag{4.6}$$

y_ρ, y_1 are the solutions of (1.1) corresponding, respectively, to (v_ρ, w_ρ) and to (v_1, w_1), z is the weak solution of

$$\begin{cases} \dfrac{\partial z}{\partial t} + Az + f'_y(x,t,y_1)z = 0 & \text{in } Q, \\ \dfrac{\partial z}{\partial n_A} + g'_y(s,t,y_1,v_1)z = g(s,t,y_1,v_1) - g(s,t,y_1,v_2) & \text{on } \Sigma, \\ z(0) = w_2 & \text{in } \Omega, \end{cases} \tag{4.7}$$

and

$$\Delta J = J'_y(y_1, v_1, w_1)z + J(y_1, v_2, w_1) - J(y_1, v_1, w_1) + \int_\Omega L'_w(x, y_1(T), w_1) w_2\, dx. \tag{4.8}$$

The proof relies on the following lemma.

LEMMA 4.2. *Let v_1, v_2 be in V_{ad} and let y be in $C(\overline{Q})$. For every $\rho \in]0,1[$, there exists a sequence of measurable subsets $(E^n_\rho)_n$ in Σ such that*

$$\mathcal{L}^N(E^n_\rho) = \rho \mathcal{L}^N(\Sigma), \tag{4.9}$$

$$\int_{E^n_\rho} (G(s,t,y,v_1) - G(s,t,y,v_2))\, dsdt = \rho \int_\Sigma (G(s,t,y,v_1) - G(s,t,y,v_2))\, dsdt, \tag{4.10}$$

$$\frac{1}{\rho} \chi_{E^n_\rho} \rightharpoonup 1 \text{ weak star in } L^\infty(\Sigma), \text{ when } n \to \infty, \tag{4.11}$$

where $\chi_{E^n_\rho}$ is the characteristic of E^n_ρ.

Remark 4.1. A statement similar to (4.1), (4.3), (4.4) is given in [30], [31], [26], and [12] (conditions (4.9), (4.11) are also stated in [12]). In [30] the proof relies on an extension of Uhl's theorem. The proofs in [26] and [12] are constructive. Since the existence of E^n_ρ, satisfying together conditions (4.9), (4.10), (4.11), is not proved, neither in [12] nor in [30], we here give a short proof of Lemma 4.2 based on the Lyapunov convexity theorem (see, for example, [14, Theorem 16.1.ii]).

Proof of Lemma 4.2. We consider a family $(\varphi_n)_n$ dense in $L^1(\Sigma)$. For $n \geq 0$, we set

$$f^n = (1, G(\cdot, y, v_1) - G(\cdot, y, v_2), \varphi_0, \varphi_1, \ldots, \varphi_n) \in (L^1(\Sigma))^{n+3}.$$

Thanks to Lyapunov's convexity theorem, for every $n \geq 0$ and every $\rho \in]0,1[$, there exists a measurable subset $E_\rho^n \subset \Sigma$ satisfying

$$\int_{E_\rho^n} f^n \, dsdt = \rho \int_\Sigma f^n \, dsdt.$$

Thus, for every $n \geq 0$, E_ρ^n satisfies (4.9), (4.10) and

$$\int_{E_\rho^n} \varphi_m \, dsdt = \rho \int_\Sigma \varphi_m \, dsdt, \tag{4.12}$$

for every $m \in \{0, \ldots, n\}$. Now, for any fixed φ in $L^1(\Sigma)$, we have

$$\begin{aligned} |\int_\Sigma \left(\frac{1}{\rho}\chi_{E_\rho^n} - 1\right) \varphi \, dsdt| &\leq |\int_\Sigma \left(\frac{1}{\rho}\chi_{E_\rho^n} - 1\right) (\varphi - \varphi_m) \, dsdt| + |\int_\Sigma \left(\frac{1}{\rho}\chi_{E_\rho^n} - 1\right) \varphi_m \, dsdt| \\ &\leq \left(\frac{1}{\rho} + 1\right) \|\varphi - \varphi_m\|_{1,\Sigma} + |\int_\Sigma \left(\frac{1}{\rho}\chi_{E_\rho^n} - 1\right) \varphi_m \, dsdt|. \end{aligned}$$

Since $(\varphi_m)_m$ is dense in $L^1(\Sigma)$, for $\epsilon > 0$ there exists $\bar{m} > 0$ such that $\|\varphi - \varphi_{\bar{m}}\|_{1,\Sigma} \leq \frac{\epsilon}{\frac{1}{\rho}+1}$.

Thanks to (4.12), for every $n \geq \bar{m}$, we have $\int_\Sigma \left(\frac{1}{\rho}\chi_{E_\rho^n} - 1\right) \varphi_{\bar{m}} \, dsdt = 0$. Thus, it follows that

$$\lim_n |\int_\Sigma \left(\frac{1}{\rho}\chi_{E_\rho^n} - 1\right) \varphi \, dsdt| = 0,$$

and the proof is complete. ∎

Proof of Theorem 4.1. The existence of E_ρ satisfying (4.1), (4.2) is an easy consequence of Lemma 4.2. The only delicate point is the proof of (4.3). This kind of result is already given in ([12, Theorem 5.2]) and in ([26, Theorem 3.3]). Since we deal with unbounded controls and nonmonotone operators, our assumptions are different from those in [12] and [26]. However, the proof of (4.3) can be adapted from the proofs given in [12] and [26].

Let ρ be in $]0,1[$ and let $(E_\rho^n)_n$ be the sequence of measurable subsets defined in Lemma 4.2. We set

$$v_\rho^n(s,t) = \begin{cases} v_1(s,t) & \text{on } \Sigma \setminus E_\rho^n, \\ v_2(s,t) & \text{on } E_\rho^n, \end{cases} \qquad w_\rho = w_1 + \rho w_2.$$

Let y_ρ^n be the solution of (1.1) corresponding to (v_ρ^n, w_ρ) and let z be the weak solution of (4.7). It is clear that $\xi_\rho^n = (y_\rho^n - y_1)/\rho - z$ is the weak solution in $C(\overline{Q}) \cap W(0,T)$ of

$$\frac{\partial \xi}{\partial t} + A\xi + a_\rho^n \xi = f_\rho^n \text{ in } Q, \quad \frac{\partial \xi}{\partial n_A} + b_\rho^n \xi = g_\rho^n + h_\rho^n \text{ on } \Sigma, \quad \xi(0) = 0 \text{ in } \Omega,$$

where

$$a_\rho^n(x,t) = \int_0^1 f_y'(x,t,(y_1+\theta(y_\rho^n - y_1))(x,t))\,d\theta,$$
$$b_\rho^n(s,t) = \int_0^1 g_y'(s,t,(y_1+\theta(y_\rho^n - y_1))(s,t),v_\rho^n(s,t))\,d\theta,$$
$$f_\rho^n = (f_y'(x,t,y_1) - a_\rho^n)z,$$
$$g_\rho^n = (g_y'(s,t,y_1,v_1) - b_\rho^n)z,$$
$$h_\rho^n = \left(1 - \frac{1}{\rho}\chi_{E_\rho^n}\right)(g(s,t,y_1,v_2) - g(s,t,y_1,v_1)),$$

and $\chi_{E_\rho^n}$ is the characteristic function of E_ρ^n. We denote by $\xi_\rho^{n,1}$ the solution in $C(\overline{Q}) \cap W(0,T)$ of

$$\frac{\partial \xi}{\partial t} + A\xi + a_\rho^n \xi = f_\rho^n \text{ in } Q, \quad \frac{\partial \xi}{\partial n_A} + b_\rho^n \xi = g_\rho^n \text{ on } \Sigma, \quad \xi(\cdot,0) = 0 \text{ in } \Omega,$$

by $\xi_\rho^{n,2}$ the solution in $C(\overline{Q}) \cap W(0,T)$ of

$$\frac{\partial \xi}{\partial t} + A\xi + a_\rho^n \xi = 0 \text{ in } Q, \quad \frac{\partial \xi}{\partial n_A} + b_\rho^n \xi = h_\rho^n \text{ on } \Sigma, \quad \xi(\cdot,0) = 0 \text{ in } \Omega,$$

and by ζ_ρ^n the solution in $C(\overline{Q}) \cap W(0,T)$ of

$$\frac{\partial \zeta}{\partial t} + A\zeta + a\zeta = 0 \text{ in } Q, \quad \frac{\partial \zeta}{\partial n_A} + b\zeta = h_\rho^n \text{ on } \Sigma, \quad \zeta(\cdot,0) = 0 \text{ in } \Omega,$$

where $a(x,t) = f_y'(x,t,y_1(x,t))$, $b(s,t) = g_y'(s,t,y_1(s,t),v_1(s,t))$. We also have

$$\frac{\partial(\xi_\rho^{n,2} - \zeta_\rho^n)}{\partial t} + A(\xi_\rho^{n,2} - \zeta_\rho^n) + a_\rho^n(\xi_\rho^{n,2} - \zeta_\rho^n) = (a - a_\rho^n)\zeta_\rho^n \qquad \text{in } Q,$$
$$\frac{\partial(\xi_\rho^{n,2} - \zeta_\rho^n)}{\partial n_A} + b_\rho^n(\xi_\rho^{n,2} - \zeta_\rho^n) = (b - b_\rho^n)\zeta_\rho^n \qquad \text{on } \Sigma,$$
$$(\xi_\rho^{n,2} - \zeta_\rho^n)(\cdot,0) = 0 \qquad \text{in } \Omega.$$

Due to ([37, Proposition 3.3]), there exists $C = C(T,\Omega,N,q,\sigma,C_0) > 0$ (independent of n and ρ) such that

$$\|\xi_\rho^{n,2} - \zeta_\rho^n\|_{C(\overline{Q})} \le C(\|a - a_\rho^n\|_{q,Q} + \|b - b_\rho^n\|_{\sigma,\Sigma})\|\zeta_\rho^n\|_{C(\overline{Q})}, \tag{4.13}$$

$$\|\xi_\rho^{n,1}\|_{C(\overline{Q})} \le C(\|f_\rho^n\|_{q,Q} + \|g_\rho^n\|_{\sigma,\Sigma}). \tag{4.14}$$

The operator $\mathcal{T}$ which associates ζ, the solution in $C(\overline{Q}) \cap W(0,T)$ of

$$\frac{\partial \zeta}{\partial t} + A\zeta + a\zeta = \varphi \text{ in } Q, \quad \frac{\partial \zeta}{\partial n_A} + b\zeta = \psi \text{ on } \Sigma, \quad \zeta(0) = 0 \text{ in } \Omega,$$

with (φ,ψ), is continuous from $L^q(Q) \times L^\sigma(\Sigma)$ into $C^{\alpha,\frac{\alpha}{2}}(\overline{Q})$ for some $0 < \alpha < 1$ (as for Corollary 3.2, this continuity result can be deduced from Chapter 3, Theorem 1.3, in [17]). Since the embedding from $C^{\alpha,\frac{\alpha}{2}}(\overline{Q})$ into $C(\overline{Q})$ is compact, $\mathcal{T}$ may also be considered as a compact operator from $L^q(Q) \times L^\sigma(\Sigma)$ into $C(\overline{Q})$.

Because of (4.11), for every $0 < \rho < 1$ the sequence $(h_\rho^n)_n$ converges to zero for the weak topology of $L^\sigma(\Sigma)$. Therefore, since $\mathcal{T}$ is compact from $L^q(Q) \times L^\sigma(\Sigma)$ into

$C(\overline{Q})$, the sequence $(\zeta_\rho^n)_n$ converges to zero in $C(\overline{Q})$. There then exists an integer depending on ρ, denoted by $n(\rho)$, such that

$$\|\zeta_\rho^{n(\rho)}\|_{C(\overline{Q})} \leq \rho. \tag{4.15}$$

Notice that $(v_\rho^{n(\rho)})_\rho$ converges to v_1 in $L^\sigma(\Sigma)$ and $(w_\rho)_\rho$ converges to w in $C(\overline{\Omega})$ as ρ tends to zero. From Theorem 3.1 it follows that $(y_\rho^{n(\rho)})_\rho$ uniformly converges to y_1 on $\overline{Q}$ as ρ tends to zero. Therefore, due to (A1) and (A2), $f_\rho^{n(\rho)}$ and $(a - a_\rho^{n(\rho)})$ both converge to zero in $L^q(Q)$ when ρ tends to zero and $g_\rho^{n(\rho)}$, $(b - b_\rho^{n(\rho)})$ both converge to zero in $L^\sigma(\Sigma)$ when ρ tends to zero. Thus, thanks to (4.13)–(4.15), we obtain

$$\lim_{\rho\to 0} \|\xi_\rho^{n(\rho)}\|_{C(\overline{Q})} \leq \lim_{\rho\to 0} \|\xi_\rho^{n(\rho),1}\|_{C(\overline{Q})} + \lim_{\rho\to 0} \|\xi_\rho^{n(\rho),2} - \zeta_\rho^{n(\rho)}\|_{C(\overline{Q})} + \lim_{\rho\to 0} \|\zeta_\rho^{n(\rho)}\|_{C(\overline{Q})} = 0.$$

Now we set $E_\rho = E_\rho^{n(\rho)}$, $v_\rho = v_\rho^{n(\rho)}$, and $\frac{1}{\rho} r_\rho = \xi_\rho^{n(\rho)}$. Conditions (4.1) to (4.3) are clearly satisfied; moreover, taking (4.2), (4.3), and the definition of (v_ρ, w_ρ) into account, we easily verify (4.4). ∎

5. Proof of Pontryagin's principle.

5.1. Penalized problem. We first give the proof of optimality conditions in qualified form (the case $\bar{\nu} = 1$ in Theorem 2.1). The proof of the nonqualified form can be obtained with slight modifications that we give in section 5.3. For notational simplicity, throughout what follows we set

$$H_\Sigma(s, t, y, v, p, 1) = H_\Sigma(s, t, y, v, p)$$

for every $(s, t, y, v, p) \in \Gamma \times [0, T] \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}$. Following [30] and [31], since $C(\overline{D})$ is separable, there exists a norm $|\cdot|_{C(\overline{D})}$, which is equivalent to the norm $\|\cdot\|_{C(\overline{D})}$ such that $(C(\overline{D}), |\cdot|_{C(\overline{D})})$ is strictly convex, and $\mathcal{M}(\overline{D})$, endowed with the dual norm of $|\cdot|_{C(\overline{D})}$ (denoted by $|\cdot|_{\mathcal{M}(\overline{D})}$), is also strictly convex (see [18, Corollary 2, p. 148 or Corollary 2, p. 167]). We define the distance function to $\mathcal{C}$ (for the new norm $|\cdot|_{C(\overline{D})}$) by

$$d_{\mathcal{C}}(\varphi) = \inf_{z\in\mathcal{C}} |\varphi - z|_{C(\overline{D})}.$$

Since $\mathcal{C}$ is convex, then $d_{\mathcal{C}}$ is convex and Lipschitz of rank 1, and we have

$$\limsup_{\substack{\rho\searrow 0,\\ \varphi'\to\varphi}} \frac{d_{\mathcal{C}}(\varphi' + \rho z) - d_{\mathcal{C}}(\varphi')}{\rho} = \max\{\langle \xi, z\rangle_{\mathcal{M}(\overline{D})\times C(\overline{D})} \mid \xi \in \partial d_{\mathcal{C}}(\varphi)\} \tag{5.1}$$

for every $\varphi, z \in C(\overline{D})$, where $\partial d_{\mathcal{C}}$ is the subdifferential in the sense of convex analysis (see [16]). Therefore, for a given $\varphi \in C(\overline{D})$ we have

$$\langle \xi, z - \varphi\rangle_{\mathcal{M}(\overline{D})\times C(\overline{D})} + d_{\mathcal{C}}(\varphi) \leq d_{\mathcal{C}}(z) \quad \text{for all } \xi \in \partial d_{\mathcal{C}}(\varphi) \text{ and all } z \in C(\overline{D}), \tag{5.2}$$

$$|\xi|_{\mathcal{M}(\overline{D})} \leq 1 \quad \text{for all } \xi \in \partial d_{\mathcal{C}}(\varphi).$$

Moreover, it is proved in ([31, Lemma 3.4]) that, since $\mathcal{C}$ is a closed convex subset of $C(\overline{D})$, for every $\varphi \notin \mathcal{C}$, and every $\xi \in \partial d_{\mathcal{C}}(\varphi)$, we have $|\xi|_{\mathcal{M}(\overline{D})} = 1$. Since $\partial d_{\mathcal{C}}(\varphi)$ is

convex in $\mathcal{M}(\overline{D})$ and $(\mathcal{M}(\overline{D}), |\cdot|_{\mathcal{M}(\overline{D})})$ is strictly convex, then if $\varphi \notin \mathcal{C}$, $\partial d_{\mathcal{C}}(\varphi)$ is a singleton and $d_{\mathcal{C}}$ is Gâteaux differentiable at φ.

Let $(\bar{y}, \bar{v}, \bar{w})$ be a solution of problem (P). Thanks to (A8), we prove in the proposition below that $(\bar{y}, \bar{v}, \bar{w})$ is also a local solution of some related penalized problems.

PROPOSITION 5.1. *For every $k > 0$, there exists $\lambda = \lambda(k)$ such that $(\bar{y}, \bar{v}, \bar{w})$ is a solution of the following problem:*

$$(P^{r,k}) \qquad \inf\{J_r(y,v,w) \mid y \in W(0,T) \cap C(\overline{Q}),\ (v,w) \in (V_{ad}(\bar{v},k) \times W_{ad}) \cap \mathcal{B}^d_{\lambda(k)}(\bar{v},\bar{w}) \text{ and } (y,v,w) \text{ satisfies (1.1)}\},$$

where $J_r(y,v,w) = J(y,v,w) + r d_{\mathcal{C}}(\phi(y))$, $\mathcal{B}^d_{\lambda(k)}(\bar{v},\bar{w}) \subset V_{ad} \times W_{ad}$ is the closed ball centered on $(\bar{v},\bar{w})$ and with radius $\lambda(k)$ (for the distance d), r only depends on the constant $\tilde{r}$ given in (A8) *and on $\overline{D}$.*

Proof. From (A8), there exist $\tilde{\varepsilon} > 0$ and $\tilde{r} > 0$ such that

$$\inf(P) = \inf\{J(y_{vw}, v, w) + \tilde{r}\gamma \mid (v,w) \in V_{ad} \times W_{ad},\ \phi(y_{vw}) \in \mathcal{C}_\gamma,\ \ \gamma \in [0,\tilde{\varepsilon}]\}.$$

Now by writing

$$\inf(P) = \inf\{\inf\{J(y_{vw}, v, w) + \tilde{r}\gamma \mid\ \phi(y_{vw}) \in \mathcal{C}_\gamma,\ \ \gamma \in [0,\tilde{\varepsilon}]\} \mid (v,w) \in V_{ad} \times W_{ad}\},$$

we have

$$\inf(P) = \inf\left\{J(y_{vw}, v, w) + \tilde{r} \inf_{z \in \mathcal{C}} \|\phi(y_{vw}) - z\|_{C(\overline{D})} \mid (v,w) \in V_{ad} \times W_{ad},\ \phi(y_{vw}) \in \mathcal{C}_{\tilde{\varepsilon}}\right\}.$$

Since the norms $|\cdot|_{C(\overline{D})}$ and $\|\cdot\|_{C(\overline{D})}$ are equivalent, there exist $r \geq \tilde{r}$ and $0 < \varepsilon \leq \tilde{\varepsilon}$ such that

$$\inf\left\{J(y_{vw}, v, w) + \tilde{r} \inf_{z \in \mathcal{C}} \|\phi(y_{vw}) - z\|_{C(\overline{D})} \mid (v,w) \in V_{ad} \times W_{ad},\ \phi(y_{vw}) \in \mathcal{C}_{\tilde{\varepsilon}}\right\}$$
$$\leq \inf\{J(y_{vw}, v, w) + r d_{\mathcal{C}}(\phi(y_{vw})) \mid (v,w) \in V_{ad} \times W_{ad},\ d_{\mathcal{C}}(\phi(y_{vw})) \leq \varepsilon\}.$$

Moreover, taking (A6) and Lemma 3.3 into account, there exists $\lambda(k) > 0$ such that

$$d_{\mathcal{C}}(\phi(y_{vw})) \leq \varepsilon \quad \text{for every} \quad (v,w) \in (V_{ad}(\bar{v},k) \times W_{ad}) \cap \mathcal{B}^d_{\lambda(k)}(\bar{v},\bar{w}).$$

Thus,

$$\inf(P) \leq \inf\left\{J(y_{vw}, v, w) + r d_{\mathcal{C}}(\phi(y_{vw})) \mid (v,w) \in (V_{ad}(\bar{v},k) \times W_{ad}) \cap \mathcal{B}^d_{\lambda(k)}(\bar{v},\bar{w})\right\}$$
$$= \inf(P^{r,k}) \leq J(\bar{y},\bar{v},\bar{w}) = \inf(P). \qquad \square$$

Now we set $J_{r,n}(y,v,w) = J(y,v,w) + r[(d_{\mathcal{C}}(\phi(y)))^2 + n^{-2}]^{\frac{1}{2}}$, and we denote by $(P^{r,k}_n)$ the problem

$$\inf\{J_{r,n}(y,v,w) \mid y \in W(0,T) \cap C(\overline{Q}),\ (v,w) \in (V_{ad}(\bar{v},k) \times W_{ad}) \cap \mathcal{B}^d_{\lambda(k)}(\bar{v},\bar{w}) \text{ and } (y,v,w) \text{ satisfies (1.1)}\}.$$

PROPOSITION 5.2. *For every $k > 0$,*

$$\inf(P^{r,k}) = \lim_{n\to\infty} \inf(P_n^{r,k}) \quad \textit{and} \quad \lim_{n\to\infty} J_{r,n}(y,v,w) = J_r(y,v,w)$$

for every $y \in C(\overline{Q})$ and every $(v,w) \in L^\sigma(\Sigma) \times C(\overline{\Omega})$. Moreover, $(\bar{y},\bar{v},\bar{w})$ is a ε^2-solution of $(P_n^{r,k})$ with $\varepsilon^2 = rn^{-1}$.

Proof. The first part of the proof is immediate if we observe that

$$J_r(y,v,w) \le J_{r,n}(y,v,w) \le J_r(y,v,w) + rn^{-1}$$

for every $(y,v,w) \in C(\overline{Q}) \times L^\sigma(\Sigma) \times C(\overline{\Omega})$. Moreover, since $(\bar{y},\bar{v},\bar{w})$ is solution of $(P^{r,k})$, with the previous inequalities, we obtain

$$\begin{aligned} J_{r,n}(\bar{y},\bar{v},\bar{w}) &\le J_r(\bar{y},\bar{v},\bar{w}) + rn^{-1} \\ &\le J_r(y,v,w) + rn^{-1} \le J_{r,n}(y,v,w) + rn^{-1}, \end{aligned}$$

for every $y \in C(\overline{Q})$ and every $(v,w) \in (V_{ad}(\bar{v},k) \times W_{ad}) \cap \mathcal{B}^d_{\lambda(k)}(\bar{v},\bar{w})$ such that (y,v,w) obeys (1.1). ∎

5.2. Proof of Theorem 2.1 (Pontryagin principle in qualified form). Let k be a positive constant. Thanks to Proposition 5.2, for every $n \ge 1$, $(\bar{y},\bar{v},\bar{w})$ is an ε_n^2-solution of $(P_n^{r,k})$, with $\varepsilon_n^2 = rn^{-1}$. For every $k > 0$, we choose $n(k)$ such that

$$\varepsilon_{n(k)} = \left(\frac{r}{n(k)}\right)^{\frac{1}{2}} \le \min\left(\frac{1}{k^{2\sigma}}, \frac{\lambda(k)}{2}\right).$$

The metric space $((V_{ad}(\bar{v},k) \times W_{ad}) \cap \mathcal{B}^d_{\lambda(k)}(\bar{v},\bar{w}), d)$ is complete and the functional $(v,w) \longmapsto J_{r,n(k)}(y_{vw},v,w)$ is continuous on this metric space. Thanks to Ekeland's variational principle, for every $k \ge 1$, there exists $(v_k,w_k) \in (V_{ad}(\bar{v},k) \times W_{ad}) \cap \mathcal{B}^d_{\lambda(k)}(\bar{v},\bar{w})$ such that

$$d((v_k,w_k),(\bar{v},\bar{w})) \le \varepsilon_{n(k)}, \tag{5.3}$$

$$J_{r,n(k)}(y_k,v_k,w_k) \le J_{r,n(k)}(y_{vw},v,w) + \varepsilon_{n(k)} d((v_k,w_k),(v,w)) \tag{5.4}$$

for every $(v,w) \in (V_{ad}(\bar{v},k) \times W_{ad}) \cap \mathcal{B}^d_{\lambda(k)}(\bar{v},\bar{w})$ (y_k and y_{vw} being the states corresponding, respectively, to (v_k,w_k) and to (v,w)).

The proof is split into five steps.

Step 1. *Approximate optimality conditions for the boundary control v_k satisfying* (5.3) *and* (5.4). For fixed v_0 in V_{ad}, we denote by v_{0k} ($k > 0$) the function in $V_{ad}(\bar{v},k)$ defined by

$$v_{0k}(s,t) = \begin{cases} v_0(s,t) & \text{if } |v_0(s,t) - \bar{v}(s,t)| \le k, \\ \bar{v}(s,t) & \text{if not.} \end{cases} \tag{5.5}$$

Applying Theorem 4.1, we deduce the existence of measurable sets E_ρ^k, such that $\mathcal{L}^N(E_\rho^k) = \rho \mathcal{L}^N(\Sigma)$,

$$y_{1,\rho}^k = y_k + \rho z_k^1 + r_\rho^k, \qquad \lim_{\rho\to 0} \frac{1}{\rho} \|r_\rho^k\|_{C(\overline{Q})} = 0, \tag{5.6}$$

$$J(y^k_{1,\rho}, v^k_{1,\rho}, w^k_{1,\rho}) = J(y_k, v_k, w_k) + \rho \Delta J^1_k + o(\rho), \tag{5.7}$$

where $v^k_{1,\rho}$ and $w^k_{1,\rho}$ are defined by

$$v^k_{1,\rho}(s,t) = \begin{cases} v_k(s,t) & \text{on } \Sigma \setminus E^k_\rho, \\ v_{0k}(s,t) & \text{on } E^k_\rho, \end{cases} \quad w^k_{1,\rho} = w_k, \tag{5.8}$$

$y^k_{1,\rho}$ is the state corresponding to $(v^k_{1,\rho}, w^k_{1,\rho})$, z^1_k is the weak solution of

$$\begin{aligned} &\frac{\partial z^1_k}{\partial t} + Az^1_k + f'_y(x,t,y_k)z^1_k = 0 && \text{in } Q, \\ &\frac{\partial z^1_k}{\partial n_A} + g'_y(s,t,y_k,v_k)z^1_k = g(s,t,y_k,v_k) - g(s,t,y_k,v_{0k}) && \text{on } \Sigma, \\ &z^1_k(\cdot,0) = 0 && \text{in } \Omega, \end{aligned}$$

and

$$\begin{aligned} \Delta J^1_k = &\int_Q F'_y(x,t,y_k(x,t))z^1_k(x,t)\,dxdt \int_\Sigma G'_y(s,t,y_k(s,t),v_k(s,t))z^1_k(s,t)\,dsdt \\ &+ \int_\Sigma [G(s,t,y_k(s,t),v_{0k}(s,t)) - G(s,t,y_k(s,t),v_k(s,t))]\,dsdt \\ &+ \int_\Omega L'_y(x,y_k(x,T),w_k(x))z^1_k(x)\,dx. \end{aligned}$$

On the other hand, we have

$$\begin{aligned} d((v^k_{1,\rho}, w^k_{1,\rho}), (\bar v, \bar w)) &\le d((v^k_{1,\rho}, w^k_{1,\rho}), (v_k, w_k)) + d((v_k, w_k), (\bar v, \bar w)) \\ &\le \mathcal{L}^N(E^k_\rho) + \varepsilon_{n(k)} \le \rho \mathcal{L}^N(\Sigma) + \varepsilon_{n(k)}. \end{aligned}$$

There then exists ρ_k such that, for every $0 < \rho < \rho_k$, we have

$$d((v^k_{1,\rho}, w^k_{1,\rho}), (\bar v, \bar w)) \le \rho \mathcal{L}^N(\Sigma) + \varepsilon_{n(k)} \le \lambda(k).$$

Therefore, for all $k > 0$ and all $0 < \rho < \rho_k$, $(v^k_{1,\rho}, w^k_{1,\rho})$ belongs to $(V_{ad}(\bar v, k) \times W_{ad}) \cap \mathcal{B}^d_{\lambda(k)}(\bar v, \bar w)$. If we set $(v,w) = (v^k_{1,\rho}, w^k_{1,\rho})$ in (5.4), it follows that

$$\limsup_{\rho \to 0} \frac{J_{r,n(k)}(y_k, v_k, w_k) - J_{r,n(k)}(y^k_{1,\rho}, v^k_{1,\rho}, w^k_{1,\rho})}{\rho} \le \varepsilon_{n(k)} \mathcal{L}^N(\Sigma). \tag{5.9}$$

Taking (5.2), (5.7), and the definition of $J_{r,n}$ into account, we get

$$-\Delta J^1_k - \langle \mu_k, \phi'(y_k) z^1_k \rangle_{\mathcal{M}(\overline{D}) \times C(\overline{D})} \le \varepsilon_{n(k)} \mathcal{L}^N(\Sigma), \tag{5.10}$$

where

$$\mu_k = \begin{cases} \dfrac{r d_{\mathcal{C}}(\phi(y_k)) \nabla d_{\mathcal{C}}(\phi(y_k))}{[d_{\mathcal{C}}(\phi(y_k))^2 + n(k)^{-2}]^{\frac{1}{2}}} & \text{if } d_{\mathcal{C}}(\phi(y_k)) > 0, \\ 0 & \text{if not.} \end{cases} \tag{5.11}$$

For every $k > 0$, we consider the weak solution p_k of

$$(5.12)\quad \begin{cases} -\dfrac{\partial p_k}{\partial t} + Ap_k + f'_y(x,t,y_k)p_k = F'_y(x,t,y_k) + [\phi'(y_k)^*\mu_k]|_Q & \text{in } Q, \\ \dfrac{\partial p_k}{\partial n_A} + g'_y(s,t,y_k,v_k)p_k = G'_y(s,t,y_k,v_k) + [\phi'(y_k)^*\mu_k]|_\Sigma & \text{on } \Sigma, \\ p_k(T) = L'_y(x,y_k(T),w_k) + [\phi'(y_k)^*\mu_k]|_{\overline{\Omega}_T} & \text{in } \Omega, \end{cases}$$

where $[\phi'(y_k)^*\mu_k]|_Q$ is the restriction of $[\phi'(y_k)^*\mu_k]$ to Q, $[\phi'(y_k)^*\mu_k]|_\Sigma$ is the restriction of $[\phi'(y_k)^*\mu_k]$ to Σ, and $[\phi'(y_k)^*\mu_k]|_{\overline{\Omega}_T}$ is the restriction of $[\phi'(y_k)^*\mu_k]$ to $\overline{\Omega}_T$. By using the Green formula of Theorem 3.4, we obtain

$$\int_Q F'_y(x,t,y_k)z_k^1\,dxdt + \int_\Sigma G'_y(s,t,y_k,v_k)z_k^1\,dsdt + \int_\Omega L'_y(x,y_k(T),w_k)z_k^1(T)\,dx$$

$$+\langle \mu_k, \phi'(y_k)z_k^1\rangle_{\mathcal{M}(\overline{D})\times C(\overline{D})}$$

$$= \int_Q p_k\left(\frac{\partial z_k^1}{\partial t} + Az_k^1 + f'_y(x,t,y_k)z_k^1\right)dxdt + \int_\Sigma p_k\left(\frac{\partial z_k^1}{\partial n_A} + g'_y(s,t,y_k,v_k)z_k^1\right)dsdt$$

$$= \int_\Sigma p_k\left[g(s,t,y_k,v_k) - g(s,t,y_k,v_{0k})\right]\,dsdt.$$

With this equality, with (5.10) and the definition of ΔJ_k^1, we get

$$(5.13)\quad \int_\Sigma [G(s,t,y_k,v_k) - p_k g(s,t,y_k,v_k)]\,dsdt$$

$$\le \int_\Sigma [G(s,t,y_k,v_{0k}) - p_k g(s,t,y_k,v_{0k})]\,dsdt + \varepsilon_{n(k)}\mathcal{L}^N(\Sigma)$$

$$\le \int_\Sigma [G(s,t,y_k,v_{0k}) - p_k g(s,t,y_k,v_{0k})]\,dsdt + \frac{1}{k^{2\sigma}}\mathcal{L}^N(\Sigma)$$

for every $k > 0$ and every $v_0 \in V_{ad}$ (where v_{0k} is defined according to v_0 in (5.5)).

Step 2. *Approximate optimality conditions for the initial control w_k satisfying* (5.3) and (5.4).

Let w_0 be in W_{ad}. We consider the sequence $(v_{2,\rho}^k, w_{2,\rho}^k)$ defined by

$$v_{2,\rho}^k = v_k, \qquad w_{2,\rho}^k = w_k + \rho(w_0 - w_k)$$

and we denote by $y_{2,\rho}^k$ the state corresponding to $(v_{2,\rho}^k, w_{2,\rho}^k)$. Since W_{ad} is convex, we see that $\{(v_{2,\rho}^k, w_{2,\rho}^k), k > 0\} \subset V_{ad}(\bar{v},k) \times W_{ad}$. Moreover, we have

$$d((v_{2,\rho}^k, w_{2,\rho}^k), (\bar{v},\bar{w})) \le \rho\|w_0 - w_k\|_{\infty,\Omega} + \varepsilon_{n(k)}$$

$$\leq \varepsilon_{n(k)} + \rho(\|w_0 - \bar{w}\|_{\infty,\Omega} + \|w_k - \bar{w}\|_{\infty,\Omega})$$

$$\leq \varepsilon_{n(k)} + \rho(\varepsilon_{n(k)} + \|w_0 - \bar{w}\|_{\infty,\Omega}).$$

As in Step 1, if ρ_k is small enough, for every $0 < \rho < \rho_k$, we have

$$d((v^k_{2,\rho}, w^k_{2,\rho}), (\bar{v}, \bar{w})) \leq \lambda(k).$$

Thus, $(v^k_{2,\rho}, w^k_{2,\rho})$ belongs to $(V_{ad}(\bar{v}, k) \times W_{ad}) \cap \mathcal{B}^d_{\lambda(k)}(\bar{v}, \bar{w})$, for every $k > 0$ and for every $0 < \rho < \rho_k$. Then Theorem 4.1 gives

$$y^k_{2,\rho} = y_k + \rho z^2_k + r^k_\rho, \qquad \lim_{\rho\to 0} \frac{1}{\rho}\|r^k_\rho\|_{C(\overline{Q})} = 0, \tag{5.14}$$

$$J(y^k_{2,\rho}, v^k_{2,\rho}, w^k_{2,\rho}) = J(y_k, v_k, w_k) + \rho\Delta J^2_k + o(\rho), \tag{5.15}$$

where z^2_k is the weak solution of

$$\begin{aligned}
&\frac{\partial z^2_k}{\partial t} + Az^2_k + f'_y(x,t,y_k)z^2_k = 0 && \text{in } Q,\\
&\frac{\partial z^2_k}{\partial n_A} + g'_y(s,t,y_k,v_k)z^2_k = 0 && \text{on } \Sigma,\\
&z^2_k(0) = w_0 - w_k && \text{in } \Omega,
\end{aligned}$$

and

$$\begin{aligned}
\Delta J^2_k = &\int_Q F'_y(x,t,y_k)z^2_k\,dxdt + \int_\Sigma G'_y(s,t,y_k,v_k)z^2_k\,dsdt\\
&+ \int_\Omega L'_y(x,y_k(T),w_k)z^2_k(T)\,dx + \int_\Omega L'_w(x,y_k(T),w_k)(w_0 - w_k)\,dx.
\end{aligned}$$

As in Step 1, from (5.4) and (5.15) we deduce that

$$\begin{aligned}
&-\Delta J^2_k - \langle \mu_k, \phi'(y_k)z^2_k\rangle_{\mathcal{M}(\overline{D})\times C(\overline{D})}\\
&\leq \limsup_{\rho\to 0} \frac{J_{r,n(k)}(y_k,v_k,w_k) - J_{r,n(k)}(y^k_{2,\rho}, v^k_{2,\rho}, w^k_{2,\rho})}{\rho}\\
&\leq \varepsilon_{n(k)}\|w_0 - w_k\|_{\infty,\Omega} \leq \varepsilon_{n(k)}(\varepsilon_{n(k)} + \|w_0 - \bar{w}\|_{\infty,\Omega}),
\end{aligned} \tag{5.16}$$

where μ_k is defined in (5.11).

If we consider the weak solution p_k of (5.12), still using the Green formula of Theorem 3.4, we obtain

$$\begin{aligned}
&\int_Q F'_y(x,t,y_k)z^2_k\,dxdt + \int_\Sigma G'_y(s,t,y_k,v_k)z^2_k\,dsdt + \int_\Omega L'_y(x,y_k(T),w_k)z^2_k(T)\,dx\\
&\qquad + \langle [\phi'(y_k)^*\mu_k]|_{\overline{Q}\setminus\overline{\Omega}_0}, z^2_k\rangle_{\mathcal{M}_b(\overline{Q}\setminus\overline{\Omega}_0)\times C_b(\overline{Q}\setminus\overline{\Omega}_0)}
\end{aligned}$$

$$= \int_Q p_k \left(\frac{\partial z_k^2}{\partial t} + A z_k^2 + f_y'(x,t,y_k) z_k^2 \right) dxdt + \int_\Sigma p_k \left(\frac{\partial z_k^2}{\partial n_A} + g_y'(s,t,y_k,v_k) z_k^2 \right) dsdt$$
$$+ \langle p_k(0), z_k^2(0) \rangle_{\mathcal{M}(\overline{\Omega})\times C(\overline{\Omega})} = \langle p_k(0), w_0 - w_k \rangle_{\mathcal{M}(\overline{\Omega})\times C(\overline{\Omega})}.$$

Taking (5.16) and the definition of ΔJ_k^2 into account, we get

$$(5.17) \quad -\int_\Omega L_w'(x, y_k(T), w_k)(w_0 - w_k)\, dx - \langle p_k(0), w_0 - w_k \rangle_{\mathcal{M}(\overline{\Omega})\times C(\overline{\Omega})}$$
$$-\langle [\phi_y'(y_k)^* \mu_k]|_{\overline{\Omega}_0}, w_0 - w_k \rangle_{\mathcal{M}(\overline{\Omega})\times C(\overline{\Omega})} \le \varepsilon_{n(k)} (\varepsilon_{n(k)} + \|w_0 - \bar{w}\|_{\infty,\Omega})$$

for every $w_0 \in W_{ad}$.

Step 3. *Convergence of sequences* $(\mu_k)_k$ *and* $(p_k)_k$. We observe that

$$|\mu_k|_{\mathcal{M}(\overline{D})} \le r d_{\mathcal{C}}(\phi(y_k)) \left[(d_{\mathcal{C}}(\phi(y_k)))^2 + n(k)^{-2} \right]^{-1/2} \le r.$$

The sequence $(\mu_k)_k$ is bounded in $\mathcal{M}(\overline{D})$, so there exist $\bar{\mu} \in \mathcal{M}(\overline{D})$ and a subsequence, still denoted by $(\mu_k)_k$, such that

$$(5.18) \qquad \mu_k \rightharpoonup \bar{\mu} \ \text{ weak}^* \text{ in } \mathcal{M}(\overline{D}).$$

Let (δ, d) be a pair fulfilling $(C_{q\sigma})$. From Theorem 3.4, we deduce

$$\|p_k\|_{L^{\delta'}(0,T;W^{1,d'}(\Omega))} \le C_4 \Big\{ \|F_y'(\cdot, y_k)\|_{1,Q} + \|G_y'(\cdot, y_k, v_k)\|_{1,\Sigma} + \|L_y'(\cdot, y_k(T), w_k)\|_{1,\Omega}$$
$$+ |\mu_k|_{\mathcal{M}(\overline{D})} \|\phi'(y_k)\|_{\mathcal{L}(C(\overline{Q});C(\overline{D}))} \Big\}.$$

(Here $\mathcal{L}(C(\overline{Q}); C(\overline{D}))$ denotes the space of linear continuous mappings from $C(\overline{Q})$ to $C(\overline{D})$.)

Since the sequences $(\mu_k)_k, (y_k)_k, (v_k)_k$, and $(w_k)_k$ are bounded, respectively, in $\mathcal{M}(\overline{D})$, $C(\overline{Q})$, $L^\sigma(\Sigma)$, and in $C(\overline{\Omega})$, the sequence $(p_k)_k$ is bounded in $L^{\delta'}(0,T;W^{1,d'}(\Omega))$. There then exist $\bar{p} \in L^{\delta'}(0,T;W^{1,d'}(\Omega))$ and a subsequence, still denoted by $(p_k)_k$, such that $(p_k)_k$ weakly converges to $\bar{p}$ in $L^{\delta'}(0,T;W^{1,d'}(\Omega))$.

Let us prove that $\bar{p}$ is the weak solution of (2.4). Let φ be in $C^1(\overline{Q})$ satisfying $\varphi(\cdot, 0) = 0$ in $\overline{\Omega}$. For every $k > 0$, we have

$$(5.19) \qquad \int_Q \left\{ p_k \frac{\partial \varphi}{\partial t} + \sum_{i,j=1}^N a_{ij} D_j p_k D_i \varphi + f_y'(\cdot, y_k) p_k \varphi \right\} dxdt$$
$$+ \int_\Sigma \left\{ p_k \frac{\partial \varphi}{\partial n_A} + g_y'(\cdot, y_k, v_k) p_k \varphi \right\} dsdt$$
$$= \int_Q F_y'(x,t,y_k)\varphi\, dxdt + \int_\Sigma G_y'(s,t,y_k,v_k)\varphi\, dsdt + \int_\Omega L_y'(x, y_k(T), w_k)\varphi(T)\, dx$$
$$+ \langle \mu_k, \phi'(y_k)\varphi \rangle_{\mathcal{M}(\overline{D})\times C(\overline{D})}.$$

Since (δ, d) satisfies $(C_{q\sigma})$, the following imbeddings are continuous:

$$L^{\delta'}(0,T;W^{1,d'}(\Omega)) \hookrightarrow L^{q'}(Q), \qquad L^{\delta'}(0,T;W^{1-\frac{1}{d'},d'}(\Gamma)) \hookrightarrow L^{\sigma'}(\Sigma).$$

Therefore, $(p_k)_k$ weakly converges to $\bar p$ in $L^{q'}(Q)$ and the sequence of traces $(p_k|_\Sigma)_k$ weakly converges to the trace $\bar p|_\Sigma$ in $L^{\sigma'}(\Sigma)$.

Moreover, since $(v_k)_k$ converges to $\bar v$ in $L^\sigma(\Sigma)$ (indeed, since $d_E(v_k,\bar v)\le \varepsilon_{n(k)}\le \frac{1}{k^{2\sigma}}$ and $|v_k-\bar v|\le k$ a.e. on Σ, we have $\int_\Sigma |v_k-\bar v|^\sigma\,dsdt\le \frac{1}{k^\sigma}$), since $(w_k)_k$ converges to $\bar w$ in $C(\overline{\Omega})$, and since $(y_k)_k$ converges to $\bar y$ in $C(\overline{Q})$, due to assumptions on f,g,F,G,L,ϕ, we have

$$\lim_k \|f'_y(\cdot,y_k)-f'_y(\cdot,\bar y)\|_{q,Q}=0,\qquad \lim_k \|F'_y(\cdot,y_k)-F'_y(\cdot,\bar y)\|_{1,Q}=0,$$

$$\lim_k \|g'_y(\cdot,y_k,v_k)-g'_y(\cdot,\bar y,\bar v)\|_{\sigma,\Sigma}=0,\qquad \lim_k \|G'_y(\cdot,y_k,v_k)-G'_y(\cdot,\bar y,\bar v)\|_{1,\Sigma}=0,$$

$$\lim_k \|L'_y(\cdot,y_k(T),w_k)-L'_y(\cdot,\bar y(T),\bar w)\|_{1,\Omega}=0,\qquad \lim_k \|\phi'(y_k)-\phi'(\bar y)\|_{\mathcal{L}(C(\overline{Q});C(\overline{D}))}=0.$$

Thus, by passing to the limit in (5.19), it follows that

$$\int_Q\left\{\bar p\frac{\partial\varphi}{\partial t}+\sum_{i,j=1}^N a_{ij}D_j\bar pD_i\varphi+f'_y(x,t,\bar y)\bar p\varphi\right\}dxdt+\int_\Sigma\left\{\bar p\frac{\partial\varphi}{\partial n_A}+g'_y(s,t,\bar y,\bar v)\bar p\varphi\right\}dsdt$$
$$=\int_Q F'_y(x,t,\bar y)\varphi\,dxdt+\int_\Sigma G'_y(s,t,\bar y,\bar v)\varphi\,dsdt+\int_\Omega L'_y(x,\bar y(T),\bar w)\varphi(T)\,dx$$
$$+\langle\bar\mu,\phi'(\bar y)\varphi\rangle_{\mathcal{M}(\overline{D})\times C(\overline{D})}$$

for every $\varphi\in C^1(\overline{Q})$ satisfying $\varphi(\cdot,0)=0$ in $\overline{\Omega}$. Therefore, $\bar p$ is the unique weak solution of (2.4). Since the weak solution of (2.4) is unique in the sense of Definition 3.1, we can deduce by classical arguments that $\bar p$ is independent of the pair (δ,d) (chosen after (5.18)) and that the original sequence $(p_k)_k$ converges weakly to $\bar p$ in $L^{\delta'}(0,T;W^{1,d'}(\Omega))$ for every (δ,d) satisfying $(C_{q\sigma})$. To pass to the limit in (5.17), we prove that

(5.20) $(p_k(0)+[\phi'_y(y_k)^*\mu_k]|_{\overline{\Omega}_0})_k \rightharpoonup \bar p(0)+[\phi'_y(\bar y)^*\bar\mu]|_{\overline{\Omega}_0}$ weakly star in $\mathcal{M}(\overline{\Omega})$.

For this, let φ be in $C(\overline{\Omega})$ and let y be the solution of

$$\frac{\partial y}{\partial t}+Ay=0\ \text{ in } Q,\quad \frac{\partial y}{\partial n_A}=0\ \text{ on }\Sigma,\quad y(0)=\varphi\ \text{ in }\Omega.$$

With the Green formula of Theorem 3.4, we have

$$\langle p_k(0)+[\phi'_y(y_k)^*\mu_k]|_{\overline{\Omega}_0},\varphi\rangle_{\mathcal{M}(\overline{\Omega})\times C(\overline{\Omega})}-\langle\bar p(0)+[\phi'_y(\bar y)^*\bar\mu]|_{\overline{\Omega}_0},\varphi\rangle_{\mathcal{M}(\overline{\Omega})\times C(\overline{\Omega})}$$

$$=\int_Q\left[\bar pf'_y(x,t,\bar y)y-p_kf'_y(x,t,y_k)y\right]dxdt+\int_\Sigma\left[\bar pg'_y(s,t,\bar y,\bar v)y-p_kg'_y(s,t,y_k,v_k)y\right]dsdt$$
$$+\langle\mu_k,\phi'_y(y_k)y\rangle_{\mathcal{M}(\overline{D})\times C(\overline{D})}-\langle\bar\mu,\phi'_y(\bar y)y\rangle_{\mathcal{M}(\overline{D})\times C(\overline{D})}.$$

Now (5.20) follows from the previous convergence results.

Step 4. *Integral Pontryagin's principle.* Notice that $(v_{0k})_k$ tends to v_0 in $L^\sigma(\Sigma)$ and $(v_k)_k$ tends to $\bar v$ in $L^\sigma(\Sigma)$. By passing to the limit when k tends to infinity in (5.13) and (5.17), and by using the convergence results stated in Step 3, we obtain

$$\int_\Sigma H_\Sigma(s,t,\bar y,\bar v,\bar p)\,dsdt\le\int_\Sigma H_\Sigma(s,t,\bar y,v_0,\bar p)\,dsdt \tag{5.21}$$

for every $v_0 \in V_{ad}$, and

$$\int_\Omega L'_w(x,\bar y(T),\bar w)(\bar w - w_0)\,dx + \langle \bar p(0) + [\phi'_y(\bar y)^*\bar\mu]|_{\overline{\Omega}_0}, \bar w - w_0\rangle_{\mathcal{M}(\overline{\Omega})\times C(\overline{\Omega})} \le 0 \tag{5.22}$$

for every $w_0 \in W_{ad}$. On the other hand, from the definition of μ_k and from (5.2), we deduce

$$\langle \mu_k, z - \phi(y_k)\rangle_{\mathcal{M}(\overline{D})\times C(\overline{D})} \le 0 \quad \text{for all } z \in \mathcal{C}.$$

By passing to the limit in this expression, we obtain (2.3).

Step 5. *Pointwise Pontryagin's principle.* The functions

$$(s,t) \longmapsto \bar v(s,t), \qquad (s,t) \longmapsto H_\Sigma(s,t,\bar y(s,t),\bar v(s,t),\bar p(s,t))$$

are measurable on Σ, and the function

$$(s,t,v) \longmapsto H_\Sigma(s,t,\bar y(s,t),v,\bar p(s,t))$$

is a Carathéodory function from $\Sigma\mathbb{R}$ into $\mathbb{R}$. Thanks to Lusin's theorem and Scorza–Dragoni's theorem, for every $\epsilon > 0$, there exist a compact subset $\Sigma_\epsilon \subset \Sigma$, continuous mappings $\varphi_0^\epsilon, \varphi_1^\epsilon$ from Σ_ϵ into $\mathbb{R}$, and a continuous mapping φ_2^ϵ from $\Sigma_\epsilon\mathbb{R}$ into $\mathbb{R}$ such that

$$\mathcal{L}^N(\Sigma \setminus \Sigma_\epsilon) \le \epsilon, \qquad \varphi_0^\epsilon(s,t) = \bar v(s,t) \ \text{ on } \Sigma_\epsilon,$$

$$\varphi_1^\epsilon(s,t) = H_\Sigma(s,t,\bar y(s,t),\bar v(s,t),\bar p(s,t)) \ \text{ on } \Sigma_\epsilon,$$

$$\varphi_2^\epsilon(s,t,v) = H_\Sigma(s,t,\bar y(s,t),v,\bar p(s,t)) \ \text{ on } \Sigma_\epsilon\mathbb{R}.$$

Since $\bar v$ is continuous on Σ_ϵ, $\bar v$ is bounded on Σ_ϵ and, for $M > \|\bar v\|_{\infty,\Sigma_\epsilon}$, the multimapping

$$(s,t) \longmapsto K_M(s,t) := K_V(s,t) \cap [-M,M]$$

has nonempty compact values for all $(s,t) \in \Sigma_\epsilon$. From a Lusin type theorem for measurable multimappings with compact values (see, for example, [2, Theorem 1.4.1]), for every integer $M > \|\bar v\|_{\infty,\Sigma_\epsilon}$, there exists a measurable subset $\Sigma_{\epsilon M} \subset \Sigma_\epsilon$ such that $\mathcal{L}^N(\Sigma_\epsilon \setminus \Sigma_{\epsilon M}) \le \frac{\epsilon}{2^M}$ and the restriction of K_M to $\Sigma_{\epsilon M}$ is continuous. Let us denote by $\widetilde{\Sigma}_{\epsilon M}$ the set of Lebesgue points in $\Sigma_{\epsilon M}$, of the characteristic function of $\Sigma_{\epsilon M}$. Now let (s_0,t_0) be in $\widetilde{\Sigma}_{\epsilon M}$ and let $v \in K_M(s_0,t_0)$. Since the multimapping K_M is continuous on $\Sigma_{\epsilon M}$ (in fact we only use the lower semicontinuity of K_M), for every integer $n > 0$, the multimapping K_M admits a measurable selection v_n and there exists an increasing function γ from $\mathbb{R}^+$ into $\mathbb{R}^+$ such that

$$|v_n(s,t) - v| \le \frac{1}{n} \ \text{ on } \ B\left((s_0,t_0),\gamma\left(\frac{1}{n}\right)\right) \cap \Sigma_{\epsilon M} \ \text{ and } \ \lim_{n\to\infty} \gamma\left(\frac{1}{n}\right) = 0,$$

where $B((s_0,t_0),\gamma(\frac{1}{n}))$ is the ball in $\mathbb{R}^N$ centered on (s_0,t_0) and of radius $\gamma(\frac{1}{n})$. We now set

$$S_{\epsilon,M,n} = \widetilde{\Sigma}_{\epsilon M} \cap B\left((s_0,t_0),\gamma\left(\frac{1}{n}\right)\right),$$

and we consider the variation

$$v_M(s,t) = \begin{cases} \bar{v}(s,t) & \text{on } \Sigma \setminus S_{\epsilon,M,1}, \\ v_n(s,t) & \text{on } S_{\epsilon,M,n} \setminus S_{\epsilon,M,n+1} \quad \text{for every } n > 0, \\ v & \text{if } (s,t) = (s_0,t_0). \end{cases}$$

It is clear that $v_M \in V_{ad}$ and that

$$\lim_{(s,t)\to(s_0,t_0)} v_M(s,t) = v.$$

If we take $v_0 = \chi_{S_{\epsilon,M,n}} v_M + (1-\chi_{S_{\epsilon,M,n}})\bar{v}$ in (5.21) (where $\chi_{S_{\epsilon,M,n}}$ is the characteristic function of $S_{\epsilon,M,n}$), it follows that

$$\frac{1}{\mathcal{L}^N(S_{\epsilon,M,n})} \int_{S_{\epsilon,M,n}} \varphi_1^\varepsilon(s,t)\, ds dt \le \frac{1}{\mathcal{L}^N(S_{\epsilon,M,n})} \int_{S_{\epsilon,M,n}} \varphi_2^\varepsilon(s,t,v_M(s,t))\, ds dt.$$

(Note that for every $n \ge 1$, $\mathcal{L}^N(S_{\epsilon,M,n}) \neq 0$ because $(s_0,t_0) \in \widetilde{\Sigma}_{\epsilon M}$.) By passing to the limit in the above inequality when n tends to infinity, and using the continuity of φ_1^ε and φ_2^ε, we obtain

$$\varphi_1^\varepsilon(s_0,t_0) = H_\Sigma(s_0,t_0,\bar{y}(s_0,t_0),\bar{v}(s_0,t_0),\bar{p}(s_0,t_0))$$

$$\le \varphi_2^\varepsilon(s_0,t_0,v) = H_\Sigma(s_0,t_0,\bar{y}(s_0,t_0),v,\bar{p}(s_0,t_0))$$

for every $(s_0,t_0) \in \widetilde{\Sigma}_{\epsilon M}$ and every $v \in K_V(s_0,t_0)$ such that $|v| \le M$.

We set

$$\widetilde{\Sigma}_\epsilon = \bigcap_{\substack{M \in N^* \\ M > \|\bar{v}\|_{\infty,\Sigma_\epsilon}}} \widetilde{\Sigma}_{\epsilon M}$$

and we observe that $\mathcal{L}^N(\Sigma \setminus \widetilde{\Sigma}_\epsilon) \le 2\epsilon$. For every $(s,t) \in \widetilde{\Sigma}_\epsilon$ and every $v \in K_V(s,t)$ we have

$$H_\Sigma(s,t,\bar{y}(s,t),\bar{v}(s,t),\bar{p}(s,t)) \le H_\Sigma(s,t,\bar{y}(s,t),v,\bar{p}(s,t)).$$

Upon setting $\widetilde{\Sigma} = \bigcup_{\epsilon>0} \widetilde{\Sigma}_\epsilon$, we have $\mathcal{L}^N(\widetilde{\Sigma}) = \mathcal{L}^N(\Sigma)$. The pointwise Pontryagin's principle is satisfied on $\widetilde{\Sigma}$ and the proof is complete. □

5.3. Proof of Pontryagin principle in nonqualified form. In this case, as in [21], [31], [42], [26], and [12], we can consider the penalized functional

$$J_n(y,v,w) = \left\{ \left[\left(J(y,v,w) - J(\bar{y},\bar{v},\bar{w}) + \frac{1}{n^2} \right)^+ \right]^2 + (d_{\mathcal{C}}(\phi(y)))^2 \right\}^{1/2}.$$

With such a choice, for every $k > 0$, $(\bar{y},\bar{v},\bar{w})$ is a $\frac{1}{n^2}$-solution of the penalized problem

$$(P_n^k) \qquad \inf\{J_n(y,v,w) \mid (y,v,w) \in C(\overline{Q}) \times V_{ad}(\bar{v},k) \times W_{ad},\ (y,v,w) \text{ satisfies (1.1)}\}.$$

As in section 5.2, for every $k > 0$, we choose $n(k)$ such that

$$\frac{1}{n(k)} \le \frac{1}{k^{2\sigma}}.$$

Due to Ekeland's principle, there exists $(v_k, w_k) \in V_{ad}(\bar{v}, k) \times W_{ad}$ such that

$$d((v_k, w_k), (\bar{v}, \bar{w})) \leq \frac{1}{n(k)},$$

$$J_{n(k)}(y_k, v_k, w_k) \leq J_{n(k)}(y_{vw}, v, w) + \frac{1}{n(k)} d((v_k, w_k), (v, w))$$

for every $(v, w) \in V_{ad}(\bar{v}, k) \times W_{ad}$ (y_k is the solution of (1.1) corresponding to (v_k, w_k)).

With calculations similar to those in [26], [42], and [12], by using diffuse perturbations, we get

$$\int_\Sigma H(s,t,y_k(s,t),v_k(s,t),p_k(s,t),\nu_k)\,dsdt$$

$$\leq \int_\Sigma H(s,t,y_k(s,t),v_{0k}(s,t),p_k(s,t),\nu_k) + \frac{1}{n(k)}\mathcal{L}^N(\Sigma)$$

for every $k > 0$ and every $v_0 \in V_{ad}$ (v_{0k} is defined in function of v_0 in (5.5)) and

$$-\int_\Omega^{\nu_k} L'_w(x, y_k(T), w_k)(w_0 - w_k)\,dx - \langle p_k(0) + [\phi'_y(y_k)^* \mu_k]|_{\overline{\Omega}_0}, w_0 - w_k \rangle_{\mathcal{M}(\overline{\Omega}) \times C(\overline{\Omega})}$$

$$\leq \frac{1}{n(k)} \left(\frac{1}{n(k)} + \|w_0 - \bar{w}\|_{\infty,\Omega} \right)$$

for every $w_0 \in W_{ad}$, where

$$\nu_k = \frac{\left(J(y_k, v_k, w_k) - J(\bar{y}, \bar{v}, \bar{w}) + \frac{1}{n(k)^2} \right)^+}{J_{n(k)}(y_k, v_k, w_k)},$$

$$\mu_k = \begin{cases} \dfrac{d_{\mathcal{C}}(\phi(y_k)) \nabla d_{\mathcal{C}}(\phi(y_k))}{J_{n(k)}(y_k, v_k, w_k)} & \text{if } \phi(y_k)) \notin \mathcal{C}, \\ 0 & \text{otherwise,} \end{cases}$$

and p_k is the weak solution of

$$\begin{aligned} &-\frac{\partial p_k}{\partial t} + Ap_k + f'_y(x,t,y_k)p_k = \nu_k F'_y(x,t,y_k) + [\phi'_y(y_k)^* \mu_k]|_Q && \text{in } Q, \\ &\frac{\partial p_k}{\partial n_A} + g'_y(s,t,y_k,v_k)p_k = \nu_k G'_y(s,t,y_k,v_k) + [\phi'_y(y_k)^* \mu_k]|_\Sigma && \text{on } \Sigma, \\ &p_k(T) = \nu_k L'_y(x, y_k(T), w_k) + [\phi'_y(y_k)^* \mu_k]|_{\overline{\Omega}_T} && \text{in } \Omega. \end{aligned}$$

By passing to the limit when k tends to infinity, as in section 5.2, we finally get the Pontryagin principle in nonqualified form with $\bar{\nu} = \lim_k \nu_k$ and $\bar{\mu}$ the weak-star limit of μ_k. To prove that $(\bar{\nu}, \bar{\mu})$ is nonzero, we remark that $\nu_k^2 + |\mu_k|^2_{\mathcal{M}(\overline{D})} = 1$. If $\bar{\nu} > 0$, the proof is complete. If $\bar{\nu} = 0$, we can prove that $|\bar{\mu}|_{\mathcal{M}(\overline{D})} > 0$ by using $\lim_k |\mu_k|_{\mathcal{M}(\overline{D})} = 1$ and $\text{int}_{C(\overline{D})} \mathcal{C} \neq \emptyset$. Indeed, if $\text{int}_{C(\overline{D})} \mathcal{C}$ is nonempty, there exists a ball $B(z; \rho) \subset \mathcal{C}$ with $\rho > 0$ (where $B(z; \rho)$ is the ball in $C(\overline{D})$, centered at z and with

radius ρ). We can choose $z_k \in B(0;\rho)$ such that $\langle \mu_k, z_k \rangle_{\mathcal{M}(\overline{D})\times C(\overline{D})} = \frac{1}{2}\rho|\mu_k|_{\mathcal{M}(\overline{D})}$. Since $z + z_k \in \mathcal{C}$, from the definition of μ_k and from (5.2), we have

$$\langle \mu_k, z + z_k - \phi(y_k) \rangle_{\mathcal{M}(\overline{D})\times C(\overline{D})} \leq 0.$$

By passing to the limit, we obtain

$$\frac{1}{2}\rho + \langle \bar{\mu}, z - \phi(\bar{y}) \rangle_{\mathcal{M}(\overline{D})\times C(\overline{D})} \leq 0;$$

thus $\bar{\mu} \neq 0$.

Acknowledgments. We thank the Associate Editor and the referees for several suggestions which improved the language of the paper.

REFERENCES

[1] F. Abergel and R. Temam, *Optimality conditions for some nonqualified problems of distributed control,* SIAM J. Control Optim., 27 (1989), pp. 1–12.
[2] N. U. Ahmed and K. L. Teo, *Optimal Control of Distributed Parameter Systems,* Elsevier North–Holland, New York, 1981.
[3] J. J. Alibert and J. P. Raymond, *Boundary control of semilinear elliptic equations with discontinuous leading coefficients and unbounded controls,* Numer. Funct. Anal. Optim., 18 (1997), pp. 235–250.
[4] D. Azé, *Some remarks on duality in state-constrained optimal control problems governed by elliptic partial differential equations,* An. Stiint. Univ. Al. I. Cuza Iasi, to appear.
[5] D. Azé and S. Bolintinéanu, *A boundary control problem with parabolic dynamic and final boundary observation,* in Proc. 2nd Catalan Days on Applied Mathematics, M. Sofonea and J. N. Corvellec, eds., Presses Universitaires de Perpignan, Perpignan, France, 1995, pp. 12–25.
[6] V. Barbu, *Analysis and Control of Nonlinear Infinite Dimensional Systems,* Academic Press, New York, 1993.
[7] M. Bergounioux, *Optimal control of parabolic problems with state constraints: A penalization method for optimality conditions,* Appl. Math. Optim., 29 (1994), pp. 285–307.
[8] J. F. Bonnans and E. Casas, *An extension of Pontryagin's principle for state constrained optimal control of semilinear elliptic equations and variational inequalities,* SIAM J. Control Optim., 33 (1995), pp. 274–298.
[9] J. V. Burke, *Calmness and exact penalization,* SIAM J. Control Optim. 29 (1991), pp. 493–497.
[10] E. Casas, *Control of an elliptic problem with pointwise state constraints,* SIAM J. Control Optim., 24 (1986), pp. 1309–1318.
[11] E. Casas, *Boundary control of semilinear elliptic equations with pointwise state constraints,* SIAM J. Control Optim., 31 (1993), pp. 993–1006.
[12] E. Casas, *Pontryagin's principle for state-constrained boundary control problems of semilinear parabolic equations,* SIAM J. Control Optim., 35 (1997), pp. 1297–1327.
[13] E. Casas and J. Yong, *Maximum principle for state constrained optimal control problems governed by quasilinear elliptic equations,* Differential Integral Equations, 8 (1995), pp. 1–18.
[14] L. Cesari, *Optimization, Theory and Applications,* Springer-Verlag, New York, 1983.
[15] F. H. Clarke, *The maximum principle under minimal hypotheses,* SIAM J. Control Optim., 14 (1976), pp. 1078–1091.
[16] F. H. Clarke, *Optimization and Nonsmooth Analysis,* John Wiley, New York, 1983.
[17] E. DiBenedetto, *Degenerate Parabolic Equations,* Springer-Verlag, New York, 1993.
[18] J. Diestel, *Geometry of Banach Spaces—Selected Topics,* Lecture Notes in Math. 485, Springer-Verlag, Berlin, 1975.
[19] I. Ekeland, *On the variational principle,* J. Math. Anal. Appl., 47 (1974), pp. 324–353.
[20] I. Ekeland, *Nonconvex minimization problems,* Bull. Amer. Math. Soc. N.S., 1 (1979), pp. 443–474.
[21] H. O. Fattorini, *A unified theory of necessary conditions for nonlinear nonconvex control systems,* Appl. Math. Optim., 15 (1987), pp. 141–185.

[22] H. O. Fattorini, *Optimal control problems with state constraints for semilinear distributed-parameter systems,* J. Optim. Theory Appl., 88 (1996), pp. 25–59.

[23] H. O. Fattorini and T. Murphy, *Optimal control problems for nonlinear parabolic boundary control systems: The Dirichlet boundary condition*, Differential Integral Equations, 7 (1994), pp. 1367–1388.

[24] H. O. Fattorini and T. Murphy, *Optimal problems for nonlinear parabolic boundary control systems,* SIAM J. Control Optim., 32 (1994), pp. 1577–1596.

[25] H. O. Fattorini and S. Sritharan, *Necessary and sufficient conditions for optimal controls in viscous flow problems,* Proc. Roy. Soc. Edinburgh Sect. A, 124 (1994), pp. 211–251.

[26] B. Hu and J. Yong, *Pontryagin maximum principle for semilinear and quasilinear parabolic equations with pointwise state constraints,* SIAM J. Control Optim., 33 (1995), pp. 1857–1880.

[27] O. A. Ladyzenskaja, V. A. Solonnikov, and N. N. Ural'ceva, *Linear and Quasilinear Equations of Parabolic Type,* Transl. Math. Monographs 23, AMS, Providence, RI, 1968.

[28] X. J. Li, *Vector-valued measure and the necessary conditions for the optimal control problems of linear systems,* J. Math. Res. Exposition, 4 (1984), pp. 51–56.

[29] X. J. Li and S. N. Chow, *Maximum principle of optimal control for functional differential systems,* J. Optim. Theory Appl., 54 (1987), pp. 335–360.

[30] X. J. Li and Y. Yao, *Maximum principle of distributed parameter systems with time lags,* in Proc. Conference on Control Theory of Distributed Parameter Systems and Applications, F. Kappel and K. Kunish, eds., Springer-Verlag, New York, 1985, pp. 410–427.

[31] X. J. Li and J. Yong, *Necessary conditions for optimal control of distributed parameter systems,* SIAM J. Control Optim., 29 (1991), pp. 895–908.

[32] X. J. Li and J. Yong, *Optimal Control Theory for Infinite Dimensional Systems,* Birkhäuser, Basel, 1995.

[33] U. Mackenroth, *Convex parabolic boundary control problems with pointwise state constraints,* J. Math. Anal. Appl., 87 (1982), pp. 256–277.

[34] J. P. Raymond, *Nonlinear boundary control of semilinear parabolic equations with pointwise state constraints,* Discrete Continuous Dynam. Systems, 3 (1997), pp. 341–370.

[35] J. P. Raymond, *Optimal control problem governed by a semilinear parabolic equation with pointwise state constraints,* in Modelling and Optimization of Distributed Parameter Systems with Applications to Engineering, K. Malanowski et al., eds., Chapman and Hall, London, 1996, pp. 216–222.

[36] J. P. Raymond, *Pontryagin's principle for state-constrained control problems with unbounded controls,* in Proc. 2nd Catalan Days on Applied Mathematics, M. Sofonea and J. N. Corvellec, eds., Presses Universitaires de Perpignan, Perpignan, France, 1995, pp. 227–237.

[37] J. P. Raymond and H. Zidani, *Hamiltonian Pontryagin's principles for control problems governed by semilinear parabolic equations,* Appl. Math. Optim., to appear.

[38] J. P. Raymond and H. Zidani, *Optimal control problem governed by a semilinear parabolic equation,* in System Modelling and Optimization, J. Dolezal and J. Fidler, eds., Chapman and Hall, 1996, pp. 211–217.

[39] F. Tröltzsch, *Optimality Conditions for Parabolic Control Problems and Applications,* Teubner-Texte zür Mathematik 62, B.G. Teubner Verlagsgesellshaft, Leipzig, 1984.

[40] Y. Yao, *Vector measure and maximum principle of distributed parameter systems,* Sci. Sinica Ser., 26 (1983), pp. 102–112.

[41] Y. Yao, *Maximum principle of semi-linear distributed systems,* in Proc. 3rd IFAC Symposium on the Control of Distributed Parameter Systems, Toulouse, France, 1982.

[42] J. Yong, *Pontryagin maximum principle for semilinear second order elliptic partial differential equations and variational inequalities with state constraints,* Differential Integral Equations, 5 (1992), pp. 1307–1334.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 6, pp. 1880–1889, November 1998

002

RENDEZVOUS ON THE LINE WHEN THE PLAYERS' INITIAL DISTANCE IS GIVEN BY AN UNKNOWN PROBABILITY DISTRIBUTION*

VIC BASTON† AND SHMUEL GAL‡

Abstract. Two players A and B are randomly placed on a line. The distribution of the distance between them is unknown except that the expected initial distance of the (two) players does not exceed some constant μ. The players can move with maximal velocity 1 and would like to meet one another as soon as possible. Most of the paper deals with the asymmetric rendezvous in which each player can use a different trajectory. We find rendezvous trajectories which are efficient against all probability distributions in the above class. (It turns out that our trajectories do not depend on the value of μ.) We also obtain the minimax trajectory of player A if player B just waits for him. This trajectory oscillates with a geometrically increasing amplitude. It guarantees an expected meeting time not exceeding 6.8μ. We show that, if player B also moves, then the expected meeting time can be reduced to 5.7μ.

The expected meeting time can be further reduced if the players use mixed strategies. We show that if player B rests, then the optimal strategy of player A is a mixture of geometric trajectories. It guarantees an expected meeting time not exceeding 4.6μ. This value can be reduced even more (below 4.42μ) if player B also moves according to a (correlated) mixed strategy. We also obtain a bound for the expected meeting time of the corresponding symmetric rendezvous problem.

Key words. rendezvous, linear search

AMS subject classifications. 90B40, 90D05, 90D12, 90D26

PII. S0363012996314130

1. Introduction.

Rendezvous search is a form of cooperative optimal search in which two or more people wish to meet as quickly as possible. Although the topic was discussed by Schelling in a book published in 1960 [15], it was not until 1994 that rendezvous search was formulated into a rigorous mathematical manner in a seminal paper by Alpern [1]. Schelling's discussion concentrates on *focal points* to which the players would move, whereas Alpern places emphasis on the symmetries of the region X in which the people are situated and formalizes the problem by specifying a particular subgroup of the isomorphism group of X. (These symmetries reflect the knowledge of the players about their location and movement in the region.) The field is an attractive one because many of the problems are mathematically challenging even though they are simple to state and understandable to a nonmathematician. Anderson and Weber [4] investigated rendezvous search on a complete graph in 1990 but a solution of the problem for a complete graph with 4 or more vertices has still not been found. Many one-dimensional rendezvous search problems also remain unsolved, and our purpose in this paper is to present some results on problems on the line. The main problem that we will investigate is the linear rendezvous problem which was described by Alpern in [1] as follows.

*Received by the editors December 27, 1996; accepted for publication (in revised form) November 24, 1997; published electronically July 17, 1998.
http://www.siam.org/journals/sicon/36-6/31413.html
†Faculty of Mathematical Studies, University of Southampton, Southampton SO9 5NH, UK (vjdb@maths.soton.ac.uk).
‡Department of Statistics, University of Haifa, Mount Carmel, Haifa 31905, Israel (rsst501@uvm.haifa.ac.il).

> Two friends have agreed to meet at noon on a certain street but have neglected to specify a specific point on the street. Assuming they know the distribution of their arrival points on the street at noon, how should they move to meet in minimum expected time?

It can be formulated mathematically in the following way.

At time $t = 0$ two players are placed on a line at a distance d from each other, where d is chosen by means of a cumulative distribution function F with support in $[0, \infty)$. The players know F but not d and neither player knows the direction of the other (nor do they have a common notion of positive direction). They can move with a maximum speed of 1 and wish to meet up with each other in the shortest possible expected time (called the *rendezvous value*).

There are in fact several different rendezvous values depending on which search strategies are permitted. We shall be primarily concerned with *asymmetric* problems in which the players are distinguishable and so are able to adopt different strategies. A strategy α is any continuous trajectory with speed not exceeding 1. A player placed at point d will have the time paths $d \pm \alpha(t)$ equiprobably. We choose the coordinate system such that the first player, say A, starts at 0 and the other, say B, starts equiprobably at $\pm d$, where d is drawn from F. If player A uses trajectory α, and B trajectory β, then their expected meeting time is the average of the meeting time of player A following trajectory α with 4 agents (representing B) following the paths: $\pm d \pm \beta(t)$; we will adopt the convention that agents 1, 2, 3, and 4 follow the paths $+d - \beta(t)$, $+d + \beta(t)$, $-d + \beta(t)$, and $-d - \beta(t)$, respectively. The *asymmetric* rendezvous value is the minimal expected time achievable by the players.

Alpern and Gal [2] obtained a number of results when the support of F is bounded or discrete and found the asymmetric rendezvous value for the case when the support of F is a single point. They also pointed out that the asymmetric rendezvous problem is closely related to the linear search problem (searching for a stationary target on the line) which was introduced by Bellman [9] and extensively studied by Anatole Beck and others (the latest references are [6] and [7]). In particular they found inequalities connecting the two problems. Note that a rendezvous search problem on the line in which one of the players has to remain stationary is just a symmetric linear search problem. (A linear search problem in which the distribution of the target is symmetric around the origin.) This problem is equivalent to finding the expected meeting time of the *wait for mummy* strategy in which one player stays in his original location and the other looks for him in an optimal way.

In contrast to the work in [2], we shall deal with general distribution functions having support in $[0, \infty)$. We consider the case where the distribution F is unknown, with the only information being that $E(d) \leq \mu$. (Such an approach has been used by Beck and Newman [8] for the linear search problem.) Our results will provide upper bounds for rendezvous values in the form $K\mu$, where K is a universal constant and μ is the mean of F. However, it should be noted that our analysis does not require the value of μ to be known. The work has a natural interpretation as a "game against nature" and can also be viewed as a version of Alpern's "adversary-rendezvous game" [1] in which the initial placement of the players is made by an opponent who wishes to maximize their rendezvous time.

In section 2 we find the minimax solution for the symmetric linear search problem. Work on the linear search problem has concentrated primarily on the searcher adopting a *pure* search strategy; we conclude the section by demonstrating that, when this restriction is removed and we allow the use of mixed strategies, the optimal so-

lution yields an expected meeting time which is significantly lower. The next section presents a strategy pair which yields an upper bound of at most 5.73μ for the asymmetric rendezvous value; in this strategy pair the players have symmetric roles in which one player takes a turn moving while the other remains stationary. The concept of correlated strategies is well established in game theory and it has recently been introduced into rendezvous search by Lim [13]. In section 4 we prove that the asymmetric rendezvous value is at most 4.42μ when the players are permitted to use correlated strategies. Symmetric rendezvous problems (ones in which both players have to play the same strategy) appear to be more difficult to analyze than symmetric ones; the symmetric problem on the line when the support of F comprises a single point (see [3]) remains unsolved, whereas the corresponding asymmetric problem has a comparatively simple solution [2]. In section 5 we find an upper bound for the symmetric rendezvous value. Many attractive problems remain open and the paper concludes with two conjectures.

2. The symmetric linear search problem. When players can adopt different strategies in trying to meet up with each other, an obvious plan is for one of them to remain stationary while the other one tries to find him. In this case the rendezvous search problem reduces to a linear search problem in which the distribution function is symmetric. The following theorem gives a best possible result for this case.

THEOREM 2.1. *For the symmetric linear search problem, there is a trajectory such that the expected meeting time is at most* $(4+2\sqrt{2})\mu \approx 6.83\mu$, *where* μ *is the mean of* F.

Furthermore, this result is best possible in the sense that the constant cannot be lowered. Moreover, the minimax search trajectory is unique up to a multiplicative constant.

Proof. Adopting a coordinate system which has the searcher's starting point as origin, the proofs of Theorem 3.1 and Lemma 3.2 in [7] show that it is sufficient to consider a search strategy of the form $((-1)^i x_i)$, where $x_i \geq 0$ and $x_i < x_{i+1}$ whenever $x_i > 0$; this search strategy represents a path beginning at 0 and consisting of the intervals $\cdots, [-x_{2i-1}, x_{2i}]$ traversed in the positive direction followed by $[-x_{2i+1},\ x_{2i}]$ traversed in the negative direction, $\cdots$. Put $s_i = \sum_{j=-\infty}^{i} x_j$, then the expected meeting time $L(F)$ is given by

$$\begin{aligned} L(F) &= \sum_{i=-\infty}^{\infty} \frac{1}{2}\int_{x_i}^{x_{i+1}} (2s_{i+1}+t)\,dF(t) + \frac{1}{2}\int_{x_i}^{x_{i+1}} (2s_i+t)\,dF(t) \\ &= \mu + \sum_{i=-\infty}^{\infty} \int_{x_i}^{x_{i+1}} (s_i+s_{i+1})\,dF(t). \end{aligned}$$

Let

$$x_i = (1+\sqrt{2})^i, \tag{2.1}$$

then $s_i = (1+\sqrt{2})x_i/\sqrt{2}$ and

$$\begin{aligned} \sum_{i=-\infty}^{\infty} \int_{x_i}^{x_{i+1}} (s_i+s_{i+1})\,dF(t) &= \sum_{i=-\infty}^{\infty} \int_{x_i}^{x_{i+1}} x_i(1+\sqrt{2})(2+\sqrt{2})/\sqrt{2}\,dF(t) \\ &= (3+2\sqrt{2}) \sum_{i=-\infty}^{\infty} \int_{x_i}^{x_{i+1}} x_i\,dF(t) \leq (3+2\sqrt{2})\mu, \end{aligned}$$

and the first part of the result follows.

We now show it is best possible. Let $\lambda = \sup_{x_j \neq 0} x_{j+1}/x_j$; we will only consider the case where λ is finite because the case $\lambda = \infty$ follows in an analogous manner. For each $\epsilon > 0$, we can choose a k such that $x_{k+1}/x_k > \lambda - \epsilon$. Putting $r_j = x_j/x_{j+1}$ if $x_j \neq 0$ and $r_j = 0$ otherwise, we have

$$\begin{aligned}\frac{s_k + s_{k+1}}{x_k} &= r_k^{-1} + 2(1 + r_{k-1} + r_{k-1}r_{k-2} + r_{k-1}r_{k-2}r_{k-3} + \cdots) \\ &> \lambda - \epsilon + 2(1 + \lambda^{-1} + \lambda^{-2} + \lambda^{-3} + \cdots) = \lambda - \epsilon + 2\lambda/(\lambda - 1).\end{aligned}$$

It is routine to check that the right-hand side is minimized as a function of λ by $\lambda = 1 + \sqrt{2}$, so we have

$$s_k + s_{k+1} \geq (3 + 2\sqrt{2})x_k.$$

Let F_η be the distribution function given by 0 if $t < x_k + \eta$, and 1 otherwise; then, by taking $\eta > 0$ sufficiently small, $L(F_\eta)$ can be made as near as we please to $(4+2\sqrt{2})\mu$. Hence the asserted constant is the best possible and the second part of the theorem is established.

The above result can also be proved using a general theorem (Theorem 7 in Chapter 6.5 of [11]). Moreover, it also follows from the general theorem that any trajectory with $s_k + s_{k+1} \leq (3 + \sqrt{2})x_k$ is equal to (2.1) up to a multiplicative constant.

Note that the corresponding value for the (usual) linear search problem is 9μ (see [8]).

COROLLARY 2.2. *For all distribution functions with support in* $[0, \infty)$, *the asymmetric linear rendezvous value is at most* $(4 + 2\sqrt{2})\mu$, *where* μ *is the mean of* F.

Proof. From Theorem 2.1 such a value can be achieved by a "wait for mummy" strategy.

THEOREM 2.3. *When the searcher can use a mixed strategy, an upper bound for the expected meeting time in the symmetric linear search problem is*

$$\left(1 + \frac{g^* + 1}{\ln g^*}\right)\mu,$$

where μ *is the mean of* F *and* g^* *is the value of* g *which minimizes* $(g + 1)/\ln g$; *the upper bound is approximately* 4.6μ *and* g^* *approximately* 3.6.

This is a best possible result in the sense that the constant cannot be lowered.

Proof. Let S be the searcher strategy given by (x_i), where $x_i = (-1)^i g^{i+u}$ and u is a random variable uniformly distributed in [0,1). For a distance x and a fixed trajectory, let $D = D(x)$ be the first geometric term numerically greater than or equal to x; then the expected time of the searcher to find an object which is initially at distance x from the searcher's starting point is given by

$$\frac{1}{2}\left(\frac{2D}{g-1} + x\right) + \frac{1}{2}\left(\frac{2Dg}{g-1} + x\right) = x + \frac{D(g+1)}{g-1}.$$

Put $x = g^{i+v}$, where $0 \leq v < 1$. Then

$$D = \begin{cases} g^{i+u} & \text{if } v \leq u < 1, \\ g^{i+1+u} & \text{if } 0 \leq u < v. \end{cases}$$

Hence the expected time for strategy S to reach x is

$$x+\frac{g+1}{g-1}\left(\int_0^v g^{i+1+u}du+\int_v^1 g^{i+u}du\right)$$
$$=x+\frac{g+1}{(g-1)\ln g}\left(g^{i+1}(g^v-1)+g^i(g-g^v)\right)=x+\frac{(g+1)g^{i+v}}{\ln g}=x\left(1+\frac{g+1}{\ln g}\right).$$

Thus S yields an expected meeting time of at most

$$\mu\left(1+\frac{g+1}{\ln g}\right).$$

The minimum of $(g+1)/\ln g$ is attained when g is approximately 3.6 which gives an upper bound of at most 4.6μ.

Note that this is the same result as for the (usual) linear search problem (see [8]). This is not surprising because the worst (actually ϵ-worst) distribution for the linear search problem is symmetric. This distribution can keep the expected meeting time for both problems to at least $(1+(g^*+1)/\ln g^*)\mu$ which shows that this constant cannot be lowered and that the strategy described in the proof is optimal.

3. The asymmetric linear rendezvous problem. It is not easy to improve the bound in Corollary 2.2. If the other player, say B, is also moving, then any reduction in the meeting time of player A and agent 1 also increases the meeting time with agent 2. Similarly, meeting agent 3 earlier causes player A to meet agent 4 later and vice versa.

THEOREM 3.1. *The asymmetric linear rendezvous value is at most* 5.73μ*, where* μ *is the mean of* F.

Proof. We describe the trajectories of the players in time intervals $[\rho_{2i},\rho_{2i+2}]$, where $-\infty<i<\infty$ and $\rho_i=(g^2+1)g^{i-1}/(g-1)$. Note that $\rho_{i+1}=\rho_i+g^{i+1}+g^{i-1}$.

In the interval $[\rho_{2i},\rho_{2i+1}]$, player A remains stationary at a point g^{2i} from his starting point while player B moves at speed 1 from a point g^{2i-1} from his starting point to the point g^{2i+1} on the opposite side of his starting point.

In the interval $[\rho_{2i+1},\rho_{2i+2}]$, player B remains stationary at a point g^{2i+1} from his starting point while player A moves at speed 1 from a point g^{2i} from his starting point to the point g^{2i+2} on the opposite side of his starting point.

Suppose the players start at distance x apart, where $g^{i-1}<x\le g^i$ and $g\ge(1+\sqrt{5})/2$. At time $\rho_{i-1}+g^{i-2}$, one player, say A, is back at his starting point and the other, say B, is at a distance g^{i-1} from his starting point. Take a coordinate system with origin being the starting point of A and positive direction being the direction of A's motion in $[\rho_{i-1},\rho_i]$; then the agents 1, 2, 3, and 4 of player B are located at the points $x-g^{i-1}$, $x+g^{i-1}$, $-x+g^{i-1}$, and $-x-g^{i-1}$, respectively. Note that the agents of B remain stationary in $[\rho_{i-1}+g^{i-2},\rho_i]$.

Agent 1. Player A will meet agent 1 at time $\rho_{i-1}+g^{i-2}+x-g^{i-1}<\rho_i$ and so at time

$$x+\frac{2g^{i-1}}{g-1}.$$

Agent 2. Player A will meet agent 2 at time $\rho_{i-1}+g^{i-2}+x+g^{i-1}\le\rho_i$ if $x\le g^i-g^{i-1}$ and at time $\rho_i+x+g^{i-1}-g^i<\rho_{i+1}$ if $x>g^i-g^{i-1}$. Thus they meet

at time

$$x + \frac{2g^i}{g-1}.$$

Agent 3. Player A will meet agent 3 at time $\rho_{i-2} + g^{i-3} + x - g^{i-2} \le \rho_{i-1}$ if $x \le g^{i-1} + g^{i-2}$ and at time $\rho_{i+1} + x + g^i + g^{i+1} < \rho_{i+2}$ if $x > g^{i-1} + g^{i-2}$ and $x \le g^i(g^2 - g)$. The last inequality holds because $g \ge (1+\sqrt{5})/2$, and so they meet at time

$$x + \frac{2}{g-1}\begin{cases} g^{i-2} & \text{if } x \le g^{i-1} + g^{i-2}, \\ g^{i+2} & \text{if } x > g^{i-1} + g^{i+2}. \end{cases}$$

Agent 4. Player A will meet agent 4 at time $\rho_i + g^{i-1} + x + g^i \le \rho_{i+1}$ if $x \le g^{i+1} - g^i$ and at time $\rho_{i+1} + x + g^i - g^{i+1}$ if $x > g^{i+1} - g^i$. Thus they meet at time

$$x + \frac{2g^{i+1}}{g-1}.$$

Hence the expected meeting time for the players when they start at distance x apart is at most

$$x + \frac{(g+1)(g^2+1)}{2(g-1)}\begin{cases} g^{i-2} & \text{if } x \le g^{i-1} + g^{i-2} \\ g^{i-1} & \text{if } x > g^{i-1} + g^{i-2} \end{cases}$$

and so at most

$$x + \frac{g^2+1}{2(g-1)} \max\left\{ gx, \frac{(g+1)x}{g} \right\} = x\left\{ 1 + \frac{g(g^2+1)}{2(g-1)} \right\}$$

because $(1+\sqrt{5})/2 \le g$. The minimum occurs when $2g + 1 = 2/(g-1)^2$, and taking $g = 1.6775$ gives a bound slightly less than 5.73.

It is easy to see that, for a particular distribution function, the expected meeting time is not increased if player B continues moving after ρ_{2i+1}, provided he arranges to be stationary at the point g^{2i+1} from his starting point during the time interval $[\rho_{2i+1} + g^{2i+1} + g^{2i-1}, \rho_{2i+2}]$. A corresponding comment holds for player A in the time interval $[\rho_{2i}, \rho_{2i+1}]$. Although this modification does not result in an improvement in our bound, it does suggest that the bound might be lowered by using strategies in which players move simultaneously for at least some of the time. Note that, when the support of F is a single point, the minimum expected meeting time is achieved by strategies in which the players are always moving; furthermore, the trajectories of the players are not symmetrical.

How much can the bound be reduced? The problem seems difficult but we can get an idea by looking at the point distribution considered in [2]. There, the expected meeting time is 2μ for the wait for mummy strategy and $13\mu/8$ for the optimum, i.e., a reduction of nearly 19%. Such an extrapolation leads to about 5.55μ in our case. While this is in no way a precise argument it may hint that the value 5.73μ we obtained in Theorem 3.1 is not far from the optimum.

4. The correlated asymmetric linear rendezvous problem. Aumann [5] has pointed out that, in game theory, it is often important whether players can use the same randomizing device for their strategies; if they can, the resulting strategies

are called correlated strategies and lead to the concept of correlated equilibria. The idea has proved fruitful and is now covered in standard textbooks on game theory ([10], [14]). Lim [13] recently introduced the concept into rendezvous search, and our analysis will address the simple case in which both players observe the same random variable and know the realization value. Note that Theorem 2.3 provides us with an upper bound for the correlated asymmetric linear rendezvous value where the random variable is uniformly distributed in [0,1]. The next theorem shows that this bound can be improved.

THEOREM 4.1. *Correlated strategies can reduce the expected meeting time of the asymmetric linear rendezvous below* 4.42μ, *where* μ *is the mean of* F.

Proof. Let $g > 1$ be fixed and let u be the realization value (known to both players) of a random variable uniformly distributed in [0,1]. Put $D(i) = g^{i+u}$. We describe the trajectories of the players in the time interval $[\sigma_{i-1}, \sigma_i]$, where $\sigma_i = (g+1)D(i)/(g-1)$.

At time σ_{i-1}, player A is at distance $D(i-1)$ from his starting point and in the interval moves at speed 1 to the point $D(i)$ on the opposite side of his starting point.

Player B chooses a direction at random as forward at $t = 0$ and is at his starting point at time σ_{i-1}; at speed 1 he moves forward, then backward, then forward, then backward for times $D(i-1)$, $D(i-1)$, $(D(i) - D(i-1))/2$, $(D(i) - D(i-1))/2$, respectively.

Note that the movement in this time interval corresponds to an optimal trajectory under the assumption that the initial distance is $D(i)$.

Suppose the players start at distance x apart where $D(i-1) < x \le D(i)$. With the given trajectories, player A will meet two of player B's agents in $[\sigma_{i-1}, \sigma_i]$ at times

$$\sigma_{i-1}+2D(i-1)+\frac{x-D(i-1)}{2} \quad \text{and} \quad \sigma_{i-1}+2D(i-1)+\frac{D(i)-D(i-1)}{2}+\frac{x-D(i-1)}{2}$$

and the other two in $[\sigma_i, \sigma_{i+1}]$ at times

$$\sigma_i + (D(i)+x)/2 \quad \text{and} \quad \sigma_i + D(i) + (D(i)+x)/2;$$

since $\sigma_i = \sigma_{i-1} + D(i-1) + D(i)$, the expected time of meeting is therefore

$$\sigma_{i-1} + x/2 + 9(D(i-1) + D(i))/8 = x/2 + D(i-1)\left\{\frac{2}{g-1} + \frac{17}{8} + \frac{9g}{8}\right\}.$$

Hence if the players start at distance d apart, where $d = g^{i+a}$ and $0 \le a \le 1$, the expected meeting time is

$$\frac{d}{2}+\left\{\frac{2}{g-1}+\frac{17}{8}+\frac{9g}{8}\right\}\left\{\int_0^a g^{i+u}\,du+\int_a^1 g^{i-1+u}\,du\right\} = \frac{d}{2}+\left\{\frac{2}{g-1}+\frac{17}{8}+\frac{9g}{8}\right\}\frac{g-1}{g\ln g}d.$$

On minimizing with respect to g, the optimal g is approximately 3.5, giving a bound of less than $4.42d$, and the result follows.

5. Symmetric rendezvous on the line. We now turn to the case where the players have to use the same strategy. Let g and a be positive constants. Consider the situation in which the players adopt a (mixed) strategy which has the properties (i)–(iii) listed below in time intervals $[\tau_n, \tau_{n+1}]$, where $\tau_n = ag^n/(g-1)$ and $-\infty < n < \infty$. Assume that the players have not met up to and including time τ_n and that, at time τ_n, they are in the same relative position to each other as they were at the start.

(i) If the players do not meet in $[\tau_n, \tau_{n+1}]$ then, at time τ_{n+1}, they are also in the same relative position to each other as they were at the start.

(ii) If their distance apart at time τ_n is greater than g^n, the players do not meet in $[\tau_n, \tau_{n+1}]$.

(iii) If their distance apart at time τ_n is at most g^n, then there is a probability $\rho > 0$ (a pure constant) that the players will meet in $[\tau_n, \tau_{n+1}]$; if they meet in $[\tau_n, \tau_{n+1}]$, their expected meeting time is $\tau_n + x/2 + \nu g^n$, where ν is a pure constant.

Suppose the players start at distance x apart, where $x \in (g^{n-1}, g^n]$. By property (ii), the players do not meet in any of the time intervals $[\tau_r, \tau_{r+1}]$ for $r \le n-1$. By property (iii), for $r \ge n$, there is a probability $(1-\rho)^{r-n}\rho$ that the players will meet for the first time in $[\tau_r, \tau_{r+1}]$. Hence provided $(1-\rho)g < 1$, the expected meeting time is

$$\sum_{r=n}^{\infty}(1-\rho)^{r-n}\rho\left\{\frac{ag^r}{g-1}+\frac{x}{2}+\nu g^r\right\} = \frac{x}{2}+\frac{\rho(a/(g-1)+\nu)}{(1-\rho)^n}\sum_{r=n}^{\infty}(1-\rho)^r g^r$$
$$= \frac{x}{2}+\frac{\rho(a/(g-1)+\nu)g^n}{1-(1-\rho)g} \le \frac{x}{2}+\frac{\rho(a/(g-1)+\nu)}{1-(1-\rho)g}gx.$$

The last expression tends to infinity as $g \to 1$ from the right and as $g \to 1/(1-\rho)$ from the left, so it has a minimum g^* in $(1, (1-\rho)^{-1})$; note that $g^*(1-\rho) < 1$. The turning points are at

$$g = \frac{\nu \pm \sqrt{a^2+\nu a\rho - a^2\rho}}{\nu + a - a\rho} \quad \text{so} \quad g^* = \frac{\nu+\sqrt{a^2+\nu a\rho-a^2\rho}}{\nu+a-a\rho}.$$

We use this analysis to prove the following theorem.

THEOREM 5.1. *The symmetric linear rendezvous value is at most* $(7+2\sqrt{10})\mu$, *where* μ *is the mean of* F.

Proof. We describe the trajectories of the players in the time interval $[\tau_n, \tau_{n+1}]$, where $\tau_n = 2g^n/(g-1)$.

At time τ_n a player is at his starting point and chooses a direction at random as forward; he then moves at speed 1 forward, then backward, then forward for times $g^n/2,\ g^n,\ g^n/2$, respectively.

Clearly the players meet in $[\tau_n, \tau_{n+1}]$ if and only if they start at distance $x \le g^n$ apart and they choose opposite directions as forward. Thus for such x, they meet with probability $1/2$ and the expected meeting time is $\tau_n + x/2 + g^n/2$. Hence we have a special case of the above analysis with $a = 2,\ \rho = 1/2 = \nu$, so $g^* = (1+\sqrt{10})/3$ giving the asserted bound of $(7+2\sqrt{10})\mu$; the value of g is approximately 1.39 and the bound slightly less than 13.33μ.

THEOREM 5.2. *Correlated strategies can reduce the expected meeting time of the symmetric linear rendezvous below* 11.4μ, *where* μ *is the mean of* F.

Proof. Let $g > 1$ be fixed and let u be the realization value (known to both players) of a random variable uniformly distributed in [0,1]. Put $D(i) = g^{i+u}$. We describe the trajectories of the players in the time interval $[\eta_n, \eta_{n+1}]$, where $\eta_n = 2D(n)/(g-1)$.

At time η_n a player is at his starting point and chooses a direction at random as forward; he then moves at speed 1 forward, then backward, then forward for times $D(n)/2,\ D(n),\ D(n)/2$, respectively.

Suppose, at time $t = 0$, the players are at distance x apart, where $D(i-1) < x \le D(i)$, then, from our earlier analysis in this section, their expected meeting time is

$$\frac{x}{2}+\frac{1/(g-1)+(1/4)}{1-g/2}D(i) = \frac{x}{2}+\frac{g+3}{2(2-g)(g-1)}D(i).$$

Hence if the players start at distance d apart, where $d = g^{i+a}$ and $0 \le a \le 1$, the expected meeting time is

$$\frac{d}{2} + \frac{g+3}{2(2-g)(g-1)}\left\{\int_0^a g^{i+u+1}du + \int_a^1 g^{i+u}du\right\} = \frac{d}{2} + \frac{g+3}{2(2-g)\ln g}d.$$

The last expression is minimized in [1,2] at approximately 1.43, giving a value for the expression slightly less than 10.9μ. The theorem now follows.

6. Conclusions. Although our bounds are given in the form $K\mu$, where μ is the mean of the distribution function, we do not need to know the value of μ in the analysis of any of our cases. Apart from Theorems 2.1 and 2.3, it would be surprising if our results were best possible. Many attractive problems remain, especially finding the minimax trajectories and optimal correlated strategies for the linear rendezvous, and we now formulate two conjectures.

The *greedy strategy pair of length* D is given by the following:

• At time $t = 0$ player A chooses a direction as forward, then moves at speed 1 forward for a time D and then backward for a time $2D$;

• At time $t = 0$ player B chooses a direction as forward, then moves at speed 1 forward, then backward, then forward, then backward for times $D/2$, $D/2$, D, and D, respectively.

The motions of the players after time $3D$ can be defined arbitrarily.

The name "greedy" originates from the fact that, if the players start at distance D apart, then at time $t = 0$ player A and agent 1 move at maximum speed towards each other; after they meet then player A and agent 2 move at maximum speed towards each other and so on.

Note that the greedy strategy pair of length D is optimal for the asymmetric linear rendezvous search problem where the players know they start at distance D apart.

Note that in a time interval $(t, t + \delta t)$ the players gain by **both** moving if the distribution of their distance is increasing in $(0, \delta t)$. If it is decreasing, then it is better for one of them to remain stationary while the other player moves. This fact has been observed in [12] (which also showed that "wait for mummy" is never optimal). Thus the following conjectures seem reasonable.

CONJECTURE 6.1. *Let F be a distribution function with support in $[0, D]$ which has a density function that is nondecreasing in $[0, D]$; then the greedy strategy pair of length D is optimal for the asymmetric linear rendezvous search problem having distribution function F.*

CONJECTURE 6.2. *Let F be a distribution function with support in $[0, D]$ which has a density function that is strictly decreasing in $[0, D]$; then there is an optimal strategy pair for the asymmetric linear rendezvous search problem which takes the following form:*

• *player A oscillates with speed 1;*

• *player B is stationary while player A discovers new points and moves at speed 1 forward, then backward for times $t/2$ and $t/2$ during any period of length t in which player A does not discover new points.*

REFERENCES

[1] S. ALPERN, *The rendezvous search problem*, SIAM J. Control Optim., 33 (1995), pp. 673–683.

[2] S. Alpern and S. Gal, *The rendezvous search problem on the line with distinguishable players*, SIAM J. Control Optim., 33 (1995), pp. 1270–1276.
[3] E. J. Anderson and S. Essegaier, *Rendezvous search on the line with distinguishable players*, SIAM J. Control Optim., 33 (1995), pp. 1637–1642.
[4] E. J. Anderson and R. R. Weber, *The rendezvous problem on discrete locations*, J. Appl. Probab., 28 (1990), pp. 839–851.
[5] R. F. Aumann, *Subjectivity and correlation in randomized strategies*, J. Math. Econom., 1 (1974), pp. 67–96.
[6] V. J. Baston and A. Beck, *Generalizations in the linear search problem*, Israel J. Math., 90 (1995), pp. 301–323.
[7] A. Beck and M. Beck, *The revenge of the linear search problem*, SIAM J. Control Optim., 30 (1992), pp. 112–122.
[8] A. Beck and D. J. Newman, *Yet more on the linear search problem*, Israel J. Math., 8 (1970), pp. 419–429.
[9] R. Bellman, *An optimal search problem*, SIAM Rev., 5 (1963), p. 274.
[10] D. Fudenberg and J. Tirole, *Game Theory*, The MIT Press, Cambridge, MA, 1991.
[11] S. Gal, *Search Games*, Academic Press, New York, 1980.
[12] S. Gal, *Rendezvous search on the line*, Oper. Res., to appear.
[13] W. S. Lim, *A rendezvous-evasion game on discrete locations with joint randomization*, Adv. in Appl. Probab., 29 (1997), pp. 1004–1017.
[14] M. Osborne and A. Rubinstein, *A Course in Game Theory*, The MIT Press, Cambridge, MA, 1991.
[15] T. C. Schelling, *The Strategy of Conflict*, Harvard University Press, Cambridge, MA, 1960.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 6, pp. 1890–1907, November 1998

003

ADAPTIVE LINEAR QUADRATIC GAUSSIAN CONTROL: THE COST-BIASED APPROACH REVISITED*

M. C. CAMPI† AND P. R. KUMAR‡

Abstract. In adaptive control, a standard approach is to resort to the so-called certainty equivalence principle which consists of generating some standard parameter estimate and then using it in the control law as if it were the true parameter. As a consequence of this philosophy, the estimation problem is decoupled from the control problem and this substantially simplifies the corresponding adaptive control scheme. On the other hand, the complete absence of dual properties makes certainty equivalent controllers run into an identifiability problem which generally leads to a strictly suboptimal performance.

In this paper, we introduce a cost-biased parameter estimator to overcome this difficulty. This estimator is applied to a linear quadratic Gaussian controller. The corresponding adaptive scheme is proven to be stable and optimal when the unknown system parameter lies in an infinite, yet compact, parameter set.

Key words. adaptive control, linear quadratic Gaussian control, self-optimizing control, cost-biased approach, certainty equivalence

AMS subject classifications. 15A15, 15A09, 15A23

PII. S0363012997317499

1. Introduction.

Consider a linear time-invariant system

$$x_{t+1} = A^\circ x_t + B^\circ u_t + w_{t+1}, \tag{1}$$

where $x_t \in \mathbf{R}^n$ is the state, $u_t \in \mathbf{R}^m$ the control variable, and w_t is a noise process of independent, Normal $N(0,1)$ random variables. The system matrices A° and B° are unknown.

Our control objective is to select the input u_t in such a way as to minimize the long-term average quadratic cost criterion

$$\limsup_{t\to\infty} \frac{1}{t} \sum_{s=1}^{t} \left[x_s^T Q x_s + u_s^T R u_s\right], \qquad Q = Q^T \geq 0, \quad R = R^T > 0. \tag{2}$$

To this purpose, we observe the state x_t and, based on this, we first generate an estimate of the system matrices A° and B° and then exploit these estimates in a certainty equivalence fashion.

*Received by the editors February 26, 1997; accepted for publication (in revised form) January 8, 1998; published electronically July 17, 1998.

http://www.siam.org/journals/sicon/36-6/31749.html

†Department of Electrical Engineering and Automation, University of Brescia, Via Branze 38, 25123 Bresica, Italy. The research of this author was supported by MURST under the 60% project "Adaptive identification, prediction and control" and the 40% project "Model identification, system control and signal processing." This research was conducted while M. C. Campi was visiting the Coordinated Science Laboratory at the University of Illinois at Urbana-Champaign in the summer of 1995.

‡Department of Electrical and Computer Engineering, and the Coordinated Science Laboratory, University of Illinois, 1308 West Main Street, Urbana, IL 61801 (prkumar@decision.csl.uiuc.edu). The research of this author was supported by U.S. Army Research Office contract DAAH-04-95-1-0090 and Joint Service Electronics Program contract N00014-96-1-0129.

A common way to generate an estimate of A° and B° is to resort to the least squares method which corresponds to minimizing the performance index

$$V_t(A,B) = \sum_{s=1}^{t} \|x_s - Ax_{s-1} - Bu_{s-1}\|^2. \tag{3}$$

It is well known, however, that the corresponding certainty equivalent adaptive control law can suffer from an identifiability problem and that this can result in a degradation of the control system performance; see [1, 2, 3, 4]. In particular, for the case where matrices A° and B° belong to a finite known set, it is shown in [2] that the least squares estimate can converge with positive probability to a false estimate, which then leads to a strictly suboptimal value of the long-term average cost criterion. For the case of controlled Markov chains, such a counterexample had earlier been exhibited in [1]. Parameter consistency is guaranteed under certain conditions which are satisfied only in specific adaptive control situations, as, e.g., studied in [5] and [6].

This inability to identify the open loop system from closed-loop measurements is one of the fundamental obstacles to self-optimizing adaptive control. To overcome this, one approach is to occasionally probe the system. This can be done by either adding dither to the control or by occasionally breaking the control loop. However, such perturbations should be of small enough magnitude or infrequent enough so that they do not in themselves add to the cost incurred. An account of this approach can be found in Chen and Guo [7, 8, 9, 10, 11, 12].

To overcome this general problem of identifiability in closed loop, a very different approach, which still preserves the certainty equivalent structure of the adaptive controller and holds out the promise of general self-optimizing controllers, was proposed in [13] for the class of controlled Markov chains. The novelty of this adaptive controller is the employment of a *cost-biased maximum likelihood* parameter estimator, rather than the usual maximum likelihood parameter estimator. This cost biasing modifies the log-likelihood criterion by incorporating an additional term which favors parameter estimates with smaller optimal costs. For controlled Markov chains with a finite parameter set, it was shown in [13] that such a cost biasing eliminates parameters with costs larger than the optimal cost from occurring as limit points of the estimator. As a consequence, the corresponding adaptive controller was proved to provide optimal performance. This result was extended in [14] to the case of general parameter sets, for controlled Markov chains with finite state spaces. Another extension to the case of a finite parameter set, but allowing for a general state space and nonlinear systems, was provided in [15]. In the reference most pertinent to this paper, [2], it was shown that the cost-biased maximum likelihood-based certainty equivalent controller yielded an optimal cost for linear systems with quadratic costs, as in (1) and (2), provided that the parameter set is finite.

The assumption that the parameter set is finite is crucial in the derivations of [2]. Indeed, it was shown in [2] that the log-likelihood ratio $V_t(A^\circ, B^\circ) - V_t(A,B)$ stays bounded for any *fixed* parameter (A,B), and, therefore, a wrong fixed parameter (A,B) can gain, at most, a finite advantage over the true parameter (A°, B°) in the standard least squares criterion. Thus, when the number of possible parameters is finite, the maximum of these finite advantages is still finite, and so a mild biasing is sufficient to prevent elements (A,B) with larger cost than the optimal cost from occurring as limit points of the parameter estimator. This mildness of the biasing is important in order not to destroy the ability of the least squares estimate to identify closed-loop dynamics. Unfortunately, this argument is no longer true when turning to

a more general setting allowing for infinitely possible true parameterizations. Indeed, in such a case, $inf_{(A,B)}[V_t(A^\circ, B^\circ) - V_t(A, B)]$ is no longer bounded and the above argument valid for the finite case fails to apply. As a consequence of this and other difficulties, the infinite parameter set case has remained so far unsolved.

It is the purpose of this paper to establish the optimality of a certainty equivalent controller based on the cost-biased maximum likelihood parameter estimator for linear quadratic Gaussian systems, in the case of compact parameter uncertainty set. The aforementioned difficulty that the log-likelihood ratio $V_t(A^\circ, B^\circ) - V_t(A, B)$ is unbounded is circumvented by resorting to a Bayesian embedding approach. In this setting, one can show that the least squares estimate converges along the directions of diverging information to the true parameter value. As a consequence, a sequence of parameters (A'_t, B'_t) can be determined with the property that it converges to the true parameter (A°, B°) and for which $inf_{(A,B)}[V_t(A'_t, B'_t) - V_t(A, B)]$ remains bounded. Loosely speaking, (A'_t, B'_t) can be used in the analysis in place of (A°, B°) and, by a careful use of continuity arguments, the optimality of the adaptive controller can be established.

The paper is organized as follows. Our adaptive control scheme is described in section 2. In section 3, the properties of the cost-biased maximum likelihood parameter estimator are worked out. Section 4 is devoted to the study of the self-tuning properties of the adaptive scheme, and its stability and optimality are established in section 5.

2. The adaptive control system. Throughout this paper, let $[A, B] \in \mathbf{R}^{n\times(n+m)}$ denote the matrix obtained by concatenating matrices $A \in \mathbf{R}^{n\times n}$ and $B \in \mathbf{R}^{n\times m}$.

In our adaptive control problem, matrices A° and B° of system (1) are unknown and belong to a known compact set Θ as precisely stated in the following assumptions.

(A.i) There is a known compact set $\Theta \subset R^{n\times(n+m)}$ such that

$$[A^\circ, B^\circ] \in \text{interior}(\Theta).$$

(A.ii) (A, B) is reachable and $(A, Q^{1/2})$ is observable, $\forall [A, B] \in \Theta$.

Given the system parameters $[A, B] \in \Theta$, the control law minimizing the cost (2) for the system $x_{t+1} = Ax_t + Bu_t + w_{t+1}$ is easily derived (see, e.g., Kumar and Varaiya [16] or Bertsekas [17] for a comprehensive presentation of linear quadratic control problems). First, one has to compute the positive semidefinite solution to the algebraic Riccati equation

$$P = A^T PA - A^T PB(B^T PB + R)^{-1}B^T PA + Q.$$

The existence and uniqueness of such a solution is a consequence of the reachability and observability assumption (A.ii). Denoting such a solution by $P(A, B)$, the control law is then given by

$$u_t = K(A, B)x_t, \tag{4}$$

where $K(A, B)$ is the linear quadratic Gaussian (LQG) optimal gain defined by

$$K(A, B) = -(B^T P(A, B)B + R)^{-1}B^T P(A, B)A. \tag{5}$$

The corresponding optimal cost is denoted by $J(A, B)$.

When one is facing an adaptive control problem, the system matrices (A°, B°) are not known and some estimates $\widehat{A}_t$ and $\widehat{B}_t$ of them are needed. Once these estimates have been generated, in the certainty equivalence approach they are simply used as if they were the true system matrices. Correspondingly, the adaptive control law is given by

$$u_t = K(\widehat{A}_t, \widehat{B}_t)x_t. \tag{6}$$

The heart of our adaptive control scheme lies in the cost-biased least squares estimator of the system matrices as described below.

Choose a deterministic sequence μ_t such that $\mu_t \to \infty$ and $\mu_t = o(\log t)$ as $t \to \infty$. The parameter estimate sequence $\{[\widehat{A}_t, \widehat{B}_t]\}$ is given by

$$[\widehat{A}_t, \widehat{B}_t] = \begin{cases} \arg\min_{[A,B]\in\Theta} \left\{ \sum_{s=1}^{t} \|x_s - Ax_{s-1} - Bu_{s-1}\|^2 + \mu_t J(A,B) \right\}, & \text{for } t \text{ even}, \\ \left[\widehat{A}_{t-1}, \widehat{B}_{t-1}\right], & \text{for } t \text{ odd} \end{cases} \tag{7}$$

(when there is more than one minimizer, any of them can be chosen).

The distinguishing feature of the criterion (7) is the term $\mu_t J(A, B)$, which introduces a mild bias in favor of parameters (A, B) with lower optimal costs. The biasing is "mild" because $\mu_t = o(\log t)$. On the other hand, it is nonnegligible because $\mu_t \to \infty$. Without this term one would simply have the usual least squares parameter estimator, with its attendant difficulty in identifying the system in closed loop.

The intuitive rationale for the cost biasing in the least squares criterion is as follows. Suppose that one simply employs a straightforward least squares parameter estimator. Then, generically, it can be shown that the least squares parameter estimates sequence $[\widehat{A}_t^{LS}, \widehat{B}_t^{LS}]$ converges to a limiting random variable $[\widehat{A}_\infty^{LS}, \widehat{B}_\infty^{LS}]$ (see [18]). Such a limiting estimate results in a limiting controller $u_t = K(\widehat{A}_\infty^{LS}, \widehat{B}_\infty^{LS})x_t$. It is natural to expect that the least squares estimator will asymptotically identify, at a minimum, the closed-loop behavior of the system. Thus, one expects that the behavior of the true system with the loop closed by $u_t = K(\widehat{A}_\infty^{LS}, \widehat{B}_\infty^{LS})x_t$ will be the same as the closed-loop estimated system, i.e., their closed-loop gains are equal:

$$A^\circ + B^\circ K(\widehat{A}_\infty^{LS}, \widehat{B}_\infty^{LS}) = \widehat{A}_\infty^{LS} + \widehat{B}_\infty^{LS} K(\widehat{A}_\infty^{LS}, \widehat{B}_\infty^{LS}).$$

This implies that the cost of running the true system (A°, B°) with the feedback gain $K(\widehat{A}_\infty^{LS}, \widehat{B}_\infty^{LS})$ is the same as the cost of running the estimated system $(\widehat{A}_\infty^{LS}, \widehat{B}_\infty^{LS})$ with the feedback $K(\widehat{A}_\infty^{LS}, \widehat{B}_\infty^{LS})$. The latter is, however, the optimal configuration for the system $x_{t+1} = \widehat{A}_\infty^{LS} x_t + \widehat{B}_\infty^{LS} u_t + w_{t+1}$, while the former is not necessarily an optimal configuration for the true system. Thus one has

$$J(\widehat{A}_\infty^{LS}, \widehat{B}_\infty^{LS}) \geq J(A^\circ, B^\circ).$$

This means that the least squares estimator has a natural tendency to return estimates with larger optimal cost than the optimal cost associated with the true system. This motivates the idea of somehow introducing a bias into the parameter estimator so that it favors parameters (A, B) with smaller values of $J(A, B)$.

Thus, one conceives of adding a term such as $\mu_t J(A, B)$ to the squared error in (7). However, one needs to choose μ_t with care. One does not want to destroy the

ability of the least squares estimator to identify the closed-loop dynamics. This is achieved by choosing μ_t small enough so that $\mu_t = o(logt)$. On the other hand, one definitely wants the $\mu_t J(A,B)$ term to assert itself, and this is achieved by choosing $\mu_t \to \infty$. Hence, we arrive at the cost-biased least squares parameter estimator (7).

Notation. For brevity, the following notation will be used throughout the paper: $P^\circ := P(A^\circ, B^\circ)$, $\widehat{P}_t := P(\widehat{A}_t, \widehat{B}_t)$, $K^\circ := K(A^\circ, B^\circ)$, $\widehat{K}_t := K(\widehat{A}_t, \widehat{B}_t)$, $J^\circ := J(A^\circ, B^\circ)$, and $\widehat{J}_t := J(\widehat{A}_t, \widehat{B}_t)$. □

3. The properties of the parameter estimates. In this section, we study the properties of the estimates $[\widehat{A}_t, \widehat{B}_t]$ returned by the estimator (7). Our main result is that the introduction of the cost-bias term $\mu_t J(A,B)$ in the identification criterion prevents parameters $[A,B]$ with cost $J(A,B)$ strictly larger than the optimal cost from occurring as limit points of $[\widehat{A}_t, \widehat{B}_t]$ (Theorem 2). In this way, our modification is proven successful in counteracting the natural tendency of least squares to return estimates with larger cost than the optimal one. In addition, we show that the estimator preserves the capability of the least squares method of identifying the control system closed-loop dynamics (Theorem 3).

We start by summarizing some known results on the least squares estimates relevant to the forthcoming developments.

Denote by $[\widehat{A}_t^{LS}, \widehat{B}_t^{LS}]$ the least squares estimate of $[A^\circ, B^\circ]$:

$$\left[\widehat{A}_t^{LS}, \widehat{B}_t^{LS}\right] := \arg\min_{[A,B]\in\mathbf{R}^{n\times(n+m)}} \sum_{s=1}^{t} \|x_s - Ax_{s-1} - Bu_{s-1}\|^2 .$$

The partial ability of the least squares estimates $(\widehat{A}_t^{LS}, \widehat{B}_t^{LS})$ to estimate a portion of the open-loop system can be stated precisely using the notion of the *excited subspace*, originally introduced in [19].

DEFINITION 1. *Defining* $v_s^T := [x_s^T \quad u_s^T]$, *the subspace*

$$\mathcal{E}^\perp := \left\{ z \in \mathbf{R}^{n+m} : \qquad z^T \sum_{s=1}^{\infty} v_s v_s^T z < \infty \right\}$$

is called the unexcited *subspace. Its orthogonal complement* $\mathcal{E}$ *is the* excited *subspace.*

Given $[A,B]$, let $[A,B]_{\mathcal{E}}$ and $[A,B]_{\mathcal{E}^\perp}$ denote the matrices in $\mathbf{R}^{n\times(n+m)}$ formed by projecting the rows of $[A,B]$ onto $\mathcal{E}$ and $\mathcal{E}^\perp$, respectively.

The main properties of the least squares estimate are stated in Theorem 1 below (the proof of point (i) can be derived as a slight modification to that of Theorem 1 in [18], whereas point (ii) follows from Theorem 2 in [20] and Theorem 2 in [21]).

THEOREM 1. *There exists a set* $N \in \mathbf{R}^{n+m}$ *with zero Lebesgue measure such that, if* $[A^\circ, B^\circ]$ *does not belong to* N, *then*

(i)

$$\lim_{t\to\infty} [\widehat{A}_t^{LS}, \widehat{B}_t^{LS}] = [\widehat{A}_\infty^{LS}, \widehat{B}_\infty^{LS}] \qquad a.s.,$$

where $[\widehat{A}_\infty^{LS}, \widehat{B}_\infty^{LS}]$ *is an almost surely (a.s.) bounded random variable.*

(ii)

$$[\widehat{A}_\infty^{LS}, \widehat{B}_\infty^{LS}]_{\mathcal{E}} = [A^\circ, B^\circ]_{\mathcal{E}} \qquad a.s.$$

In particular, point (ii) asserts that the asymptotic estimation error is confined to the unexcited subspace. This is not surprising since the uncertainty in the excited

directions is overcome by the information, which diverges with time. This turns out to be a crucial property in the derivation of several results concerning our adaptive scheme.

Throughout, we assume that $[A^\circ, B^\circ]$ does not belong to N.

Our first result on the cost-biased estimate $[\widehat{A}_t, \widehat{B}_t]$ proves that it abandons the region with costs larger than the optimal cost, for t large enough. A key role is played by the composite estimate

$$[A'_t, B'_t] := [\widehat{A}_t^{LS}, \widehat{B}_t^{LS}]_{\mathcal{E}} + [A^\circ, B^\circ]_{\mathcal{E}^\perp}.$$

THEOREM 2.

$$\limsup_{t\to\infty} \hat{J}_t \le J^\circ \qquad a.s.$$

Proof. Define

$$V_t(A,B) := \sum_{s=1}^{t} \|x_s - Ax_{s-1} - Bu_{s-1}\|^2,$$
$$D_t(A,B) := V_t(A,B) + \mu_t J(A,B).$$

Note for future use that

$$V_t(A,B) - V_t(\widehat{A}_t^{LS}, \widehat{B}_t^{LS}) = \sum_{s=1}^{t} \left\| \left\{ [A,B] - [\widehat{A}_t^{LS}, \widehat{B}_t^{LS}] \right\} v_{s-1} \right\|^2.$$

Indeed, recalling that the minimizer of $V_t(A,B)$ is given by $[\widehat{A}_t^{LS}, \widehat{B}_t^{LS}] = (\sum_{s=1}^t x_s v_{s-1}^T)(\sum_{s=1}^t v_{s-1} v_{s-1}^T)^{-1}$, one has

$$\begin{aligned}
V_t(A,B) - V_t(\widehat{A}_t^{LS}, \widehat{B}_t^{LS}) - \sum_{s=1}^{t} \left\| \left\{ [A,B] - [\widehat{A}_t^{LS}, \widehat{B}_t^{LS}] \right\} v_{s-1} \right\|^2 \\
= \sum_{s=1}^{t} \|x_s\|^2 + \sum_{s=1}^{t} v_{s-1}^T [A,B]^T [A,B] v_{s-1} - 2\sum_{s=1}^{t} v_{s-1}^T [A,B]^T x_s \\
- \sum_{s=1}^{t} \|x_s\|^2 - \sum_{s=1}^{t} v_{s-1}^T [\widehat{A}_t^{LS}, \widehat{B}_t^{LS}]^T [\widehat{A}_t^{LS}, \widehat{B}_t^{LS}] v_{s-1} + 2\sum_{s=1}^{t} v_{s-1}^T [\widehat{A}_t^{LS}, \widehat{B}_t^{LS}]^T x_s \\
- \sum_{s=1}^{t} v_{s-1}^T [A,B]^T [A,B] v_{s-1} - \sum_{s=1}^{t} v_{s-1}^T [\widehat{A}_t^{LS}, \widehat{B}_t^{LS}]^T [\widehat{A}_t^{LS}, \widehat{B}_t^{LS}] v_{s-1} \\
+ 2\sum_{s=1}^{t} v_{s-1}^T [A,B]^T [\widehat{A}_t^{LS}, \widehat{B}_t^{LS}] v_{s-1} \\
= -2\sum_{s=1}^{t} v_{s-1}^T [A,B]^T x_s - 2Trace\left\{ [\widehat{A}_t^{LS}, \widehat{B}_t^{LS}] \left(\sum_{s=1}^{t} v_{s-1} v_{s-1} \right) [\widehat{A}_t^{LS}, \widehat{B}_t^{LS}]^T \right\} \\
+ 2\sum_{s=1}^{t} v_{s-1}^T [\widehat{A}_t^{LS}, \widehat{B}_t^{LS}]^T x_s + 2Trace\left\{ [\widehat{A}_t^{LS}, \widehat{B}_t^{LS}] \left(\sum_{s=1}^{t} v_{s-1} v_{s-1} \right) [A,B]^T \right\}
\end{aligned}$$

$$
\begin{aligned}
&= -2\sum_{s=1}^{t} v_{s-1}^T [A,B]^T x_s - 2Trace\left\{\sum_{s=1}^{t} x_s v_{s-1}^T [\widehat{A}_t^{LS}, \widehat{B}_t^{LS}]^T\right\} \\
&+ 2\sum_{s=1}^{t} v_{s-1}^T [\widehat{A}_t^{LS}, \widehat{B}_t^{LS}]^T x_s + 2Trace\left\{\sum_{s=1}^{t} x_s v_{s-1}^T [A,B]^T\right\} \\
&= 0.
\end{aligned}
$$

For every $[A,B] \in S_\epsilon := \{[A,B] \in \Theta \ : \ J(A,B) \geq J^\circ + \epsilon\}, \epsilon > 0$, the following chain of inequalities holds true:

$$
\begin{aligned}
D_t(A,B) - D_t(A'_t,B'_t) &\geq V_t(\widehat{A}_t^{LS}, \widehat{B}_t^{LS}) + \mu_t J(A,B) \\
&\quad - V_t(A'_t,B'_t) - \mu_t J(A'_t,B'_t) \\
&\geq -\sum_{s=1}^{t} \left\| \left\{ [A'_t,B'_t] - [\widehat{A}_t^{LS}, \widehat{B}_t^{LS}] \right\} v_{s-1} \right\|^2 \\
&\quad + \mu_t \left\{ J^\circ + \epsilon - J(A'_t,B'_t) \right\}. \qquad (8)
\end{aligned}
$$

Recalling that $J(A,B) = TraceP(A,B)$ (see [16] or [17]), and that $P(\cdot,\cdot)$ is a continuous function of the entries of matrices A and B for any $[A,B] \in \Theta$ (see [22]), we can conclude that $J(\cdot,\cdot)$ is continuous in $[A^\circ, B^\circ]$. Since $[A'_t,B'_t] \to [A^\circ,B^\circ]$ (which follows from (ii) of Theorem 1), we therefore have

$$
J^\circ + \epsilon - J(A'_t,B'_t) \to \epsilon \qquad \text{a.s.}
$$

Thus, the second term on the right-hand side of (8) tends to infinity as $t \to \infty$. On the other hand, by the very definition of unexcited subspace and $[A'_t,B'_t]$, the first term stays bounded. Therefore, the right-hand side of (8) is diverging, uniformly in $[A,B] \in S_\epsilon$. That is, $D_t(A,B)$ is strictly larger than $D_t(A'_t,B'_t)$ for any $[A,B] \in S_\epsilon$ when t is large enough. Finally, by noting that $[A'_t,B'_t] \in \Theta$ for t large enough, the conclusion is drawn that $[\widehat{A}_t, \widehat{B}_t]$ leaves set S_ϵ in finite time. In view of the arbitrariness of $\epsilon > 0$, the proof is complete. □

We now introduce C_δ as the set of parameters $[A,B]$ such that the gain of the corresponding optimal closed-loop system differs from the gain of the true system with the loop closed by $K(A,B)$ by at least δ in norm, i.e.,

$$
C_\delta := \{[A,B] \in \Theta \ : \ \|[A^\circ + B^\circ K(A,B)] - [A + BK(A,B)]\| \geq \delta\}.
$$

We now prove that the estimate $[\widehat{A}_t, \widehat{B}_t]$ can visit C_δ only rarely, and so our cost-biased estimator (7) still possesses good closed-loop identification properties.

THEOREM 3.

$$
\sum_{s=1}^{t} 1([\widehat{A}_s, \widehat{B}_s] \in C_\delta) = O(\mu_t) \qquad a.s., \qquad \forall \delta > 0.
$$

Proof. We first prove that

$$
\sum_{s=1}^{t} \left\| \left\{ [A'_t,B'_t] - [\widehat{A}_t, \widehat{B}_t] \right\} v_{s-1} \right\|^2 = O(\mu_t), \qquad t \text{ even} \qquad \text{a.s.} \qquad (9)
$$

Indeed,

$$\sum_{s=1}^{t} \left\| \left\{ [A_t', B_t'] - [\widehat{A}_t, \widehat{B}_t] \right\} v_{s-1} \right\|^2$$

$$\leq 2 \sum_{s=1}^{t} \left\| \left\{ [A^\circ, B^\circ]_{\mathcal{E}^\perp} - [\widehat{A}_t^{LS}, \widehat{B}_t^{LS}]_{\mathcal{E}^\perp} \right\} v_{s-1} \right\|^2$$
$$+ 2 \sum_{s=1}^{t} \left\| \left\{ [\widehat{A}_t^{LS}, \widehat{B}_t^{LS}] - [\widehat{A}_t, \widehat{B}_t] \right\} v_{s-1} \right\|^2 .$$

The first term is bounded because of the definition of unexcited subspace. As for the second term, it can be handled as follows:

$$\sum_{s=1}^{t} \left\| \left\{ [\widehat{A}_t^{LS}, \widehat{B}_t^{LS}] - [\widehat{A}_t, \widehat{B}_t] \right\} v_{s-1} \right\|^2 = V_t(\widehat{A}_t, \widehat{B}_t) - V_t(\widehat{A}_t^{LS}, \widehat{B}_t^{LS})$$
$$= D_t(\widehat{A}_t, \widehat{B}_t) - D_t(A_t', B_t') + \mu_t \left\{ J(A_t', B_t') - \widehat{J}_t \right\}$$
$$+ \left\{ V_t(A_t', B_t') - V_t(\widehat{A}_t^{LS}, \widehat{B}_t^{LS}) \right\}.$$

The last term equals $\sum_{s=1}^{t} ||\{[A_t', B_t'] - [\widehat{A}_t^{LS}, \widehat{B}_t^{LS}]\} v_{s-1}||^2$ and is bounded, whereas, by noting that $[A_t', B_t'] \in \Theta$ for t large enough, the first term is less than or equal to zero in the limit. Result (9) then follows from the fact that $J(A_t', B_t') - \widehat{J}_t$ is bounded (remember that $J(\cdot, \cdot)$ is a continuous function on Θ and Θ is a compact set).

Note now that the matrix

$$[A^\circ + B^\circ K(A, B)] - [\bar{A} + \bar{B} K(A, B)]$$

is continuous as a function of $[A, B] \in \Theta$ and $[\overline{A}, \overline{B}] \in \Theta$ (this follows from the expression (5) of the gain $K(A, B)$ and the continuity of $P(A, B)$ in Θ (see [22])). Therefore, $\forall [\widetilde{A}, \widetilde{B}] \in C_\delta$, there exists a neighborhood $N(\widetilde{A}, \widetilde{B})$ of $[\widetilde{A}, \widetilde{B}]$ and a nonzero matrix H such that

$$\left([A^\circ + B^\circ K(A, B)] - [\overline{A} + \overline{B} K(A, B)] \right)^T \left([A^\circ + B^\circ K(A, B)] - [\overline{A} + \overline{B} K(A, B)] \right)$$
$$\geq H^T H, \qquad \forall [A, B], [\overline{A}, \overline{B}] \in N(\widetilde{A}, \widetilde{B}). \tag{10}$$

The set of all these neighborhoods constitutes a cover of C_δ, from which a finite subcover $\{N_j\}_{j=1}^{q}$ can be extracted. The thesis of the theorem can then be recast as

$$\sum_{s=1}^{t} 1([\widehat{A}_s, \widehat{B}_s] \in N_j) = O(\mu_t) \qquad \text{a.s.,} \qquad \forall j \in [1, q]. \tag{11}$$

Equation (11) will be proven by contradiction. To this purpose, set

$$\#_{j,t} := \sum_{\substack{s=1 \\ s \text{ even}}}^{t} 1([\widehat{A}_s, \widehat{B}_s] \in N_j)$$

and assume that there exist $\bar{j} \in [1, q]$ and a sequence of even time points $\{t_k\}$ such that $[\widehat{A}_{t_k}, \widehat{B}_{t_k}] \in N_{\bar{j}} \forall k$, and

$$\lim_{k\to\infty} \frac{1}{\mu_{t_k}} \#_{\bar{j},t_k} = \infty. \tag{12}$$

We prove that (12) implies

$$\liminf_{k\to\infty} \frac{1}{\#_{\bar{j},t_k}} \sum_{\substack{s=1 \\ s \text{ even}}}^{t_k} \alpha_{\bar{j},s+1} > 0, \tag{13}$$

where

$$\alpha_{\bar{j},s+1} := (\|Hx_{s+1}\|^2 \wedge 1) 1([\widehat{A}_s, \widehat{B}_s] \in N_{\bar{j}}) \tag{14}$$

(H is the matrix introduced in (10) associated with $N_{\bar{j}}$) and, in turn, this contradicts (9).

For the proof of (13), define $\mathcal{F}_s := \sigma(w_1, \dots, w_s)$ and note first that

$$\begin{aligned} E[\|Hx_{s+1}\|^2 \wedge 1 \mid \mathcal{F}_s] &= E[\|H(A^\circ x_s + B^\circ u_s) + Hw_{s+1}\|^2 \wedge 1 \mid \mathcal{F}_s] \\ &\geq \text{Prob}(\|H(A^\circ x_s + B^\circ u_s) + Hw_{s+1}\| \geq 1 \mid \mathcal{F}_s) \\ &\geq 1 - \text{Prob}\left(\|H(A^\circ x_s + B^\circ u_s)\| - 1 < \|Hw_{s+1}\| \right. \\ &\qquad \left. < \|H(A^\circ x_s + B^\circ u_s\| + 1 \mid \mathcal{F}_s)\right) \\ &\geq 1 - \sup_\alpha \text{Prob}(\alpha - 1 < \|Hw_{s+1}\| < \alpha + 1) \\ &\geq c, \end{aligned}$$

for a suitable constant $c > 0$, the last inequality following from the fact that $H \neq 0$. We therefore have

$$\frac{1}{\#_{\bar{j},t_k}} \sum_{\substack{s=1 \\ s \text{ even}}}^{t_k} E[\alpha_{\bar{j},s+1} \mid \mathcal{F}_s] \geq \frac{1}{\#_{\bar{j},t_k}} \sum_{\substack{s=1 \\ s \text{ even}}}^{t_k} 1([\widehat{A}_s, \widehat{B}_s] \in N_{\bar{j}}) \cdot c = c. \tag{15}$$

On the other hand,

$$\begin{aligned} &\sum_{\substack{s=1 \\ s \text{ even}}}^{t_k} \left\{\alpha_{\bar{j},s+1} - E[\alpha_{\bar{j},s+1} \mid \mathcal{F}_s]\right\} \\ &= \sum_{\substack{s=1 \\ s \text{ even}}}^{t_k} 1([\widehat{A}_s, \widehat{B}_s] \in N_{\bar{j}}) \left\{(\|Hx_{s+1}\|^2 \wedge 1) - E[\|Hx_{s+1}\|^2 \wedge 1 \mid \mathcal{F}_s]\right\} \\ &= o\left(\sum_{\substack{s=1 \\ s \text{ even}}}^{t_k} 1([\widehat{A}_s, \widehat{B}_s] \in N_{\bar{j}})\right), \end{aligned} \tag{16}$$

on the set where

$$\sum_{\substack{s=1 \\ s \text{ even}}}^{t_k} 1([\widehat{A}_s, \widehat{B}_s] \in N_{\bar{j}}) = \infty \tag{17}$$

(see [23]). Since (17) is satisfied if (12) holds, equations (15) and (16) prove that (13) follows from (12).

We now prove that (13) contradicts (9).

The convergence result $[A'_t, B'_t] \to [A^\circ, B^\circ]$ (see Theorem 1) implies that

$$\begin{aligned}\left([A'_t + B'_t K(A,B)] - [\overline{A} + \overline{B}K(A,B)]\right)^T \left([A'_t + B'_t K(A,B)] - [\overline{A} + \overline{B}K(A,B)]\right) \\ \geq \left(\frac{1}{2}H\right)^T \left(\frac{1}{2}H\right) \qquad \forall [A,B], [\overline{A},\overline{B}] \in N_{\overline{j}},\end{aligned}$$

for t sufficiently high (see (10)). In view of this, the following chain of inequalities can be derived when (12), and, consequently, inequality (13) hold true:

$$\begin{aligned}\infty &= \lim_{k\to\infty} \frac{1}{\mu_{t_k}} \#_{\overline{j},t_k} \cdot \liminf_{k\to\infty} \frac{1}{\#_{\overline{j},t_k}} \sum_{\substack{s=1 \\ s \text{ even}}}^{t_k} \alpha_{\overline{j},s+1} \\ &\leq \lim_{k\to\infty} \frac{1}{\mu_{t_k}} \sum_{\substack{s=1 \\ s \text{ even}}}^{t_k-2} \alpha_{\overline{j},s+1} \\ &\leq \lim_{k\to\infty} \frac{1}{\mu_{t_k}} \sum_{\substack{s=1 \\ s \text{ even}}}^{t_k-2} \|Hx_{s+1}\|^2 \cdot 1([\widehat{A}_s, \widehat{B}_s] \in N_{\overline{j}}) \\ &\leq \lim_{k\to\infty} \frac{1}{\mu_{t_k}} \sum_{\substack{s=1 \\ s \text{ even}}}^{t_k-2} 4\| \left\{[A'_{t_k} + B'_{t_k}\widehat{K}_s] - [\widehat{A}_{t_k} + \widehat{B}_{t_k}\widehat{K}_s]\right\} x_{s+1}\|^2 \cdot 1([\widehat{A}_s, \widehat{B}_s] \in N_{\overline{j}}) \\ &\leq \lim_{k\to\infty} \frac{1}{\mu_{t_k}} \sum_{\substack{s=1 \\ s \text{ even}}}^{t_k-2} 4\| \left([A'_{t_k}, B'_{t_k}] - [\widehat{A}_{t_k}, \widehat{B}_{t_k}]\right) v_{s+1}\|^2.\end{aligned}$$

This contradicts (9). Thus, (12) is false with probability 1, and so (11) is proven. □

4. The self-tuning property. A key issue in the analysis of any adaptive control method consists of determining whether it is able to generate, at least asymptotically, control laws close to the optimal control law for the true system. The objective of the present section is to prove that this is indeed the case for our adaptive scheme, except for very rare time instants. This result will play a crucial role in the next section where we address stability and optimality issues.

THEOREM 4.

$$\sum_{s=1}^{t} 1(\|\widehat{K}_s - K^\circ\| > \rho) = O(\mu_t) \qquad a.s., \qquad \forall \rho > 0.$$

Proof. Since Θ is compact,

$$\sup_{[A,B]\in\Theta} \lambda_{\max}[A + BK(A,B)] < 1.$$

This implies that $A^\circ + B^\circ K(A,B)$ is stable for $[A,B]$ belonging to the closed set $\overline{C_\delta^c}$ (where the overbar indicates closure and the superscript "c" indicates the complement of the set), for δ small enough.

Denote by $J(A, B; K)$ the cost for the system $x_{t+1} = Ax_t + Bx_t + w_{t+1}$ controlled by $u_t = Kx_t$, whenever the corresponding closed-loop system is stable. It is known that (see [16] or [17])

$$J(A, B; K) = TraceP(A, B; K), \tag{18}$$

where $P(A, B; K)$ is the unique positive semidefinite solution of the Lyapunov equation

$$P = K^T RK + [A + BK]^T P[A + BK] + Q. \tag{19}$$

From this, it is easy to verify that $J(A^\circ, B^\circ; K(A, B))$ is a continuous function of $[A, B] \in \overline{C_\delta^c}$. On the other hand, the optimal gain K° for the true system (1) is unique within the class of stabilizing gains:

$$J(A^\circ, B^\circ; K) > J^\circ, \qquad \forall K \neq K^\circ, \qquad K \text{ stabilizing.}$$

Therefore, there exists $\nu(\rho) > 0$ such that every gain $K = K(A, B)$, $[A, B] \in \overline{C_\delta^c}$, for which

$$J(A^\circ, B^\circ; K) \leq J^\circ + \nu(\rho)$$

also satisfies the bound

$$\|K - K^\circ\| \leq \rho. \tag{20}$$

Note now that since $A+BK(A, B)$ is close to $A^\circ+B^\circ K(A, B)$ when $[A, B] \in \overline{C_\delta^c}$, δ small, from (19), we have

$$\sup_{[A,B]\in C_\delta^c} \|P(A^\circ, B^\circ; K(A, B)) - P(A, B; K(A, B))\| \to 0, \qquad \delta \to 0,$$

and, in view of (18),

$$\sup_{[A,B]\in C_s^c} |J(A^\circ, B^\circ; K(A, B)) - J(A, B; K(A, B))| \to 0, \qquad \delta \to 0.$$

Fix $\delta(\rho)$ such that

$$\sup_{[A,B]\in C_{\delta(\rho)}^c} \|J(A^\circ, B^\circ; K(A, B)) - J(A, B; K(A, B))\| \leq \frac{1}{2}\nu(\rho). \tag{21}$$

Finally,

$$\begin{aligned}
&\limsup_{t\to\infty} \frac{1}{\mu_t} \sum_{s=1}^{t} 1(\|\widehat{K}_s - K^\circ\| > \rho) \\
&\qquad \leq \limsup_{t\to\infty} \frac{1}{\mu_t} \sum_{s=1}^{t} 1(J(A^\circ, B^\circ; \widehat{K}_s) - J^\circ > \nu(\rho)) \qquad \text{(using (20))} \\
&\qquad \leq \limsup_{t\to\infty} \frac{1}{\mu_t} \sum_{s=1}^{t} 1\left(|J(A^\circ, B^\circ; \widehat{K}_s) - \widehat{J}_s| > \frac{1}{2}\nu(\rho)\right) \\
&\qquad\qquad \text{(using Theorem 2)} \\
&\qquad \leq \limsup_{t\to\infty} \frac{1}{\mu_t} \sum_{s=1}^{t} 1\left([\widehat{A}_s, \widehat{B}_s] \in C_{\delta(\rho)}\right) \qquad \text{(using (21))} \\
&\qquad < \infty \qquad \text{(using Theorem 3).} \quad \square
\end{aligned}$$

5. Stability and optimality. According to Theorem 4, the adaptive gain $\widehat{K}_s$ is close to the optimal gain K° except at very rare time instants, the number of which grows at most as μ_t. At these exceptional time points, the closed-loop system may be unstable. However, due to their rare occurrence, we establish that they cannot endanger the stability of the adaptive closed-loop control system. The corresponding stability result is given in Theorem 5. The proof of Theorem 5 relies heavily on the results of [24] concerning stability of rarely destabilized time-varying systems. It is very similar to that of Theorem 12 in [2] and is provided here only for the sake of completeness.

THEOREM 5.

$$\limsup_{t\to\infty} \frac{1}{t}\sum_{s=1}^{t}[\|x_s\|^p + \|u_s\|^p] < \infty \qquad a.s., \qquad \forall p > 0.$$

Proof. We start by noting that, since $A^\circ + B^\circ K^\circ$ is a stable matrix, there exists a suitable norm on $\mathbf{R}^n$ such that, under the corresponding induced matrix norm, $\|A^\circ + B^\circ K^\circ\| < 1$ (see, e.g., [25]). Throughout this proof all the norm symbols refer to this particular norm.

It is easy to verify that the following inequality holds true for any integer n and real numbers a, b, and $\epsilon > 0$,

$$(a+b)^{2^n} \le (1+\epsilon^2)^{2^n-1}a^{2^n} + (1+\epsilon^{-2})^{2^n-1}b^{2^n}. \tag{22}$$

Taking into account the relation $x_{t+1} = A^\circ x_t + B^\circ \widehat{K}_t x_t + w_{t+1}$, from (22) we obtain

$$\|x_{t+1}\|^{2^n} \le (1+\epsilon^2)^{2^n-1}\|A^\circ + B^\circ \widehat{K}_t\|^{2^n}\|x_t\|^{2^n} + (1+\epsilon^{-2})^{2^n-1}\|w_{t+1}\|^{2^n},$$

for any integer n and positive real ϵ.

Now fix $\bar{n}$ such that $2^{\bar{n}} \ge p$ and choose $\bar{\epsilon} > 0$ such that $(1+\bar{\epsilon}^2)^{2^{\bar{n}}-1}\|A^\circ + B^\circ K^\circ\|^{2^{\bar{n}}} < 1$. Further, select ρ in such a way that

$$a := \sup_{K\,:\,\|K-K^\circ\|\le\rho} (1+\bar{\epsilon}^2)^{2^{\bar{n}}-1}\|A^\circ + B^\circ K\|^{2^{\bar{n}}} < 1$$

and also let

$$b := \sup_{[A,B]\in\Theta} (1+\bar{\epsilon}^2)^{2^{\bar{n}}-1}\|A^\circ + B^\circ K(A,B)\|^{2^{\bar{n}}}.$$

Then

$$\|x_{t+1}\|^{2^{\bar{n}}} \le \gamma_t\|x_t\|^{2^{\bar{n}}} + (1+\bar{\epsilon}^{-2})^{2^{\bar{n}}-1}\|w_{t+1}\|^{2^{\bar{n}}}, \tag{23}$$

where

$$\gamma_t = \begin{cases} a, & \text{if } \|\widehat{K}_t - K^\circ\| \le \rho, \\ b, & \text{otherwise.} \end{cases}$$

We now apply Theorem 2 in [24] to (23) (see also Remark 1 in the same paper). By noting that $\sum_{s=1}^{t} 1(\|\widehat{K}_s - K^\circ\| > \rho) = O(\mu_t)$ (Theorem 4) and that $\mu_t = o(\log t)$, from that theorem we can conclude that

$$\limsup_{t\to\infty} \frac{1}{t}\sum_{s=1}^{t}\|x_s\|^{2^{\bar{n}}} < \infty \qquad \text{a.s.}$$

This implies that (recall that $2^{\bar{n}} \geq p$)

$$\limsup_{t\to\infty} \frac{1}{t} \sum_{s=1}^{t} \|x_s\|^p < \infty \qquad \text{a.s.}$$

Since $\|u_s\| \leq \sup_{[A,B]\in\Theta} \|K(A,B)\| \|x_s\|$, we also have

$$\limsup_{t\to\infty} \frac{1}{t} \sum_{s=1}^{t} \|u_s\|^p < \infty \qquad \text{a.s.}$$

This proves the stability result. □

We are now in a position to prove the optimality of the adaptive scheme, namely, that the incurred cost equals the optimal cost that could be obtained if the true system parameter were known at the start.

THEOREM 6.

$$\limsup_{t\to\infty} \frac{1}{t} \sum_{s=1}^{t} \left[x_s^T Q x_s + u_s^T R u_s\right] = J^\circ \qquad \textit{a.s.}$$

Proof. The dynamic programming equation for model $x_{s+1} = \widehat{A}_s x_s + \widehat{B}_s u_s + w_{s+1}$ is (see [16])

$$\begin{aligned}
&\widehat{J}_s + x_s^T \widehat{P}_s x_s \\
&\quad = x_s^T Q x_s + u_s^T R u_s + E[(\widehat{A}_s x_s + \widehat{B}_s u_s + w_{s+1})^T \widehat{P}_s (\widehat{A}_s x_s + \widehat{B}_s u_s + w_{s+1}) \mid \mathcal{F}_s] \\
&\quad = x_s^T Q x_s + u_s^T R u_s + E[x_{s+1}^T \widehat{P}_s x_{s+1} \mid \mathcal{F}_s] \\
&\quad + \left\{(\widehat{A}_s x_s + \widehat{B}_s u_s)^T \widehat{P}_s (\widehat{A}_s x_s + \widehat{B}_s u_s) - (A^\circ x_s + B^\circ u_s)^T \widehat{P}_s (A^\circ x_s + B^\circ u_s)\right\}.
\end{aligned}$$

From this,

$$\begin{aligned}
&\underbrace{\frac{1}{t}\sum_{s=1}^{t} \widehat{J}_s}_{A} + \underbrace{\frac{1}{t}\sum_{s=1}^{t} \left\{x_s^T \widehat{P}_s x_s - E[x_{s+1}^T \widehat{P}_{s+1} x_{s+1} \mid \mathcal{F}_s]\right\}}_{B} \\
&\qquad = \frac{1}{t}\sum_{s=1}^{t} [x_s^T Q x_s + u_s^T R u_s] + \underbrace{\frac{1}{t}\sum_{s=1}^{t} E[x_{s+1}^T (\widehat{P}_s - \widehat{P}_{s+1}) x_{s+1} \mid \mathcal{F}_s]}_{C} \\
&\qquad + \underbrace{\frac{1}{t}\sum_{s=1}^{t} \left\{(\widehat{A}_s x_s + \widehat{B}_s u_s)^T \widehat{P}_s (\widehat{A}_s x_s + \widehat{B}_s u_s) - (A^\circ x_s + B^\circ u_s)^T \widehat{P}_s (A^\circ x_s + B^\circ u_s)\right\}}_{D}.
\end{aligned} \tag{24}$$

Let us study separately the different terms appearing in this expression.

(A) From Theorem 2 we have

$$\limsup_{t\to\infty} \frac{1}{t} \sum_{s=1}^{t} \widehat{J}_s \leq J^\circ.$$

(B)

$$\frac{1}{t}\sum_{s=1}^{t}\left\{x_s^T\widehat{P}_s x_s - E[x_{s+1}^T\widehat{P}_{s+1}x_{s+1}\mid\mathcal{F}_s]\right\}$$

$$= \frac{1}{t}x_1^T\widehat{P}_1x_1 - \frac{1}{t}x_{t+1}^T\widehat{P}_{t+1}x_{t+1} + \frac{1}{t}\sum_{s=1}^{t}\left\{x_{s+1}^T\widehat{P}_{s+1}x_{s+1} - E[x_{s+1}^T\widehat{P}_{s+1}x_{s+1}\mid\mathcal{F}_s]\right\}.$$

The first term obviously tends to zero. As for the second one, note that, $\widehat{P}_{t+1} \le \sup_{[A,B]\in\Theta} P(A,B)$ being bounded, it tends to zero provided that $\|x_t\|^2/t \to 0$. The fact that this is the case can be proven by contradiction. Suppose that there exists a time sequence $\{t_k\}$ and a real number $\alpha > 0$ such that $\|x_{t_k}\|^2 > \alpha t_k, \forall k$. Then $\limsup_{t\to\infty}\frac{1}{t}\sum_{s=1}^{t}\|x_s\|^4 \ge \limsup_{k\to\infty}\frac{1}{t_k}\|x_{t_k}\|^4 \ge \limsup_{k\to\infty}\frac{1}{t_k}\alpha^2 t_k^2 = \infty$. This contradicts Theorem 5. In the third term,

$$\{\alpha_{s+1}\} := \{x_{s+1}^T\widehat{P}_{s+1}x_{s+1} - E[x_{s+1}^T\widehat{P}_{s+1}x_{s+1}\mid\mathcal{F}_s]\}$$

is a martingale difference. Therefore, $\frac{1}{t}\sum_{s=1}^{t}\alpha_{s+1}\to 0$, provided that

$$\sum_{s=1}^{\infty} s^{-2}E[\alpha_{s+1}^2\mid\mathcal{F}_s] < \infty$$

(see [26]). Since $\widehat{P}_{s+1}$ is bounded, it is easily seen that this last condition is implied by $\sum_{s=1}^{\infty}s^{-2}[\|x_s\|^4+\|u_s\|^4] < \infty$. Again, this conclusion can be drawn by contradiction from Theorem 5. In fact, if this conclusion were false, sequence $s^{-1/2}[\|x_s\|^4+\|u_s\|^4]$ would be unbounded and, therefore, there would exist a sequence of times $\{t_k\}$ such that $[\|x_{t_k}\|^4+\|u_{t_k}\|^4] \ge t_k^{1/2}\forall k$. From this, $\limsup_{k\to\infty}\frac{1}{t}\sum_{s=1}^{t}[\|x_s\|^4+\|u_s\|^4]^4 \ge \limsup_{k\to\infty}\frac{1}{t_k}[\|x_{t_k}\|^4+\|u_{t_k}\|^4]^4 \ge \limsup_{k\to\infty}\frac{1}{t_k}t_k^2 = \infty$, and this is in contradiction with Theorem 5. In conclusion,

$$\lim_{t\to\infty}\frac{1}{t}\sum_{s=1}^{t}\left\{x_s^T\widehat{P}_s x_s - E[x_{s+1}^T\widehat{P}_{s+1}x_{s+1}\mid\mathcal{F}_s]\right\} = 0 \qquad \text{a.s.}$$

(C) We start by proving that

$$\lim_{t\to\infty}\frac{1}{t}\sum_{s=1}^{t}\|\widehat{P}_s - \widehat{P}_{s+1}\|^2 = 0 \qquad \text{a.s.} \tag{25}$$

Since P° satisfies the equation

$$P^\circ = K^{\circ T}RK^\circ + [A^\circ + B^\circ K^\circ]^T P^\circ[A^\circ + B^\circ K^\circ] + Q,$$

and $\widehat{P}_s$ satisfies the equation

$$\widehat{P}_s = \widehat{K}_s^T R\widehat{K}_s + [\widehat{A}_s + \widehat{B}_s\widehat{K}_s]^T\widehat{P}_s[\widehat{A}_s + \widehat{B}_s\widehat{K}_s] + Q,$$

P° is close to $\widehat{P}_s$ when K° is close to $\widehat{K}_s$ and $A^\circ + B^\circ K^\circ$ is close to $\widehat{A}_s + \widehat{B}_s\widehat{K}_s$. In view of Theorems 3 and 4, the total of the numbers of time points in which this does not happen is $O(\mu_t)$. Therefore,

$$\sum_{s=1}^{t} 1(\|\widehat{P}_s - P^\circ\| > \rho) = O(\mu_t) \qquad \text{a.s.}, \qquad \forall \rho > 0.$$

Equation (25) then easily follows from

$$\begin{aligned}\frac{1}{t}\sum_{s=1}^{t}\|\widehat{P}_s - \widehat{P}_{s+1}\|^2 &\le \frac{2}{t}\sum_{s=1}^{t}\left[\|\widehat{P}_s - P^\circ\|^2 + \|\widehat{P}_{s+1} - P^\circ\|^2\right] \\ &\le \frac{4}{t}\sum_{s=1}^{t+1}\|\widehat{P}_s - P^\circ\|^2 1(\|\widehat{P}_s - P^\circ\| > \rho) + \frac{4(t+1)}{t}\rho^2 \\ &\to 4\rho^2,\end{aligned}$$

since ρ is an arbitrary positive real number.

Notice now that, by the Schwarz inequality,

$$\frac{1}{t}\sum_{s=1}^{t}|x_{s+1}^T(\widehat{P}_s - \widehat{P}_{s+1})x_{s+1}| \le \left(\frac{1}{t}\sum_{s=1}^{t}\|\widehat{P}_s - \widehat{P}_{s+1}\|^2\right)^{1/2}\left(\frac{1}{t}\sum_{s=1}^{t}\|x_{s+1}\|^4\right)^{1/2}.$$

Therefore, $t^{-1}\sum_{s=1}^{t}\|x_{s+1}\|^4$ being bounded (Theorem 5), (25) implies

$$\lim_{t\to\infty}\frac{1}{t}\sum_{s=1}^{t}x_{s+1}^T(\widehat{P}_s - \widehat{P}_{s+1})x_{s+1} = 0 \qquad \text{a.s.} \tag{26}$$

Finally, the conclusion

$$\lim_{t\to\infty}\frac{1}{t}\sum_{s=1}^{t}E[x_{s+1}^T(\widehat{P}_s - \widehat{P}_{s+1})x_{s+1} \mid \mathcal{F}_s] = 0 \qquad \text{a.s.}$$

is drawn from (26) by observing that

$$\{\beta_{s+1}\} := \{x_{s+1}^T(\widehat{P}_s - \widehat{P}_{s+1})x_{s+1} - E[x_{s+1}^T(\widehat{P}_s - \widehat{P}_{s+1})x_{s+1} \mid \mathcal{F}_s]\}$$

is a martingale difference for which, by calculations resembling those developed in point (B), $\sum_{s=1}^{\infty} s^{-2}E[\beta_{s+1}^2 \mid \mathcal{F}_s] < \infty$.

(D) Since

$$\|P^TP - R^TR\| \le \|P - R\|(\|P\| + \|R\|), \qquad \forall P, R \in \mathbb{R}^{n\times n},$$

we have

$$\begin{aligned}&|(\widehat{A}_sx_s + \widehat{B}_su_s)^T\widehat{P}_s(\widehat{A}_sx_s + \widehat{B}_su_s) - (A^\circ x_s + B^\circ u_s)^T\widehat{P}_s(A^\circ x_s + B^\circ u_s)| \\ &\quad= |x_s^T(\widehat{A}_s + \widehat{B}_s\widehat{K}_s)^T\widehat{P}_s(\widehat{A}_s + \widehat{B}_s\widehat{K}_s)x_s - x_s^T(A^\circ + B^\circ\widehat{K}_s)^T\widehat{P}_s(A^\circ + B^\circ\widehat{K}_s)x_s| \\ &\quad\le \|x_s\|^2\|\widehat{P}_s\|\,\|(\widehat{A}_s + \widehat{B}_s\widehat{K}_s) - (A^\circ + B^\circ\widehat{K}_s)\|(\|\widehat{A}_s + \widehat{B}_s\widehat{K}_s\| + \|A^\circ + B^\circ\widehat{K}_s\|).\end{aligned}$$

Also, $\|\widehat{P}_s\|$ is uniformly bounded over time. The same holds for $(\|\widehat{A}_s+\widehat{B}_s\widehat{K}_s\|+\|A^\circ+B^\circ\widehat{K}_s\|)$. Furthermore, using the Schwarz inequality,

$$\frac{1}{t}\sum_{s=1}^{t}\|x_s\|^2\|(\widehat{A}_s+\widehat{B}_s\widehat{K}_s)-(A^\circ+B^\circ\widehat{K}_s)\|$$

$$\leq\left(\frac{1}{t}\sum_{s=1}^{t}\|x_s\|^4\right)^{1/2}\left(\frac{1}{t}\sum_{s=1}^{t}\|(\widehat{A}_s+\widehat{B}_s\widehat{K}_s)-(A^\circ+B^\circ\widehat{K}_s)\|^2\right)^{1/2}.$$

By Theorem 5 the first term is bounded. In light of Theorem 3, the second term can be handled analogously to the calculations for $t^{-1}\sum_{s=1}^{t}\|\widehat{P}_s-\widehat{P}_{s+1}\|^2$ in point (C), thus yielding

$$\lim_{t\to\infty}\frac{1}{t}\sum_{s=1}^{t}\|(\widehat{A}_s+\widehat{B}_s\widehat{K}_s)-(A^\circ+B^\circ\widehat{K}_s)\|^2=0\qquad\text{a.s.}$$

This suffices to prove that $D\to 0$, a.s.

By inserting all the partial results in (24) we finally obtain

$$\limsup_{t\to\infty}\frac{1}{t}\sum_{s=1}^{t}\left[x_s^TQx_s+u_s^TTu_s\right]\leq J^\circ\qquad\text{a.s.}$$

Since J° is the optimal cost for the true system, this proves the theorem. □

6. Concluding remarks. In an adaptive control context, the minimization of a given cost function is made difficult by the general identifiability problem stemming from the natural tendency of classical identification methods to return estimates with the corresponding optimal cost larger than the optimal cost for the true system. A way out of this problem is to employ a more fine-grained estimation scheme which exploits the properties of the set to which the estimates converge. Such a scheme has been presented and analyzed in this paper for the linear quadratic Gaussian control problem.

The results of this paper need to be extended in several directions to provide a fuller theory of optimal adaptation:

- *The presented scheme is nonrecursive.* However, one can conceive of somehow recursively minimizing our identification performance index so as to retain its asymptotic identification properties. This must be further investigated.
- *We assume full state observations.* This limitation needs to be removed.
- *Our adaptive scheme is, to some extent, tailored to linear quadratic Gaussian control.* In particular, a central role in the analysis is played by the uniqueness of the optimal gain in linear quadratic Gaussian control problems. It would be of interest to investigate how the biasing idea applies to other control strategies. An additional point is concerning the Gaussianity of the noise. This assumption is exploited in proving that the least squares estimate converges and that it tends to the true value in the excited subspace. In an attempt to remove the Gaussianity assumption one can use a weighted least squares algorithm, as suggested in [12], guaranteeing estimate convergence. In doing so, however, consistency in the excited subspace is lost and this may pose a difficulty in the derivation of many results.

• *Assumption $\mu_t = o(\log t)$ may be very conservative.* It is mainly motivated by the stability analysis and it is possible that our results still hold with μ_t growing at a faster rate. This and other choices made in the definition of our algorithm may be further investigated.

All the above problems suggest interesting research opportunities and a promise of self-optimizing adaptive control for nonlinear stochastic systems.

REFERENCES

[1] V. Borkar and P. P. Varaiya, *Adaptive control of Markov chains,* I: *Finite parameter set*, IEEE Trans. Automat. Control, 24 (1979), pp. 953–958.

[2] P. R. Kumar, *Optimal adaptive control of linear-quadratic-Gaussian systems*, SIAM J. Control Optim., 21 (1983), pp. 163–178.

[3] A. Becker, P. R. Kumar, and C. Z. Wei, *Adaptive control with the stochastic approximation algorithm: Geometry and convergence*, IEEE Trans. Automat. Control, 30 (1985), pp. 330–338.

[4] W. Lin, P. R. Kumar, and T. I. Seidman, *Will the self-tuning approach work for general cost criteria?*, Systems Control Lett., 6 (1985), pp. 77–85.

[5] T. E. Duncan and B. Pasik-Duncan, *Adaptive control of continuous-time linear stochastic systems*, Math. Control Signals Systems, 3 (1990), pp. 45–60.

[6] T. E. Duncan, P. Mandl, and B. Pasik-Duncan, *Control theory methods for consistency in some least squares identification problems,* IEEE Trans. Automat. Control, 38 (1993), pp. 1289–1292.

[7] H. F. Chen and L. Guo, *Convergence rate of least squares identification and adaptive control for stochastic systems*, Internat. J. Control, 44 (1986), pp. 1459–1476.

[8] H. F. Chen and L. Guo, *Asymptotically optimal adaptive control with consistent parameter estimates*, SIAM J. Control Optim., 25 (1987), pp. 558–575.

[9] H. F. Chen and L. Guo, *Optimal adaptive control and consistent parameter estimates for ARMAX model with quadratic cost*, SIAM J. Control Optim., 25 (1987), pp. 845–867.

[10] H. F. Chen and L. Guo, *A robust stochastic adaptive controller*, IEEE Trans. Automat. Control, 33 (1988), pp. 1035–1043.

[11] H. Chen and L. Guo, *Identification and Stochastic Adaptive Control*, Birkhäuser, Basel, 1991.

[12] L. Guo, *Self-convergence of weighted least-squares with applications to stochastic adaptive control*, IEEE Trans. Automat. Control, 41 (1996), pp. 79–89.

[13] P. R. Kumar and A. Becker, *A new family of optimal adaptive controllers for Markov chains*, IEEE Trans. Automat. Control, 27 (1982), pp. 137–146.

[14] W. Lin and P. R. Kumar, *Stochastic control of a queue with two servers of different rates*, in Analysis and Optimization of Systems, Lecture Notes in Control and Inform. Sci., A. Bensoussan and J. L. Lions, eds., Springer-Verlag, 1982, Chap. 44.

[15] P. R. Kumar, *Simultaneous identification and adaptive control of unknown systems over finite parameter sets*, IEEE Trans. Automat. Control, 28 (1983), pp. 68–76.

[16] P. R. Kumar and P. P. Varaiya, *Stochastic Systems: Estimation, Identification and Adaptive Control*, Prentice–Hall, Englewood Cliffs, NJ, 1986.

[17] D. P. Bertsekas, *Dynamic Programming: Deterministic and Stochastic Models*, Prentice–Hall, Englewood Cliffs, NJ, 1987.

[18] P. R. Kumar, *Convergence of least-squares parameter estimate based adaptive control schemes*, IEEE Trans. Automat. Control, 35 (1990), pp. 416–424.

[19] S. Bittanti, P. Bolzern, and M.C. Campi, *Recursive least squares identification algorithms with incomplete excitation: Convergence analysis and application to adaptive control*, IEEE Trans. Automat. Control, 35 (1990), pp. 1371–1373.

[20] H. Rootzen and J. Sternby, *Consistency in least squares estimation: A Bayesian approach*, Automatica, 20 (1984), pp. 471–475.

[21] M. C. Campi, *The problem of pole-zero cancellation in transfer function identification and application to adaptive stabilization*, Automatica, 32 (1996), pp. 849–857.

[22] D. Delchamps, *Analytic feedback control and the algebraic Riccati equation*, IEEE Trans. Automat. Control, 29 (1984), pp. 1031–1033.

[23] Y. Chow and H. Teicher, *Probability Theory: Independence, Interchangeability, Martingales*, 2nd ed., Springer-Verlag, New York, 1988.

[24] P. R. Kumar and T. I. Seidman, *Stability in the sense of bounded average power*, IMA J. Math. Control Inform., 2 (1985), pp. 109–122.
[25] J. M. Ortega, *Numerical Analysis: A Second Course*, Academic Press, New York, 1972.
[26] P. Hall and C. C. Heyde, *Martingale Limit Theory and Its Application*, Academic Press, Toronto, 1980.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 6, pp. 1908–1936, November 1998

004

MINIMIZING AND STATIONARY SEQUENCES OF CONSTRAINED OPTIMIZATION PROBLEMS*

CHIN-CHENG CHOU†, KUNG-FU NG‡, AND JONG-SHI PANG§

Abstract. We study some fundamental asymptotic properties associated with the well-posedness of constrained optimization problems. Emphasis is placed on the relation between minimizing and stationary sequences and their characterizations in terms of a set of asymptotic Karush–Kuhn–Tucker (KKT) optimality conditions. Unlike the subdifferential approach used by Auslender, Cominetti, and Crouzeix, we use a residual function approach that is closely tied to the theory of error bounds; this approach handles constraints explicitly and allows the effective treatment of infeasible, stationary sequences. The asymptotic KKT conditions provide an asymptotic optimality certificate for inequality constrained programs. Specializations of the results to convex quadratically constrained convex quadratic spline minimization problems and convex programs with Hölderian minima are discussed. An application of the results to the convergence of a family of Newton-type iterative descent methods involving singular quasi-Newton matrices for solving a constrained minimization problem is also presented.

Key words. well-posed optimization problems, minimizing sequences, stationary sequences, approximate optimality conditions, convex analysis, residual functions, error bounds, weak sharp minima, quadratic splines, convergence of algorithms

AMS subject classifications. 90C30, 90C33

PII. S0363012995292548

1. Introduction.

This paper studies some fundamental asymptotic aspects of the finite-dimensional, constrained, differentiable optimization problem:

$$\begin{aligned} &\text{minimize} && \theta(x) \\ &\text{subject to} && x \in X, \end{aligned} \tag{1}$$

where $\theta : \Re^n \to \Re$ is a continuously differentiable function and X is a nonempty closed subset of $\Re^n$. We write

$$\theta_{\inf} \equiv \inf_{x \in X} \theta(x) \geq -\infty.$$

Throughout the paper we do not assume, unless explicitly stated, that $\theta_{\inf}$ is finite or a global minimizer of (1) exists. Thus the theory developed herein is applicable to problems (1) whose optimum objective values are not necessarily attained.

*Received by the editors September 29, 1995; accepted for publication (in revised form) September 5, 1997; published electronically July 29, 1998. This research was carried out while the first and third authors were visiting the Department of Mathematics, Chinese University of Hong Kong, at the invitation of the second author. This visit was made possible by financial support from the Institute of Mathematical Sciences, Chinese University of Hong Kong, and the Research Grant Council of Hong Kong (Direct and Earmarked grants). Research of the third author was also partially supported by National Science Foundation grant CCR-9213739 and Office of Naval Research grant N00014-93-1-0228.

http://www.siam.org/journals/sicon/36-6/29254.html

†Département de Mathématiques, Université de Perpignan, 66025 Perpignan, France (chou@syspo.univ-perp.fr).

‡Department of Mathematics, The Chinese University of Hong Kong, Shatin, New Territory, Hong Kong (kfng@math.cuhk.hk).

§Department of Mathematical Sciences, The Johns Hopkins University, Baltimore, MD 21218-2682 (jsp@vicp1.mts.jhu.edu).

The asymptotic properties studied in this work are closely tied to the theory of well-posedness for constrained optimization problems [8, 21]. According to the introductory discussion in these two books, the well-posedness concept for minimization problems originates with A.N. Tykhonov [36]. Whereas the original Tykhonov well-posedness concept was concerned with an unconstrained optimization problem, the extension to the constrained case was introduced by Levitin and Polyak [17], who were interested in the convergence analysis of numerical methods for solving constrained optimization problems. Central to the Levitin–Polyak theory of well-posedness for the problem (1) is the concept of a minimizing sequence. Specifically, a sequence $\{x^k\} \subset \Re^n$ is said to be *Levitin–Polyak minimizing* (or in short, LP minimizing) for (1) if

(i) it is *asymptotically feasible*; i.e.,

$$\lim_{k\to\infty} \operatorname{dist}(x^k, X) = 0,$$

where $\operatorname{dist}(x, X)$ is the distance function from a vector $x \in \Re^n$ to the set X measured in the Euclidean norm; and

(ii) it is *asymptotically optimal*; i.e.,

$$\lim_{k\to\infty} \theta(x^k) = \theta_{\inf}.$$

Based on this fundamental definition, various well-posedness concepts can be defined for the problem (1); see [8, 21, 31]. Invariably, all of these concepts assume that the optimization problem in question has a global minimizer which may or may not be unique. In contrast, many results obtained in the present paper do not require that (1) attains its minimum. This is an important point of departure of our work from the classical well-posedness studies.

Independently of the classical Tykhonov–Levitin–Polyak theory of well-posedness, a closely related theory of well-behaved convex functions is developed in the papers by Auslender, Crouzeix, and Cominetti [3, 2, 1]. Cast in the framework of extended-valued convex functions [33], the latter theory studies the class of well-behaved convex functions and their role in the convergence theory of iterative minimization methods. Specifically, a proper closed convex function $f : \Re^n \to \Re \cup \{\infty\}$ is said to be well-behaved if for all sequences $\{x^k\}$ and $\{a^k\}$ such that $\lim_{k\to\infty} a^k = 0$ and $a^k \in \partial f(x^k)$ for all k, where ∂f denotes the subdifferential of f, we have $\lim_{k\to\infty} f(x^k) = \inf\{f(x) : x \in \Re^n\}$. Note that a well-behaved convex function is not required to attain its global minimum. Characterizations of well-behaved proper closed convex functions are obtained in [3]. A large subclass of these functions, denoted $\mathcal{R}$, is introduced in [2] which consists of those proper closed convex functions f such that the origin is an element of the relative interior of the domain of the subdifferential of the conjugate f^* of f. Some nice properties of functions in this subclass are obtained; among these, a function in the class $\mathcal{R}$ must have a nonempty set of global minimizers. In a later section, we will discuss how this subdifferential approach is related to the Levitin–Polyak approach, which handles constraints explicitly.

In an interesting paper [15], Lemaire refined the analysis in [3, 2, 1] and introduced several additional asymptotic concepts that tie together the theory of well-behaved convex functions in a Banach space with that of well-posed optimization problems. Since an important goal of Lemaire's study was to connect these two theories, he had assumed that his problems all attained their global minima. Further details of Lemaire's results are given in sections 4.1 and 4.3.

This paper has two major goals. One is to study the detailed connection between an LP minimizing sequence and a "stationary sequence"; the other is to investigate how these sequences can be characterized in terms of some asymptotic KKT-type optimality conditions. We use two classes of constrained optimization problems to illustrate our results: one class involves the minimization of a "convex quadratic spline function" subject to convex quadratic inequality constraints; the other class consists of convex programs with "Hölderian minima." An application of the derived results to the convergence of a family of Newton-type iterative descent methods involving singular quasi-Newton matrices for solving the problem (1) is also presented.

Unlike the subdifferential approach employed by Auslender and Crouzeix [3] and Lemaire [15], our definition of a stationary sequence (see the next section) is based on the concept of a residual function that is derived from the theory of error bounds. There are multiple reasons to introduce a residual-based concept of asymptotic stationarity. One is that this approach handles constraints more effectively (than the subdifferential approach) and allows the treatment of infeasible sequences. Another motivation is that residuals are computable quantities often used to define stopping rules in the practical implementation of iterative methods; thus it is reasonable to develop an asymptotic theory that is based on these residuals.

Apart from the difference in the employed approaches, an important topic included in our study that has not been treated by Auslender, Cominetti, Crouzeix and Lemaire is the converse question: Are LP minimizing sequences necessarily stationary? As we shall see, the answer to this question is in the negative in general; it turns out that the well-known ε-variational principle by Ekeland [9] has a key role to play in the analysis of this question. In addition to this question, we shall consider an alternate definition of a minimizing sequence in the case where the problem (1) actually has a global minimizer; we shall investigate when this alternate notion and the notion of an LP minimizing sequence are equivalent.

The dominant role of the (exact) KKT optimality conditions for the study of inequality constrained nonlinear programs is well known. Inspired by such a role, we will introduce a set of asymptotic KKT conditions as a characterization of an LP minimizing sequence for these programs. To the best of our knowledge, these asymptotic optimality conditions have never been formally considered in optimization theory.

Although the motivation of studying minimizing sequences stems from a computational consideration, one should be cautious in directly applying the results obtained herein to sequences generated by specific algorithms [35]. The principal reason for such caution is that, by the way they are generated, the latter sequences typically possess additional properties that are not taken into account in a general study of this type. In particular the convergence results in the papers [14, 38, 40] that pertain to specific descent methods should not be inferred as immediate consequences of the results in this work and the theory of well-posedness. Instead, the latter theory is developed in order to identify key properties of optimization problems that will render these problems to behave well when solved by numerical methods of a broad nature. In such an algorithmic context, there has been some initial success of the theory [3, 2]. The present paper will add to this success by demonstrating that a condition important for LP minimizing sequences is necessary and sufficient for the convergence of the (singular) Newton-type descent methods for solving the constrained minimization problem (1). This result provides a new piece of evidence showing that by focusing on the class of "well-behaved" constrained optimization problems, convergence of some

well-known algorithms can be established under assumptions that are less restrictive than those usually made in the literature.

The rest of this paper is divided into several sections. In the next section we introduce several asymptotic stationarity concepts and clarify their relationships. Sections 3 and 4 treat the main topic of this paper, namely, the connection between minimizing and stationary sequences. Section 5 discusses some approximate optimality systems and how they are related to LP minimizing sequences. Two results that summarize the various properties studied in the paper will be presented there. Finally, in section 6, we establish the convergence of the (singular) Newton-type methods for solving the problem (1), using the arguments established in the previous sections.

In addition to the results obtained herein, the accompanying paper [13] studies minimizing sequences of merit functions for nonlinear complementarity problems and variational inequalities; a distinguishing feature of these functions is that they are typically nonconvex but are nonnegative.

2. Stationary sequences. Associated with the constrained optimization problem (1), we have defined the concept of an LP minimizing sequence. Another central concept in the asymptotic analysis of this problem is that of a stationary sequence. For an unconstrained problem which has $X = \Re^n$ and a differentiable objective function θ, the concept of a stationary sequence can be defined easily. Namely, a sequence $\{x^k\}$ is stationary if $\{\nabla\theta(x^k)\}$ converges to zero. Nevertheless, for a problem with constraints which has X being a proper subset of $\Re^n$, there are several closely related definitions. In this section, we shall present the various definitions of asymptotic stationarity and clarify their interrelationships.

2.1. Residual-based asymptotic stationarity. In order to define the concept of a stationary sequence of (1), we introduce the residual function

$$R_N(x) \equiv x - \Pi_X(x - \nabla\theta(x)), \quad x \in \Re^n,$$

where Π_X is the Euclidean projector onto the set X. Clearly, when $X = \Re^n$, $R_N(x)$ reduces to $\nabla\theta(x)$. Moreover it follows from the well-known variational principle of nonlinear programming that if x is a local minimum of (1), then $R_N(x) = 0$; conversely, if θ is convex, then every zero of R_N is a global minimum of (1).

The residual function $R_N(x)$ plays an important role in the error bound theory for variational inequalities and complementarity problems; see [28, 29]. In [27], the term "natural residual" was coined for this function. Borrowing this terminology (which explains the subscript N in the function R_N), we say that a sequence $\{x^k\} \subset \Re^n$ is *naturally stationary* (or in short, N-stationary) for the constrained optimization problem (1) if $\lim_{k\to\infty} R_N(x^k) = 0$. This definition does not require the sequence $\{x^k\}$ to be feasible to (1). Nevertheless it is easy to see that

$$\lim_{k\to\infty} R_N(x^k) = 0 \implies \lim_{k\to\infty} \operatorname{dist}(x^k, X) = 0.$$

Thus if $\{x^k\}$ is N-stationary, then it must be asymptotically feasible. We note a fundamental property of the residual vector $R_N(x)$. Namely, for all vectors $x \in \Re^n$ and $y \in X$,

$$(y - x + R_N(x))^T(\nabla\theta(x) - R_N(x)) \geq 0; \tag{2}$$

this inequality is a consequence of the variational principle for the Euclidean projector.

An alternative definition of a stationary sequence is based on the normal map for the problem (1). Like the function R_N, the normal map is also fundamental for variational inequalities and complementarity problems [32]. Specifically, associated with the pair $(\nabla\theta, X)$, the normal map $R_{\mathcal{N}} : \Re^n \to \Re^n$ is defined as

$$R_{\mathcal{N}}(z) \equiv \nabla\theta \circ \Pi_X(z) + z - \Pi_X(z), \quad z \in \Re^n.$$

We say that a sequence $\{x^k\} \subset \Re^n$ is *normally stationary* (or in short, $\mathcal{N}$-stationary) for the constrained optimization problem (1) if

(i) $\{x^k\}$ is asymptotically feasible, and

(ii) there exists a sequence $\{z^k\} \subset \Re^n$ such that $\lim_{k\to\infty} R_{\mathcal{N}}(z^k) = 0$ and $\Pi_X(x^k) = \Pi_X(z^k)$ for each k.

Unlike the concept of N-stationarity, in which the asymptotic feasibility of the sequence $\{x^k\}$ is an easy consequence of the limit condition $\lim_{k\to\infty} R_N(x^k) = 0$, we give below an example which shows that it is possible for a sequence $\{x^k\}$ to satisfy condition (ii) in the definition of $\mathcal{N}$-stationarity and fail condition (i) in the same definition. Thus for such a sequence $\{x^k\}$, we cannot expect it to be LP minimizing because asymptotic feasibility is part of the requirement of the LP minimizing property.

Example 1. Let $\theta(x_1, x_2) \equiv \cos^2(x_1 x_2)$ and

$$X \equiv \{\, (x_1, x_2) \in \Re^2 \,:\, (x_1, x_2) \leq 0,\ x_1 x_2 \geq \pi/2 \,\}.$$

We have $\theta_{\inf} = 0$. Consider the sequence $\{x^k\}$ with $x^k \equiv (1, -k\pi/2)$ for all k. Clearly,

$$\theta(x^k) = \begin{cases} 0 & \text{if } k \text{ is odd,} \\ 1 & \text{if } k \text{ is even,} \end{cases}$$

and $\operatorname{dist}(x^k, X) \geq 1$ for all k. Thus $\{x^k\}$ is neither asymptotically feasible nor asymptotically minimizing. Yet the projected vector $\bar{x}^k = (\bar{x}^k_1, \bar{x}^k_2)$ must satisfy $\bar{x}^k_1 \bar{x}^k_2 = \pi/2$ and $\nabla\theta(\bar{x}^k_1, \bar{x}^k_2) = (0, 0)$. Thus with $z^k \equiv \bar{x}^k$, we have $R_{\mathcal{N}}(z^k) = 0$ for all k. Consequently condition (ii) in the definition of $\mathcal{N}$-stationarity holds for the sequence $\{x^k\}$, but condition (i) fails. Thus for this sequence $\{x^k\}$, the corresponding projected sequence $\{\bar{x}^k\}$ is LP minimizing, N-stationary, and $\mathcal{N}$-stationary; yet the sequence $\{x^k\}$ itself does not satisfy any of the asymptotic properties defined so far.

The above example illustrates an important feature of the natural residual function $R_N(x)$ which is not shared by the normal residual function $R_{\mathcal{N}}(z)$. Namely, the former residual function itself captures both the asymptotic feasibility and minimizing property of an infinite sequence, whereas the latter residual function alone is not sufficient to handle infeasible sequences.

Since residual functions are practical tools employed in termination rules of iterative methods, it would be useful for us to say a few words about the potential utility of the two residual functions $R_N(x)$ and $R_{\mathcal{N}}(z)$ in this regard. In an iterative algorithm for solving the constrained optimization problem (1), one generates an infinite sequence of iterates $\{x^k\}$ which is not always feasible to the problem; sometimes an auxiliary sequence $\{z^k\}$ is also obtained in the computational process. If this is the situation, then both residual functions $R_N(x^k)$ and $R_{\mathcal{N}}(z^k)$ can legitimately be used in tests for terminating the iterative process (although as we have seen from the above example, one needs to be somewhat cautious in using $R_{\mathcal{N}}(z^k)$ as the sole termination indicator). If no auxiliary sequence $\{z^k\}$ is readily available, then it may be

computationally difficult to employ the residual function $R_{\mathcal{N}}$ to verify the asymptotic stationarity of $\{x^k\}$; in this case, $R_N(x^k)$ is the natural choice.

For later purposes, it is useful to state a relation between the two maps $R_N(x)$ and $R_{\mathcal{N}}(z)$ as follows. For $\bar{x} \equiv \Pi_X(z)$, we have

$$R_N(\bar{x}) \;=\; \Pi_X(\bar{x} - \nabla\theta(\bar{x}) + R_{\mathcal{N}}(z)) - \Pi_X(\bar{x} - \nabla\theta(\bar{x})),$$

which implies, by the nonexpansiveness of the projection,

$$\| R_N(\bar{x}) \| \;\leq\; \| R_{\mathcal{N}}(z) \|.$$

Consequently, for any two sequences $\{\bar{x}^k\}$ and $\{z^k\}$ such that $\bar{x}^k \equiv \Pi_X(z^k)$, we have

$$\lim_{k\to\infty} R_{\mathcal{N}}(z^k) \;=\; 0 \;\Longrightarrow\; \lim_{k\to\infty} R_N(\bar{x}^k) \;=\; 0. \tag{3}$$

2.2. Subdifferential-based asymptotic stationarity. Since every constrained optimization problem can be equivalently stated as an unconstrained problem with the use of the indicator function of the feasible set, one can also define the concept of asymptotic stationarity using the subdifferential approach of Auslender and Crouzeix [3] (this concept was not formally defined in the reference). In order to introduce this definition, we recall such standard notation as $\partial\phi$ and $\mathrm{dom}(\partial\phi)$ for the subdifferential and its domain of an (extended-valued) convex function ϕ; we also use $\partial_\varepsilon\phi$ for the ε-subdifferential of ϕ [16]. If ϕ is the indicator function (denoted $\mathbf{I}_S$) of a closed convex set $S \subseteq \Re^n$ (that is, $\mathbf{I}_S(x)$ is equal to 0 if $x \in S$ and equal to ∞ if $x \notin S$), then $\partial_\varepsilon\phi(x)$ coincides with the set of ε-normals to the set S at the vector $x \in S$; that is,

$$\partial_\varepsilon \mathbf{I}_S(x) \;=\; \{\, v \in \Re^n \,:\, v^T(y - x) \;\leq\; \varepsilon \text{ for all } y \in S\}.$$

Consider the problem (1), where we assume that θ is convex. Let

$$\phi(x) \;\equiv\; \theta(x) + \mathbf{I}_X(x) \quad \forall\, x \in \Re^n. \tag{4}$$

We say that a sequence $\{x^k\} \subset \Re^n$ is *AC-stationary* (AC for Auslender and Crouzeix) if

(i) $\{x^k\}$ is asymptotically feasible, and
(ii) the projected sequence $\{\bar{x}^k\}$, where $\bar{x}^k \equiv \Pi_X(x^k)$ for all k, has the property that for each k there exists $a^k \in \partial\phi(\bar{x}^k)$ and the sequence of subgradients $\{a^k\}$ converges to zero.

(Note: Since $\mathrm{dom}(\partial\phi) \subseteq X$, we need to use the projected sequence in condition (ii) in order to allow for the possibility that the original sequence $\{x^k\}$ is not feasible.) We say that the sequence $\{x^k\} \subset \Re^n$ is AC_ε-stationary if

(i) $\{x^k\}$ is asymptotically feasible, and
(ii) the projected sequence $\{\bar{x}^k\}$ has the property that for some sequence of nonnegative scalars $\{\varepsilon_k\}$ converging to zero, there exists for each k a vector $a^k \in \partial_{\varepsilon_k}\phi(\bar{x}^k)$, and the sequence of ε-subgradients $\{a^k\}$ converges to zero, where $\partial_\varepsilon\phi$ denotes the ε-subdifferential of ϕ.

2.3. Connections. As mentioned in the Introduction, Auslender, Crouzeix, and Lemaire used the subdifferential approach to analyze the connection between an asymptotically stationary sequence and an LP minimizing sequence. Their analysis dealt principally with feasible sequences $\{\bar{x}^k\} \subset X$, satisfying requirement (ii) in the definition of an AC-stationary sequence; we broaden this treatment by permitting the sequence $\{x^k\}$ to be infeasible.

We say that a function $F : \Re^n \to \Re^m$ is *uniformly continuous* near a sequence $\{x^k\}$ if, for every $\varepsilon > 0$, there exists a $\delta > 0$ such that for all k and y,

$$\|y - x^k\| \leq \delta \;\Rightarrow\; \|F(y) - F(x^k)\| \leq \varepsilon.$$

Uniform continuity is known to have an important role to play in the asymptotic well-posedness theory of optimization problems; see [8]. There is no exception in our work.

In what follows, we present a result that clarifies the relationship between the asymptotic stationarity concepts defined so far. Specifically, this result states that AC-stationarity is equivalent to $\mathcal{N}$-stationarity, whereas these two concepts are not necessarily equivalent to N-stationarity without some restrictions.

PROPOSITION 2.1. *Let $\theta : \Re^n \to \Re$ be a convex, continuously differentiable function and X a closed convex subset of $\Re^n$. Let $\{x^k\}$ be an arbitrary sequence of vectors in $\Re^n$. The following statements are valid.*

(a) *The sequence $\{x^k\}$ is AC-stationary if and only if it is $\mathcal{N}$-stationary.*
(b) *If $\{x^k\}$ is AC-stationary and each $x^k \in X$, then $\{x^k\}$ is N-stationary.*
(c) *If $\{x^k\}$ is AC-stationary and $\nabla\theta$ is uniformly continuous near $\{x^k\}$, then $\{x^k\}$ is N-stationary.*
(d) *If $\{x^k\}$ is N-stationary, $\{\nabla\theta(x^k)\}$ is bounded, and $\nabla\theta$ is uniformly continuous near $\{x^k\}$, then $\{x^k\}$ is AC_ε-stationary.*

Proof. Write $\bar{x}^k \equiv \Pi_X(x^k)$. Let ϕ be defined by (4). Since θ is differentiable, it follows that

$$\partial\phi(x) \;=\; \nabla\theta(x) + \partial\mathbf{I}_X(x), \quad \text{for all } x \in \Re^n.$$

To prove (a), let $\{x^k\}$ be AC-stationary. Let $\{a^k\}$ be a sequence of vectors converging to zero such that $a^k \in \partial\phi(\bar{x}^k)$ for each k. For each k, let $b^k \in \partial\mathbf{I}_X(\bar{x}^k)$ be such that $a^k = \nabla\theta(\bar{x}^k) + b^k$. Define $z^k \equiv \bar{x}^k + b^k$. It is then easy to verify that $\Pi_X(z^k) = \bar{x}^k$ and $R_{\mathcal{N}}(z^k) = a^k$. This establishes the "only if" statement in (a). The converse can be proved easily be reversing the argument. The details are omitted. Thus (a) holds.

To prove (b) and (c), let $\{x^k\}$ be AC-stationary. Let $\{z^k\}$ be the auxiliary sequence as stated in condition (ii) of AC-stationarity. By (3), it follows that $\lim_{k\to\infty} R_N(\bar{x}^k) = 0$. If each x^k belongs to X, then $x^k = \bar{x}^k$ and (b) follows readily. Instead, if $\nabla\theta$ is uniformly continuous near $\{x^k\}$, then since $\lim_{k\to\infty} \|x^k - \bar{x}^k\| = 0$, it follows that $\lim_{k\to\infty} R_N(x^k) = 0$ also, establishing (c).

To prove (d), let $\{x^k\}$ be N-stationary. It has been noted this implies $\{x^k\}$ is asymptotically feasible. Moreover, the uniform continuity of $\nabla\theta$ near $\{x^k\}$ implies that $\lim_{k\to\infty} R_N(\bar{x}^k) = 0$ and $\{\nabla\theta(\bar{x}^k)\}$ is bounded. Hence by letting

$$\varepsilon_k \;\equiv\; |\, R_N(\bar{x}^k)^T \nabla\theta(\bar{x}^k)\,|,$$

the sequence $\{\varepsilon_k\}$ converges to zero. It remains to show that $R_N(\bar{x}^k) \in \partial_{\varepsilon_k}\phi(\bar{x}^k)$, or equivalently, for all $y \in X$,

$$(R_N(\bar{x}^k) - \nabla\theta(\bar{x}^k))^T(y - \bar{x}^k) \;\leq\; \varepsilon_k.$$

By (2), we have

$$(R_N(\bar{x}^k) - \nabla\theta(\bar{x}^k))^T(y - \bar{x}^k) \;\leq\; R_N(\bar{x}^k)^T(\nabla\theta(\bar{x}^k) - R_N(\bar{x}^k)) \;\leq\; \varepsilon_k,$$

as desired. ∎

3. Minimizing ⇒ stationary. We are now ready to investigate the detailed relationships between an LP minimizing sequence and an asymptotically stationary sequence for the constrained optimization problem (1). Throughout the analysis, we let $\{x^k\} \subset \Re^n$ be an arbitrary sequence satisfying $\theta(x^k) > \theta_{\inf}$ for all k. We do *not* assume that $\{x^k\}$ is bounded.

This section deals with the issue stated in its heading. Specifically, we wish to answer the question, If $\{x^k\}$ is an LP minimizing sequence, is it necessarily an N-stationary sequence? For this part of the analysis, we need $\theta_{\inf}$ to be finite but do not need the convexity of θ. Thus the assumption that $\theta_{\inf} > -\infty$ is made throughout this section. A remark about the assumption is made at the closing of the section.

We begin by giving an example to show that a minimizing sequence is not necessarily stationary. This example illustrates the difference between an unbounded minimizing sequence and a bounded minimizing sequence and provides the motivation for the remaining study. Since this example deals with an unconstrained optimization problem (with $X = \Re^2$), there is no distinction between $\mathcal{N}$-stationarity and N-stationarity; in this case, a sequence $\{x^k\}$ is stationary if $\lim_{k\to\infty} \nabla\theta(x^k) = 0$.

Example 2. Consider an unconstrained optimization problem with the objective function given by

$$\theta(x_1, x_2) \equiv e^{x_1^2 - x_2}, \quad (x_1, x_2) \in \Re^2.$$

The function θ is convex and differentiable with an infimum value of zero that is not attained. Consider the sequence $\{x^k\}$ defined by

$$x^k = (x_1^k, x_2^k) \equiv (k, k^2 + \tfrac{1}{2}\log k).$$

It is trivial to check that

$$\theta(x^k) = 1/\sqrt{k} \quad \text{and} \quad \nabla\theta(x^k) = \begin{pmatrix} 2\sqrt{k} \\ -1/\sqrt{k} \end{pmatrix}.$$

Thus $\{x^k\}$ is minimizing but not stationary. Alternatively, consider the perturbed sequence $\{y^k\}$ defined by

$$y^k = (y_1^k, y_2^k) \equiv (k - 1/\sqrt{k}, k^2 + \tfrac{1}{2}\log k).$$

It is easy to verify that $\{y^k\}$ is both minimizing and stationary; moreover, $\{x^k - y^k\} \to 0$.

The above example illustrates an interesting phenomenon; namely, although the sequence $\{x^k\}$ is not stationary, we have identified a nearby sequence $\{y^k\}$ that is both minimizing and stationary. This phenomenon is not incidental, it is actually a fact as stated in the proposition below. The proof of this proposition uses the well-known ε-variational principle due to Ekeland [9]. Our original proof of the result used a smooth variant of this principle obtained recently by Deville, Godefroy, and Zizler [7]. The present proof is inspired by a referee's comment.

PROPOSITION 3.1. *Let X be a nonempty closed convex subset of $\Re^n$, and let $\theta : \Re^n \to \Re$ be a continuously differentiable function with θ_{inf} finite. Let $\{x^k\} \subset \Re^n$ be an LP minimizing sequence of θ on X. If θ is uniformly continuous near $\{x^k\}$, then there exists a nearby feasible sequence $\{y^k\} \subset X$ satisfying*

$$\text{(i)} \lim_{k\to\infty}(x^k - y^k) = 0, \quad \text{(ii)} \lim_{k\to\infty} \theta(y^k) = \theta_{inf}, \quad \textit{and} \quad \text{(iii)} \lim_{k\to\infty} R_N(y^k) = 0.$$

Proof. For each k, let $\bar{x}^k \equiv \Pi_X(x^k)$; then $\lim_{k\to\infty} \|x^k - \bar{x}^k\| = 0$ because $\{x^k\}$ is asymptotically feasible. Since θ is uniformly continuous near $\{x^k\}$, it follows that $\{\bar{x}^k\}$ is a feasible LP minimizing sequence. Take an arbitrary sequence of positive scalars $\{\varepsilon_k\}$ such that

$$\lim_{k\to\infty} \varepsilon_k = 0 \quad \text{and} \quad \theta(\bar{x}^k) < \theta_{\inf} + \varepsilon_k \quad \forall\, k.$$

Let $\phi(x)$ be defined by (4). We clearly have

$$\phi_{\inf} \equiv \inf_{x\in\Re^n} \phi(x) = \theta_{\inf},$$

and for each k, $\phi(\bar{x}^k) < \phi_{\inf} + \varepsilon_k$. By Ekeland's variational principle, there exists a vector y^k such that with

$$g_k(y) \equiv \sqrt{\varepsilon_k}\, \| y - y^k \|, \quad y \in \Re^n,$$

(a) $\phi + g_k$ attains its global minimum at y^k, (b) $\phi(y^k) \le \phi(\bar{x}^k)$, and (c) $\|\bar{x}^k - y^k\| \le \sqrt{\varepsilon_k}$. Clearly, $\lim_{k\to\infty}(x^k - y^k) = 0$. Moreover, $y^k \in X$ because ϕ takes the value ∞ outside X. Indeed y^k is a global minimizer of the problem

$$\begin{aligned} \text{minimize} \quad & \theta(x) + g_k(x) \\ \text{subject to} \quad & x \in X. \end{aligned}$$

Thus $\lim_{k\to\infty} \theta(y^k) = \theta_{\inf}$. For any $y \in X$, the directional derivative of $\theta + g_k$ at y^k along $y - y^k$ must be nonnegative. So there exists $w^k \in \partial(\theta + g_k)(y^k)$ such that $(w^k)^T(y - y^k) \ge 0$. By [6, Propositions 2.3.3 and 2.1.1], w^k can be written as $\nabla\theta(y^k) + z^k$ with $\|z^k\| \le \sqrt{\varepsilon_k}$. Since $y^k = \Pi_X(y^k - w^k)$, the global nonexpansiveness of the Euclidean projector and the fact that $w^k - \nabla\theta(y^k) = z^k \to 0$ imply

$$\lim_{k\to\infty} \left[\, y^k - \Pi_X(y^k - \nabla\theta(y^k)) \,\right] = 0;$$

that is,

$$\lim_{k\to\infty} R_N(y^k) = 0$$

as desired. ∎

We remark that the only place in the above proof where the uniform continuity of θ is used is to establish that the projected sequence $\{\bar{x}^k\}$ is also LP minimizing. In particular, this assumption can be dropped if the given sequence $\{x^k\}$ is already feasible to (1). As the following example shows, in general, if $\{x^k\}$ is LP minimizing, the projected sequence $\{\bar{x}^k\}$ is not necessarily LP minimizing if θ is not uniformly continuous near $\{x^k\}$. (See also Example 1, which gives a sequence $\{x^k\}$ for which the projected sequence $\{\bar{x}^k\}$ is LP minimizing but $\{x^k\}$ itself is not.)

Example 3. Let $\theta(x_1, x_2) \equiv \sin(x_1 x_2)$ and

$$X \equiv \{(x_1, x_2) \in \Re^2 : (x_1, x_2) \le 0,\, x_1 x_2 \ge \pi/2\}.$$

We have $\theta_{\inf} = -1$. Consider the sequence $\{x^k\}$ with $x^k \equiv (1/k, -k\pi/2)$ for all k. Since the distance between x^k and the vector $y^k \equiv (-1/k, -k\pi/2) \in X$ approaches zero as $k \to \infty$, it follows that $\{x^k\}$ is asymptotically feasible; thus $\{x^k\}$ is an LP minimizing sequence for the constrained program (1). Since each projected vector

$\bar{x}^k = (\bar{x}_1^k, \bar{x}_2^k)$ must satisfy $\bar{x}_1^k \bar{x}_2^k = \pi/2$, it follows that the sequence $\{\bar{x}^k\}$ is not LP minimizing.

An immediate consequence of Proposition 3.1 is the following result, which does not require a proof.

THEOREM 3.2. *Let X be a nonempty closed convex subset of $\Re^n$, and let $\theta : \Re^n \to \Re$ be a continuously differentiable function with θ_{inf} finite. Let $\{x^k\} \subset \Re^n$ be an LP minimizing sequence of θ on X. If θ and $\nabla\theta$ are uniformly continuous near $\{x^k\}$, then $\{x^k\}$ is an N-stationary sequence.*

The finiteness assumption of $\theta_{\inf}$ is essential for Ekeland's variational principle to be applicable in the proof of Proposition 3.1. Indeed Theorem 3.2 fails to hold without this assumption. A trivial counterexample is the one-dimensional unconstrained problem with $\theta(x) \equiv x$, $x \in X \equiv \Re$. Clearly, $\theta_{\inf} = -\infty$. Since the derivative is equal to the constant 1, there is no stationary sequence; yet any sequence that tends to $-\infty$ is minimizing.

4. Stationary $\Rightarrow$ minimizing. We next turn to the converse of Theorem 3.2. Although this issue has been treated to a reasonable extent in the papers [3, 2, 1], our treatment offers a fresh perspective and additional insights for the results. Consistent with the residual approach, we stress the importance of error bounds for the level sets of (1) in our treatment. For this part of the analysis, we need the convexity of θ but do not need the finiteness of $\theta_{\inf}$.

4.1. The role of error bounds. As evidenced in the work of Auslender and Crouzeix, a key element in the treatment of the issue stated in this section's heading is the theory of error bounds for the level sets of the problem (1). In light of the recent advances in this theory as described in the survey [29], we find it useful to give a summary of the relevant error bound results that are useful for our purpose here.

First, we introduce a concept for the problem (1). Specifically, we say that this problem has *H-metrically regular level sets*, or in short, (1) is H-metrically regular (H for Hölderian) if for every scalar $\lambda > \theta_{\inf}$, there exist positive scalars c and γ (possibly depending on λ) with $\gamma < 1$ such that

$$\operatorname{dist}(x, L(\lambda)) \leq c\, r_\gamma(x) \quad \forall\, x \in X, \tag{5}$$

where $L(\lambda)$ is the λ-level set of (1); that is,

$$L(\lambda) \equiv \{\, x \in X : \theta(x) \leq \lambda \,\},$$

and $r_\gamma(x)$ is the following residual for $L(\lambda)$:

$$r_\gamma(x) \equiv \max\left([\, (\theta(x) - \lambda)_+ \,]^\gamma,\ (\theta(x) - \lambda)_+ \right).$$

The function $r_\gamma(x)$ is a computable measure of the violation of the constraints by vectors $x \in X$ that fail to be in the λ-level set. (A remark: $L(\lambda)$ is nonempty for all $\lambda > \theta_{\inf}$.) In a nutshell, the H-metric regularity of (1) stipulates that error bounds of a Hölderian type [29] hold for the λ-level sets of this problem for all $\lambda > \theta_{\inf}$. Note that we do not require the exponent γ to be the same for all λ. If error bounds of a Lipschitzian type hold for all of these level sets (i.e., (5) holds with $\gamma = 1$ for all $\lambda > \theta_{\inf}$), then we say that (1) is L(ipschitzian)-metrically regular.

The H- or L-metric regularity of (1) does not require this problem to attain a finite optimum objective value. If the set of optimal solutions

$$X_{\text{opt}} \equiv \{\, x \in X : \theta(x) \leq \theta_{\inf} \,\}$$

is nonempty, several additional concepts can be defined. Specifically, (1) is said to have *weak sharp minima* if $X_{\rm opt}$ is nonempty and a Lipschitzian error bound holds for $X_{\rm opt}$, i.e., there exists a constant $c > 0$ such that

$$\mathrm{dist}(x, X_{\rm opt}) \leq c\,(\,\theta(x) - \theta_{\rm inf}\,) \quad \forall\, x \in X.$$

This concept was introduced in Michael Ferris's Ph.D. dissertation [10], and its roles are investigated extensively in [5, 11]. An obvious generalization of this definition is the following. We say that (1) has *Hölderian minima* if $X_{\rm opt}$ is nonempty and an Hölderian error bound holds for $X_{\rm opt}$; that is, there exist positive constants c and γ such that (5) holds for $\lambda = \theta_{\rm opt}$.

The concept of ψ-sharp minima has played an important role in the Tykhonov–Levitin–Polyak well-posedness theory of optimization problems. A continuous function $\psi : [0,\infty) \to [0,\infty)$ is said to be a *forcing function* if for every sequence $\{t_k\} \subset [0,\infty)$,

$$\lim_{k\to\infty} \psi(t_k) = 0 \implies \lim_{k\to\infty} t_k = 0.$$

We say that the problem (1) has ψ-sharp minima if $X_{\rm opt}$ is nonempty and

$$\theta(x) \geq \theta_{\rm inf} + \psi(\mathrm{dist}(x, X_{\rm opt})) \quad \forall\, x \in X,$$

where ψ is a given forcing function. Adopting Lemaire's terminology [15] to our setting, we say that (1) has *well-conditioned minima* if $X_{\rm opt}$ is nonempty and there exists a forcing function ψ such that (1) has ψ-sharp minima. If the forcing function satisfies the stronger property that for every sequence $\{t_k\} \subset (0,\infty)$,

$$\lim_{k\to\infty} \frac{\psi(t_k)}{t_k} = 0 \implies \lim_{k\to\infty} t_k = 0,$$

we use the terminology "ψ-very-sharp minima" and "very-well-conditioned minima," respectively.

Clearly, weak sharp minima is a special case of well-conditioned minima with the forcing function being a positive multiple of the identity function; more generally, Hölderian minima can also be shown to imply well-conditioned minima with a forcing function ψ being the inverse of the strictly increasing function $s \in [0,\infty) \mapsto c\,(s+s^\gamma) \in [0,\infty)$, where c and γ are the constants in the Hölderian error bound for $X_{\rm opt}$.

Admittedly, the concept of ψ-sharp minima is significantly broader than that of Hölderian minima. Nevertheless, it is generally not easy to identify the forcing function ψ (thus to verify that (1) has well-conditioned sharp minima). The theory of error bounds as summarized in [29], with the postulate of a specific family of ψ functions as given above, offers a practical way of verifying this sharp property of $X_{\rm opt}$.

In what follows, we return to L- and H-metric regularity of the problem (1). In the remainder of this section, we do not assume the nonemptiness of $X_{\rm opt}$ unless otherwise stated.

Inspired by a "strong Slater" condition proposed recently by Mangasarian [26], we present a necessary and sufficient condition for the problem (1) to be L-metrically regular. In addition to [29], we refer the reader to [18] for a systematic treatment of error bounds for convex inequality systems.

LEMMA 4.1. *Let $\theta : \Re^n \to \Re$ be a convex function and X be a closed convex subset of $\Re^n$. For a scalar $c > 0$, the following two conditions are equivalent.*

(a) *For all* $x \in X$ *with* $\theta(x) > \theta_{inf}$, *there exists* $\hat{x} \in X$ *satisfying* $\theta(\hat{x}) < \theta(x)$ *and*

$$\| \hat{x} - x \| \leq c\,(\theta(x) - \theta(\hat{x})).$$

(b) *For all* $\lambda > \theta_{inf}$,

$$dist(x, L(\lambda)) \leq c\,(\theta(x) - \lambda)_+ \quad \forall\ x \in X.$$

In particular, if (1) *has weak sharp minima, then* (1) *is L-metrically regular.*

Proof. (a) $\Rightarrow$ (b). Let $\lambda > \theta_{\inf}$ and $x \in X$ be given. Without loss of generality, we may assume that $\theta(x) > \lambda$. Consider the problem of projecting x onto $L(\lambda)$:

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\,(z - x)^T(z - x) \\ \text{subject to} & z \in L(\lambda). \end{array}$$

Let $\bar{x} \equiv \Pi_{L(\lambda)}(x)$. We must have $\theta(\bar{x}) = \lambda$. Let $\hat{x} \in X$ be the vector such that $\theta(\hat{x}) < \theta(\bar{x})$ and

$$\| \hat{x} - \bar{x} \| \leq c\,(\theta(\bar{x}) - \theta(\hat{x})).$$

By results from convex analysis [33], there exist a nonnegative scalar η and a vector $a \in \partial\theta(\bar{x})$ such that for all $z \in X$,

$$(z - \bar{x})^T(\bar{x} - x + \eta\, a) \geq 0. \tag{6}$$

Letting $z = \hat{x}$, we deduce

$$\eta\, a^T(\hat{x} - \bar{x}) \geq -(\hat{x} - \bar{x})^T(\bar{x} - x).$$

By the definition of a, we have

$$\theta(\hat{x}) - \theta(\bar{x}) \geq a^T(\hat{x} - \bar{x}),$$

which implies

$$\eta\,(\theta(\hat{x}) - \theta(\bar{x})) \geq -(\hat{x} - \bar{x})^T(\bar{x} - x).$$

Consequently, it follows that

$$\eta \leq c\,\| \bar{x} - x \|.$$

Letting $z = x$ in (6), we deduce

$$\| \bar{x} - x \|^2 \leq \eta\, a^T(x - \bar{x}) \leq \eta\,(\theta(x) - \theta(\bar{x})).$$

Since $x \neq \bar{x}$ and $\theta(\bar{x}) = \lambda < \theta(x)$, we obtain

$$\| \bar{x} - x \| \leq c\,(\theta(x) - \lambda)_+.$$

Thus (b) holds.

(b) $\Rightarrow$ (a). Let $x \in X$ with $\theta(x) > \theta_{\inf}$ be given. Choose λ such that $\theta(x) > \lambda > \theta_{\inf}$. Let $\hat{x}$ be the Euclidean projection of x onto the level set $L(\lambda)$. Then $\hat{x} \in X$ and $\theta(\hat{x}) = \lambda < \theta(x)$. By (b), we have

$$\| \hat{x} - x \| = \operatorname{dist}(x, L(\lambda)) \leq c\,(\theta(x) - \lambda)_+ = c\,(\theta(x) - \theta(\hat{x})).$$

Thus (a) holds.

If (1) has weak sharp minima, then clearly (a) holds with $\hat{x} \in X_{\rm opt}$ satisfying $\| x - \hat{x} \| = {\rm dist}(x, X_{\rm opt})$. Thus (b) follows. □

We review some terminology of piecewise smooth functions; see Li [19]. We say that a function $\theta : \Re^n \to \Re$ is *piecewise quadratic* if θ is continuous and there exist finitely many convex polyhedra P_i, $i = 1, \ldots, p$ for some positive integer p, whose union is $\Re^n$ such that θ is a quadratic function on each P_i; the latter quadratic functions are called the *pieces* of θ. A vector function $F : \Re^n \to \Re^m$ is *piecewise linear* if F is continuous and there exist finitely many convex polyhedra S_i, $i = 1, \ldots, q$, for some positive integer q, whose union is $\Re^n$ such that F is an affine function on each S_i; these affine functions are called the *pieces* of F. It is well known [12] that a piecewise linear function on $\Re^n$ is globally Lipschitz continuous. Clearly, if θ is a differentiable piecewise quadratic real-valued function, then the gradient map $\nabla\theta$ is a piecewise linear vector-valued function. We say that a real-valued function θ defined on $\Re^n$ is *convex piecewise quadratic*, or CPQ in short, if θ is convex and piecewise quadratic (thus the pieces of θ must be convex quadratic functions). A *convex quadratic spline* is a differentiable CPQ function. A simple one-dimensional convex quadratic spline is the function $t \in \Re \mapsto (\max(0, t))^2 \in \Re_+$.

Lemma 4.2 below identifies two sufficient conditions for the problem (1) to be H-metrically regular. The first condition assumes that this problem has an analytic objective function and its feasible set is compact and defined by finitely many analytic inequalities; the conclusion follows easily from the error bound theory of analytic inequality systems [23]. This part of the result illustrates the fact that H-metric regularity holds trivially for the broad class of "analytic programs" with compact feasible regions; this conclusion will not be used later. In contrast, the second condition of Lemma 4.2 will serve two important objectives. One, it further illustrates that H-metric regularity is a condition that will be easily satisfied by another class of constrained optimization problems. As a result, we can readily establish that this class of problems is well behaved. Second, the proof of Lemma 4.2 under the second condition is based on an Hölderian error bound for a general convex quadratic inequality system established in [37]; it is a significant extension of the result of Luo and Luo [22] for a convex quadratic inequality system which requires a Slater condition.

LEMMA 4.2. *Suppose that the pair (θ, X) satisfies either one of the two conditions below:*

(a) *θ is an analytic function and X is a compact set defined by finitely many analytic inequalities;*

(b) *$\theta(x)$ is a CPQ function and*

$$X \equiv \{\, x \in \Re^n : g_i(x) \leq 0,\ i = 1, \ldots, m \,\},$$

where each g_i is a convex quadratic function and m is a given positive integer. Then the problem (1) *has H-metrically regular level sets. Furthermore if X_{opt} is nonempty, then an Hölderian error bound holds for X_{opt}.*

Proof. We prove the lemma under condition (b); the same proof given below can be applied when (a) holds by using the error bound results in [23]. Let $\{P_i : i = 1 \ldots, p\}$ be the family of convex polyhedra whose union is $\Re^n$ such that θ is equal to a quadratic function, which we denote q_i, on each P_i. Let $\lambda > \theta_{\rm inf}$ be given. Clearly,

$$L(\lambda) = \bigcup_{i=1}^{p} L_i(\lambda),$$

where each

$$L_i(\lambda) \equiv \{\, x \in X \cap P_i : q_i(x) \leq \lambda \,\}$$

is the solution set of a system of finitely many convex quadratic inequalities ($L_i(\lambda)$, if nonempty, need not have a nonempty interior). By the error bound in [37], for each i for which $L_i(\lambda)$ is nonempty, there exists positive constant c_i and γ_i with $\gamma_i < 1$ such that

$$\operatorname{dist}(x, L_i(\lambda)) \leq c_i \max\left([\, (q_i(x) - \lambda)_+ \,]^{\gamma_i},\ (q_i(x) - \lambda)_+ \right) \quad \forall\, x \in X \cap P_i.$$

Since

$$\operatorname{dist}(x, L_i(\lambda)) \geq \operatorname{dist}(x, L(\lambda))$$

and q_i coincides with θ on P_i, it follows that

$$\operatorname{dist}(x, L(\lambda)) \leq c_i \max\left([\, (\theta(x) - \lambda)_+ \,]^{\gamma_i},\ (\theta(x) - \lambda)_+ \right) \quad \forall\, x \in X \cap P_i.$$

Since the union of the P_i is $\Re^n$, it follows that by letting

$$c \equiv \max(\, c_i : i = 1, \ldots, p \,) \quad \text{and} \quad \gamma \equiv \min(\, \gamma_i : i = 1, \ldots, p \,),$$

we obtain

$$\operatorname{dist}(x, L(\lambda)) \leq c \max\left([\, (\theta(x) - \lambda)_+ \,]^{\gamma},\ (\theta(x) - \lambda)_+ \right) \quad \forall\, x \in X,$$

as desired. The above proof is clearly applicable to X_{opt} if this set is nonempty. ∎

By the main result, Theorem 3.1 in [37], one can show that if X is a convex polyhedron, the exponent γ in the error bounds for the level sets $L(\lambda)$ can be chosen to be equal to $1/2$ for all $\lambda \geq \theta_{\text{inf}}$. This conclusion extends a result of Li [19, Corollary 2.8] which pertains to $\lambda = \theta_{\text{inf}}$. An extension of this conclusion is presented in Proposition 4.10 in section 4.3, which was proved by Li recently [20].

4.2. The implication and special cases. We are now ready to state the principal result regarding the topic of this section. Although the main idea of proof is borrowed from [3], the result itself is a useful refinement of this work in several respects. First, our result gives a full treatment of infeasible sequences; second, it is based on residual functions which are closely tied to the practical implementation of iterative methods; third, the assumption of H-metric regularity offers an effortless demonstration that an important class of constrained optimization problems is well behaved; see Corollary 4.5 and Theorem 5.3.

THEOREM 4.3. *Let X be a nonempty closed convex subset of $\Re^n$, and let $\theta : \Re^n \to \Re$ be a continuously differentiable convex function. Let $\{x^k\} \subset \Re^n$ be an arbitrary sequence of vectors satisfying any one of the following three conditions:*

(i) *$\{\nabla\theta(x^k)\}$ is bounded, $\nabla\theta$ is uniformly continuous near $\{x^k\}$, and $\{x^k\}$ is N-stationary;*

(ii) *θ is uniformly continuous near $\{x^k\}$ and $\{x^k\}$ is $\mathcal{N}$-stationary;*

(iii) *each x^k is feasible to (1) and $\{x^k\}$ is $\mathcal{N}$-stationary.*

If (1) is H-metrically regular, then $\{x^k\}$ is an LP minimizing sequence of θ on X.

Proof. We first establish that the conclusion holds under (iii). In essence, if (1) is L-metrically regular, then the conclusion follows immediately from the work of Auslender and Crouzeix [3]. Our proof below is a slight extension of their argument

in the case where the level sets have Hölderian (instead of Lipschitzian) error bounds. For completeness, we give the detailed proof.

So we assume that each x^k is feasible to (1) and the sequence $\{x^k\}$ is $\mathcal{N}$-stationary. It suffices to show $\lim_{k\to\infty}\theta(x^k) = \theta_{\inf}$. Assume for the sake of contradiction that this is not true. Let the scalar λ be such that $\liminf_{k\to\infty}\theta(x^k) > \lambda > \theta_{\inf}$. Let c and γ be such that (5) holds. Since $L(\lambda)$ is a nonempty closed convex set, it follows that for each k, there exists $y^k \in L(\lambda)$ such that $\operatorname{dist}(x^k, L(\lambda)) = \|x^k - y^k\|$; moreover since the line segment joining x^k and y^k lies in X, we must have $\theta(y^k) = \lambda$. Since θ is convex, by the gradient inequality,

$$\lambda \;=\; \theta(y^k) \;\geq\; \theta(x^k) + \nabla\theta(x^k)^T(y^k - x^k).$$

Since $\{x^k\}$ is $\mathcal{N}$-stationary, there exists a sequence $\{z^k\}$ such that $x^k = \Pi_X(z^k)$ for all k and

$$\lim_{k\to\infty} R_{\mathcal{N}}(z^k) \;=\; 0.$$

By the definition of $R_{\mathcal{N}}(z^k)$, it follows that

$$(y^k - x^k)^T(\nabla\theta(x^k) - R_{\mathcal{N}}(z^k)) \;\geq\; 0, \tag{7}$$

which implies

$$\nabla\theta(x^k)^T(y^k - x^k) \;\geq\; R_{\mathcal{N}}(z^k)^T(y^k - x^k).$$

Thus

$$\theta(x^k) - \lambda \;\leq\; -R_{\mathcal{N}}(z^k)^T(y^k - x^k).$$

since $\theta(x^k) > \lambda$, by the Cauchy–Schwarz inequality and (5), we deduce

$$\theta(x^k) - \lambda \;\leq\; c\,\|\,R_{\mathcal{N}}(z^k)\,\|\,\max\left(\theta(x^k) - \lambda,\,(\theta(x^k) - \lambda)^{\gamma}\right).$$

Dividing by $\theta(x^k) - \lambda$, we obtain

$$1 \;\leq\; c\,\|R_{\mathcal{N}}(z^k)\|\,\max\left(1, (\theta(x^k) - \lambda)^{\gamma-1}\right).$$

Since $\{R_{\mathcal{N}}(z^k)\}$ converges to zero and $\liminf_{k\to\infty}\theta(x^k) > \lambda$, we obtain a contradiction from the above expression by passing to the limit $k \to \infty$.

To prove (ii), let $\bar{x}^k \equiv \Pi_X(x^k)$. Then $\{\bar{x}^k\}$ satisfies the conditions in (iii). By the above proof, we deduce

$$\lim_{k\to\infty} \theta(\bar{x}^k) \;=\; \theta_{\inf}. \tag{8}$$

Since θ is uniformly continuous near $\{x^k\}$, the same limit holds for the sequence $\{x^k\}$.

Finally, assume the conditions in (i). We have noted that $\{x^k\}$ must be asymptotically feasible to (1); let $\bar{x}^k \equiv \Pi_X(x^k)$. The sequence $\{\bar{x}^k\}$ is also N-stationary because $\nabla\theta$ is uniformly continuous near $\{x^k\}$.

At this point, there are two ways to finish the proof. Both require us to show that θ is uniformly continuous near $\{x^k\}$. So we show this first. By the mean-value theorem, for any vector $y \in \Re^n$, there exists a scalar $\tau_k \in [0,1]$ such that

$$\begin{aligned}\theta(y) - \theta(x^k) &= \nabla\theta(x^k + \tau_k(y - x^k))^T(y - x^k)\\ &= [\nabla\theta(x^k + \tau_k(y - x^k)) - \nabla\theta(x^k)]^T(y - x^k) + \nabla\theta(x^k)^T(y - x^k),\end{aligned}$$

which easily establishes the uniform continuity of θ near $\{x^k\}$ under the assumptions in (i).

To complete the proof, one way is to apply Proposition 2.1(c) to conclude that $\{x^k\}$ is AC_ε-stationary; hence so is $\{\bar{x}^k\}$. By following the argument in Proposition 2.1 in [3] (which uses the theorem of Bronstedt and Rockafellar [4] about ε-subgradients), we can deduce that (8) holds. Thus the uniform continuity of θ near $\{x^k\}$ completes the proof.

Alternatively, we can follow our proof above and use instead of (7) the inequality

$$(y^k - \bar{x}^k + R_N(\bar{x}^k))^T(\nabla\theta(\bar{x}^k) - R_N(\bar{x}^k)) \geq 0,$$

which follows from (2). Either way, the desired conclusion of the theorem holds under (i). □

The above proof reveals a general fact that is worth further discussion; namely, if θ is a continuously differentiable function with $\{\nabla\theta(x^k)\}$ being bounded and $\nabla\theta$ uniformly continuous near $\{x^k\}$, then θ is uniformly continuous near $\{x^k\}$. When θ is a quadratic function (not necessarily convex), then $\nabla\theta(x)$ is an affine function in x; thus $\nabla\theta$ is uniformly continuous near any sequence. In this case, θ is uniformly continuous near a sequence $\{x^k\}$ if and only if $\{\nabla\theta(x^k)\}$ is bounded. Indeed it suffices to prove the "only if" part. In turn, this proof is rather easy because, by means of an orthogonal transformation of variables, we may assume without loss of generality that θ is a separable quadratic function; under this assumption, the desired assertion is easily seen to be valid.

Combining the above discussion with the globally Lipschitz continuous property of a piecewise linear function, we state a useful property of a differentiable piecewise quadratic function. The following lemma requires no proof.

LEMMA 4.4. *Let $\theta : \Re^n \to \Re$ be a differentiable piecewise quadratic function. The following two statements hold:*

(a) *$\nabla\theta$ is globally Lipschitz continuous;*

(b) *if $\{\nabla\theta(x^k)\}$ is bounded, then θ is uniformly continuous near $\{x^k\}$.*

Based on Lemmas 4.4 and 4.2(b), we establish a corollary of Theorem 4.3 that pertains to the following convex quadratically constrained quadratic spline program (CQQSP):

$$\begin{aligned} &\text{minimize} && \theta(x) \\ &\text{subject to} && G_i(x) \equiv \tfrac{1}{2}x^T C_i x + a_i^T x + b_i \leq 0, \quad i = 1, \ldots, m, \end{aligned} \tag{9}$$

where θ is a convex quadratic spline, each C_i is an $n \times n$ symmetric positive semidefinite matrix, each a_i is an n-vector, and each b_i is a scalar. A special case of the CQQSP is the convex quadratically constrained quadratic program (CQQP) in which the objective function θ is a convex quadratic function. The corollary below illustrates an important benefit of weakening the Auslender–Crouzeix assumption of L-metric regularity to H-metric regularity.

COROLLARY 4.5. *Assume that the above CQQSP is feasible and that θ_{inf} is finite. Let $\{x^k\}$ be an arbitrary sequence of vectors such that $\{\nabla\theta(x^k)\}$ is bounded. Then $\{x^k\}$ is N-stationary for* (9) *if and only if it is LP minimizing.*

Proof. By the boundedness of $\{\nabla\theta(x^k)\}$, Lemma 4.4 implies that both θ and $\nabla\theta$ are uniformly continuous near $\{x^k\}$. Consequently, if $\{x^k\}$ is LP minimizing for (9), then Theorem 3.2 implies that $\{x^k\}$ is N-stationary.

Conversely assume that $\{x^k\}$ is N-stationary for the CQQSP. Condition (i) in

Theorem 4.3 is thus satisfied. Moreover by Lemma 4.2, the CQQSP is H-metrically regular. Theorem 4.3 completes the proof. □

Using Lemma 4.1, we can easily establish the following result which identifies another sufficient condition under which LP minimizing sequences and N-stationary sequences are equivalent. As with the other results in this subsection, the result below does not require $X_{\rm opt}$ to be nonempty.

PROPOSITION 4.6. *Let $\theta : \Re^n \to \Re$ be a convex, continuously differentiable function, and let X be a closed convex subset of $\Re^n$. Suppose $\theta_{inf} > -\infty$ and there exists a constant $c > 0$ such that for all $x \in X$ with $\theta(x) > \theta_{inf}$, there exists $\hat{x} \in X$ satisfying $\theta(\hat{x}) < \theta(x)$ and*

$$\| \hat{x} - x \| \leq c\,(\theta(x) - \theta(\hat{x})).$$

Let $\{x^k\}$ be an arbitrary sequence such that $\{\nabla\theta(x^k)\}$ is bounded and $\nabla\theta$ is uniformly continuous near $\{x^k\}$. Then $\{x^k\}$ is LP minimizing for (1) *if and only if it is N-stationary.*

Proof. This follows easily from Theorems 3.2 and 4.3 and Lemma 4.1. □

An immediate consequence of the above theorem is that if the program (1) has weak sharp minima, then for any sequence $\{x^k\}$ such that $\{\nabla\theta(x^k)\}$ is bounded and $\nabla\theta$ is uniformly continuous near $\{x^k\}$, the sequence $\{x^k\}$ is LP minimizing for (1) if and only if it is N-stationary. This conclusion will be generalized in the next subsection in which an expanded study is presented for the program (1) under the assumption that $X_{\rm opt}$ is nonempty.

4.3. When optimal solutions exist. The boundedness of the sequence of gradients $\{\nabla\theta(x^k)\}$ has played an important role in several results in the last subsection. In what follows, we derive some results that pertain to this boundedness issue. In particular, these results lend support to the presumption that this boundedness assumption is not too restrictive in order for the results in the last subsection to hold.

Consider the case where the problem (1) attains its global minimum; i.e., assume $X_{\rm opt} \neq \emptyset$. In this case, we say that a sequence $\{x^k\} \subset \Re^n$ is *near the optimal set* if $\lim_{k\to\infty} {\rm dist}(x^k, X_{\rm opt}) = 0$. This notion requires $X_{\rm opt}$ to be nonempty. The next result shows that if θ is uniformly continuous, then the property of "near the optimal set" implies that of LP minimizing. This implication does not require θ to be differentiable.

PROPOSITION 4.7. *Assume $X_{opt} \neq \emptyset$. If $\{x^k\} \subset \Re^n$ is a sequence near the optimal set of* (1) *and θ is uniformly continuous near $\{x^k\}$, then $\{x^k\}$ is LP minimizing for* (1).

Proof. Since $X_{\rm opt} \subseteq X$, it follows that

$${\rm dist}(x^k, X) \leq {\rm dist}(x^k, X_{\rm opt}) \quad \forall k.$$

Thus the sequence $\{x^k\}$ is asymptotically feasible. Moreover if $y^k \equiv \Pi_{X_{\rm opt}}(x^k)$, then

$$\lim_{k\to\infty} \theta(y^k) = \lim_{k\to\infty} \theta(x^k),$$

by the uniform continuity of θ near $\{x^k\}$. Hence $\{x^k\}$ is LP minimizing for (1), as desired. □

The following example shows that the uniform continuity assumption in the above proposition is essential for the result to hold.

Example 4. Let θ and X be as given in Example 3. Let

$$x^k \equiv -\begin{pmatrix} 1/k \\ k\pi/2 \end{pmatrix}, \quad y^k \equiv -\begin{pmatrix} 3/k \\ k\pi/2 \end{pmatrix} \quad \text{for } k = 0, 1, 2, \ldots.$$

Each $y^k \in X_{\text{opt}}$ and $\lim_{k\to\infty} \|x^k - y^k\| = 0$. Thus the sequence $\{x^k\}$ is near the optimal set. Yet it is clear that $\{x^k\}$ is not LP minimizing because $\theta(x^k) = 1$ for all k.

Next we show that under the condition of Hölderian minima, the converse of the above proposition holds.

PROPOSITION 4.8. *Assume that* (1) *has Hölderian minima. If* $\{x^k\} \subset \Re^n$ *is an LP minimizing sequence of* (1)*, and if* θ *is uniformly continuous near* $\{x^k\}$*, then* $\{x^k\}$ *is near the optimal set.*

Proof. Let $\bar{x}^k \equiv \Pi_X(x^k)$ for each k. Since θ is uniformly continuous near $\{x^k\}$, the projected sequence $\{\bar{x}^k\}$ is also LP minimizing. We have

$$\text{dist}(x^k, X_{\text{opt}}) \le \|x^k - \bar{x}^k\| + \text{dist}(\bar{x}^k, X_{\text{opt}});$$

the right-hand sum clearly approaches zero as $k \to \infty$ because $\{x^k\}$ is asymptotically feasible and the Hölderian error bound for X_{opt} implies that $\text{dist}(\bar{x}^k, X_{\text{opt}}) \to 0$ as k tends to ∞. Thus $\{x^k\}$ is near the optimal set. □

An immediate consequence of the above proposition is that if θ is uniformly continuous on $\Re^n$ and (1) has weak sharp minima, then an arbitrary sequence $\{x^k\} \subset \Re^n$ is LP minimizing if and only if it is near the optimal set. The CQQSP is another instance where these two sequential properties are equivalent; see Theorem 5.3.

It has been shown by Mangasarian [24] that for a differentiable convex program, i.e., for (1) where θ is a continuously differentiable convex function and X is a convex set, the gradient of θ is equal to a constant on X_{opt}. (Actually, this result was extended to a nondifferentiable function in the reference; nevertheless, consistent with the treatment of this paper, we restrict the discussion to differentiable functions.) Based on this fact, we establish a related property of a sequence near the optimal set; this property can be used to partially atone for the boundedness assumption of the gradient sequence $\{\nabla\theta(x^k)\}$ in Theorem 4.3. In particular, under the assumptions of Propositions 4.8 and 4.9, if these gradients are unbounded, then $\{x^k\}$ cannot be an LP minimizing sequence for (1).

PROPOSITION 4.9. *Let* X *be a nonempty closed convex subset of* $\Re^n$*, and let* $\theta : \Re^n \to \Re$ *be a continuously differentiable convex function such that* $X_{opt} \neq \emptyset$*. Let* $g \equiv \nabla\theta(\bar{x})$ *for any* $\bar{x} \in X_{opt}$*. If* $\{x^k\}$ *is a sequence near the optimal set of* (1) *and* $\nabla\theta$ *is uniformly continuous near* $\{x^k\}$*, then* $\lim_{k\to\infty} \nabla\theta(x^k)$ *exists and equals* g*.*

Proof. Let $y^k \equiv \Pi_{X_{\text{opt}}}(x^k)$. Since $\nabla\theta(y^k) = g\,\forall k$, the desired conclusion follows easily. □

Under the assumption $X_{\text{opt}} \neq \emptyset$, several of our results are closely related to the work of Lemaire [15], which is cast in the framework of an extended-valued convex function (such as the function ϕ in (4)) in a Banach space. Phrased in our terminology, Theorem 3.1 in this reference says that the following statements are equivalent.

(i) Every feasible, LP-minimizing sequence is near the optimal set.
(ii) The function ϕ defined in (4) has well-conditioned minima in the sense of Lemaire's Definition 2.1.
(iii) The function ϕ has very well-conditioned minima in the sense of Lemaire's Definition 2.2.
(iv) Every feasible AC-stationary sequence is near the optimal set.

There are two obvious differences between Lemaire's work and ours. The first difference lies in the treatment of infeasible sequences. In our case, the given sequence $\{x^k\}$ is not necessarily feasible; this is a broadening of Lemaire's treatment. The other difference is that Lemaire did not consider the question of whether a sequence that is near the optimal set is necessarily LP minimizing. As we see from Example 4 above, the answer to this question in general is in the negative. Although easy, the issue considered in Proposition 4.9 has not been dealt with before.

To end this section, we give a result for the program (1) under a convexity assumption and the existence of Hölderian minima for this convex program. This result, which is obtained by Li [20], is related in spirit to some results in [3], like Proposition 2.8 therein; cf. also Lemma 4.1 in section 4.1.

PROPOSITION 4.10. *Let $\theta : \Re^n \to \Re$ be a convex function, and let X be a closed convex subset of $\Re^n$. Suppose that, for some scalar $\lambda_0 \geq \theta_{inf}$, $L(\lambda_0)$ is nonempty and there exist constants $c > 0$ and $\gamma \in (0,1)$ such that*

$$\operatorname{dist}(x, L(\lambda_0)) \leq c \max\left([\,(\theta(x) - \lambda_0)_+\,]^\gamma,\ (\theta(x) - \lambda_0)_+ \right) \quad \forall\, x \in X.$$

Then for all $\lambda > \lambda_0$,

$$\operatorname{dist}(x, L(\lambda)) \leq c \max\left([\,(\theta(x) - \lambda)_+\,]^\gamma,\ (\theta(x) - \lambda)_+ \right) \quad \forall\, x \in X.$$

In particular, if (1) has Hölderian minima, then (1) has H-metrically regular level sets.

Proof. Let $x \in X$ and $\lambda > \lambda_0$ be given. Without loss of generality, we may assume that $\theta(x) > \lambda$. Let $\bar{x} \in \lambda_0$ be such that

$$\operatorname{dist}(x, L(\lambda_0)) = \| x - \bar{x} \|.$$

We must have $\theta(\bar{x}) = \lambda_0$. Let $\tau \in (0,1)$ be such that

$$\theta(x^\tau) = \lambda, \quad \text{where} \quad x^\tau \equiv \tau\, x + (1 - \tau)\, \bar{x}.$$

By the convexity of θ, it follows easily that

$$\lambda = \theta(x^\tau) \leq \tau\, \theta(x) + (1 - \tau)\, \theta(\bar{x}),$$

which yields

$$(1-\tau)\,(\theta(x)-\lambda_0)_+ \leq (\theta(x)-\lambda)_+ \quad \text{and} \quad [\,(1-\tau)\,(\theta(x)-\lambda_0)_+\,]^\gamma \leq [\,(\theta(x)-\lambda)_+\,]^\gamma.$$

Consequently, since $x^\tau \in L(\lambda)$, we have

$$\begin{aligned}
\operatorname{dist}(x, L(\lambda)) &\leq \| x - x^\tau \| \\
&= (1 - \tau)\, \| x - \bar{x} \| = (1 - \tau) \operatorname{dist}(x, L(\lambda_0)) \\
&\leq (1 - \tau)\, c \max\left([\,(\theta(x) - \lambda_0)_+\,]^\gamma,\ (\theta(x) - \lambda_0)_+ \right) \\
&\leq c \max\left([\,(1 - \tau)\,(\theta(x) - \lambda_0)_+\,]^\gamma,\ (\theta(x) - \lambda)_+ \right) \\
&\leq c \max\left([\,(\theta(x) - \lambda)_+\,]^\gamma,\ (\theta(x) - \lambda)_+ \right),
\end{aligned}$$

where the next-to-last inequality holds because $\tau \in (0,1)$ and $\gamma \in (0,1)$. Finally, with $\lambda_0 \equiv \theta_{\inf}$, the last assertion of the proposition is obvious. □

5. Asymptotic optimality systems. We return to the general situation where we do not assume the nonemptiness of $X_{\rm opt}$.

When the set X possesses additional structure, we can use some asymptotic optimality systems to characterize the LP minimizing property of a sequence. This section is divided into two subsections. In the first one we consider the case where X is a closed convex cone; in the second one we assume that X is defined by a finite differentiable inequality system. In both subsections we will focus only on feasible sequences. From Theorems 3.2 and 4.3 we see that by postulating some uniform continuity assumptions on θ or $\nabla\theta$ near a given (infeasible) sequence, results that are derived for the projected (feasible) sequence will easily yield corresponding results for the given (infeasible) sequence.

5.1. Approximate complementarity conditions. Consider the problem (1) where X is a closed convex cone. In this case it follows that if x is a local minimum of (1), then the following complementarity system holds:

$$X \ni x \perp \nabla\theta(x) \in X^*, \tag{10}$$

where $u \perp v$ means $u^T v = 0$ and

$$X^* \equiv \{y \in \Re^n : y^T v \geq 0 \text{ for all } v \in X\}$$

is the dual cone of X. Conversely, if θ is convex, then every vector x satisfying (10) is a global minimum of (1). The goal in this subsection is to study the case where instead of a single vector x, we are given a sequence $\{x^k\} \subset X$ (possibly unbounded), and we want to characterize the LP minimizing property of this sequence in terms of some approximate complementarity conditions. The following theorem is the main result in this regard.

THEOREM 5.1. *Let X be a nonempty closed convex cone in $\Re^n$, and let $\theta : \Re^n \to \Re$ be a continuously differentiable function. Let $\{x^k\} \subset X$ be an arbitrary feasible sequence.*

(a) *Assume $\theta_{inf} > -\infty$. If $\{x^k\}$ is LP minimizing for* (1) *and $\nabla\theta$ is uniformly continuous near this sequence, then there exists a sequence $\{w^k\} \subset X^*$ such that*

$$\lim_{k\to\infty} (\nabla\theta(x^k) - w^k) = 0 \quad \textit{and} \quad \lim_{k\to\infty} (x^k)^T w^k = 0. \tag{11}$$

(b) *Conversely, if θ is convex,* (1) *is H-metrically regular, and there exists a sequence $\{w^k\} \subset X^*$ satisfying* (11) *, then $\{x^k\}$ is LP minimizing for* (1).

Proof. We first prove (a). By the proof of Proposition 3.1, we deduce the existence of two sequences of vectors $\{y^k\} \subset X$ and $\{w^k\} \subset \Re^n$ such that for each k, $y^k = \Pi_X(y^k - w^k)$, and

$$\lim_{k\to\infty} (w^k - \nabla\theta(y^k)) = 0, \quad \lim_{k\to\infty} (y^k - x^k) = 0, \quad \text{and} \quad \lim_{k\to\infty} \theta(y^k) = \theta_{\rm inf}. \tag{12}$$

Since X is a convex cone, we have

$$X \ni y^k \perp w^k \in X^*.$$

We have

$$\nabla\theta(x^k) - w^k = (\nabla\theta(x^k) - \nabla\theta(y^k)) + (\nabla\theta(y^k) - w^k),$$

which clearly approaches zero as $k \to \infty$, by the uniform continuity of $\nabla\theta$ near $\{x^k\}$. It remains to show $\lim_{k\to\infty}(x^k)^T w^k = 0$. We have

$$\begin{aligned}(x^k)^T w^k &= (x^k - y^k)^T w^k + (y^k)^T w^k \\ &= (x^k - y^k)^T (w^k - \nabla\theta(y^k)) \\ &\quad + (x^k - y^k)^T \nabla\theta(x^k) + (x^k - y^k)^T (\nabla\theta(y^k) - \nabla\theta(x^k)).\end{aligned}$$

In view of (12) and by the uniform continuity of $\nabla\theta$ near $\{x^k\}$, it suffices to show

$$\lim_{k\to\infty} (x^k - y^k)^T \nabla\theta(x^k) = 0. \tag{13}$$

We have

$$\theta(y^k) - \theta(x^k) = (y^k - x^k)^T \nabla\theta(x^k) + o(\|y^k - x^k\|),$$

where $o(t)$ is a quantity which approaches zero as $t \downarrow 0$. Since $\{y^k\}$ is also LP minimizing for (1), the limit of the left-hand side is equal to zero as $k \to \infty$; thus (13) follows. This establishes (a).

To prove (b), we follow the proof of Theorem 4.3 and assume for the sake of contradiction that there is a scalar λ satisfying $\liminf_{k\to\infty} \theta(x^k) > \lambda > \theta_{\inf}$. Let c and γ be the scalars in the error bound for the level set $L(\lambda)$, and let $y^k \in X$ be such that $\theta(y^k) = \lambda$ and $\operatorname{dist}(x^k, L(\lambda)) = \|x^k - y^k\|$. As in previous arguments, we have

$$\theta(x^k) - \lambda \leq -\nabla\theta(x^k)^T (y^k - x^k).$$

The right-hand side is equal to

$$\begin{aligned}\nabla\theta(x^k)^T (x^k - y^k) &= (w^k)^T (x^k - y^k) + (\nabla\theta(x^k) - w^k)^T (x^k - y^k) \\ &\leq |(w^k)^T x^k| + \|\nabla\theta(x^k) - w^k\| \, \|y^k - x^k\|,\end{aligned}$$

where the inequality holds because $(w^k)^T y^k = 0$. Thus we deduce

$$1 \leq c \, \| \nabla\theta(x^k) - w^k \| \, \max\left(1, (\theta(x^k) - \lambda)^{\gamma - 1}\right) + \frac{|\,(w^k)^T x^k\,|}{\theta(x^k) - \lambda},$$

which again is a contradiction upon passing to the limit $k \to \infty$. □

It is natural to ask how the approximate complementarity system, (11) plus $\{w^k\} \subset X^*$, is related the limiting residual condition, $\lim_{k\to\infty} R_N(x^k) = 0$, used in the last sections; in particular, whether they are equivalent for an arbitrary feasible sequence $\{x^k\}$. Our preliminary analysis suggests a negative answer. Since this issue is not of primary importance in the remainder of the paper, we will not pursue it further.

The conditions $\{w^k\} \subset X^*$ and $\lim_{k\to\infty}(\nabla\theta(x^k) - w^k) = 0$ imply

$$\lim_{k\to\infty} \operatorname{dist}(\nabla\theta(x^k), X^*) = 0.$$

Nevertheless the vector w^k is not necessarily the closest point in X^* to the vector $\nabla\theta(x^k)$, as illustrated in the example below.

Example 5. Consider the convex program in 4 variables:

$$\begin{array}{ll} \text{minimize} & \theta(x) \equiv (x_{12}-1)^2 + (x_{21}-1)^2 + 1.5\,x_{22}^2 \\ \text{subject to} & x \equiv \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix} \text{ is symmetric positive semidefinite.} \end{array}$$

Thus the constraint set is the cone of 2×2 symmetric positive semidefinite matrices. It is easy to check that

$$x^\varepsilon \equiv \begin{pmatrix} 1/\varepsilon^2 & \varepsilon+1 \\ \varepsilon+1 & \varepsilon \end{pmatrix}, \quad \varepsilon \downarrow 0,$$

constitutes an LP minimizing sequence and $\theta_{\inf} = 0$; moreover this optimal objective value is not attained. We have

$$\nabla\theta(x^\varepsilon) \equiv \varepsilon \begin{pmatrix} 0 & 2 \\ 2 & 3 \end{pmatrix},$$

which is not positive semidefinite. With

$$w^\varepsilon \equiv \begin{pmatrix} \varepsilon^3 & 0 \\ 0 & 3\varepsilon \end{pmatrix},$$

we can easily verify that the sequence $\{w^\varepsilon\}$ satisfies the conditions in Theorem 5.1. However the Frobenius projection of the matrix $\nabla\theta(x^\varepsilon)$ onto the cone of 2×2 symmetric positive semidefinite matrices can be calculated to be

$$\frac{4\varepsilon}{5} \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix},$$

which is not equal to w^ε; moreover the above matrix is not asymptotically perpendicular to x^ε as $\varepsilon \downarrow 0$.

5.2. Approximate KKT conditions. Consider the nonlinear program

$$\begin{array}{ll} \text{minimize} & \theta(x) \\ \text{subject to} & G(x) \leq 0, \end{array} \tag{14}$$

where $\theta : \Re^n \to \Re$ and $G : \Re^n \to \Re^m$ are continuously differentiable functions. We assume that each G_i is convex and that a Slater point exists for the constraints, i.e., there exists a vector $\hat{x} \in \Re^n$ satisfying $G_i(\hat{x}) < 0$ for all i. In this setting it follows that if x is a local minimum of (14), then there exist multipliers μ_i, $i = 1, \ldots, m$, such that the following KKT conditions hold:

$$\begin{aligned} \mathcal{L}(x,\mu) \equiv \nabla\theta(x) + \sum_{i=1}^m \mu_i \nabla G_i(x) = 0, \\ 0 \geq G(x) \perp \mu \geq 0; \end{aligned}$$

conversely if θ is also convex and (x, μ) satisfies the latter KKT system, then x is a global minimum of (14). In what follows, we generalize these basic results in nonlinear

programming to the context of an LP minimizing sequence. For this purpose we recall an asymptotic constraint qualification (ACQ) stated in Luo and Luo [22] that is essentially due to Mangasarian [25]:
(ACQ) The scalar

$$\sup_{x,\mu,I}\left\{\|\mu_I\| \,:\, G(x) \le 0,\, \mu_I > 0,\, G_I(x) = 0,\, \left\|\sum_{i\in I}\mu_i\,\nabla G_i(x)\right\| = 1,\right.$$
$$\left.\nabla G_i(x) \text{ linearly independent},\, i \in I \subseteq \{1,\ldots,m\}\right\}$$

is finite.

THEOREM 5.2. *Let $\theta : \Re^n \to \Re$ and $G : \Re^n \to \Re^m$ be continuously differentiable functions. Assume that each G_i is convex and a Slater point exists. Let $\{x^k\}$ be an arbitrary sequence of feasible vectors to* (14).

(a) *Assume $\theta_{inf} > -\infty$ and ACQ holds. If $\{x^k\}$ is LP minimizing for* (14)*, $\nabla\theta$ and each ∇G_i are uniformly continuous near $\{x^k\}$, and $\{\nabla\theta(x^k)\}$ is bounded, then there exists a sequence of bounded vectors $\{\mu^k\} \subset \Re^m_+$ such that*

$$\lim_{k\to\infty} v^k = 0 \quad \textit{and} \quad \lim_{k\to\infty} (\mu^k)^T G(x^k) = 0, \tag{15}$$

where $v^k \equiv \mathcal{L}(x^k, \mu^k)$.

(b) *Conversely if θ is convex,* (1) *is H-metrically regular, and there exists a sequence of bounded vectors $\{\mu^k\} \subset \Re^m_+$ such that* (15) *holds, then $\{x^k\}$ is LP minimizing for* (14).

Proof. Let X denote the feasible set of (14). We first prove (a). Similar to the proof of Theorem 5.1, we deduce the existence of two sequences of vectors $\{y^k\} \subset X$ and $\{w^k\} \subset \Re^n$ such that for each k, $y^k = \Pi_X(y^k - w^k)$ and (12) holds. In particular, y^k is an optimal solution of the following minimization problem:

$$\begin{aligned} \text{minimize} \quad & \tfrac{1}{2}\,\|\, y - (\, y^k - w^k\,)\,\|^2 \\ \text{subject to} \quad & y \in X. \end{aligned}$$

Since a Slater point exists for X, it follows that for each k there exists multipliers $\{\mu^k\} \subset \Re^m$ such that

$$w^k + \sum_{i=1}^m \mu_i^k \nabla G_i(y^k) = 0,$$
$$0 \ge G(y^k) \perp \mu^k \ge 0.$$

By ACQ, it follows that for some constant $c > 0$, each multiplier μ^k can be chosen to satisfy

$$\|\mu^k\| \le c\,\|\, w^k\,\|.$$

Since $\{\nabla\theta(x^k)\}$ is bounded and $\{\nabla\theta(x^k) - w^k\}$ converges to zero (see the proof of Theorem 5.1), it follows that $\{w^k\}$ is bounded; thus so is $\{\mu^k\}$. We have

$$v^k = \nabla\theta(x^k) - w^k + \sum_{i=1}^m \mu_i^k\,(\nabla G_i(x^k) - \nabla G_i(y^k)),$$

which easily implies $\lim_{k\to\infty} v^k = 0$.

To complete the proof of (i), it remains to show $\lim_{k\to\infty}(\mu^k)^T G(x^k) = 0$. We have

$$0 \geq (\mu^k)^T G(x^k) \geq (\mu^k)^T G(y^k) + \sum_{i=1}^{m} \mu_i^k \nabla G_i(y^k)^T (x^k - y^k) = (w^k)^T (y^k - x^k).$$

Since the last term tends to zero as k tends to infinity, the proof of (a) is completed.

The proof of (b) is by contradiction. Since this is very similar to the proofs of Theorem 4.3 and part (b) of Theorem 5.1, we omit the details. □

We conclude this section by giving two results that combine all the essential concepts presented in the paper. The first result pertains to the CQQSP (9) where θ is a convex quadratic spline; the other result is for the convex program (14) with Hölderian sharp minima.

THEOREM 5.3. *Assume that the CQQSP* (9) *has a strictly feasible solution and that the objective function $\theta(x)$ is bounded below on the feasible set X. Then θ_{inf} is finite and attained. Moreover for an arbitrary feasible sequence $\{x^k\} \subset X$, the following statements are equivalent:*

(a) *$\{x^k\}$ is LP minimizing for the CQQSP* (9)*;*

(b) *there exists a sequence of bounded multipliers $\{\mu^k\} \subset \Re^m_+$ such that* (15) *holds, where $v^k \equiv \mathcal{L}(x^k, \mu^k)$;*

(c) *$\{x^k\}$ is near the optimal set of the CQQSP;*

(d) *$\{\nabla\theta(x^k)\}$ is bounded and $\{x^k\}$ is N-stationary for the CQQSP.*

Proof. Let $\{P_i : i = 1, \ldots, p\}$ be the family of convex polyhedra whose union is equal to $\Re^n$, and let $\{q_i(x) : i = 1, \ldots, p\}$ be the corresponding family of convex quadratic pieces of θ. By the theory of ℓ_p-programming [34], it follows that for each i for which $X \cap P_i$ is nonempty, the CQQP below attains its finite optimal objective value:

$$\begin{aligned} &\text{minimize} && q_i(x) \\ &\text{subject to} && x \in X \cap P_i. \end{aligned}$$

The smallest of these p optimal objective values is clearly equal to θ_{inf}; moreover X_{opt} is obviously nonempty. By Lemma 4.2, the CQQSP has Hölderian minima and H-metrically regular level sets.

(a) ⇒ (c). This follows easily because $\{x^k\}$ is assumed feasible and the CQQSP has Hölderian minima.

(c) ⇒ (a). Assume (c). By Proposition 4.9, the sequence $\{\nabla\theta(x^k)\}$ is bounded. Since $\nabla\theta$ is uniformly continuous near $\{x^k\}$, it follows that θ is uniformly continuous near $\{x^k\}$. Hence (a) follows by Proposition 4.7.

(a) ⇒ (b). Assume (a). By Lemma 3.5 in [22], (ACQ) holds for the feasible set X. By the equivalence of (a) and (c) and the above argument, the sequence $\{\nabla\theta(x^k)\}$ is bounded. Hence (b) follows from Theorem 5.2.

(b) ⇒ (a). This follows easily from Theorem 5.2.

(d) ⇔ (a). This is the content of Corollary 4.5. □

THEOREM 5.4. *Let $\theta : \Re^n \to \Re$ and $G : \Re^n \to \Re^m$ be continuously differentiable functions. Assume that θ and each G_i are convex, a Slater point exists for the feasible set*

$$X \equiv \{\, x \in \Re^n : G(x) \leq 0 \,\},$$

the ACQ holds for X, and the problem (14) *has Hölderian minima. Let $\{x^k\}$ be an arbitrary sequence of feasible vectors to* (14). *Suppose that $\nabla\theta$ and each ∇G_i are uniformly continuous near $\{x^k\}$. The following statements are equivalent:*

(a) *$\{x^k\}$ is LP minimizing for the program* (14);
(b) *there exists a sequence of bounded multipliers $\{\mu^k\} \subset \Re^m_+$ such that* (15) *holds, where $v^k \equiv \mathcal{L}(x^k, \mu^k)$;*
(c) *$\{x^k\}$ is near X_{opt};*
(d) *$\{\nabla\theta(x^k)\}$ is bounded and $\{x^k\}$ is N-stationary.*

Proof. This follows from the same argument as in the previous theorem, except that Proposition 4.10 is used in place of Lemma 4.2. □

Remark. When X is an arbitrary closed convex set, statements (a), (c), and (d) of Theorem 5.4 remain equivalent without the Slater assumption and the ACQ.

6. Convergence of an iterative algorithm. In this last section, we consider a family of iterative methods for solving the constrained minimization problem (1). Our goal is to show that the H-metric regularity assumption that has played such an important role in the study of LP minimizing sequences is key to the convergence of these methods.

Consider the minimization problem (1) where θ is continuously differentiable and X is closed and convex. We present below an iterative descent method for solving this problem. The method generates a sequence of feasible vectors $\{x^k\} \subset X$ with decreasing objective function values $\{\theta(x^k)\}$. The generation of each iterate x^{k+1} is by solving a convex subprogram with a quadratic objective function which yields a feasible descent direction d^k for (1) at x^k, followed by an Armijo line search on θ starting at x^k and moving along d^k. Since the convergence analysis does not require X to be a polyhedron, we do not make this polyhedral assumption on X; thus the direction subprogram is not necessarily a quadratic program. In practice, this algorithm is perhaps restricted to a linearly constrained nonlinear program in order for the subprograms to be solved effectively.

A descent algorithm.

Step 0 (initialization). Let $\rho, \sigma \in (0,1)$ be given scalars. Let $\delta > 0$ be an arbitrary constant. Let $x^0 \in X$ be a given vector, and let B_0 be a symmetric positive semidefinite matrix. Set $k = 0$.

Step 1 (direction generation). Solve the convex program in the variable $d \in \Re^n$:

$$\begin{array}{lll} & \text{minimize} & \nabla\theta(x^k)^T d + \frac{1}{2}\, d^T B_k d \\ (16) & \text{subject to} & x^k + d \in X \\ & \text{and} & \| d \| \leq \delta. \end{array}$$

Let d^k be an arbitrary globally optimal solution, which must exist and satisfy $\nabla\theta(x^k)^T d^k \leq 0$.

If $\nabla\theta(x^k)^T d^k = 0$, stop because x^k is a stationary point of (1); thus x^k is a globally optimal solution if θ is convex. If $\nabla\theta(x^k)^T d^k < 0$, continue.

Step 2 (Armijo line search). Let m_k be the smallest nonnegative integer m such that

$$\theta(x^k + \rho^m d^k) - \theta(x^k) \leq \sigma\, \rho^m\, \nabla\theta(x^k)^T d^k.$$

Let $\tau_k \equiv \rho^{m_k}$ and set $x^{k+1} \equiv x^k + \tau_k\, d^k$.

Step 3 (termination check). Test x^{k+1} to determine if it satisfies a prescribed stopping rule. If so, stop; x^{k+1} is a desired approximate solution of (1). Otherwise

choose a symmetric positive semidefinite matrix B_{k+1} and return to Step 1 with k replaced by $k+1$.

We make several remarks about the above algorithm. First, $d = 0$ is a feasible solution to the subprogram (16) because $x^k \in X$. Moreover a globally optimal solution to this subprogram must exist because it has a nonempty compact feasible region. Such an optimal solution is not necessarily unique because the matrix B_k is not assumed to be positive definite. In the algorithm the direction d^k can be any optimal solution of (16). The assertion about the iterate x^k in the case where $\nabla\theta(x^k)^T d^k = 0$ can easily be proved; see [30] for a proof, which we omit. The integer m_k is well defined; the justification is standard. Finally, each iterate x^{k+1} clearly belongs to X.

The following is the main convergence result for the above algorithm. In the theorem, the infimum value $\theta_{\inf}$ is not assumed to be finite; moreover the sequence $\{x^k\}$ is not assumed to be bounded.

THEOREM 6.1. *Let $\theta : \Re^n \to \Re$ be a continuously differentiable convex function, and let X be a closed convex subset of $\Re^n$. Let $\{B_k\}$ be a sequence of symmetric positive semidefinite matrices. Let $\{x^k\}$ be an infinite sequence of vectors generated by the iterative descent algorithm. Assume that*

(a) *$\{B_k\}$ is bounded;*

(b) *$\nabla\theta$ is uniformly continuous near $\{x^k\}$;*

(c) *for each $\lambda > \theta_{inf}$, there exist scalars $c > 0$ and $\gamma \in (0,1)$ such that for all k,*

$$dist(x^k, L(\lambda)) \;\le\; c \max\left((\theta(x^k) - \lambda)_+, [(\theta(x^k) - \lambda)_+]^\gamma \right).$$

Then $\lim_{k\to\infty} \theta(x^k) = \theta_{inf}$.

Before proving this theorem we make several remarks about the assumptions. The first remark concerns the sequence $\{B_k\}$. We assume that each matrix B_k is only semidefinite and not definite. This is a significant departure from many convergence results of this type which assume that there exist constants $\alpha > \beta > 0$ such that

$$\beta\, v^T v \;\le\; v^T B_k v \;\le\; \alpha\, v^T v \quad \forall k \text{ and } v;$$

see, e.g., the recent note [38]. Our assumption is equivalent to the existence of the positive upper bound α but allows the lower bound β to be zero. Wu and Wu [39], focusing on the case where each B_k is identically equal to zero, established the above theorem without conditions (b) and (c). In a private discussion, S. Wu told the authors that he was not able to generalize his result to nonzero matrices B_k. Condition (c) is of course the error bound assumption that is central to the issue of whether a stationary sequence is LP minimizing. (We remind the reader that this condition holds, for instance, in the case of the CQQSP and a convex program with Hölderian minima.) In the present context, this condition can easily be seen to be necessary for the conclusion of the theorem to hold. Thus, condition (c) is actually necessary for the claimed convergence of the sequence $\{\theta(x^k)\}$. We suspect that condition (b) is not essential to the theorem; however, the proof below makes use of this condition. Finally we note that simple examples can be constructed to show that the theorem is false without assumption (a).

Proof of Theorem 6.1. Clearly $\{\theta(x^k)\}$ is a decreasing sequence of real numbers. Thus the limit $\lim_{k\to\infty} \theta(x^k)$ exists and is not smaller than $\theta_{\inf}$. We may assume by way of contradiction that this limit is greater than $\theta_{\inf}$. Let λ be such that

$$\lim_{k\to\infty} \theta(x^k) \;>\; \lambda \;>\; \theta_{\inf}.$$

By assumption, it follows that, for all k, there exists a vector $\bar{x}^k \in L(\lambda)$ satisfying $\theta(\bar{x}^k) = \lambda$ and

$$\| x^k - \bar{x}^k \| \leq c \max\left((\theta(x^k) - \lambda), (\theta(x^k) - \lambda)^\gamma\right).$$

Thus the sequence $\{x^k - \bar{x}^k\}$ is bounded. Moreover, defining the scalar

$$\delta_k \equiv \begin{cases} 1 & \text{if } \| x^k - \bar{x}^k \| \leq 1, \\ \| x^k - \bar{x}^k \|^{-1} & \text{otherwise,} \end{cases}$$

we see that the vector $d \equiv \delta_k(\bar{x}^k - x^k)$ is feasible to the subprogram (16). Hence, by the variational principle for this problem, we obtain

$$[\, \delta_k\, (\bar{x}^k - x^k) - d^k\,]^T (\nabla\theta(x^k) + B_k d^k) \geq 0.$$

By the gradient inequality, we have, for each k,

$$\begin{aligned} \lambda - \theta(x^k) &= \theta(\bar{x}^k) - \theta(x^k) \geq (\bar{x}^k - x^k)^T \nabla\theta(x^k) \\ &\geq \delta_k^{-1}\, (d^k)^T (\nabla\theta(x^k) + B_k d^k) - (\bar{x}^k - x^k)^T B_k d^k. \end{aligned} \tag{17}$$

Since $\{x^k - \bar{x}^k\}$ is bounded, it follows that

$$0 < \inf_k \delta_k \leq \sup_k \delta_k \leq 1.$$

Thus $\{\delta_k^{-1}\}$ is bounded. For each k, we have

$$\nabla\theta(x^k)^T d^k + \frac{1}{2}(d^k)^T B_k d^k \leq 0 \tag{18}$$

and

$$\theta(x^{k+1}) - \theta(x^k) \leq \sigma\, \tau_k\, \nabla\theta(x^k)^T d^k \leq 0.$$

It follows that

$$\lim_{k\to\infty} \tau_k\, \nabla\theta(x^k)^T d^k = 0.$$

If $\inf_k \tau_k > 0$, then we deduce

$$\lim_{k\to\infty} \nabla\theta(x^k)^T d^k = 0. \tag{19}$$

Assume that $\inf_k \tau_k = 0$. Let $\{\tau_k : k \in \kappa\}$ be a subsequence of $\{\tau_k\}$ that converges to zero. Thus

$$\lim_{k(\in\kappa)\to\infty} m_k = \infty.$$

We claim that

$$\lim_{k(\in\kappa)\to\infty} \nabla\theta(x^k)^T d^k = 0. \tag{20}$$

Since $\{d^k\}$ is bounded and $\nabla\theta$ is uniformly continuous near $\{x^k\}$, it follows that

$$\lim_{k(\in\kappa)\to\infty} \frac{\theta(x^k + \rho^{m_k - 1} d^k) - \theta(x^k) - \rho^{m_k - 1}\, \nabla\theta(x^k)^T d^k}{\rho^{m_k - 1}} = 0.$$

By the definition of m_k, we have

$$\theta(x^k + \rho^{m_k-1} d^k) - \theta(x^k) > \sigma \rho^{m_k-1} \nabla\theta(x^k)^T d^k.$$

Using the fact that $\sigma \in (0,1)$, we easily deduce from the last two expressions that (20) must hold.

Consequently, we have shown that regardless of the infimum value of the sequence $\{\tau_k\}$, (20) must hold for an infinite set κ. Without loss of generality, we may assume that (19) holds. By (18), and the positive semidefiniteness of each B_k, it follows that the sequence of scalars $\{(d^k)^T B_k d^k\}$ converges to zero. Furthermore, since $\{B_k\}$ is bounded (which implies that the eigenvalues of B_k are bounded above), it can easily be shown, by diagonalizing each B_k, that the sequence of vectors $\{B_k d^k\}$ converges to the zero vector.

Since $\{x^k - \bar{x}^k\}$ and $\{\delta_k^{-1}\}$ are bounded, passing to the limit $k \to \infty$ in (17), we deduce that the left-hand side converges to a negative limit, whereas the right-hand side converges to zero. This is a contradiction. □

Acknowledgments. The authors are grateful to Dr. Asen Dontchev for pointing out the connection between the subject of this paper and the theory of well-posedness of constrained optimization problems and for bringing to their attention the references [8, 21]. We are also grateful to Professor Zhi-Quan Luo, who drew our attention to the reference [34]. We acknowledge some fruitful discussions with Dr. Shiquan Wu on the topic of section 6. Three referees of the paper have made many constructive and insightful comments on previous versions of the paper that have significantly contributed to the present revision. One of them supplied us the reference [15], which had prompted us to include the discussion relating Lemaire's work at our second revision. Finally the authors are indebted to Professor Wu Li at Old Dominion University for his contribution of Proposition 4.10, which has allowed us to establish Theorem 5.4 under the Hölderian minima assumption. (In a previous version of this paper, the theorem was established under the (more restrictive) assumption of weak sharp minima, and the proof used Lemma 4.1 instead.)

REFERENCES

[1] A. Auslender, *Convergence of stationary sequences for variational inequalities with maximal monotone operators*, Appl. Math. Optim., 28 (1993), pp. 161–172.

[2] A. Auslender, R. Cominetti, and J.-P. Crouzeix, *Convex functions with unbounded level sets and applications to duality theory*, SIAM J. Optim., 3 (1993), pp. 669–687.

[3] A. Auslender and J.-P. Crouzeix, *Well-behaved asymptotical convex functions*, Analyse nonlinéare, 6 (1989), pp. 101–121.

[4] A. Bronstedt and R.T. Rockafellar, *On the subdifferentiability of convex functions*, Proc. Amer. Math. Soc., 16 (1965), pp. 605–611.

[5] J.V. Burke and M.C. Ferris, *Weak sharp minima in mathematical programming*, SIAM J. Control Optim., 31 (1993), pp. 1340–1359.

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

[7] R. Deville, G. Godefroy, and V. Zizler, *A smooth variational principle with applications to Hamilton-Jacobi equations in infinite dimensions*, J. Funct. Anal., 111 (1993), pp. 197–212.

[8] A.L. Dontchev and T. Zolezzi, *Well-Posed Optimization Problems*, Lecture Notes in Math. 1543, Springer-Verlag, Berlin, 1991.

[9] I. Ekeland, *On the variational principle*, J. Math. Anal. Appl., 47 (1974), pp. 324–353.

[10] M.C. Ferris, *Weak Sharp Minima and Penalty Functions in Mathematical Programming*, Technical report 779, Computer Science Department, University of Wisconsin, Madison, WI, 1988.

[11] M.C. Ferris, *Finite termination of the proximal point algorithm*, Math. Programming, 50 (1991), pp. 359–366.

[12] T. Fujisawa and E.S. Kuh, *Piecewise linear theory of nonlinear networks*, SIAM J. Appl. Math., 22 (1972), pp. 307–328.
[13] M. Fukushima and J.S. Pang, *Minimizing and stationary sequences of merit functions for nonlinear complementarity problems and variational inequalities*, in Complementarity and Variational Problems: State of the Art, M.C. Ferris and J.S. Pang, eds., SIAM, Philadelphia, 1997, pp. 91–104.
[14] K.C. Kiwiel and K. Murty, *Convergence of the steepest descent method for minimizing quasiconvex functions*, J. Optim. Theory Appl., 89 (1996), pp. 221–226.
[15] B. Lemaire, *Bonne position, conditionnement, et bon comportement asymptotique*, Exposé 5, Seminaire D'Analyse Convexe, Université de Montpellier, 1992.
[16] J.P. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms*, Vols. I and II, Springer-Verlag, Berlin, 1993.
[17] E.S. Levitin and B.T. Polyak, *Convergence of minimizing sequences in conditional extremum problems*, Soviet Math. Dokl., 7 (1966), pp. 764–767.
[18] A.S. Lewis and J.S. Pang, *Error bounds for convex inequality systems*, in Generalized Convexity, Generalized Monotonicity: Recent Results, in Proceedings of the Fifth Symposium on Generalized Convexity, Luminy, June 1996, J.P. Crouzeix, J.-E. Martinez-Legaz, and M. Volle, eds., Kluwer Academic Publishers, Dordrecht, 1997, pp. 75–110.
[19] W. Li, *Error bounds for piecewise convex quadratic programs and applications*, SIAM J. Control Optim., 33 (1995), pp. 1510–1529.
[20] W. Li, *private e-mail communication*, 1997.
[21] R. Lucchetti and J. Revalski, eds., *Recent Developments in Well-Posed Variational Problems*, Kluwer Academic Publishers, Dordrecht, 1995.
[22] X.D. Luo and Z.Q. Luo, *Extension of Hoffman's error bound to polynomial systems*, SIAM J. Optim., 4 (1994), pp. 383–392.
[23] Z.Q. Luo and J.S. Pang, *Error bounds for analytic systems and their applications*, Math. Programming, 67 (1994), pp. 1–28.
[24] O.L. Mangasarian, *A simple characterization of solution sets of convex programs*, Oper. Res. Lett., 7 (1988), pp. 21–26.
[25] O.L. Mangasarian, *A condition number for differentiable convex inequalities*, Math. Oper. Res., 10 (1985), pp. 175–179.
[26] O.L. Mangasarian, *Error Bounds for Nondifferentiable Convex Inequalities Under a Strong Slater Constraint Qualification*, Mathematical Programming Technical Report 96-04, Computer Science Department, University of Wisconsin, Madison, WI, 1996.
[27] O.L. Mangasarian and J. Ren, *New improved error bounds for the linear complementarity problem*, Math. Programming, 66 (1994) pp. 241–257.
[28] J.S. Pang, *A posteriori error bounds for the linearly constrained variational inequality problem*, Math. Oper. Res., 12 (1987), pp. 474–484.
[29] J.S. Pang, *Error bounds in mathematical programming*, Math. Programming Ser. B, 79 (1997), pp. 299–332.
[30] J.S. Pang, S.P. Han, and N. Rangaraj, *Minimization of locally Lipschitzian functions*, SIAM J. Optim., 1 (1991), pp. 57–82.
[31] J.P. Revalski, *Various aspects of well-posedness of optimization problems*, in Recent Developments in Well-Posed Variational Problems, Kluwer Academic Publishers, Dordrecht, 1995, pp. 229–256.
[32] S.M. Robinson, *Normal maps induced by linear transformations*, Math. Oper. Res., 17 (1992), pp. 691–714.
[33] R.T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[34] T. Terlaky, *On ℓ_p-programming*, European J. Oper. Res., 22 (1985), pp. 70–100.
[35] M.J. Todd, *On convergence properties of algorithms for unconstrained minimization*, IMA J. Numer. Anal., 9 (1989), pp. 435–441.
[36] A.N. Tykhonov, *On the stability of the functional optimization problem*, USSR Comput. Math. Math. Phys., 6 (1966), pp. 28–33.
[37] T. Wang and J.S. Pang, *Global error bounds for convex quadratic inequality systems*, Optimization, 31 (1994), pp. 1–12.
[38] Z. Wei, L. Qi, and H. Jiang, *Some Convergence Properties of Descent Methods*, manuscript, School of Mathematics, University of New South Wales, Sydney, 1995.
[39] F. Wu and S. Wu, *A modified Frank-Wolfe algorithm and its convergence properties*, Acta Math. Appl. Sinica, 11 (1995), pp. 286–291.
[40] S. Wu, *Convergence properties of descent methods for unconstrained minimization*, Optimization, 26 (1992), pp. 229–237.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 6, pp. 1937–1939, November 1998

005

ON THE OPTIMALITY OF THE DISCRETE KARHUNEN–LOÈVE EXPANSION*

ARNE DÜR†

Abstract. A short proof of the optimality of the discrete Karhunen–Loève expansion as the best linear approximation in the quadratic mean is presented.

Key words. Karhunen–Loève expansion

AMS subject classifications. 93E24, 94A12

PII. S0363012997315750

Let $\mathbf{R}$ denote the field of real numbers, and view vectors in $\mathbf{R^n}$ as column vectors. For $x, y \in \mathbf{R^n}$, let $\langle x, y\rangle = \sum_{i=1}^{n} x_i y_i$ denote the scalar product, and let $\|x\| = \sqrt{\langle x, x\rangle}$ denote the Euclidean norm. Then the Karhunen–Loève expansion is defined as follows.

THEOREM 0.1. *Let $X = (X_1, \ldots, X_n)^T$ be a real-valued random-vector with finite second moments, and let m be a fixed natural number not greater than n.*

Let a and A denote the expectation vector or the covariance matrix of X, respectively, and let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ be the eigenvalues of A with corresponding orthonormal eigenvectors $w_1, \ldots, w_n$.

Then, for any vectors $v_0, v_1, \ldots, v_m$ and any linear functionals $\phi_1, \ldots, \phi_m : \mathbf{R^n} \to \mathbf{R}$, the expectation

$$E\left(\left\|X - \left(v_0 + \sum_{i=1}^{m} \phi_i(X-a)v_i\right)\right\|^2\right)$$

is minimal if $v_0 = a$, $v_i = w_i$, and $\phi_i(x) = \langle x, w_i\rangle$ for $i = 1, \ldots, m$. Hence,

$$Y := a + \sum_{i=1}^{m} \langle X - a, w_i\rangle w_i$$

is the best m-dimensional linear approximation to X in the quadratic mean and is called the Karhunen–Loève expansion of X of order m. The approximation error is

$$E(\|X - Y\|^2) = \lambda_{m+1} + \cdots + \lambda_n.$$

An application of the discrete Karhunen–Loève expansion to image processing can be found in [4, p. 382]. In this book this optimality property of the discrete Karhunen–Loève expansion is stated without proof. For the author, the proof given in [3] is unclear. In [1], the optimality of the Karhunen–Loève transform among all unitary transforms is proved by Lagrangian multipliers but neglecting the orthogonality constraints among the basis vectors. Therefore, the author suggests the following short proof. Another optimality property of the Karhunen–Loève expansion is discussed in [2].

*Received by the editors January 29, 1997; accepted for publication (in revised form) December 18, 1997; published electronically July 29, 1998.
http://www.siam.org/journals/sicon/36-6/31575.html
†Institut für Mathematik, Universität Innsbruck, Technikerstraße 25, 6020 Innsbruck, Austria (arne.duer@uibk.ac.at).

Proof. For $i = 1, \ldots, m$, write $\phi_i(x) = F_i x$ with row vector F_i. Then $G := \sum_{i=1}^m v_i F_i$ is a quadratic matrix of rank $\leq m$. Let $H := I - G$, where I is the identity matrix, and let $Z := X - a$ be the centralized random vector. Then

$$X - \left(v_0 + \sum_{i=1}^m \phi_i(X-a) v_i \right) = \left(I - \sum_{i=1}^m v_i F_i \right) Z - (v_0 - a) = HZ - (v_0 - a)$$

and, since $E(Z) = 0$,

$$E\left(\left\| X - \left(v_0 + \sum_{i=1}^m \phi_i(X-a) v_i \right) \right\|^2 \right) = E(\|HZ\|^2) + \|v_0 - a\|^2.$$

As $E(\|HZ\|^2)$ does not depend on v_0, the optimal choice for v_0 is $v_0 = a$. Moreover

$$E(\|HZ\|^2) = \sum_{i,j=1}^n E(Z_i Z_j)(H^T H)_{ji} = tr(AH^T H).$$

If $v_i = w_i$ and $\phi_i(x) = \langle x, w_i \rangle = w_i^T x$ for $i = 1, \ldots, m$, then

$$H = I - \sum_{i=1}^m w_i w_i^T = \sum_{i=m+1}^n w_i w_i^T$$

is the orthogonal projection onto the subspace spanned by $w_{m+1}, \ldots, w_n$; thus $H^T H = H$ and

$$\mathrm{tr}(AH^T H) = \sum_{i=m+1}^n \mathrm{tr}(A w_i w_i^T) = \sum_{i=m+1}^n \lambda_i \mathrm{tr}(w_i w_i^T) = \sum_{i=m+1}^n \lambda_i.$$

For general v_i and ϕ_i, we invoke Theorem 1 of [6] or [5] to obtain

$$\mathrm{tr}(AH^T H) \geq \sum_{i=1}^n \lambda_i \mu_{n+1-i},$$

where $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n$ denote the eigenvalues of $H^T H$. We end the proof by showing that

$$\mu_1 \geq \cdots \geq \mu_{n-m} \geq 1,$$

which implies

$$\sum_{i=1}^n \lambda_i \mu_{n+1-i} \geq \sum_{i=m+1}^n \lambda_i.$$

Let U be the nullspace of the matrix G. As G has rank $k \leq m$, U has dimension $n - k \geq n - m$. But for all nonzero unit vectors $u \in U$ we have $Hu = (I - G)u = u$, and hence $\|Hu\|^2 = 1$. Now the Courant–Fischer minimax theorem [7, p. 100] implies that $\mu_{n-k} \geq 1$, and hence $\mu_{n-m} \geq 1$. □

REFERENCES

[1] A. N. Akansu and R. A. Haddad, *Multiresolution Signal Decomposition*, Academic Press, New York, 1992.
[2] V. R. Algazi and D. J. Sakrison, *On the optimality of the Karhunen-Loève expansion*, IEEE Trans. Inform. Theory, 15 (1969), pp. 319–320.
[3] Y. T. Chien and K. S. Fu, *On the generalized Karhunen-Loève expansion*, IEEE Trans. Inform. Theory, 13 (1967), pp. 518–520.
[4] D. F. Elliott and K. R. Rao, *Fast Transforms: Algorithms, Analyses, Applications*, Academic Press, New York, 1982.
[5] L. Mirsky, *On the trace of matrix products*, Math. Nachr. 20 (1959), pp. 171–174.
[6] H. Richter, *Zur Abschätzung von Matrizennormen*, Math. Nach. 18 (1958), pp. 178–187.
[7] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 6, pp. 1940–1961, November 1998

006

INTEGRAL CONTROL OF INFINITE-DIMENSIONAL LINEAR SYSTEMS SUBJECT TO INPUT SATURATION*

H. LOGEMANN†, E. P. RYAN†, AND S. TOWNLEY‡

Abstract. Closing the loop around an exponentially stable single-input single-output regular linear system, subject to a globally Lipschitz and nondecreasing actuator nonlinearity and compensated by an integral controller, is shown to ensure asymptotic tracking of constant reference signals, provided that (a) the steady-state gain of the linear part of the plant is positive, (b) the positive integrator gain is sufficiently small, and (c) the reference value is feasible in a very natural sense. The class of actuator nonlinearities under consideration contains standard nonlinearities important in control engineering such as saturation and deadzone.

Key words. regular infinite-dimensional systems, integral control, actuator nonlinearities, input saturation, robust tracking, operator Riccati equations

AMS subject classifications. 93C10, 93C20, 93C25, 93D09, 93D10, 93D21

PII. S0363012996314142

1. Introduction.

The synthesis of low-gain integral (I) and proportional-plus-integral (PI) controllers for uncertain stable plants has received considerable attention in the last 20 years. The following principle is well known (see Davison [5], Lunze [20], and Morari [24]): closing the loop around a stable, finite-dimensional, continuous-time, single-input, single-output plant with transfer function $\mathbf{G}(s)$, compensated by a pure integral controller k/s (see Fig. 1.1), will result in a stable closed-loop system which achieves asymptotic tracking of arbitrary constant reference signals, provided that $|k|$ is sufficiently small and $k\mathbf{G}(0) > 0$. Therefore, if a plant is known to be stable and if the sign of $\mathbf{G}(0)$ is known (this information can be obtained from plant step response data), then the problem of tracking by low-gain integral control reduces to that of tuning the gain parameter k. Such a controller design approach ("tuning regulator theory" [5]) has been successfully applied in process control; see, for example, Coppus, Sha, and Wood [3] and Lunze [19].

An analogous result holds for finite-dimensional multivariable systems under suitable assumptions on $\mathbf{G}(0)$; see [5, 20] and [24]. Moreover, the result has been extended by Logemann, Bontsema, and Owens [13], Logemann and Owens [14], Logemann and Townley [17], Pohjolainen [27, 28], and Pohjolainen and Lätti [29] to various classes of (abstract) infinite-dimensional systems and by Jussila and Koivo [9] and Koivo and Pohjolainen [11] to differential delay systems. Furthermore, the problem of tuning the integrator gain adaptively has been addressed recently in a number of papers; see Cook [2] and Miller and Davison [22, 23] for the finite-dimensional case and Logemann and Townley [16, 17, 18] for the infinite-dimensional case.

In this paper we present results which show that the above principle remains true if the plant to be controlled is a single-input, single-output, infinite-dimensional, linear

*Received by the editors December 27, 1996; accepted for publication (in revised form) September 26, 1997; published electronically August 3, 1998. This research was supported by the Human Capital and Mobility programme, project number CHRX-CT93-0402, and by NATO grant CRG 950179.
http://www.siam.org/journals/sicon/36-6/31414.html

†Department of Mathematical Sciences, University of Bath, Bath BA2 7AY, UK (hl@maths.bath.ac.uk, epr@math.bath.ac.uk).

‡Department of Mathematics (also with the Centre for Systems and Control Engineering, School of Engineering), University of Exeter, North Park Road, Exeter EX4 4QE, UK (townley@maths.ex.ac.uk).

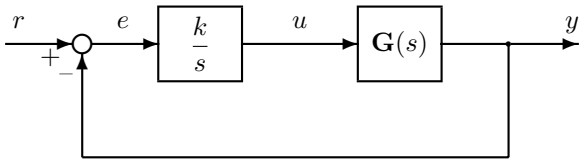


FIG. 1.1. *Low-gain control system.*

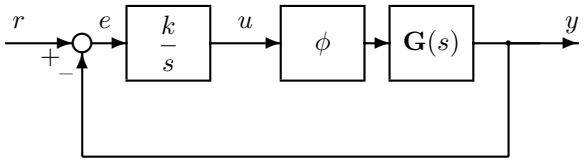


FIG. 1.2. *Low-gain control with input nonlinearity.*

system subject to an input nonlinearity (see Fig. 1.2). More precisely, we prove that, for an exponentially stable system with $\mathbf{G}(0) > 0$, there exists a number $K > 0$ such that, for all nondecreasing globally Lipschitz nonlinearities ϕ with Lipschitz constant λ and all $k \in (0, K/\lambda)$, the output $y(t)$ of the closed-loop system shown in Fig. 1.2 converges to r as $t \to \infty$, provided that $[\mathbf{G}(0)]^{-1} r \in \operatorname{clos}(\operatorname{im} \phi)$. The number K is the supremum of the set of all numbers $k > 0$ such that the function

$$1 + k\mathrm{Re}\,\frac{\mathbf{G}(s)}{s}$$

is positive real. The essence of our approach is to invoke a particular coordinate transformation and perform a Liapunov-type analysis on the transformed system. A parametrized operator Riccati equation plays a central role in the latter analysis, which further develops an idea presented in Townley and Kamstra [34].

The linear, infinite-dimensional part of the plant in Fig. 1.2 is assumed to be regular. The class of regular linear infinite-dimensional systems, introduced by Weiss [35, 36, 37, 38, 39], is rather general. It includes most distributed parameter systems and all time-delay systems (retarded and neutral) which are of interest in applications. Although there exist well-posed abstract infinite-dimensional systems which are not regular, the authors are of the opinion that any physically motivated, *well-posed*, linear, time-invariant control system is regular. We emphasize that our assumptions on the actuator nonlinearity allow for standard nonlinearities occurring in control engineering such as saturation and deadzone.

To our knowledge some of the results in this paper are new even for the finite-dimensional case. While Desoer and Lin [6] consider the low-gain tracking problem for a class of nonlinear finite-dimensional systems, their framework does not include input saturation.

The paper is organized as follows. Definitions and fundamental facts pertaining to regular systems are assembled in section 2. Section 3 contains the main result of the paper as outlined above. Examples and simulations illustrating our results are given in section 4. The proofs of three technical lemmas are given in the appendix.

Notation.

- For $\alpha \in \mathbb{R}$, set $\mathbb{C}_\alpha := \{s \in \mathbb{C} \,|\, \mathrm{Re}s > \alpha\}$.

- For $\alpha \in \mathbb{R}$ and H a Hilbert space, we define the exponentially weighted L^2-space $L^2_\alpha(\mathbb{R}_+, H) := \{f \in L^2_{loc}(\mathbb{R}_+, H) \,|\, f(\cdot)\exp(-\alpha\,\cdot) \in L^2(\mathbb{R}_+, H)\}$.
- If A is a linear operator, then the domain, spectrum, and resolvent set of A are denoted by $\mathrm{dom}\,(A)$, $\sigma(A)$, and $\varrho(A)$, respectively.
- The set of all linear bounded operators from H_1 to H_2 (where H_1, H_2 are Hilbert spaces) is denoted by $\mathcal{B}(H_1, H_2)$. We write $\mathcal{B}(H)$ for $\mathcal{B}(H, H)$.
- The Laplace transform is denoted by $\mathfrak{L}$.

2. Preliminaries on regular systems. In this section we give some background on well-posed linear systems; the reader is referred to Weiss [35, 36, 37, 38, 39] for full details.

First, we introduce some further notation. For any Hilbert space H and any $\tau \geq 0$, $\mathbf{R}_\tau$ denotes the right shift by τ on $L^2_{loc}(\mathbb{R}_+, H)$. The truncation operator $\mathbf{P}_\tau : L^2_{loc}(\mathbb{R}_+, H) \to L^2(\mathbb{R}_+, H)$ is given by

$$(\mathbf{P}_\tau u)(t) = \begin{cases} u(t) & \text{if} \quad t \in [0, \tau], \\ 0 & \text{if} \quad t > \tau. \end{cases}$$

For $u, v \in L^2_{loc}(\mathbb{R}_+, H)$ and $\tau \geq 0$, the *τ-concatenation* $u \overset{\tau}{\diamondsuit} v$ is defined by

$$u \overset{\tau}{\diamondsuit} v = \mathbf{P}_\tau u + \mathbf{R}_\tau v\,.$$

The fundamental concept of a well-posed linear system was introduced by Weiss [39]; an equivalent definition can be found in Salamon [33].

DEFINITION 2.1. *Let U, X, and Y be real Hilbert spaces. A* well-posed *linear system with* state-space X, input-space U, *and* output-space Y *is a quadruple* $\mathbf{\Sigma} = (\mathbf{T}, \mathbf{\Phi}, \mathbf{\Psi}, \mathbf{F})$, *where*

(1) $\mathbf{T} = (\mathbf{T}_t)_{t\geq 0}$ *is a C_0-semigroup of bounded linear operators on X,*

(2) $\mathbf{\Phi} = (\mathbf{\Phi}_t)_{t\geq 0}$ *is a family of bounded linear operators from $L^2(\mathbb{R}_+, U)$ to X such that*

$$\mathbf{\Phi}_{\tau+t}(u \overset{\tau}{\diamondsuit} v) = \mathbf{T}_t\mathbf{\Phi}_\tau u + \mathbf{\Phi}_t v$$

for all $u, v \in L^2(\mathbb{R}_+, U)$, and all $\tau, t \geq 0$,

(3) $\mathbf{\Psi} = (\mathbf{\Psi}_t)_{t\geq 0}$ *is a family of bounded linear operators from X to $L^2(\mathbb{R}_+, Y)$ such that*

$$\mathbf{\Psi}_{\tau+t}x_0 = \mathbf{\Psi}_\tau x_0 \overset{\tau}{\diamondsuit} \mathbf{\Psi}_t\mathbf{T}_\tau x_0$$

for all $x_0 \in X$ and all $\tau, t \geq 0$, and $\mathbf{\Psi}_0 = 0$,

(4) $\mathbf{F} = (\mathbf{F}_t)_{t\geq 0}$ *is a family of bounded linear operators from $L^2(\mathbb{R}_+, U)$ to $L^2(\mathbb{R}_+, Y)$ such that*

$$\mathbf{F}_{\tau+t}(u \overset{\tau}{\diamondsuit} v) = \mathbf{F}_\tau u \overset{\tau}{\diamondsuit} (\mathbf{\Psi}_t\mathbf{\Phi}_\tau u + \mathbf{F}_t v)\,,$$

$u, v \in L^2(\mathbb{R}_+, U)$ and all $\tau, t \geq 0$, and $\mathbf{F}_0 = 0$.

Let an input $u \in L^2_{loc}(\mathbb{R}_+, U)$ and an initial state $x_0 \in X$ be given. The state $x(t) = x(t; x_0, u)$ of $\mathbf{\Sigma}$ at time $t \geq 0$ and the output $y(\cdot) = y(\cdot\,; x_0, u)$ of $\mathbf{\Sigma}$ are defined by

$$x(t) = \mathbf{T}_t x_0 + \mathbf{\Phi}_t\mathbf{P}_t u\,, \tag{2.1}$$

$$\mathbf{P}_t y = \mathbf{\Psi}_t x_0 + \mathbf{F}_t\mathbf{P}_t u\,. \tag{2.2}$$

The state trajectory $x(\cdot)$ is a continuous function from $\mathbb{R}_+$ to X, and the output $y(\cdot)$ is in $L^2_{loc}(\mathbb{R}_+, Y)$.

We say that $\mathbf{\Sigma}$ is *exponentially stable* if the semigroup $\mathbf{T}$ is exponentially stable, i.e.,

$$\omega(\mathbf{T}) := \lim_{t\to\infty} \frac{1}{t} \log \|\mathbf{T}_t\| < 0\,.$$

If $\mathbf{\Sigma}$ is exponentially stable, then the operators $\mathbf{\Phi}_t$ and $\mathbf{\Psi}_t$ are uniformly bounded. It is clear that there exist unique operators $\mathbf{\Psi}_\infty : X \to L^2_{loc}(\mathbb{R}_+, Y)$ and $\mathbf{F}_\infty : L^2_{loc}(\mathbb{R}_+, U) \to L^2_{loc}(\mathbb{R}_+, Y)$ such that, for all $\tau \geq 0$,

$$\mathbf{\Psi}_\tau = \mathbf{P}_\tau \mathbf{\Psi}_\infty\,, \quad \mathbf{F}_\tau = \mathbf{P}_\tau \mathbf{F}_\infty\,.$$

It follows easily that $\mathbf{P}_\tau \mathbf{F}_\infty = \mathbf{P}_\tau \mathbf{F}_\infty \mathbf{P}_\tau$ for all $\tau \geq 0$, i.e., $\mathbf{F}_\infty$ is a *causal* operator. Moreover, if $\mathbf{\Sigma}$ is exponentially stable, then $\mathbf{\Psi}_\infty$ is a bounded operator from X into $L^2(\mathbb{R}_+, Y)$ and $\mathbf{F}_\infty$ maps $L^2(\mathbb{R}_+, U)$ boundedly into $L^2(\mathbb{R}_+, Y)$.

The generator of $\mathbf{T}$ is denoted by A. Let X_1 be the space $\mathrm{dom}\,(A)$ endowed with the graph norm. The norm on X is denoted by $\|\cdot\|$, while $\|\cdot\|_1$ denotes the graph norm. Let X_{-1} be the completion of X with respect to the norm $\|x\|_{-1} = \|(sI - A)^{-1}x\|$, where $s \in \varrho(A)$ is fixed. We have $X_1 \subset X \subset X_{-1}$, and the canonical injections are bounded and dense. The semigroup $\mathbf{T}$ can be restricted to a C_0-semigroup on X_1 and extended to a C_0-semigroup on X_{-1}. The exponential growth constant is the same on all three spaces. The generator on X_{-1} is an extension of A to X (which is bounded as an operator from X to X_{-1}). We shall use the same symbol $\mathbf{T}$ (respectively, A) for the original semigroup (respectively, its generator) and the associated restrictions and extensions. With this convention, we may write $A \in \mathcal{B}(X, X_{-1})$. Considered as a generator on X_{-1}, the domain of A is X.

By a representation theorem due to Salamon [33] (see also Weiss [37, 38]) there exist unique operators $B \in \mathcal{B}(U, X_{-1})$ and $C \in \mathcal{B}(X_1, Y)$ (the *control operator* and the *observation operator* of $\mathbf{\Sigma}$, respectively) such that, for all $t \geq 0$, $u \in L^2_{loc}(\mathbb{R}_+, U)$, and $x_0 \in X_1$,

$$\mathbf{\Phi}_t \mathbf{P}_t u = \int_0^t \mathbf{T}_{t-\tau} B u(\tau)\, d\tau \quad \text{and} \quad (\mathbf{\Psi}_\infty x_0)(t) = C\mathbf{T}_t x_0\ .$$

B is called *bounded* if $B \in \mathcal{B}(U, X)$ (and *unbounded* otherwise), whereas C is called *bounded* if it can be extended continuously to X (and *unbounded* otherwise). If $\mathbf{T}$ is exponentially stable, then there exist constants $\alpha, \beta > 0$ such that, for all $t \geq 0$, $u \in L^2(\mathbb{R}_+, U)$, and $x_0 \in X_1$,

$$\|\mathbf{\Phi}_t \mathbf{P}_t u\| = \left\| \int_0^t \mathbf{T}_{t-\tau} B u(\tau)\, d\tau \right\| \leq \alpha \|u\|_{L^2(0,t;\,U)}\,, \tag{2.3}$$

$$\|(\mathbf{\Psi}_\infty x_0)(\cdot)\|_{L^2(0,t;\,Y)} = \left(\int_0^t \|C\mathbf{T}_\tau x_0\|^2\, d\tau \right)^{1/2} \leq \beta \|x_0\|\,. \tag{2.4}$$

As in [38], the *Lebesgue extension* of C is defined by

$$C_L x_0 = \lim_{t\to 0} C \frac{1}{t} \int_0^t \mathbf{T}_\tau x_0\, d\tau\,,$$

where $\operatorname{dom}(C_L)$ is the set of all those $x_0 \in X$ for which the above limit exists. Clearly $X_1 \subset \operatorname{dom}(C_L) \subset X$ and, for any $x_0 \in X$, we have $\mathbf{T}_t x_0 \in \operatorname{dom}(C_L)$ for almost every (a.e.) $t \geq 0$. Furthermore,

$$(\mathbf{\Psi}_\infty x_0)(t) = C_L \mathbf{T}_t x_0 \quad \text{for a.e. } t \geq 0\,.$$

It can be shown (see Weiss [36, 38]) that, if $\alpha > \omega(\mathbf{T})$, $x_0 \in X$, and $u \in L^2_\alpha(\mathbb{R}_+, U)$, then $\mathbf{\Psi}_\infty x_0 \in L^2_\alpha(\mathbb{R}_+, Y)$, $\mathbf{F}_\infty u \in L^2_\alpha(\mathbb{R}_+, Y)$, and there exists a unique holomorphic $\mathbf{G} : \mathbb{C}_{\omega(\mathbf{T})} \to \mathcal{B}(U, Y)$ such that, for all $s \in \mathbb{C}_\alpha$,

$$\mathbf{G}(s)(\mathfrak{L}u)(s) = [\mathfrak{L}(\mathbf{F}_\infty u)](s)\,.$$

In particular, $\mathbf{G}$ is bounded on $\mathbb{C}_\alpha$ for all $\alpha > \omega(\mathbf{T})$. The function $\mathbf{G}$ is called the *transfer function* of $\mathbf{\Sigma}$.

$\mathbf{\Sigma}$ and its transfer function $\mathbf{G}$ are said to be *regular* if, for any $u \in U$, the limit

$$\lim_{s \to \infty,\, s \in \mathbb{R}} \mathbf{G}(s)u = Du$$

exists. It follows, from the principle of uniform boundedness, that $D \in \mathcal{B}(U, Y)$. The operator D is called the *feedthrough operator* of $\mathbf{\Sigma}$. If $\mathbf{\Sigma}$ is regular, then for any $x_0 \in X$ and $u \in L^2_{loc}(\mathbb{R}_+, U)$ the functions $x(\cdot)$ and $y(\cdot)$, defined by (2.1) and (2.2), satisfy the equations

$$\dot{x}(t) = Ax(t) + Bu(t)\,, \quad x(0) = x_0\,, \tag{2.5}$$

$$y(t) = C_L x(t) + Du(t) \tag{2.6}$$

for a.e. $t \geq 0$ (in particular $x(t) \in \operatorname{dom}(C_L)$ for a.e. $t \geq 0$). The derivative on the left-hand side of (2.5) has to be understood in X_{-1}. In other words, if we consider the initial value problem (2.5) in the space X_{-1}, then for any $x_0 \in X$ and $u \in L^2_{loc}(\mathbb{R}_+, U)$ the classical solution of (2.5) is given by the variation of parameters formula

$$x(t) = \mathbf{T}_t x_0 + \int_0^t \mathbf{T}_{t-\tau} B u(\tau)\, d\tau\,.$$

It has been demonstrated in [36] that if $\mathbf{\Sigma}$ is regular, then $(sI - A)^{-1}BU \subset \operatorname{dom}(C_L)$ for all $s \in \varrho(A)$ and the transfer function $\mathbf{G}$ can be expressed in the following way:

$$\mathbf{G}(s) = C_L(sI - A)^{-1}B + D \qquad \text{for all } s \in \mathbb{C}_{\omega(\mathbf{T})}\,,$$

which is familiar from finite-dimensional systems theory. The operators A, B, C, and D are called the *generating operators* of $\mathbf{\Sigma}$.

The following lemma will be needed in section 3. Certainly, it should be well known. However, since we could not find it in the literature, we include the proof.

LEMMA 2.1. *Suppose that $\mathbf{\Sigma} = (\mathbf{T}, \mathbf{\Phi}, \mathbf{\Psi}, \mathbf{F})$ is exponentially stable. Then the following statements hold:*

(1) *There exist $\alpha, \beta > 0$ such that, for any $x_0 \in X$ and any $u \in L^2(\mathbb{R}_+, U)$, the solution $x(\cdot)$ of the initial-value problem* (2.5) *satisfies*

$$\|x\|_{L^2(\mathbb{R}_+, X)} \leq \alpha \|u\|_{L^2(\mathbb{R}_+, U)} + \beta \|x_0\|.$$

(2) *If $u \in L^\infty(\mathbb{R}_+, U)$ and $\lim_{t \to \infty} u(t) = u_\infty$ exists, then for any $x_0 \in X$, $x(\cdot)$ defined by* (2.5) *satisfies*

$$\lim_{t \to \infty} \|x(t) + A^{-1}Bu_\infty\| = 0\,.$$

Proof. By the exponential stability we may assume, without loss of generality, that $x_0 = 0$. Consequently, we have $x(t) = \int_0^t \mathbf{T}_{t-\tau} B u(\tau)\, d\tau$ for all $t \geq 0$. Let $H^2(\mathbb{C}_0, X)$ denote the usual Hardy space of holomorphic functions defined on $\mathbb{C}_0$ with values in X. Appealing to the Paley–Wiener theorem, statement (1) will follow if we can show that there exists $\alpha > 0$ such that, for all $u \in L^2(\mathbb{R}_+, U)$,

$$\|\mathfrak{L}x\|_{H^2(\mathbb{C}_0,X)} \leq \alpha \|\mathfrak{L}u\|_{H^2(\mathbb{C}_0,U)}\,. \tag{2.7}$$

To this end, set $\omega_0 := \omega(\mathbf{T})$ and recall from [35] that for any $\omega > \omega_0$ there exists $M_\omega > 0$ such that, for all $s \in \mathbb{C}_\omega$,

$$\|(sI - A)^{-1} B\|_{\mathcal{B}(U,X)} \leq \frac{M_\omega}{\sqrt{\operatorname{Re}\, s - \omega}}\,. \tag{2.8}$$

It is clear that $s \mapsto (sI - A)^{-1}B$ is a holomorphic $\mathcal{B}(U, X_{-1})$-valued function: using the resolvent identity, it follows that it is also holomorphic as a $\mathcal{B}(U, X)$-valued function. The Laplace transform $\mathfrak{L}x$ of x satisfies

$$(\mathfrak{L}x)(s) = (sI - A)^{-1} B (\mathfrak{L}u)(s) \quad \text{for all } s \in \mathbb{C}_{\omega_0}\,. \tag{2.9}$$

By hypothesis, $\omega_0 < 0$ and $\mathfrak{L}u \in H^2(\mathbb{C}_0, X)$. Therefore, choosing $\omega_1 \in (\omega_0, 0)$ and combining (2.8) and (2.9) we see that (2.7) holds with, for example, $\alpha = M_{\omega_1}/\sqrt{|\omega_1|}$. This establishes statement (1).

To prove statement (2), we proceed as follows. Choose $t^* > 0$ such that $\|\mathbf{T}_t\| \leq 1/2$ for all $t \geq t^*$, let (t_n) be a sequence of real numbers satisfying

$$t^* \leq t_{n+1} - t_n \leq 2t^*\,,$$

and set $\beta = \sup\{\|\mathbf{T}_t\| \,|\, 0 \leq t \leq 2t^*\}$. For $t \geq t_n$ we have

$$x(t) = \mathbf{T}_{t-t_n} x(t_n) + \int_{t_n}^t \mathbf{T}_{t-\tau} B u(\tau)\, d\tau\,,$$

and so, by exponential stability, (2.3), and statement (1) above, there exists $\alpha > 0$ such that, for all $n \in \mathbb{N}$,

$$\|x(t)\| \leq \beta \|x(t_n)\| + \alpha\sqrt{2t^*}\|u\|_{L^\infty(t_n,t_{n+1})} \qquad \text{if } t \in [t_n, t_{n+1}] \tag{2.10}$$

and

$$\|x(t_{n+1})\| \leq \frac{1}{2}\|x(t_n)\| + \alpha\sqrt{2t^*}\|u\|_{L^\infty(t_n,t_{n+1})}\,. \tag{2.11}$$

We first consider the case when $u_\infty = 0$. Then

$$\lim_{n\to\infty} \|u\|_{L^\infty(t_n,t_{n+1})} = 0 \tag{2.12}$$

and (2.11) implies that

$$\lim_{n\to\infty} \|x(t_n)\| = 0\,. \tag{2.13}$$

Combining (2.10), (2.12), and (2.13) shows that $\lim_{t\to\infty} \|x(t)\| = 0$. Finally, if $u_\infty \neq 0$, then, by writing $u(t) = (u(t) - u_\infty) + u_\infty$, it is clear that it suffices to show that

$$\lim_{t\to\infty} \left\| \int_0^t \mathbf{T}_\tau B u_\infty\, d\tau + A^{-1} B u_\infty \right\| = 0\,. \tag{2.14}$$

Setting $z(t) = \int_0^t \mathbf{T}_\tau B u_\infty \, d\tau$ we have that

$$\lim_{t\to\infty} \|\dot{z}(t)\|_{-1} = \lim_{t\to\infty} \|\mathbf{T}_t B u_\infty\|_{-1} = 0\,. \tag{2.15}$$

The function $z(\cdot)$ is the classical solution of the initial-value problem $\dot{z}(t) = Az(t) + Bu_\infty$, $z(0) = 0$, considered in X_{-1}, and so we may write

$$z(\cdot) + A^{-1}Bu_\infty = A^{-1}\dot{z}(\cdot)\,. \tag{2.16}$$

Since $A^{-1} \in \mathcal{B}(X_{-1}, X)$, (2.14) follows from (2.15) and (2.16). □

3. Integral control in the presence of nonlinearities. In the following, let (A, B, C, D) be the generating operators of a linear, single-input, single-output regular system with state space X and transfer function $\mathbf{G}$. Suppose that the system is subject to an input nonlinearity ϕ, where $\phi : \mathbb{R} \to \mathbb{R}$ is locally Lipschitz. Denoting the constant reference signal by r, an application of the integrator

$$u(t) = u_0 + k \int_0^t [r - C_L x(\tau) - D\phi(u(\tau))]\, d\tau\,,$$

where k is a real parameter (see Fig. 1.2), leads to the following nonlinear system of differential equations:

$$\dot{x} = Ax + B\phi(u)\,, \quad x(0) = x_0 \in X, \tag{3.1}$$

$$\dot{u} = k[r - C_L x - D\phi(u)]\,, \quad u(0) = u_0 \in \mathbb{R}\,. \tag{3.2}$$

For $a \in (0, \infty]$, a continuous function

$$[0, a) \to X \times \mathbb{R}\,, \quad t \mapsto (x(t), u(t))$$

is called a *solution* of (3.1)–(3.2) if $(x(\cdot), u(\cdot))$ is absolutely continuous as an $(X_{-1} \times \mathbb{R})$-valued function, $x(t) \in \operatorname{dom}(C_L)$ for a.e. $t \in [0, a)$, $(x(0), u(0)) = (x_0, u_0)$, and the differential equations (3.1) and (3.2) are satisfied a.e. on $[0, a)$. Of course, the derivative on the left-hand side on (3.1) has to be understood in X_{-1}.[1]

An application of a well-known result on abstract Cauchy problems (see Pazy [26, Thm. 2.4, p. 107]), shows that a continuous $(X \times \mathbb{R})$-valued function $(x(\cdot), u(\cdot))$ is a solution of (3.1)–(3.2) if and only if it satisfies the following integrated version of (3.1)–(3.2):

$$x(t) = \mathbf{T}_t x_0 + \int_0^t \mathbf{T}_{t-\tau} B\phi(u(\tau))\, d\tau\,, \tag{3.3}$$

$$u(t) = u_0 + k \int_0^t [r - C_L x(\tau) - D\phi(u(\tau))]\, d\tau\,. \tag{3.4}$$

The next result shows that (3.1)–(3.2) has a unique solution.

PROPOSITION 3.1. *For any pair $(x_0, u_0) \in X \times \mathbb{R}$ of initial conditions there exists a unique solution $(x(\cdot), u(\cdot))$ of (3.1)–(3.2) defined on a maximal interval $[0, a_{max})$. If $a_{max} < \infty$, then*

$$\limsup_{t\to a_{max}} \|(x(t), u(t))\| = \infty\,. \tag{3.5}$$

[1] Being a Hilbert space, $X_{-1} \times \mathbb{R}$ is reflexive. Hence any absolutely continuous $(X_{-1} \times \mathbb{R})$-valued function is a.e. differentiable and can be recovered from its derivative by integration; see [1, Thm. 3.1, p. 10].

If ϕ is globally Lipschitz, then $a_{max} = \infty$.

For the proof of the above result it will be useful to consider the following initial-value problem for u:

$$\dot{u} = k[r - \mathbf{\Psi}_\infty x_0 - \mathbf{F}_\infty \phi(u)], \quad u(0) = u_0. \tag{3.6}$$

Clearly, (3.6) [2] is obtained from (3.2) on noting that $C_L x(t) + D\phi(u(t)) = (\mathbf{\Psi}_\infty x_0)(t) + (\mathbf{F}_\infty \phi(u))(t)$. An absolutely continuous function $u : [0, a) \to \mathbb{R}$ is a *solution* of (3.6) if $u(0) = u_0$ and the differential equation in (3.6) is satisfied a.e. on $[0, a)$.

LEMMA 3.2. *Let $x_0 \in X$. For any initial condition $u_0 \in \mathbb{R}$ there exists a unique solution $u(\cdot)$ of (3.6) defined on a maximal interval $[0, a_{max})$. If $a_{max} < \infty$, then*

$$\limsup_{t \to a_{max}} |u(t)| = \infty. \tag{3.7}$$

If ϕ is globally Lipschitz, then $a_{max} = \infty$.

The proof of this lemma is relegated to the appendix.

Proof of Proposition 3.1. Let $u : [0, a_{max}) \to \mathbb{R}$ be the unique maximal solution of (3.6) (whose existence is guaranteed by Lemma 3.2), and *define* $x(\cdot)$ to be the unique solution of

$$\dot{x} = Ax + B\phi(u), \quad x(0) = x_0.$$

Then $(x(\cdot), u(\cdot))$ is the unique solution of equations (3.1)–(3.2), which satisfies equation (3.5) if $a_{max} < \infty$. Moreover, it follows trivially from Lemma 3.2 that $a_{max} = \infty$ if ϕ is globally Lipschitz. □

Henceforth, let $\mathcal{M}$ denote the set of all bounded measures on $[0, \infty)$. A measure $\mu \in \mathcal{M}$ can be written in the form

$$\mu(dt) = a(t)dt + \sum_{i=0}^{\infty} a_i \delta_{t_i}(dt) + \mu_s(dt),$$

where $a(\cdot) \in L^1(0, \infty)$, $\sum_{i=0}^{\infty} a_i \delta_{t_i}$, and μ_s, respectively, represent the absolutely continuous, the discrete, and the singular parts of μ. In particular, δ_{t_i} denotes the unit point mass at $t_i \geq 0$ and the a_i are real numbers such that $\sum_{i=0}^{\infty} |a_i| < \infty$.

Furthermore, for $\lambda > 0$, let $\mathcal{N}(\lambda)$ denote the set of all nondecreasing globally Lipschitz nonlinearities $\phi : \mathbb{R} \to \mathbb{R}$ with Lipschitz constant λ. Finally, if $\mathbf{G}$ is holomorphic and bounded on $\mathbb{C}_\alpha$ for some $\alpha < 0$ and $\mathbf{G}(0) > 0$, then it is easy to show that

$$1 + k\mathrm{Re}\,\frac{\mathbf{G}(s)}{s} \geq 0 \qquad \text{for all } s \in \mathbb{C}_0 \tag{3.8}$$

for all sufficiently small $k > 0$; see Lemma 3.10 in [17]. We define

$$K := \sup\{k > 0 \,|\, (3.8) \text{ holds}\}.$$

The main result of this section is the following theorem.

THEOREM 3.3. *Let $\lambda > 0$ and $\phi \in \mathcal{N}(\lambda)$. Assume that $\mathbf{T}_t$ is exponentially stable, $\mathbf{G}(0) > 0$, $k \in (0, K/\lambda)$, and $r \in \mathbb{R}$ is such that*

$$\phi_r := [\mathbf{G}(0)]^{-1} r \in \mathrm{clos}\,(\mathrm{im}\,\phi). \tag{3.9}$$

If C is bounded, then for all $(x_0, u_0) \in X \times \mathbb{R}$ the unique solution $(x(\cdot), u(\cdot))$ of (3.1)–(3.2) exists on $[0, \infty)$ and satisfies

[2] Strictly speaking, to make sense of (3.6) we have to give a meaning to $\mathbf{F}_\infty v$ when v is a continuous function defined on a *finite* interval $[0, a)$ (recall that $\mathbf{F}_\infty$ operates on the space of locally square-integrable functions defined on the *infinite* interval $[0, \infty)$). This can easily be done using the causality of $\mathbf{F}_\infty$. Moreover, by slight abuse of notation, the expression $\phi(u)$ on the right-hand side of (3.6) denotes the function $t \mapsto \phi(u(t))$.

(1) $\lim_{t\to\infty} \phi(u(t)) = \phi_r$,
(2) $\lim_{t\to\infty} \|x(t) + A^{-1}B\phi_r\| = 0$,
(3) $\lim_{t\to\infty}(r - y(t)) = 0$, *where* $y(t) = Cx(t) + D\phi(u(t))$,
(4) *if* $\phi_r \in \operatorname{im}\phi$, *then*

$$\lim_{t\to\infty} \operatorname{dist}(u(t), \phi^{-1}(\phi_r)) = 0\,, \tag{3.10}$$

(5) *if* $\phi_r \in \operatorname{int}(\operatorname{im}\phi)$, *then* $u(\cdot)$ *is bounded.*

If C *is unbounded, then the statements* (1), (2), (4), *and* (5) *remain true provided that* $\mathfrak{L}^{-1}(\mathbf{G}) \in \mathcal{M}$ *and statement* (3) *remains true provided that* $x_0 \in \operatorname{dom}(A)$ *and* $\mathfrak{L}^{-1}(\mathbf{G}) \in \mathcal{M}$.

In particular, statement (4) says that $u(t)$ converges as $t \to \infty$ if the set $\phi^{-1}(\phi_r)$ is a singleton, which, in turn, is true if ϕ_r is not a critical value of ϕ.

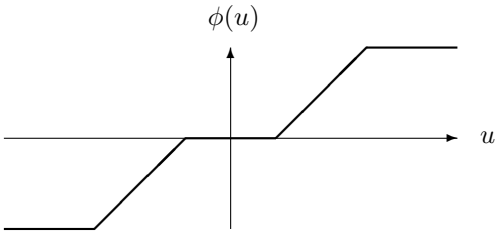


FIG. 3.1. *Nonlinearity with saturation and deadzone.*

The conditions imposed in Theorem 3.3 on ϕ are satisfied by saturation and deadzone nonlinearities and combinations of the two, as shown in Fig. 3.1. The assumption that $\mathfrak{L}^{-1}(\mathbf{G}) \in \mathcal{M}$ is not very restrictive and seems to be satisfied in all practical examples of systems with H^∞-transfer functions (in applications one usually has $\mu_s = 0$). If C is unbounded and $x_0 \notin \operatorname{dom}(A)$, then statement (3) does not hold in general. However, in that case, as an inspection of the proof of Theorem 3.3 will show, the error $e(\cdot) = r - y(\cdot)$ admits a decomposition $e = e_1 + e_2$, where $e_1 \in L^2_\alpha(\mathbb{R}_+, \mathbb{R})$ for some $\alpha < 0$ and e_2 is a continuous function satisfying $\lim_{t\to\infty} e_2(t) = 0$. Thus, while the error does not necessarily converge asymptotically to 0 as $t \to \infty$, it is small for large t in the sense that for all $\delta, \varepsilon > 0$ there exists $T > 0$ such that

$$\operatorname{meas}(\{t \geq T \,|\, |e(t)| \geq \delta\}) \leq \varepsilon\,,$$

where meas denotes the Lebesgue measure. In applying Theorem 3.3 it is important to know the constant K or at least a lower bound for K. In principle, K can be obtained from frequency-response experiments performed on the linear part of the plant; see [15] for details.

For the proof of Theorem 3.3 two lemmas are required, the proofs of which can be found in the appendix.

LEMMA 3.4. *Suppose that* $\mathbf{T}_t$ *is exponentially stable and* $\mathbf{G}(0) > 0$. *Define*

$$\mathbf{H}(s) = \frac{1}{s}(\mathbf{G}(s) - \mathbf{G}(0))\,.$$

If $0 < 2\kappa < K$, *then*

$$\|\mathbf{H}(1 + \kappa\mathbf{H})^{-1}\|_\infty < \frac{1}{\kappa} \tag{3.11}$$

and there exists $P \in \mathcal{B}(X)$, with $P = P^* \geq 0$ and such that the Riccati equation

$$\begin{aligned}\langle A_\kappa x_1, Px_2\rangle + \langle Px_1, A_\kappa x_2\rangle + \kappa^2\langle C_L x_1, C_L x_2\rangle \\ + \langle (A^{-1}B)^* Px_1, (A^{-1}B)^* Px_2\rangle = 0\end{aligned} \tag{3.12}$$

is satisfied for all $x_1, x_2 \in \mathrm{dom}(A_\kappa) = \mathrm{dom}(A)$, where $A_\kappa := A - \kappa A^{-1}BC_L$.

LEMMA 3.5. *Let $\phi : \mathbb{R} \to \mathbb{R}$ be locally Lipschitz and (ε_n) be any sequence with $\varepsilon_n > 0$ and $\lim_{n\to\infty} \varepsilon_n = 0$. Define the function $\phi^\diamond : \mathbb{R} \to \mathbb{R}$ by*

$$\phi^\diamond(\xi) = \limsup_{n\to\infty} \frac{\phi(\xi + \varepsilon_n) - \phi(\xi)}{\varepsilon_n}\,.$$

Then $\phi^\diamond \in L^\infty_{loc}(-\infty, \infty)$ ($\phi^\diamond \in L^\infty(-\infty, \infty)$ if ϕ is globally Lipschitz) and $\phi^\diamond \circ u$ is Lebesgue measurable for all Lebesgue measurable functions $u : [0, \infty) \to \mathbb{R}$. If u is absolutely continuous, so is $\phi \circ u$ and

$$\frac{d}{dt}(\phi \circ u)(t) = \phi^\diamond(u(t))\dot{u}(t) \quad \textit{for a.e. } t \in [0, \infty)\,.$$

Proof of Theorem 3.3. By Proposition 3.1, there exists a unique solution of (3.1)–(3.2) on $[0, \infty)$. We denote this solution by $(x(\cdot), u(\cdot))$ and introduce new variables by defining

$$z(t) := x(t) + A^{-1}B\phi(u(t))\,, \quad v(t) := \phi(u(t)) - \phi_r \quad \text{for all } t \geq 0.$$

By regularity it follows that $z(t) \in \mathrm{dom}\,(C_L)$ for a.e. $t \in [0, \infty)$. Moreover, by Lemma 3.5, $\dot{v}(t) = \phi^\diamond(u(t))\dot{u}(t)$ for a.e. $t \in [0, \infty)$. Therefore, an easy calculation yields

$$\dot{z} = Az - k\phi^\diamond(u)A^{-1}B(C_L z + \mathbf{G}(0)v)\,, \quad z(0) = z_0 := x_0 + A^{-1}B\phi(u_0), \tag{3.13}$$

$$\dot{v} = -k\phi^\diamond(u)(C_L z + \mathbf{G}(0)v)\,, \quad v(0) = v_0 := \phi(u_0) - \phi_r\,. \tag{3.14}$$

The derivative on the left-hand side of (3.13) and (3.14) has to be understood in X_{-1}. Notice that, since ϕ is nondecreasing, $\phi^\diamond(\xi) \geq 0$ for all $\xi \in \mathbb{R}$. We observe that, while in these new variables we still have an unbounded operator $A^{-1}BC_L$, the operator $A^{-1}B$ is in $\mathcal{B}(\mathbb{R}, X)$. We will investigate the stability properties of (3.13) and (3.14) using a Liapunov approach.

Since $0 < k\lambda < K$, it follows that there exists $\mu > \lambda/2$ such that $0 < 2\mu k < K$, and therefore, by Lemma 3.4,

$$\|\mathbf{H}(1 + \mu k\mathbf{H})^{-1}\|_\infty < \frac{1}{\mu k}\,.$$

By the same lemma, the Riccati equation (3.12) with $\kappa = \mu k$ has a solution $P \in \mathcal{B}(X)$ satisfying $P = P^* \geq 0$. Set

$$\tilde{P} = \begin{pmatrix} P & 0 \\ 0 & \mu k\mathbf{G}(0) \end{pmatrix},$$

and define

$$\tilde{A}_k = \begin{pmatrix} A - \mu kA^{-1}BC_L & -\mu kA^{-1}B\mathbf{G}(0) \\ -\mu kC_L & -\mu k\mathbf{G}(0) \end{pmatrix}, \quad \tilde{B} = \begin{pmatrix} A^{-1}B \\ 1 \end{pmatrix}, \quad \tilde{C} = (C_L \;\; \mathbf{G}(0))\,,$$

where $\mathrm{dom}(\tilde{A}_k) = \mathrm{dom}(A) \times \mathbb{R}$. The operator $\tilde{A}_k$ generates a C_0-semigroup. Using (3.12), it is easy to show that

$$\langle \tilde{A}_k \tilde{x}_1, \tilde{P}\tilde{x}_2\rangle + \langle \tilde{P}\tilde{x}_1, \tilde{A}_k\tilde{x}_2\rangle + \mu^2 k^2 \langle \tilde{C}\tilde{x}_1, \tilde{C}\tilde{x}_2\rangle + \langle \tilde{B}^*\tilde{P}\tilde{x}_1, \tilde{B}^*\tilde{P}\tilde{x}_2\rangle = 0 \tag{3.15}$$

is satisfied for all $\tilde{x}_1, \tilde{x}_2 \in \mathrm{dom}(\tilde{A}_k)$.

Setting $\tilde{z}(\cdot) = (z(\cdot), v(\cdot))$, (3.13) and (3.14) can be reformulated as

$$\dot{\tilde{z}} = \tilde{A}_k \tilde{z} + k(\mu - \phi^\diamond(u))\tilde{B}\tilde{C}\tilde{z}\,, \quad \tilde{z}(0) = \tilde{z}_0 := \begin{pmatrix} z_0 \\ v_0 \end{pmatrix}, \tag{3.16}$$

where the derivative on the left-hand side has to be understood in $X_{-1} \times \mathbb{R}$. For an intermediate step in the Liapunov analysis we need differentiability in $X \times \mathbb{R}$, and therefore, we will use an approximation argument. To this end let $T > 0$ be fixed but arbitrary, and choose $(w_n) \subset W^{1,2}(0,T;\mathbb{R})$ and $(\tilde{z}_0^n) \subset \mathrm{dom}(\tilde{A}_k)$ such that

$$\lim_{n\to\infty} \|k(\mu - \phi^\diamond(u))\tilde{C}\tilde{z} - w_n\|_{L^2(0,T)} = 0\,, \quad \lim_{n\to\infty} \|\tilde{z}_0 - \tilde{z}_0^n\|_{X\times\mathbb{R}} = 0\,. \tag{3.17}$$

Consider the system

$$\dot{\eta}(t) = \tilde{A}_k \eta(t) + \tilde{B} w_n(t)\,, \qquad \eta(0) = \tilde{z}_0^n\,. \tag{3.18}$$

$$\xi(t) = \tilde{C}\eta(t)\,. \tag{3.19}$$

The abstract initial-value problem (3.18) has a strong solution $\tilde{z}_n$ on $[0,T]$ in the sense that $\tilde{z}_n(0) = \tilde{z}_0^n$ and (3.18) is satisfied for a.e. $t \in [0,T]$ (see Pazy [26, Cor. 2.10, p. 109]). Using (3.17) we obtain

$$\lim_{n\to\infty} \|\tilde{z} - \tilde{z}_n\|_{L^2(0,T)} = 0\,; \qquad \lim_{n\to\infty} \|\tilde{z}(t) - \tilde{z}_n(t)\|_{X\times\mathbb{R}} = 0 \quad \text{for all } t \in [0,T]\,. \tag{3.20}$$

Setting $\xi_n(t) = \tilde{C}\tilde{z}_n(t)$, it follows from the regularity of (3.18) that

$$\lim_{n\to\infty} \|\tilde{C}\tilde{z} - \xi_n\|_{L^2(0,T)} = 0\,. \tag{3.21}$$

Differentiating the function

$$\tau \mapsto V_n(\tau) = \langle \tilde{z}_n(\tau), \tilde{P}\tilde{z}_n(\tau)\rangle$$

shows that, for a.e. $\tau \in [0,T]$,

$$\dot{V}_n(\tau) = \langle \tilde{z}_n(\tau), \tilde{P}\tilde{A}_k\tilde{z}_n(\tau)\rangle + \langle \tilde{A}_k\tilde{z}_n(\tau), \tilde{P}\tilde{z}_n(\tau)\rangle + 2\langle \tilde{B}w_n(\tau), \tilde{P}\tilde{z}_n(\tau)\rangle\,. \tag{3.22}$$

If $t \in [0,T]$, then integrating (3.22) from 0 to t, taking limits as $n \to \infty$, invoking (3.15), (3.17), (3.20), and (3.21), and setting

$$V(\tau) = \langle \tilde{z}(\tau), \tilde{P}\tilde{z}(\tau)\rangle$$

we obtain

$$V(t) - V(0) = -\int_0^t \mu^2 k^2 (\tilde{C}\tilde{z})^2 - \int_0^t (\tilde{B}^*\tilde{P}\tilde{z})^2 + 2\int_0^t \langle \tilde{B}k(\mu - \phi^\diamond(u))\tilde{C}\tilde{z}, \tilde{P}\tilde{z}\rangle\,.$$

Completing the square gives

$$V(t) - V(0) = -\int_0^t [\mu^2 k^2 - k^2(\phi^\diamond(u) - \mu)^2](\tilde{C}\tilde{z})^2 - \int_0^t [k(\phi^\diamond(u) - \mu)\tilde{C}\tilde{z} + \tilde{B}^*\tilde{P}\tilde{z}]^2\,,$$

and hence

$$V(t) - V(0) = -k^2 \int_0^t [2\mu\phi^\diamond(u) - (\phi^\diamond)^2(u)](\tilde{C}\tilde{z})^2 - \int_0^t [k(\phi^\diamond(u) - \mu)\tilde{C}\tilde{z} + \tilde{B}^*\tilde{P}\tilde{z}]^2 \,, \tag{3.23}$$

which holds for all $t \in [0, T]$. Since $T > 0$ was arbitrary, it follows that (3.23) holds for all $t \geq 0$. Therefore, using (3.23) and the definition of $\tilde{C}$,

$$k^2 \int_0^t (2\mu\phi^\diamond(u) - (\phi^\diamond)^2(u))(C_L z + \mathbf{G}(0)v)^2 \leq V(0) < \infty \quad \text{for all } t \geq 0 \,. \tag{3.24}$$

Now recall that $2\mu > \lambda$ and $\|\phi^\diamond(u)\|_{L^\infty(\mathbb{R}_+)} \leq \lambda$, so that

$$2\mu\phi^\diamond(u) - (\phi^\diamond)^2(u) > \varepsilon(\phi^\diamond)^2(u)$$

for some $\varepsilon > 0$. Therefore, (3.24) gives

$$\varepsilon k^2 \int_0^t (\phi^\diamond)^2(u)(C_L z + \mathbf{G}(0)v)^2 \leq V(0) < \infty \qquad \text{for all } t \geq 0 \,.$$

It follows that

$$\phi^\diamond(u)(C_L z + \mathbf{G}(0)v) \in L^2(\mathbb{R}_+) \,. \tag{3.25}$$

Using this in (3.13) and appealing to the fact that A, $A^{-1}B$, and C are the generating operators of a stable regular system we may conclude that

$$C_L z \in L^2(\mathbb{R}_+) \,. \tag{3.26}$$

Hence, by (3.25) and the boundedness of $\phi^\diamond(u)$,

$$\phi^\diamond(u)v \in L^2(\mathbb{R}_+), \tag{3.27}$$

and thus

$$(C_L z)\phi^\diamond(u)v \in L^1(\mathbb{R}_+) \,. \tag{3.28}$$

Using (3.24), (3.26)–(3.28), and the boundedness of $\phi^\diamond(u)$ it follows that

$$\phi^\diamond(u)v^2 \in L^1(\mathbb{R}_+) \,. \tag{3.29}$$

Multiplying (3.14) by $v(t)$, integrating, and then using (3.28) and (3.29) shows that

$$\lim_{t\to\infty} v^2(t) = v_0^2 + 2 \lim_{t\to\infty} \int_0^t v\dot{v} = \nu$$

for some $\nu \in [0, \infty)$. By continuity of $v(\cdot)$ it follows that

$$\lim_{t\to\infty} v(t) = \sqrt{\nu} \qquad \text{or} \qquad \lim_{t\to\infty} v(t) = -\sqrt{\nu} \,.$$

In the following we distinguish two cases: bounded and unbounded observation.

Let us first consider the case of bounded C. In order to prove statement (1), we have to show that $\nu = 0$. Seeking a contradiction, suppose that $\nu > 0$. Assuming

that $\lim_{t\to\infty} v(t) = \sqrt{\nu}$ (the case $\lim_{t\to\infty} v(t) = -\sqrt{\nu}$ can be dealt with in an entirely analogous fashion), we obtain that

$$\phi_\infty := \lim_{t\to\infty} \phi(u(t)) > \phi_r \,. \tag{3.30}$$

By Lemma 2.1, part (2), it follows that

$$\lim_{t\to\infty} \|x(t) + A^{-1}B\phi_\infty\| = 0 \,. \tag{3.31}$$

Using the boundedness of C it follows from (3.2), (3.30), and (3.31) that

$$\lim_{t\to\infty} \dot{u}(t) = k(r + CA^{-1}B\phi_\infty - D\phi_\infty) = k\mathbf{G}(0)(\phi_r - \phi_\infty) < 0 \,,$$

and so

$$\lim_{t\to\infty} u(t) = -\infty \,. \tag{3.32}$$

Since ϕ is nondecreasing we obtain

$$\phi_\infty = \lim_{t\to\infty} \phi(u(t)) = \inf(\operatorname{im}\phi) \le \phi_r \,,$$

contradicting (3.30). Therefore, $\nu = 0$, and consequently $\lim_{t\to\infty} \phi(u(t)) = \phi_r$, which is statement (1). Statement (2) follows now from Lemma 2.1, part (2), and statement (3) is a consequence of statements (1) and (2).

To prove statement (4), let $\phi_r \in \operatorname{im}\phi$. Seeking a contradiction, suppose that the claim is not true. Then there exists a sequence of positive numbers (t_n) with $\lim_{n\to\infty} t_n = \infty$ and $\varepsilon > 0$ such that

$$\operatorname{dist}(u(t_n), \phi^{-1}(\phi_r)) \ge \varepsilon \,. \tag{3.33}$$

If the sequence $(u(t_n))$ is bounded, we may assume, without loss of generality, that it converges to a finite limit u_∞. By continuity of ϕ and statement (1) we have that $\phi(u_\infty) = \phi_r$, and thus $u_\infty \in \phi^{-1}(\phi_r)$. This contradicts (3.33). So, suppose that $(u(t_n))$ is unbounded. Without loss of generality, we may then assume that $\lim_{n\to\infty} u(t_n) = \infty$. By monotonicity and statement (1) it follows that $\phi_r = \sup\phi$. Since $\phi_r \in \operatorname{im}\phi$ there exists ξ^* such that

$$\phi(\xi^*) = \phi_r = \sup\phi = \max\phi \,.$$

By monotonicity of ϕ we have

$$\phi(\xi) = \phi_r = \max\phi \qquad \text{for all } \xi \ge \xi^* \,.$$

In particular, we see that $u(t_n) \in \phi^{-1}(\phi_r)$ for all sufficiently large n, contradicting (3.33).

To prove statement (5) assume that $\phi_r \in \operatorname{int}(\operatorname{im}\phi)$. Again seeking a contradiction, suppose that the claim is not true. Then there exists a sequence of positive numbers (t_n) with $\lim_{n\to\infty} t_n = \infty$ and $\lim_{n\to\infty} |u(t_n)| = \infty$. Without loss of generality, we may assume that $\lim_{n\to\infty} u(t_n) = \infty$. By monotonicity it then follows that

$$\phi_r = \lim_{n\to\infty} \phi(u(t_n)) = \sup\phi \,,$$

contradicting the hypothesis $\phi_r \in \operatorname{int}(\operatorname{im}\phi)$.

Now let us consider the case of unbounded C with $\mathfrak{L}^{-1}(\mathbf{G}) \in \mathcal{M}$. We will again be seeking a contradiction, and hence assume that $\nu > 0$. It is clear that (3.30) and (3.31) still hold. It only remains to show that (3.32) is also true in this case. To this end, write (3.2) in the form

$$\dot{u} = k[r - C_L \mathbf{T}_t x_0 - \mathfrak{L}^{-1}(\mathbf{G}) \star \phi(u)]\,. \tag{3.34}$$

Since $\lim_{t\to\infty} \phi(u(t)) = \phi_\infty$ and $\mathfrak{L}^{-1}(\mathbf{G}) \in \mathcal{M}$ it follows that $\lim_{t\to\infty}(\mathfrak{L}^{-1}(\mathbf{G}) \star \phi(u))(t) = \mathbf{G}(0)\phi_\infty$ (see [8, Thm. 6.1, part (ii), p. 96]). Therefore, by (3.30) there exists $\delta > 0$ and $T > 0$ such that

$$\mathbf{G}(0)\phi_r - (\mathfrak{L}^{-1}(\mathbf{G}) \star \phi(u))(t) \le -\delta \qquad \text{for all } t \ge T\,. \tag{3.35}$$

Integrating (3.34) from T to t and using (3.35) gives

$$u(t) \le u(T) + k\left[\int_T^t |C_L \mathbf{T}_\tau x_0|\, d\tau - \delta(t - T)\right]\,. \tag{3.36}$$

By exponential stability of $\mathbf{T}_t$ we have that the map $t \mapsto C_L \mathbf{T}_t x_0$ is in $L^2_\alpha(\mathbb{R}_+, \mathbb{R})$ for some $\alpha < 0$, and hence in $L^1(\mathbb{R}_+, \mathbb{R})$. As a consequence, (3.36) yields

$$\lim_{t\to\infty} u(t) = -\infty\,,$$

which is (3.32). Statements (2), (4), and (5) then follow as in the case of bounded C. Finally, write $y(t)$ in the form

$$y(t) = C_L \mathbf{T}_t x_0 + (\mathfrak{L}^{-1}(\mathbf{G}) \star \phi(u))(t)\,.$$

Under the assumption that $x_0 \in \operatorname{dom}(A)$ and $\mathfrak{L}^{-1}(\mathbf{G}) \in \mathcal{M}$, we obtain

$$\lim_{t\to\infty} y(t) = \mathbf{G}(0) \lim_{t\to\infty} \phi(u(t))\,.$$

Combining this with statement (1) yields statement (3). □

One of the conditions imposed in Theorem 3.3 is that $[\mathbf{G}(0)]^{-1} r \in \operatorname{clos}(\operatorname{im}\phi)$. The following proposition shows that this condition is necessary for solvability of the tracking problem.

PROPOSITION 3.6. *Let $r \in \mathbb{R}$, and suppose that $\phi : \mathbb{R} \to \mathbb{R}$ is continuous, $\mathbf{T}_t$ is exponentially stable, and $\mathbf{G}(0) \neq 0$. If there exist an initial condition $x_0 \in X$ and a continuous function $u : [0, \infty) \to \mathbb{R}$ such that $\phi(u(\cdot))$ is bounded and*

$$\lim_{t\to\infty} [C_L x(t) + D\phi(u(t))] = r\,,$$

where $x(t) = \mathbf{T}_t x_0 + \int_0^t \mathbf{T}_{t-\tau} B\phi(u(\tau))\, d\tau$, then $\phi_r = [\mathbf{G}(0)]^{-1} r \in \operatorname{clos}(\operatorname{im}\phi)$.

The proof of the above proposition requires some preparation. Recall the concept of an ω-limit point (and ω-limit set $\Omega(\psi)$) of a continuous function $\psi : [0, \infty) \to \mathbb{R}$. A point ψ^* is an *ω-limit point* of ψ if there exists an increasing sequence $(t_n) \subset [0, \infty)$ such that $t_n \to \infty$ and $\psi(t_n) \to \psi^*$ as $n \to \infty$. The set $\Omega(\psi)$ of all ω-limit points is the *ω-limit set* of ψ.

The following lemma is probably standard; however, we were unable to locate it in the literature and so include a proof for completeness.

LEMMA 3.7. *Let $\psi : [0,\infty) \to \mathbb{R}$ be continuous and bounded. Then*

$$\lim_{s\to 0,\, s>0} [s(\mathfrak{L}\psi)(s)] = \omega \quad \Longrightarrow \quad \omega \in \Omega(\psi)\,.$$

Proof. It suffices to prove the result in the case $\omega = 0$ (if $\omega \neq 0$, then simply replace ψ by $\psi_\omega : t \mapsto \psi(t) - \omega$). It is well known that $\Omega(\psi)$ is compact and is approached by $\psi(t)$ as $t \to \infty$ (see, for example, [10, p. 113]). Seeking a contradiction, suppose $0 \notin \Omega(\psi)$. Then there exists $\varepsilon > 0$ and $T > 0$ such that for all $t \geq T$, $|\psi(t)| \geq \varepsilon$. Since ψ is continuous, we may restrict our attention, without loss of generality, to the case $\psi(t) \geq \varepsilon$ for all $t \geq T$. Then, for all $s \in (0,\infty)$, we have

$$(\mathfrak{L}\psi)(s) = \int_0^\infty e^{-st}\psi(t)\,dt \geq \int_0^T e^{-st}\psi(t)\,dt + \varepsilon \int_T^\infty e^{-st}\,dt \tag{3.37}$$

$$= \int_0^T e^{-st}\psi(t)\,dt + \frac{\varepsilon e^{-sT}}{s}\,, \tag{3.38}$$

whence the contradiction

$$0 = \lim_{s\to 0,\, s>0} s(\mathfrak{L}\psi)(s) \geq \varepsilon > 0\,.$$

□

Proof of Proposition 3.6. For $\delta \in (0, \pi/2)$ define the open sector $\mathcal{S}(\delta) \subset \mathbb{C}_0$ by

$$\mathcal{S}(\delta) := \{\rho e^{i\alpha} \mid \rho \in (0,\infty),\, \alpha \in (-\delta, \delta)\}\,.$$

Setting $\psi(t) = \phi(u(t))$ and $y(t) = C_L x(t) + D\psi(t)$ we obtain

$$(\mathfrak{L}y)(s) = \mathbf{G}(s)(\mathfrak{L}\psi)(s) + C(sI - A)^{-1}x_0\,,$$

and so by the final-value theorem (see [7, Satz 34.2] or [25, Thm. 14, p. 95])

$$r = \lim_{t\to\infty} y(t) = \lim_{s\to 0,\, s\in\mathcal{S}(\delta)} s(\mathfrak{L}y)(s) = \lim_{s\to 0,\, s\in\mathcal{S}(\delta)} s\mathbf{G}(s)(\mathfrak{L}\psi)(s)\,.$$

Since $\mathbf{G}(0) \neq 0$ it follows using Lemma 3.7 that

$$\phi_r = [\mathbf{G}(0)]^{-1} r = \lim_{s\to 0,\, s\in\mathcal{S}(\delta)} s(\mathfrak{L}\psi)(s) \in \Omega(\psi) \subset \operatorname{clos}(\operatorname{im}\phi)\,. \qquad \square$$

A result similar to Proposition 3.6 was stated without proof by Miller and Davison [22] in a finite-dimensional context. However, their approach (as outlined by Miller [21]) does not extend to infinite-dimensional regular systems.

4. Example: Controlled diffusion process with output delay. Consider a diffusion process (with diffusion coefficient $a > 0$ and with Dirichlet boundary conditions), on the one-dimensional spatial domain $[0,1]$, with scalar nonlinear pointwise control action (applied at point $x_b \in (0,1)$ via a nonlinearity ϕ with Lipschitz constant $\lambda > 0$) and delayed (delay $h \geq 0$) pointwise scalar observation (output at point $x_c \in (0,1)$, $x_c \geq x_b$.). We formally write this single-input, single-output system as

$$z_t(t,x) = az_{xx}(t,x) + \delta(x - x_b)\phi(u(t)), \quad y(t) = z(t-h, x_c),$$
$$z(t,0) = 0 = z(t,1) \qquad \text{for all } t > 0\,.$$

For simplicity, we assume zero initial conditions as follows:

$$z(t,x) = 0 \qquad \text{for all } (t,x) \in [-h,0] \times [0,1]\,.$$

With input $\phi(u(\cdot))$ and output $y(\cdot)$, this example qualifies as a regular linear system with transfer function given by

$$\mathbf{G}(s) = \frac{e^{-sh}\sinh\left(x_b\sqrt{(s/a)}\right)\sinh\left((1-x_c)\sqrt{(s/a)}\right)}{a\sqrt{(s/a)}\sinh\sqrt{(s/a)}}\,.$$

In this case, a detailed analysis (see [15] for related investigations) yields

$$\begin{aligned} K &:= \sup\{k > 0 \,|\, (3.8) \text{ holds}\} \\ &= \frac{1}{|\mathbf{G}'(0)|} = \frac{6a^2}{x_b(1-x_c)(6ha + 1 - x_b^2 - (1-x_c)^2)}. \end{aligned}$$

Therefore, by Theorem 3.3, for each $k \in (0, K/\lambda)$, the integral control

$$u(t) = k\int_0^t [r - y(t)]\,dt$$

guarantees asymptotic tracking of every constant reference signal r satisfying

$$\frac{r}{\mathbf{G}(0)} = \frac{ar}{x_b(1-x_c)} \in \operatorname{clos}(\operatorname{im}\phi)\,.$$

For purposes of illustration, we adopt the following values:

$$a = 0.1, \quad x_b = \frac{1}{3}, \quad x_c = \frac{2}{3}, \quad h = 1, \quad r = 1.$$

We consider a nonlinearity ϕ of saturation type, defined as follows:

$$u \mapsto \phi(u) := \begin{cases} 1, & u \geq 1, \\ u, & u \in (0,1), \\ 0, & u \leq 0 \end{cases}$$

in which case $\lambda = 1$ and

$$K = \frac{243}{620}\ (\approx 0.3919).$$

For $r = 1$, we have

$$\frac{r}{\mathbf{G}(0)} = \frac{a}{x_b(1-x_c)} = 0.9 \in [0,1] = \operatorname{clos}(\operatorname{im}\phi)\,.$$

In each of the following three cases of admissible controller gains

$$\text{(i) } k = 0.39, \quad \text{(ii) } k = 0.26, \quad \text{(iii) } k = 0.13,$$

Fig. 4.1 depicts the output behavior of the system under integral control, while Fig. 4.2 depicts the corresponding control input. These figures were generated using SIMULINK Simulation Software within MATLAB wherein a truncated eigenfunction expansion, of order 10, was adopted to model the diffusion process.

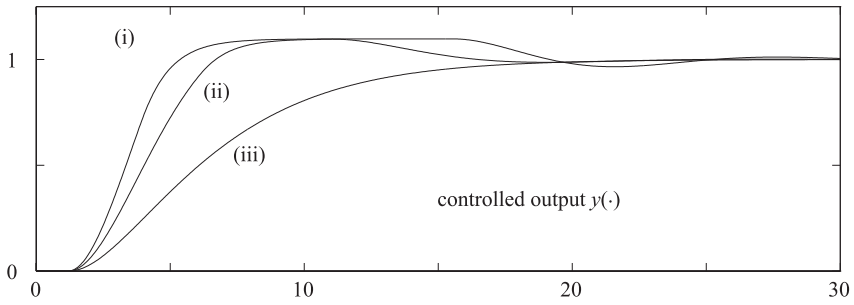


FIG. 4.1. *Controlled output.*

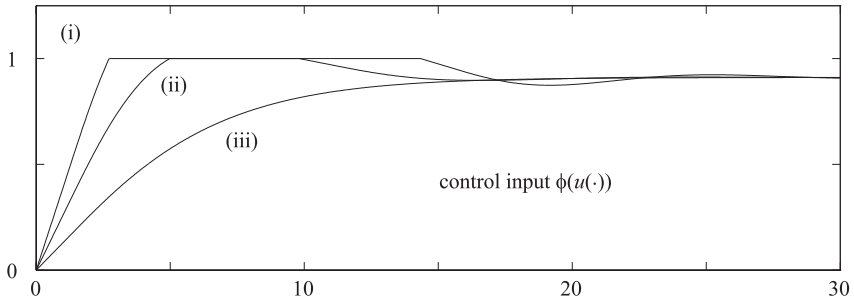


FIG. 4.2. *Control input.*

Appendix.

Proof of Lemma 3.2. In proving Lemma 3.2, we will study an initial-value problem which is slightly more general than (3.6). Let $\alpha \geq 0$, and let $w \in C([0,\alpha],\mathbb{R})$. Consider the initial-value problem

$$\dot{u}(t) = k[r - (\mathbf{\Psi}_\infty x_0)(t) - (\mathbf{F}_\infty \phi(u))(t)]\,, \quad t \geq \alpha\,, \tag{A.1}$$

$$u(t) = w(t)\,, \quad t \in [0,\alpha]\,. \tag{A.2}$$

LEMMA A.1. *Let $x_0 \in X$. For any initial function $w \in C([0,\alpha],\mathbb{R})$ there exists $\varepsilon > 0$ and a unique function $u \in C([0,\alpha+\varepsilon],\mathbb{R})$ with $u(t) = w(t)$ for all $t \in [0,\alpha]$ and such that u is absolutely continuous on $[\alpha,\alpha+\varepsilon]$ and* (A.1) *is satisfied for a.e. $t \in [\alpha,\alpha+\varepsilon]$.*

Proof. Without loss of generality, we may assume that $k = 1$. For $\delta > 0$ and $\eta > \|w\|_\infty$, define

$$\begin{aligned}\mathcal{C}_{\delta,\eta} = \{u \in C([0,\alpha+\delta],\mathbb{R}) \mid\ & |u(t) - w(t)| \leq \eta \quad \text{if } 0 \leq t \leq \alpha\,; \\ & |u(t) - w(\alpha)| \leq \eta \quad \text{if } \alpha \leq t \leq \alpha+\delta\}\,.\end{aligned}$$

Choosing $\eta > \|w\|_\infty$ guarantees that $\mathcal{C}_{\delta,\eta}$ contains the zero function. Using the causality of $\mathbf{F}_\infty$, the boundedness of the operators $\mathbf{P}_t\mathbf{F}_\infty$, and the Lipschitz continuity of ϕ, it is clear that, for given numbers $\delta > 0$ and $\eta > \|w\|_\infty$, there exists $\lambda > 0$ such that, for all $\varepsilon \in (0,\delta]$ and all $u, v \in \mathcal{C}_{\varepsilon,\eta}$,

$$\int_\alpha^{\alpha+\varepsilon} |\mathbf{F}_\infty\phi(u) - \mathbf{F}_\infty\phi(v)|^2 \leq \lambda^2 \int_0^{\alpha+\varepsilon} |u - v|^2\,.$$

Using Hölder's inequality we obtain the estimate

$$\int_\alpha^{\alpha+\varepsilon} |\mathbf{F}_\infty\phi(u) - \mathbf{F}_\infty\phi(v)| \le \lambda\sqrt{\varepsilon}\left(\int_0^{\alpha+\varepsilon} |u-v|^2\right)^{1/2}, \tag{A.3}$$

which holds for all $u, v \in \mathcal{C}_{\varepsilon,\eta}$, and all $\varepsilon \in (0,\delta]$. Moreover, if $v = 0$, then we may conclude that, for all $u \in \mathcal{C}_{\varepsilon,\eta}$ and all $\varepsilon \in (0,\delta]$,

$$\int_\alpha^{\alpha+\varepsilon} |\mathbf{F}_\infty\phi(u)| \le \int_\alpha^{\alpha+\varepsilon} |(\mathbf{F}_\infty\phi(0))(\tau)|\,d\tau + \lambda\sqrt{\varepsilon}\left(\int_0^{\alpha+\varepsilon} |u|^2\right)^{1/2}. \tag{A.4}$$

Set $f(t) = r - (\mathbf{\Psi}_\infty x_0)(t)$, and choose $\rho > 0$ such that

$$\int_\alpha^{\alpha+\rho} (|f(\tau)| + |(\mathbf{F}_\infty\phi(0))(\tau)|)\,d\tau \le \frac{\eta}{2}. \tag{A.5}$$

Now choose $\varepsilon > 0$ such that

$$\varepsilon \le \delta, \quad \varepsilon \le \rho, \quad \varepsilon < \frac{1}{\lambda}, \quad \varepsilon \le \frac{1}{4(\alpha+\rho)}\left(\frac{\eta}{\lambda\max\{\|w\|_\infty, |w(\alpha)|+\eta\}}\right)^2. \tag{A.6}$$

Define the operator Γ by

$$\begin{aligned}(\Gamma u)(t) &= w(t), \qquad 0 \le t \le \alpha,\\ (\Gamma u)(t) &= w(\alpha) + \int_\alpha^t f(\tau)\,d\tau - \int_\alpha^t (\mathbf{F}_\infty\phi(u))(\tau)\,d\tau, \qquad t \ge \alpha,\end{aligned}$$

and set

$$\tilde{\mathcal{C}}_{\varepsilon,\eta} := \{u \in \mathcal{C}_{\varepsilon,\eta} \mid u(t) = w(t) \quad \text{if } 0 \le t \le \alpha\}.$$

Clearly, $\tilde{\mathcal{C}}_{\varepsilon,\eta}$ is a complete metric space, and the lemma follows if we can show that Γ is a contraction on $\tilde{\mathcal{C}}_{\varepsilon,\eta}$.

We first show that $\Gamma(\tilde{\mathcal{C}}_{\varepsilon,\eta}) \subset \tilde{\mathcal{C}}_{\varepsilon,\eta}$. Using (A.4)–(A.6) we obtain, for all $u \in \mathcal{C}_{\varepsilon,\eta}$ and all $t \in [\alpha, \alpha+\varepsilon]$,

$$\begin{aligned}|(\Gamma u)(t) - w(\alpha)| &\le \lambda\sqrt{\varepsilon}\left(\int_0^{\alpha+\varepsilon} |u(\tau)|^2\,d\tau\right)^{1/2} + \frac{\eta}{2}\\ &\le \frac{\eta}{2} + \lambda\sqrt{\varepsilon(\alpha+\rho)}\max\{\|w\|_\infty, |w(\alpha)|+\eta\}\\ &\le \eta,\end{aligned}$$

which shows that $\Gamma(\tilde{\mathcal{C}}_{\varepsilon,\eta}) \subset \tilde{\mathcal{C}}_{\varepsilon,\eta}$. It remains to show that Γ is a contraction on $\tilde{\mathcal{C}}_{\varepsilon,\eta}$. To this end, let $u, v \in \tilde{\mathcal{C}}_{\varepsilon,\eta}$. Using (A.3) we obtain

$$\sup_{0\le\tau\le\alpha+\varepsilon} |(\Gamma u)(\tau) - (\Gamma v)(\tau)| \le \lambda\sqrt{\varepsilon}\left(\int_\alpha^{\alpha+\varepsilon} |u-v|^2\right)^{1/2} \le \varepsilon\lambda \sup_{0\le\tau\le\alpha+\varepsilon} |u(\tau) - v(\tau)|.$$

By (A.6) we have that $\varepsilon\lambda < 1$, showing that Γ is a contraction on $\tilde{\mathcal{C}}_{\varepsilon,\eta}$. ∎

Proof of Lemma 3.2. We proceed in several steps.

Step 1. Existence and uniqueness on a small interval.

An application of Lemma A.1 with $\alpha = 0$ shows that there exists an $\varepsilon > 0$ such that (3.6) has a unique solution on the interval $[0, \varepsilon)$.

Step 2. Extended uniqueness.

Let u_i be a solution of (3.6) on the interval $[0, a_i)$, $i = 1, 2$. We claim that $u_1(t) = u_2(t)$ for all $t \in [0, a)$, where $a = \min(a_1, a_2)$. Seeking a contradiction, assume that there exists $t \in (0, a)$ such that $u_1(t) \neq u_2(t)$. Defining

$$t^* = \inf\{t \in (0, a) \,|\, u_1(t) \neq u_2(t)\}\,,$$

it follows that $t^* > 0$ (by Step 1), $t^* < a$ (by assumption), and $u_1(t^*) = u_2(t^*)$ (by continuity of u_1 and u_2). Clearly, the initial-value problem

$$\begin{aligned} \dot{u}(t) &= k[r - (\mathbf{\Psi}_\infty x_0)(t) - (\mathbf{F}_\infty \phi(u))(t)]\,, \quad t \geq t^*\,, \\ u(t) &= u_1(t)\,, \quad t \in [0, t^*]\,, \end{aligned}$$

is solved by u_1 and u_2. This implies (by Lemma A.1) that there exists an $\varepsilon > 0$ such that $u_1(t) = u_2(t)$ for all $t \in [0, t^* + \varepsilon)$, which contradicts the definition of t^*.

Step 3. Continuation of solutions.

Let u be a solution of (3.6) on the interval $[0, a)$, $a < \infty$. In order to prove that u can be extended to a maximal solution (which satisfies (3.7) if $a_{max} < \infty$), it is sufficient to show that u can be continued to the right (beyond a) if u is bounded on $[0, a)$. Now $u(t) = (\Gamma u)(t)$ for all $t \in [0, a)$, where Γ is the operator defined in the proof of Lemma A.1 with $\alpha = 0$. It is clear that $\lim_{t \to a-} (\Gamma u)(t) = \gamma$ exists and is finite. Consequently, $\lim_{t \to a-} u(t) = \gamma$, and hence setting $u(a) = \gamma$ makes u into a continuous function on $[0, a]$. Finally, Lemma A.1 shows that the initial value problem

$$\begin{aligned} \dot{v} &= k[r - \mathbf{\Psi}_\infty x_0 - \mathbf{F}_\infty \phi(v)]\,, \quad t \geq a\,, \\ v(t) &= u(t)\,, \quad t \in [0, a]\,, \end{aligned}$$

has a unique solution u^* on $[0, a + \varepsilon)$ for some $\varepsilon > 0$. By the causality of the map $\mathbf{F}_\infty \phi$, the function u^* is a solution of (3.6) on $[0, a + \varepsilon)$, i.e., u^* is a continuation of u.

Step 4. Global existence if ϕ is globally Lipschitz.

Assume that ϕ is globally Lipschitz. Seeking a contradiction suppose that $a_{max} < \infty$. Let u be the solution of (3.6) defined on $[0, a_{max})$. Multiplying (3.6) by u and estimating we obtain that, for all $\tau \in [0, a_{max})$,

$$u(\tau)\dot{u}(\tau) \leq k[r^2 + (\mathbf{\Psi}_\infty x_0)^2(\tau) + u^2(\tau) + |(\mathbf{F}_\infty \phi(u))(\tau) u(\tau)|\,]\,. \tag{A.7}$$

Integrating (A.7) from 0 to t and combining the estimate

$$\int_0^t |(\mathbf{F}_\infty \phi(u))u| \,\leq \int_0^t |\mathbf{F}_\infty(\phi(u) - \phi(0))|\,|u| \,+ \frac{1}{2}\left(\int_0^t (\mathbf{F}_\infty \phi(0))^2 \,+ \int_0^t u^2\right),$$

the Cauchy–Schwarz inequality, and the global Lipschitz property of ϕ, it can be readily shown that there exists positive constants α and β such that, for all $t \in [0, a_{max})$,

$$u^2(t) \leq \alpha + \beta \int_0^t u^2(\tau)\, d\tau\,.$$

An application of Gronwall's lemma then shows that $u^2(t) \leq \alpha e^{\beta t}$ for all $t \in [0, a_{max})$. Hence u is bounded on $[0, a_{max})$, which by Step 3 is in contradiction to the maximality of a_{max}. ∎

Proof of Lemma 3.4. Since $0 < 2\kappa < K$, it follows that there exists $\varepsilon > 0$ such that

$$1 + 2\kappa \mathrm{Re}\, \frac{\mathbf{G}(s)}{s} \geq \varepsilon \qquad \text{for all } s \in \mathbb{C}_0 \,.$$

Hence

$$1 + 2\kappa \mathrm{Re}\, \frac{\mathbf{G}(i\omega)}{i\omega} \geq \varepsilon \qquad \text{for all } \omega \in \mathbb{R},\, \omega \neq 0 \,,$$

and thus

$$1 + 2\kappa \mathrm{Re}\mathbf{H}(i\omega) \geq \varepsilon \qquad \text{for all } \omega \in \mathbb{R} \,. \tag{A.8}$$

By considering

$$e^{-(1+2\kappa \mathrm{Re}\mathbf{H}(s))} = \left| e^{-(1+2\kappa \mathbf{H}(s))} \right| ,$$

applying the maximum modulus theorem, and using the fact that $\mathbf{H}(s) \to 0$ as $|s| \to \infty$ in $\mathbb{C}_0$, it follows from (A.8) that

$$1 + 2\kappa \mathrm{Re}\mathbf{H}(s) \geq \varepsilon \qquad \text{for all } s \in \mathbb{C}_0 \,.$$

Therefore, for all $s \in \mathbb{C}_0$,

$$\varepsilon + \kappa^2 \mathbf{H}(s)\bar{\mathbf{H}}(s) \leq (1 + \kappa \mathbf{H}(s))(1 + \kappa \bar{\mathbf{H}}(s)) \,.$$

Consequently, for all $s \in \mathbb{C}_0$,

$$\mathbf{H}(s)(1 + \kappa \mathbf{H}(s))^{-1} \bar{\mathbf{H}}(s)(1 + \kappa \bar{\mathbf{H}}(s))^{-1} < \frac{1}{\kappa^2} \,,$$

yielding (3.11).

By using the identity $s(sI - A)^{-1} = A(sI - A)^{-1} + I$, we easily obtain

$$\mathbf{H}(s) = \frac{1}{s} \left(\mathbf{G}(s) - \mathbf{G}(0) \right) = C_L (sI - A)^{-1} A^{-1} B \,.$$

Consider the state-space system given by the triple $(A, A^{-1}B, C_L)$. For any $T > 0$, the input-to-state map of this system maps $L^2(0,T)$ boundedly into X_1. Consequently, the triple $(A, A^{-1}B, C_L)$ defines a Pritchard–Salamon system with respect to the spaces X_1 and X; see Curtain et al. [4] or Pritchard and Townley [31]. Now, (3.11) means in particular that the closed-loop system obtained from $\mathbf{H}$ by negative output feedback with gain κ is input-output stable. By the equivalence of input-output and exponential stability (see [4] or [32]), we may conclude that the semigroup generated by A_κ, with $0 < 2\kappa < K$, is exponentially stable. Moreover, combining Theorem 2.4 in Pritchard and Townley [30] (or, alternatively, Theorem 1 in Logemann [12]) and (3.11), it follows that the structured complex stability radius of A_κ with respect to the weightings $A^{-1}B$ and C_L is greater than κ. Therefore, an application of Proposition 1.5 in [31] shows that the Riccati equation (3.12) has a self-adjoint positive-semidefinite solution $P \in \mathcal{B}(X)$ such that (3.12) holds for all $x_1, x_2 \in \mathrm{dom}(A_\kappa)$. □

Proof of Lemma 3.5. It is clear that $\phi^\diamond \in L^\infty_{loc}(-\infty, \infty)$ if ϕ is locally Lipschitz and that $\phi^\diamond \in L^\infty(-\infty, \infty)$ if ϕ is globally Lipschitz. Moreover, as the limsup of a

sequence of Borel functions, $\phi^\diamond$ is a Borel function. Consequently, $\phi^\diamond \circ u$ is Lebesgue measurable for all Lebesgue measurable functions u. Let u be absolutely continuous. Setting $v = \phi \circ u$, it follows from the Lipschitz continuity of ϕ and the absolute continuity of u that v is absolutely continuous. If $t \in \mathbb{R}$ is such that u is differentiable at t, then we have

$$v(t+h) - v(t) = \phi(u(t) + h\dot{u}(t)) - \phi(u(t)) + \phi(u(t+h)) - \phi(u(t) + h\dot{u}(t))\,. \tag{A.9}$$

Moreover, by Lipschitz continuity of ϕ, there exists a constant $L > 0$ such that, for all sufficiently small $|h|$,

$$\left|\frac{1}{h}\left[\phi(u(t+h)) - \phi(u(t) + h\dot{u}(t))\right]\right| \le L\left|\frac{1}{h}[u(t+h) - u(t)] - \dot{u}(t)\right|. \tag{A.10}$$

Let $\mathcal{D} \subset \mathbb{R}$ be the set of all points t such that both u and v are differentiable at t. Then $\mathcal{D}$ is of full measure, and combining (A.9) and (A.10) yields

$$\lim_{h\to 0}\frac{1}{h}[v(t+h) - v(t)] = \lim_{h\to 0}\frac{1}{h}[\phi(u(t) + h\dot{u}(t)) - \phi(u(t))] \quad \text{for all } t \in \mathcal{D}.$$

Therefore, for every $t \in \mathcal{D}$,

$$\dot{v}(t) = 0 \qquad \text{if } \dot{u}(t) = 0\,, \tag{A.11}$$

$$\dot{v}(t) = \lim_{h\to 0}\frac{\phi(u(t) + h\dot{u}(t)) - \phi(u(t))}{h\dot{u}(t)}\,\dot{u}(t)$$

$$= \phi'(u(t))\dot{u}(t) \qquad \text{if } \dot{u}(t) \neq 0\,. \tag{A.12}$$

In particular, if $t \in \mathcal{D}_0 := \{t \in \mathcal{D} \,|\, \dot{u}(t) \neq 0\}$, then ϕ is differentiable at $u(t)$. For $t \in \mathcal{D}_0$ we have, of course, $\phi^\diamond(u(t)) = \phi'(u(t))$, and thus it follows from (A.11) and (A.12) that

$$\dot{v}(t) = \phi^\diamond(u(t))\dot{u}(t) \qquad \text{for a.e. } t \in [0,\infty)\,. \qquad \square$$

REFERENCES

[1] V. Barbu, *Analysis and Control of Nonlinear Infinite-Dimensional Systems*, Academic Press, Boston, 1993.

[2] P.A. Cook, *Controllers with universal tracking properties,* in Proc. Internat. IMA Conf. on Control: Modelling, Computation, Information, Manchester, 1992.

[3] G.W.M. Coppus, S.L. Sha, and R.K. Wood, *Robust multivariable control of a binary distillation column,* IEE Proceedings, Pt. D, 130 (1983), pp. 201–208.

[4] R.F. Curtain, H. Logemann, S. Townley, and H. Zwart, *Well-posedness, stabilizability and admissibility for Pritchard-Salamon systems,* Math. Systems, Estimation and Control, 7 (1997), pp. 439–476.

[5] E.J. Davison, *Multivariable tuning regulators: The feedforward and robust control of a general servomechanism problem,* IEEE Trans. Automat. Control, 21 (1976), pp. 35–47.

[6] C.A. Desoer and C.-A. Lin, *Tracking and disturbance rejection of MIMO nonlinear systems with PI controller,* IEEE Trans. Automat. Control, 30 (1985), pp. 861–867.

[7] G. Doetsch, *Einführung in Theorie und Anwendung der Laplace-Transformation*, 3. Auflage, Birkhäuser-Verlag, Basel, Switzerland, 1976.

[8] G. Gripenberg, S.-O. Londen, and O. Staffans, *Volterra Integral and Functional Equations*, Cambridge University Press, Cambridge, 1990.

[9] T.T. Jussila and H.N. Koivo, *Tuning of multivariable PI-controllers for unknown delay-differential systems,* IEEE Trans. Automat. Control, 32 (1987), pp. 364–368.

[10] H.W. Knobloch and F. Kappel, *Gewöhnliche Differentialgleichungen*, B.G. Teubner, Stuttgart, 1974.

[11] H.N. Koivo and S. Pohjolainen, *Tuning of multivariable PI-controllers for unknown systems with input delay,* Automatica, 21 (1985), pp. 81–91.
[12] H. Logemann, *Circle criteria, small-gain conditions and internal stability for infinite-dimensional systems,* Automatica, 27 (1991), pp. 677–690.
[13] H. Logemann, J. Bontsema, and D.H. Owens, *Low-gain control of distributed parameter systems with unbounded control and observation,* Control Theory Adv. Tech., 4 (1988), pp. 429–446.
[14] H. Logemann and D.H. Owens, *Low-gain control of unknown infinite-dimensional systems: A frequency-domain approach,* Dynam. Stability Systems, 4 (1989), pp. 13–29.
[15] H. Logemann, E.P. Ryan, and S. Townley, *Integral control of linear systems with actuator nonlinearities: Lower bounds for the maximal regulating gain,* IEEE Trans. Automat. Control, to appear.
[16] H. Logemann and S. Townley, *Discrete-time low-gain control of uncertain infinite-dimensional systems,* IEEE Trans. Automat. Control, 42 (1997), pp. 22–37.
[17] H. Logemann and S. Townley, *Low-gain control of uncertain regular linear systems,* SIAM J. Control Optim., 35 (1997), pp. 78–116.
[18] H. Logemann and S. Townley, *Adaptive integral control of time-delay systems,* IEE Proc. Control Theory Appl., 144 (1997), pp. 531–536.
[19] J. Lunze, *Experimentelle Erprobung einer Einstellregel für PI-Mehrgrößenregler bei der Herstellung von Ammoniumnitrat-Harnstoff-Lösung,* Messen Steuern Regeln, 30 (1987), pp. 2–6.
[20] J. Lunze, *Robust Multivariable Feedback Control*, Prentice–Hall, London, 1988.
[21] D.E. Miller, *Private communication,* 1996.
[22] D.E. Miller and E.J. Davison, *An adaptive tracking problem with a control input constraint,* Automatica, 29 (1993), pp. 877–887.
[23] D.E. Miller and E.J. Davison, *The self-tuning robust servomechanism problem,* IEEE Trans. Automat. Control, 34 (1989), pp. 511–523.
[24] M. Morari, *Robust stability of systems with integral control,* IEEE Trans. Automat. Control, 30 (1985), pp. 574–577.
[25] R. Pallu de la Barrière, *Optimal Control Theory*, Dover Publications, New York, 1980.
[26] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.
[27] S. Pohjolainen, *Robust controller for systems with exponentially stable strongly continuous semigroups,* J. Math. Anal. Appl., 111 (1985), pp. 622–636.
[28] S. Pohjolainen, *Robust multivariable PI-controllers for infinite-dimensional systems,* IEEE Trans. Automat. Control, 27 (1982), pp. 17–30.
[29] S. Pohjolainen and I. Lätti, *Robust controller for boundary control systems,* Internat. J. Control, 38 (1983), pp. 1189–1197.
[30] A.J. Pritchard and S. Townley, *Robustness of linear systems,* J. Differential Equations, 77 (1989), pp. 254–286.
[31] A.J. Pritchard and S. Townley, *Robustness optimization for uncertain infinite-dimensional systems with unbounded inputs,* IMA J. Math. Control Inform., 8 (1991), pp. 121–133.
[32] R. Rebarber, *Conditions for the equivalence of internal and external stability for distributed parameter systems,* IEEE Trans. Automat. Control, 38 (1993), pp. 994–998.
[33] D. Salamon, *Realization theory in Hilbert space,* Math. Systems Theory, 21 (1989), pp. 147–164.
[34] S. Townley and M. Kamstra, *Integral control with saturating input,* in Proceedings of UKACC Control '96, pp. 805–808.
[35] G. Weiss, *Two conjectures on the admissibility of control operators,* in Control and Estimation of Distributed Parameter Systems, F. Kappel, K. Kunisch, and W. Schappacher, eds., Birkhäuser-Verlag, Basel, 1991, pp. 367–378.
[36] G. Weiss, *Transfer functions of regular linear systems, part* I: *Characterization of regularity,* Trans. Amer. Math. Soc., 342 (1994), pp. 827–854.
[37] G. Weiss, *Admissibility of unbounded control operators,* SIAM J. Control Optim., 27 (1989), pp. 527–545.
[38] G. Weiss, *Admissible observation operators for linear semigroups,* Israel J. Math., 65 (1989), pp. 17–43.
[39] G. Weiss, *The representation of regular linear systems on Hilbert spaces,* in Distributed Parameter Systems, F. Kappel, K. Kunisch, and W. Schappacher, eds., Birkhäuser-Verlag, Basel, 1989 pp. 401–416.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 6, pp. 1962–1986, November 1998

007

ON THE STABILIZATION OF A FLEXIBLE BEAM WITH A TIP MASS*

FRANCIS CONRAD† AND ÖMER MORGÜL‡

Abstract. We study the stability of a flexible beam that is clamped at one end and free at the other; a mass is also attached to the free end of the beam. To stabilize this system we apply a boundary control force at the free end of the beam. We prove that the closed-loop system is well-posed and is exponentially stable. We then analyze the spectrum of the system for a special case and prove that the spectrum determines the exponential decay rate for the considered case.

Key words. flexible structures, infinite dimensional systems, boundary control, stability, semigroup theory

AMS subject classifications. 93C20, 93D15, 35B35, 35P10

PII. S0363012996302366

1. Introduction. In this paper, we study the stability of a flexible beam that is clamped at one end and is free at the other end; a mass is also attached to the free end. The equations of motion for this system are given by

$$u_{tt} + u_{xxxx} = 0, \qquad 0 < x < 1, \qquad t \geq 0, \tag{1.1}$$

$$u(0,t) = u_x(0,t) = u_{xx}(1,t) = 0, \qquad t \geq 0, \tag{1.2}$$

$$-u_{xxx}(1,t) + mu_{tt}(1,t) = w(t), \qquad t \geq 0, \tag{1.3}$$

where $m > 0$ is the tip mass and $w(t)$ is the boundary control force applied at the free end of the beam; a subscript letter denotes the partial derivation with respect to that variable. For simplicity, and without loss of generality, the length of the beam, the mass per unit length, and the flexural rigidity of the beam are chosen to be unity. Our problem is to find a feedback control law for $w(t)$ so that the solutions of the resulting closed-loop system decay uniformly to zero. This can be achieved with a highly unbounded feedback law; see (2.1).

The model given by (1.1)–(1.3) is a variant of the SCOLE model in the sense that one has neglected the moment of inertia at $x = 1$, which has been studied in the past; see, e.g., [1], [9], [14], [15]. It is known that for such types of models the feedback law

$$w(t) = -\alpha u_t(1,t), \qquad \alpha > 0, \qquad t \geq 0, \tag{1.4}$$

is sufficient for strong (i.e., *asymptotic*) stability but not sufficient for uniform (i.e., *exponential*) stability; see [9], where arbitrarily slow decay is proven by using asymptotic estimates of the eigenvalues. In fact, as shown in [14], the control law given by

*Received by the editors April 22, 1996; accepted for publication (in revised form) July 21, 1997; published electronically August 3, 1998. This work was supported by INRIA Lorraine Project Numath.

http://www.siam.org/journals/sicon/36-6/30236.html

†Département de Mathématiques, Université Nancy 1, U.M.R CNRS 9973, B.P. 239 54506, Vandœuvre-lés Nancy, France (Francis.Conrad@iecn.u-nancy.fr).

‡Department of Electrical and Electronics Engineering, Bilkent University, 06533, Bilkent, Ankara, Turkey (morgul@bilkent.edu.tr).

(1.4) may be considered as a compact perturbation of the uncontrolled system. It is well known that such compact perturbations are not sufficient to provide uniform stabilization; see [6], [17], [20]. Hence, to obtain uniform stability one has to choose "stronger" feedback terms, such as u_{xxxt} (see [13], [14]), where the lack of uniform stability for the SCOLE model with usual feedback laws (e.g., velocity feedback; see (1.4)) was proven by using the compactness argument, and uniform decay of the energy was obtained by means of higher-order feedback for rather smooth initial data. Also in [15], decay estimates for a flexible cable with a tip mass were given. Let us mention that these papers study the asymptotic or uniform decay for hybrid systems by using energy multipliers; thus the decay is qualitative, and one cannot conclude on the optimality of the decay rate. In [3] a flexible beam with rate control on the bending moment was considered, the uniform decay was proven by using the estimates of the resolvent operator on the imaginary axis, and a careful analysis of the eigenvalues and eigenfunctions was given (similar to the one given in [12] but for a harder problem). In [1] a three-dimensional model for the SCOLE system, including the moment of inertia at $x = 1$, is considered, and then a feedback law similar to (1.4) and another feedback law based on optimal control techniques are studied. As stated above, these results also show the asymptotical or uniform decay of energy for the system considered, but do not give the optimality of the decay rate.

In this paper we investigate the uniform stability of the system given by (1.1)–(1.3). The paper is organized as follows. In the next section we prove the well-posedness and the uniform stability of (1.1)–(1.3) with a proper choice for $w(t)$ for a norm weaker than the one used in [14] by introducing a specific change of variables. Then we study the spectrum of the system for a particular case and prove that for the considered case the spectrum determines the exponential decay rate for almost all $\alpha > 0$. We also show that in case $m = 0$ in (1.1)–(1.4) (i.e., the case of the cantilevered beam with a boundary force control), for almost all $\alpha > 0$, the spectrum determines the exponential growth rate (see Appendix). Finally we give some concluding remarks.

2. Stability results. For the system given by (1.1)–(1.3) we propose the following linear feedback control law for $w(t)$:

$$w(t) = -\alpha u_t(1,t) + \beta u_{xxxt}(1,t), \qquad t \geq 0, \tag{2.1}$$

where α and β are positive constants.

We define the auxiliary function η as

$$\eta(t) = -u_{xxx}(1,t) + \frac{m}{\beta} u_t(1,t), \qquad t \geq 0. \tag{2.2}$$

Upon substituting (2.1) and (2.2) into (1.3), the latter becomes

$$\beta\dot{\eta}(t) + \eta(t) + \left(\alpha - \frac{m}{\beta}\right) u_t(1,t) = 0, \qquad t \geq 0, \tag{2.3}$$

where a dot represents the time derivative. We note that a similar control law has been applied to the stabilization of a cable with a tip mass, see [10].

Let us introduce the following spaces:

$$\mathcal{V} = \{u : [0,1] \to \mathbf{R} | u \in H^2(0,1), \quad u(0) = u_x(0) = 0\}, \tag{2.4}$$

$$\mathcal{H} = \{(u\ v\ \eta)^T | u \in \mathcal{V},\ v \in L^2(0,1),\ \eta \in \mathbf{R}\}, \tag{2.5}$$

where the superscript T stands for the transpose; the spaces $L^2(0,1)$ and $H^k(0,1)$ are defined as

$$L^2(0,1) = \left\{ y : [0,1] \to \mathbf{R} | \int_0^1 y^2 dx < \infty \right\}, \tag{2.6}$$

$$H^k(0,1) = \{ y : [0,1] \to \mathbf{R} | y,\ y^{(1)}, \ldots, y^{(k)} \in L^2(0,1) \}. \tag{2.7}$$

In $\mathcal{H}$ we define the following inner-product:

$$< y, \tilde{y} >_{\mathcal{H}} = \int_0^1 (u_{xx}\tilde{u}_{xx} + v\tilde{v})dx + \ K\eta\tilde{\eta}, \tag{2.8}$$

where $y = (u\ v\ \eta)^T \in \mathcal{H}$, $\tilde{y} = (\tilde{u}\ \tilde{v}\ \tilde{\eta})^T \in \mathcal{H}$, $K > 0$ is chosen as

$$K = \frac{\beta^2}{m + \alpha\beta}. \tag{2.9}$$

The reason for this choice will become clear later. Next we define the unbounded operator $A : D(A) \subset \mathcal{H} \to \mathcal{H}$ as follows:

$$A \begin{pmatrix} u \\ v \\ \eta \end{pmatrix} = \begin{pmatrix} v \\ -u_{xxxx} \\ -\frac{1}{\beta}\eta - \frac{1}{\beta}(\alpha - \frac{m}{\beta})v(1) \end{pmatrix}, \tag{2.10}$$

where the domain $D(A)$ of the operator A is defined as

$$D(A) = \Big\{ (u\ v\ \eta\ )^T | u \in H^4(0,1) \cap \mathcal{V}, v \in \mathcal{V},\ \eta \in \mathbf{R}, \\ u_{xx}(1) = 0, \quad \eta = -u_{xxx}(1) + \frac{m}{\beta}v(1) \Big\}. \tag{2.11}$$

With the previous notation, (1.1)–(1.2) and (2.3) can be written formally as

$$\dot{y} = Ay, \qquad y(0) \in \mathcal{H}, \tag{2.12}$$

where $y = (u\ v\ \eta)^T$, η is defined by (2.2), and $v = u_t$.

THEOREM 2.1. *The operator A, defined by* (2.10) *and* (2.11)*, generates a C_0 semigroup of contractions on $\mathcal{H}$. (For the terminology on the semigroup theory, the reader is referred to* [11].)

Proof. We apply the Lumer–Phillips theorem; see, e.g., [11, p. 14]. First, for any $y = (u\ v\ \eta)^T \in D(A)$,

$$\begin{aligned} \langle Ay,\ y \rangle_{\mathcal{H}} &= \int_0^1 (u_{xx}v_{xx} - vu_{xxxx})dx - \frac{K}{\beta}\eta\left(\eta + \left(\alpha - \frac{m}{\beta}\right)v(1)\right) \\ &= -\frac{K}{\beta}u_{xxx}^2(1) - \frac{Km\alpha}{\beta^2}v^2(1), \end{aligned} \tag{2.13}$$

where to derive the last equation we integrated by parts twice and used (1.2), (2.2), and (2.9). Note that due to the particular choice of K given by (2.9), the term multiplying $v(1)u_{xxx}(1)$ in (2.13) vanishes. It follows from (2.13) that the operator A is dissipative.

Next we show that the range of the operator $\lambda I - A : D(A) \subset \mathcal{H} \to \mathcal{H}$ is onto for $\lambda > 0$; that is, for any given $z = (f\ g\ h)^T \in \mathcal{H}$, we have to find $y = (u\ v\ \eta)^T \in D(A)$ so that

$$(\lambda I - A)y = z, \tag{2.14}$$

which is equivalent to the following set of equations:

$$\lambda u - v = f, \tag{2.15}$$

$$\lambda v + u_{xxxx} = g, \tag{2.16}$$

$$\left(\lambda + \frac{1}{\beta}\right)\eta + \frac{1}{\beta}\left(\alpha - \frac{m}{\beta}\right)v(1) = h. \tag{2.17}$$

Upon substituting (2.15) into (2.16), the latter becomes

$$\lambda^2 u + u_{xxxx} = \lambda f + g. \tag{2.18}$$

By using (2.15) and (2.2) in (2.17), the latter becomes

$$-\left(\lambda + \frac{1}{\beta}\right)u_{xxx}(1) + \frac{\lambda(\alpha + m\lambda)}{\beta}u(1) = h + \frac{m\lambda + \alpha}{\beta}f(1)\ . \tag{2.19}$$

Therefore to prove that $\lambda I - A$ is onto, we have to prove the existence of a solution for the following set of equations:

$$\lambda^2 u + u_{xxxx} = f^*, \tag{2.20}$$

$$u(0) = u_x(0) = u_{xx}(1) = 0, \tag{2.21}$$

$$-u_{xxx}(1) + cu(1) = h^*, \tag{2.22}$$

where f^*, h^*, and c are given by

$$f^* = \lambda f + g \in L^2(0,1), \qquad h^* = \frac{\beta}{\lambda\beta + 1}h + \frac{m\lambda + \alpha}{\lambda\beta + 1}f(1) \in \mathbf{R}, \tag{2.23}$$
$$c = \frac{\lambda(m\lambda + \alpha)}{\lambda\beta + 1} > 0.$$

The existence, as well as the uniqueness and continuous dependence, of a solution of (2.20)–(2.23) with respect to (f^*, h^*) can be considered as standard. One way to prove it is to use the weak formulation of (2.20)–(2.23), which is

$$\int_0^1 u_{xx}\varphi_{xx}dx + \lambda^2\int_0^1 u\varphi dx + cu(1)\varphi(1) = \int_0^1 f^*\varphi dx + h^*\varphi(1), \qquad u \in \mathcal{V}, \forall\varphi \in \mathcal{V}. \tag{2.24}$$

Since $c > 0$, the left-hand side of (2.24) is a coercive bilinear form of φ and u. Then the existence and uniqueness of a $u \in \mathcal{V}$ satisfying (2.24) follow from the well-known

Lax–Milgram theorem; see e.g. [19, p. 26]. By standard regularity $u \in H^4(0,1)$ and by using particular φ, one recovers the boundary conditions in u. Then v given by (2.15) and η given by (2.17) are unique and $(u\ v\ \eta)^T \in D(A)$. This shows that the operator $\lambda I - A$ is onto for $\lambda > 0$, and the proof of the theorem now follows from the Lumer–Phillips theorem. □

Remark 1. It follows from Theorum 2.1 that for $(u_0\ v_0\ \eta_0)^T \in D(A)$, the problem (2.12) has a strong solution $(u(t)\ v(t)\ \eta(t))^T \in C^1(\mathbf{R}_+, \mathcal{V} \times L^2(0,1) \times \mathbf{R}) \cap C^0(\mathbf{R}_+, D(A))$. Thus $\eta(t) = -u_{xxx}(1,t) + \frac{m}{\beta} u_t(1,t)$ is differentiable, but $u_{xxx}(1,t)$ and $u_t(1,t)$ are not guaranteed to be separably differentiable. This will be the case if $(u_0\ v_0\ \eta_0)^T \in D(A^2)$.

Next we prove that the semigroup generated by the operator A decays exponentially to zero.

THEOREM 2.2. *Let $T(t)$ be the C_0 semigroup of contractions generated by the operator A on $\mathcal{H}$. Then there exist positive constants M and δ such that the following holds:*

$$\|T(t)\|_{\mathcal{L}(\mathcal{H})} \le M e^{-\delta t}, \qquad t \ge 0, \tag{2.25}$$

where the norm used is the norm induced by the inner-product given by (2.8).

Proof. We first define the following function:

$$V(t) = tE(t) + \int_0^1 x\ u_t(x,t)\ u_x(x,t)dx, \tag{2.26}$$

where the "energy" $E(t)$ is given by

$$E(t) = \frac{1}{2}\|z(t)\|^2_{\mathcal{H}} = \frac{1}{2}\int_0^1 (u_t^2(x,t) + u_{xx}^2(x,t))dx + \frac{K}{2}\eta^2(t), \tag{2.27}$$

$z(t) = (u(\cdot,t)\ u_t(\cdot,t)\ \eta(t))^T \in \mathcal{H}$ is the solution of (2.12), and K is given by (2.9). Assume that $z(0) \in D(A)$; then by semigroup property we have $z(t) = T(t)z(0) \in D(A)\ \forall t \ge 0$. Hence, in view of (2.13), we have

$$\dot{E}(t) = < Az(t),\ z(t) >_{\mathcal{H}} = -\frac{K}{\beta} u_{xxx}^2(1,t) - \frac{Km\alpha}{\beta^2} u_t^2(1,t)\ \le\ 0. \tag{2.28}$$

Next, by using Cauchy–Schwarz and Poincaré's inequalities, it can easily be shown that the following holds for a positive constant C:

$$(t - C)E(t) \le V(t) \le (t + C)E(t), \qquad t \ge 0. \tag{2.29}$$

(One can take $C = 1$ or even $C = 1/\sqrt{2}$.) By differentiating (2.26) with respect to time and by using (1.1), we obtain

$$\begin{aligned} \dot{V}(t) = E(t) + t\dot{E}(t) &+ \int_0^1 x\ u_{xt}(x,t)\ u_t(x,t)dx \\ &- \int_0^1 x\ u_x(x,t)\ u_{xxxx}(x,t)dx. \end{aligned} \tag{2.30}$$

Using integration by parts and (1.2), we obtain

$$\int_0^1 x\ u_x(x,t)\ u_{xxxx}(x,t)dx = u_x(1,t)\ u_{xxx}(1,t) + \frac{3}{2}\int_0^1 u_{xx}^2(x,t)dx, \tag{2.31}$$

$$\int_0^1 x\ u_{xt}(x,t)\ u_t(x,t)dx = \frac{1}{2}u_t^2(1,t) - \frac{1}{2}\int_0^1 u_t^2(x,t)\ dx. \tag{2.32}$$

By using (1.2), we obtain

$$u_x^2(1,t) \le \int_0^1 u_{xx}^2(x,t)dx. \tag{2.33}$$

We also have the following inequalities:

$$u_x(1,t)u_{xxx}(1,t) \le \delta_1 u_x^2(1,t) + \frac{1}{\delta_1}u_{xxx}^2(1,t), \tag{2.34}$$

$$\eta^2(t) \le 2u_{xxx}^2(1,t) + 2\frac{m^2}{\beta^2}u_t^2(1,t), \tag{2.35}$$

where $\delta_1 > 0$ is an arbitrary constant. By using (2.28) and (2.31)–(2.35) in (2.30), we obtain

$$\begin{aligned}\dot{V}(t)\ \le -(1-\delta_1)\int_0^1 u_{xx}^2(x,t)dx - \left[\frac{K}{\beta}t - K - \frac{1}{\delta_1}\right]u_{xxx}^2(1,t)\\ - \left[\frac{Km\alpha}{\beta^2}t - \frac{1}{2} - \frac{Km^2}{\beta^2}\right]u_t^2(1,t).\end{aligned} \tag{2.36}$$

By choosing $\delta_1 < 1$, the integral term in (2.36) is negative. Hence there exists a constant $T \ge 0$, which depends only on the constants $K,\ m,\ \alpha,\ \beta,$ and δ_1 such that the following holds:

$$\dot{V}(t) \le 0, \qquad t \ge T. \tag{2.37}$$

Now, from (2.29) and (2.37) we obtain the following:

$$E(t) \le \frac{T+C}{t-C}E(0), \qquad t > \max\{C,T\}. \tag{2.38}$$

Note that $E(t) = \frac{1}{2}\|z(t)\|_{\mathcal{H}}^2 = \frac{1}{2}\|T(t)z(0)\|_{\mathcal{H}}^2$; hence from (2.38) it follows that $\|T(t)\|_{\mathcal{L}(\mathcal{H})} < 1$ for $t > 0$ sufficiently large. Hence it follows from the semigroup property that the exponential decay, i.e., (2.25), holds. □

Remark 2. From (2.25) and (2.27) we conclude that both the "energy" associated with the flexible beam (i.e., the integral terms in (2.27)) and η defined by (2.2) decay exponentially to zero. However, we cannot conclude that the same holds separately for the tip mass velocity $u_t(1,t)$ and $u_{xxx}(1,t)$. If we assume that $z(0) \in D(A)$, then we also have for the graph norm

$$\|T(t)z(0)\|_{D(A)} \le Me^{-\delta t}\|z(0)\|_{D(A)}.$$

In this case, $T(t)z(0)$ decays exponentially to zero in $H^4(0,1) \times H^2(0,1) \times \mathbf{R}$. Since, similar to (2.33), we have

$$u_t^2(1,t) \le \int_0^1 u_{xt}^2(x,t)dx,$$

we obtain exponential decay of the tip mass velocity $u_t(1,t)$ and $u_{xxx}(1,t)$ uniformly for all smooth initial data $z(0) \in D(A)$ bounded in $D(A)$ for the graph norm.

3. Analysis of the spectrum. In this section we calculate the spectrum of the operator A for a special case and claim that the spectrum determines the optimal exponential decay rate given by (2.25) for the considered case. Our method is to prove that a system of eigenvectors of A forms a Riesz basis in $\mathcal{H}$. To obtain this result we compare the flexible beam with a tip mass to the flexible beam without a tip mass for the spectral properties. Here we have to work in the complexified Hilbert spaces $\mathcal{V}$, $L^2(0,1)$ and $\mathcal{H}$. For convenience we do not change the notation for these spaces.

Let $\lambda \in \mathbf{C}$ be an eigenvalue of A and let $y = (u\ v\ \eta)^T \in D(A)$ be a corresponding eigenvector. To find y we have to solve (2.14), and hence (2.15)–(2.17) for $z = (f\ g\ h)^T = 0$. Using (2.15) in (2.16), the latter, together with the boundary conditions, becomes

$$\lambda^2 u + u_{xxxx} = 0, \tag{3.1}$$

$$u(0) = u_x(0) = u_{xx}(1) = 0. \tag{3.2}$$

Similarly, by using (2.15) and (2.2) in (2.17), the latter becomes (cf. (2.19))

$$-\left(\lambda + \frac{1}{\beta}\right) u_{xxx}(1) + \frac{\lambda(\alpha + m\lambda)}{\beta} u(1) = 0. \tag{3.3}$$

By solving (3.1)–(3.3) one can find u. Then v and η can be found from (2.15) and (2.2), respectively.

The solutions of (3.1), together with the first two boundary conditions in (3.2), can be found as (for $0 \le x \le 1$)

$$u(x) = c_1(\cosh \tau x - \cos \tau x) + c_2(\sinh \tau x - \sin \tau x), \qquad \lambda = i\tau^2, \tag{3.4}$$

where c_1 and c_2 are constants to be determined by the remaining boundary conditions, cosh and sinh are the hyperbolic cosine and sine functions, respectively, and τ is one square root of λ/i. The choice of the sign is not important since by using $-\tau$ instead of τ nothing changes except the signs of the eigenvectors associated with λ.

By using (3.4) in (3.3) and the last boundary condition in (3.2), we obtain

$$\tau^2(\cosh \tau + \cos \tau)c_1 + \tau^2(\sinh \tau + \sin \tau)c_2 = 0, \tag{3.5}$$

$$\begin{aligned}[-q_1(\lambda)\tau^3(\sinh \tau - \sin \tau) + q_2(\lambda)(\cosh \tau - \cos \tau)]c_1 \\ +[-q_1(\lambda)\tau^3(\cosh \tau + \cos \tau) + q_2(\lambda)(\sinh \tau - \sin \tau)]c_2 = 0,\end{aligned} \tag{3.6}$$

where

$$q_1(\lambda) = \lambda + \frac{1}{\beta}, \qquad q_2(\lambda) = \frac{\lambda(m\lambda + \alpha)}{\beta}.$$

By writing (3.5)–(3.6) in matrix form and taking the determinant of the coefficient matrix, it can easily be shown that (3.5)–(3.6) admit nontrivial solutions for c_1 and c_2 if and only if λ (hence τ) satisfies the following equation with λ necessarily nonzero:

$$-\tau^3 q_1(\lambda)(1 + \cosh \tau \cos \tau) + q_2(\lambda)(\sinh \tau \cos \tau - \cosh \tau \sin \tau) = 0. \tag{3.7}$$

The solutions of (3.7) give the eigenvalues of A; the corresponding eigenvectors can be found from (3.4)–(3.6), (2.15), and (2.2).

In what follows we analyze the spectrum of A for the case $\alpha = \frac{m}{\beta}$. From (2.3) or (2.10) it is clear that this choice leads to simplifications in the system (1.1)–(1.3) or (2.12), especially for the asymptotic behavior since the system is then uncoupled, except for the initial conditions. Then (3.7) can be written in the following form:

$$\left(\lambda + \frac{1}{\beta}\right)[-\tau^3(1+\cosh\tau\cos\tau) + \alpha\lambda(\sinh\tau\cos\tau - \cosh\tau\sin\tau)] = 0. \tag{3.8}$$

From (3.8) it follows that $\lambda_* = -\frac{1}{\beta}$ is an eigenvalue of A. To find the remaining eigenvalues of A let us define the function $f(\cdot)$ given by

$$f(\tau) = -\tau^3(1+\cosh\tau\cos\tau) + \alpha\lambda(\sinh\tau\cos\tau - \cosh\tau\sin\tau), \tag{3.9}$$

which is just the remaining factor of (3.8) after the division by the term $(\lambda + \frac{1}{\beta})$. Hence the remaining eigenvalues of A are precisely the (nonzero) roots of this factor:

$$-\tau^3(1+\cosh\tau\cos\tau) + \alpha\lambda(\sinh\tau\cos\tau - \cosh\tau\sin\tau) = 0. \tag{3.10}$$

It is known that (3.10) is just the characteristic equation for the system given by (1.1)–(1.4) with $m = 0$, i.e., the clamped-free (cantilevered) beam with boundary force controller at the free end; see, e.g., [12]. Moreover the eigenvectors of A corresponding to the roots of (3.10) are also related to the eigenvectors of the cantilevered beam in a simple way. For these reasons we will briefly study the spectral properties of the cantilevered beam in the following subsection.

3.1. Spectral analysis of the cantilevered beam. We consider the Euler–Bernoulli beam with boundary force control:

$$u_{tt} + u_{xxxx} = 0, \qquad 0 < x < 1, \quad t \geq 0, \tag{3.11}$$

$$u(0,t) = u_x(0,t) = u_{xx}(1,t) = 0, \quad u_{xxx}(1,t) = \alpha u_t(1,t), \quad t \geq 0, \tag{3.12}$$

where $\alpha > 0$. Note that this system is the same as (1.1)–(1.4) with $m = 0$.

We define the following spaces:

$$V = \{v \in \mathbf{H}^2(0,1); \quad v(0) = v_x(0) = 0\}, \tag{3.13}$$

$$D(B) = \{(u\ v)^T | u \in \mathbf{H}^4(0,1) \cap V, v \in V, u_{xx}(1) = 0,\ u_{xxx}(1) = \alpha v(1)\}. \tag{3.14}$$

The operator B for the problem (3.11)–(3.12) is

$$B\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} v \\ -u_{xxxx} \end{pmatrix}, \qquad (u\ v)^T \in D(B). \tag{3.15}$$

The system given by (3.11), (3.12) can be written formally as

$$\dot{z}(t) = Bz(t), \quad z(0) \in V \times \mathbf{L}^2(0,1), \tag{3.16}$$

where $z = (u(\cdot,t)\ u_t(\cdot,t))^T$, and the domain of B is given by (3.14).

We first state the following well-known result.

LEMMA 3.1. *Consider the system given by* (3.16).

i : *B generates an exponentially stable C_0 semigroup of contractions in $V \times \mathbf{L}^2(0,1)$.*

ii : *B has compact resolvent for $\lambda > 0$.*

iii : *The eigenvalues of B are countable and isolated. Moreover each eigenvalue has finite algebraic multiplicity.*

Proof. For **i** and **ii**, see [3]. Then **iii** follows from **ii**; see, e.g., [8, p. 187], [5, p. 2292]. □

Writing $z = (u\ v)^T$ and $Bz = \lambda z$, we get the following well-known characteristic equation:

$$f(\tau) = -\tau^3(1 + \cosh\tau\cos\tau) + i\alpha\tau^2(\sinh\tau\cos\tau - \cosh\tau\sin\tau) = 0, \tag{3.17}$$

where $\lambda = i\tau^2$. Note that $\lambda = 0$ is not an eigenvalue of B. Hence the roots of (3.17) are precisely the eigenvalues of B, and by Lemma 3.1, (3.17) has only countably many roots; moreover each root is isolated and has finite algebraic multiplicity. Eigenvectors corresponding to $\lambda = i\tau^2$ can be taken as $(\varphi_1\ \lambda\varphi_1)^T$, where

$$\begin{aligned}\varphi_1(\tau,x) = {} & (\ \cosh\tau + \cos\tau)(\sinh\tau x - \sin\tau x) \\ & - (\sinh\tau + \sin\tau)(\cosh\tau x - \cos\tau x).\end{aligned} \tag{3.18}$$

All eigenvalues are geometrically simple. For the algebraic multiplicity we have the following result.

LEMMA 3.2. *Consider the operator B on $V \times \mathbf{L}^2(0,1)$ given by* (3.15)*, where $D(B)$ is given by* (3.14)*. Let λ be an eigenvalue of B and set $\lambda = i\tau^2$. Then the algebraic multiplicity of λ is* 1 *if and only if $f(\tau) = 0$ and $f'(\tau) \neq 0$ (i.e., if and only if τ is a simple root of* (3.17)).

Proof. The algebraic multiplicity of λ is greater than 1 if and only if $\mathrm{Ker(B - \lambda I)^2 \backslash Ker(B - \lambda I) \neq \emptyset}$, i.e., there exists $(\psi_1\ \psi_2)^T$ which satisfies

$$(B - \lambda I)\begin{pmatrix} \psi_1 \\ \psi_2 \end{pmatrix} = \begin{pmatrix} \varphi_1 \\ \lambda\varphi_1 \end{pmatrix}, \tag{3.19}$$

which is equivalent to the following set of equations:

$$\psi_2 - \lambda\psi_1 = \varphi_1, \tag{3.20}$$

$$-\psi_{1xxxx} - \lambda\psi_2 = \lambda\varphi_1, \tag{3.21}$$

$$\psi_1(0) = \psi_{1x}(0) = \psi_{1xx}(1) = 0, \quad \psi_{1xxx}(1) = \alpha\psi_2(1). \tag{3.22}$$

By eliminating ψ_2, we obtain the following set of equations:

$$-\psi_{1xxxx} - \lambda^2\psi_1 = 2\lambda\varphi_1, \tag{3.23}$$

$$\psi_1(0) = \psi_{1x}(0) = \psi_{1xx}(1) = 0, \quad \psi_{1xxx}(1) = \alpha\lambda\psi_1(1) + \alpha\varphi_1(1). \tag{3.24}$$

The general solution of (3.23) satisfying the first three conditions of (3.24) is given for all λ by $\psi_1 = \frac{d\varphi_1}{d\lambda} + z$, where $\frac{d\varphi_1}{d\lambda} = \frac{1}{2i\tau}\frac{d\varphi_1}{d\tau}$ and z satisfies

$$z_{xxxx} + \lambda^2 z = 0, \tag{3.25}$$

$$z(0) = z_x(0) = z_{xx}(1) = 0. \tag{3.26}$$

The last condition of (3.24) becomes

$$z_{xxx}(1) - \alpha\lambda z(1) = \alpha\varphi_1(1) + \alpha\lambda\frac{d\varphi_1}{d\lambda}(1) - \left(\frac{d\varphi_1}{d\lambda}\right)_{xxx}(1) = \mu. \tag{3.27}$$

By computing μ defined in (3.27) from φ_1 and $\frac{d\varphi_1}{d\lambda}$ we get

$$\mu = \frac{d}{d\lambda}[\alpha\lambda\varphi_1(1) - \varphi_{1xxx}(1)] = 2\frac{df(\tau)}{d\lambda} = \frac{1}{i\tau}f'(\tau). \tag{3.28}$$

We conclude that the algebraic multiplicity of λ is larger than 1 if and only if (3.25)–(3.27) admit a solution z. Multiplying (3.25) by φ_1, integrating by parts, and using the boundary conditions on z and φ_1, we obtain

$$z_{xxx}(1)\varphi_1(1) - z(1)\varphi_{1xxx}(1) = [z_{xxx}(1) - \alpha\lambda z(1)]\varphi_1(1) = 0. \tag{3.29}$$

Since φ_1 is an eigenfunction of B, it could easily be shown that $\varphi_1(1) \neq 0$ (otherwise one obtains a contradiction; see, e.g., [4, p. 429]). Hence (3.25)–(3.27) admit a solution if and only if $\mu = 0$, in which case we could choose $z = \varphi_1$. Hence λ is algebraically simple if and only if $f(\tau) = 0$ and $f'(\tau) \neq 0$, i.e., if and only if λ is a simple root of (3.17). □

By Lemma 3.1, B has at most countably many and isolated eigenvalues. Let $\lambda_n = i\tau_n^2$, $n \in \mathbf{Z}$, be the roots of (3.17). The corresponding eigenvectors of B can be given as

$$F_{nr} = \begin{pmatrix} \varphi_1(\tau_n, x) \\ \lambda_n\varphi_1(\tau_n, x) \end{pmatrix}, \tag{3.30}$$

where φ_1 is given by (3.18).

THEOREM 3.3. *Consider the operator B on $V \times \mathbf{L}^2(0,1)$ given by* (3.15), *where $D(B)$ is given by* (3.14).

i. *For any $\alpha > 0$, all eigenvalues of B with sufficiently large modulus are algebraically simple.*

ii. *For almost all $\alpha > 0$, the eigenvalues of B are algebraically simple.*

iii. *If all eigenvalues are algebraically simple, then the set of eigenvectors $\{F_{nr}, n \in \mathbf{Z}\}$ is a Riesz basis for $V \times \mathbf{L}^2(0,1)$, provided that the normalization of eigenvectors is suitable.*

Proof. The proof requires detailed and lengthy calculations and is given in the appendix. In this proof we compare the set of eigenfunctions of B for $\alpha = 0$, denoted by $\{G_{nr}, n \in \mathbf{Z}\}$ with $\{F_{nr}, n \in \mathbf{Z}\}$, and show that these two sets are quadratically close. Since the former set is a Riesz basis for $V \times \mathbf{L}^2(0,1)$, we then conclude that the same is true for the latter set. □

3.2. Spectral analysis of the operator A. We now consider the operator A given by (2.10) for the case $\alpha = m/\beta$. The eigenvalues of A are given by (3.8). From (3.8) it follows that $\lambda_* = -\frac{1}{\beta}$ is an eigenvalue of A. To find the corresponding eigenfunction, we again set $\lambda_* = i\tau_*^2$ and rewrite (3.5) as ($\tau_* \neq 0$):

$$(\cosh\tau_* + \cos\tau_*)c_1 + (\sinh\tau_* + \sin\tau_*)c_2 = 0. \tag{3.31}$$

Note that since τ_* is a solution of (3.8), (3.6) is linearly dependent on (3.5) and hence will not be used to determine c_1 and c_2.

In (3.31) the coefficients c_1 and c_2 cannot be zero simultaneously. This follows easily since τ_* is not a purely imaginary number (note that $\lambda_* = -\frac{1}{\beta} = i\tau_*^2$). So the natural choice for c_1 and c_2 given by (3.5)–(3.6) is:

$$c_1 = -(\sinh \tau_* + \sin \tau_*), \tag{3.32}$$

$$c_2 = (\cosh \tau_* + \cos \tau_*). \tag{3.33}$$

Therefore an eigenfunction F_* corresponding to λ_* is

$$F_* = \begin{pmatrix} u_* \\ v_* \\ \eta_* \end{pmatrix}, \tag{3.34}$$

where

$$\begin{aligned} u_*(x) = \varphi_1(\tau_*, x) = &(\cosh \tau_* + \cos \tau_*)(\sinh \tau_* x - \sin \tau_* x) \\ &-(\sinh \tau_* + \sin \tau_*)(\cosh \tau_* x - \cos \tau_* x), \end{aligned} \tag{3.35}$$

$$v_* = \lambda_* u_*(x), \tag{3.36}$$

$$\eta_* = 2f(\tau_*), \tag{3.37}$$

where $f(\cdot)$ and φ_1 are given by (3.17) and (3.18), respectively. The remaining eigenvalues of A are precisely the (nonzero) roots of (3.10). From the preceding section it follows that these eigenvalues are the roots of (3.17), and hence the eigenvalues of B, i.e., the eigenvalues of the cantilevered beam without a tip mass. By Lemma 3.1, (3.17) admits countably many distinct roots $\lambda_n = i\tau_n^2$, $n \in \mathbf{Z}$, $\mathcal{R}e\ \{\lambda_n\} < 0$. We set

$$\begin{aligned} u_n(x) = \varphi_1(\tau_n, x) = &(\cosh \tau_n + \cos \tau_n)(\sinh \tau_n x - \sin \tau_n x) \\ &-(\sinh \tau_n + \sin \tau_n)(\cosh \tau_n x - \cos \tau_n x), \end{aligned} \tag{3.38}$$

$$v_n = \lambda_n u_n(x), \tag{3.39}$$

$$\eta_n = 2f(\tau_n) = 0, \tag{3.40}$$

where φ_1 is given in (3.18). As before, since $\mathcal{R}e\{\lambda_n\} < 0$ implies that τ_n is not a purely imaginary number, the constant factors in (3.18) cannot vanish simultaneously. Then

$$F_n = \begin{pmatrix} u_n \\ v_n \\ 0 \end{pmatrix} \tag{3.41}$$

is an eigenvector for A associated with the eigenvalue λ_n. Note that $(u_n\ v_n)^T$ is an eigenvector of B (i.e., of the cantilevered beam) associated with the same eigenvalue.

Assume now that all the eigenvalues λ_n of the cantilevered beam are algebraically simple. By Lemma 3.2, this assumption can be written as

$$f(\tau_n) = 0, \qquad f'(\tau_n) \neq 0, \tag{3.42}$$

where $\lambda = i\tau^2$ and for $f(\tau) = 0$ we have

$$f'(\tau) = -\tau^2 \left(1 + \frac{\tau^2}{i\alpha}\right)(1 + \cosh\tau\cos\tau) - 2i\alpha\tau^2 \sinh\tau \sin\tau. \tag{3.43}$$

Under the assumption given by (3.42), we now compute the algebraic multiplicity of all the eigenvalues $(\lambda_*, \lambda_n, n \in \mathbf{Z})$ of A. We note that the algebraic simplicity of λ_n as an eigenvalue of B does not imply the algebraic simplicity of λ_n as an eigenvalue of A. We have to distinguish two cases: $\eta_* = 2f(\tau_*) \neq 0$, in which case $\lambda_* \neq \lambda_n \ \forall n \in \mathbf{Z}$, or $\eta_* = 0$, in which case $\lambda_* = \lambda_N$ for some $N \in \mathbf{Z}$.

An easy computation shows that $\eta_* = 0$ if and only if

$$\alpha = \frac{\beta_*(2 + \cosh 2\beta_* + \cos 2\beta_*)}{\sinh 2\beta_* - \sin 2\beta_*}, \tag{3.44}$$

where $\beta_* = 1/\sqrt{2\beta}$. Hence the case $\eta_* = 0$ is just an exceptional one in the sense that (α, β) have to belong to the curve defined by (3.44). For instance, if α is sufficiently small, η_* is always nonzero.

Let $\tilde{\lambda}$ be an eigenvalue of A, and let $(\tilde{u}\ \tilde{v}\ \tilde{\eta})^T$ be the corresponding eigenvector. Let us study when the algebraic multiplicity of $\tilde{\lambda}$ is equal to one or not.

Obviously $\mathrm{Ker}(\mathrm{A} - \tilde{\lambda}\mathrm{I})^2 \setminus \mathrm{Ker}(\mathrm{A} - \tilde{\lambda}\mathrm{I}) \neq \emptyset$ if and only if there exists $(u\ v\ \eta)^T \in D(A)$ such that

$$A \begin{pmatrix} u \\ v \\ \eta \end{pmatrix} - \tilde{\lambda} \begin{pmatrix} u \\ v \\ \eta \end{pmatrix} = \begin{pmatrix} \tilde{u} \\ \tilde{v} \\ \tilde{\eta} \end{pmatrix}, \tag{3.45}$$

which is equivalent to the following:

$$v = \tilde{\lambda}u + \tilde{u}, \tag{3.46}$$

$$-u_{xxxx} - \tilde{\lambda}v = \tilde{v} = \tilde{\lambda}\tilde{u}, \tag{3.47}$$

$$-\left(\tilde{\lambda} + \frac{1}{\beta}\right)\eta = \tilde{\eta}, \tag{3.48}$$

where $(u\ v\ \eta)^T \in D(A)$. Equations (3.46)–(3.48) have a solution if and only if the equations

$$-u_{xxxx} - \tilde{\lambda}^2 u = 2\tilde{\lambda}\tilde{u}, \tag{3.49}$$

$$-\left(\tilde{\lambda} + \frac{1}{\beta}\right)\eta = \tilde{\eta}, \tag{3.50}$$

$$u(0) = u_x(0) = u_{xx}(1) = 0, \tag{3.51}$$

$$\eta = -u_{xxx}(1) + \alpha\tilde{\lambda}u(1) + \alpha\tilde{u}(1) \tag{3.52}$$

admit a solution.

LEMMA 3.4. *Let $\alpha = m/\beta$ and consider the operator A given by* (2.10). *Let α be such that the eigenvalues of the operator B given by* (3.15) *are algebraically simple* (*note that this is true for almost all $\alpha > 0$ by Theorem* 3.3). *Let (λ_*, F_*) be the eigenvalue-eigenvector pair of A given by $\lambda_* = 1/\beta$ and* (3.34), *respectively, and let (λ_n, F_n), $n \in \mathbf{Z}$ be the remaining eigenvalue-eigenvector pairs of A, where λ_n is a root of* (3.10) *and F_n is given by* (3.41).

i. *If $\eta_* \neq 0$, then all eigenvalues of A are algebraically simple.*

ii. *If $\eta_* = 0$, then the algebraic multiplicity of λ_* is exactly* 2 *and all the eigenvalues $\lambda_n \neq \lambda_*$ are algebraically simple.*

Proof. **i.** Let $\eta_* \neq 0$, which implies $\lambda_* \neq \lambda_n,\ n \in \mathbf{Z}$. Then, for $\tilde{\lambda} = \lambda_*$, (3.50) implies $\eta_* = 0$, which is a contradiction. Thus λ_* is algebraically simple. Choose now $\tilde{\lambda} = \lambda_n$ for $n \in \mathbf{Z}$, and for simplicity, denote by $\lambda = i\tau^2$ the eigenvalue λ_n. Then $\tilde{\eta} = \eta_n = 0$, and since $(\tilde{\lambda} + \frac{1}{\beta}) \neq 0$, we get $\eta = 0$ so that (3.49)–(3.52) reduces to

$$-u_{xxxx} - \lambda^2 u = 2\lambda u_n, \tag{3.53}$$

$$u(0) = u_x(0) = u_{xx}(1) = 0, \tag{3.54}$$

$$u_{xxx}(1) = \alpha\lambda u(1) + \alpha u_n(1). \tag{3.55}$$

Then, proceeding exactly as in Lemma 3.2, we obtain that (3.53)–(3.55) has a solution if and only if $f'(\tau) = 0$ (cf. (3.23), (3.24)). By Lemma 3.2 this implies that λ_n is not algebraically simple as an eigenvalue of B, which is a contradiction. Hence by Lemma 3.2 we see that λ_n is also algebraically simple as an eigenvalue of A.

ii. For the case $\eta_* = 0$, by the argument given above, all the λ_n such that $\lambda_n \neq \lambda_*$ are also algebraically simple.

Let $\lambda_* = \lambda_N$ for some $N \in \mathbf{Z}$, which is denoted by λ for simplicity. Then (3.49)–(3.52) reduces to

$$-u_{xxxx} - \lambda^2 u = 2\lambda u_*, \tag{3.56}$$

$$u(0) = u_x(0) = u_{xx}(1) = 0, \tag{3.57}$$

$$-u_{xxx}(1) + \alpha\lambda u(1) + \alpha u_*(1) = \eta. \tag{3.58}$$

Now proceeding again as in Lemma 3.2, but replacing the right-hand side of (3.27) by $\mu - \eta$, we obtain that (3.56)–(3.58) has a solution if and only if

$$\eta = \frac{f'(\tau)}{i\tau}, \tag{3.59}$$

and hence is nonzero by Lemma 3.2 if λ is an algebraically simple eigenvalue of B. Consequently it is always possible to compute $\eta_{**} \neq 0$ in a unique way such that (3.59) is true for $\eta = \eta_{**}$, and then one has a (nonunique) solution $u = u_{**}$ of (3.56)–(3.58) with $v = v_{**} = \lambda_* u_{**} + u_*$ such that (3.46)–(3.48) is satisfied. Thus in case $\eta_* = 0$, λ_* has algebraic multiplicity at least two, $F_* = (u_*\ v_*\ 0)^T$ is an eigenvector

of A, and $F_{**} = (u_{**}\ v_{**}\ \eta_{**})^T$ with $\eta_{**} \neq 0$ is a generalized eigenvector (but not an eigenvector) of A. In fact, λ_* has algebraic multiplicity exactly two, since for $w = (u\ v\ \eta)^T \in D(A)$, $w \in \mathrm{Ker(A} - \lambda_*\mathrm{I)}^3 \setminus \mathrm{Ker(A} - \lambda_*\mathrm{I)}^2$ implies $(A - \lambda_* I)w \in \mathrm{Ker(A} - \lambda_*\mathrm{I)}^2 \setminus \mathrm{Ker(A} - \lambda_*\mathrm{I)}$; thus $(A - \lambda_* I)w = (u_{**}\ v_{**}\ \eta_{**})^T$. But this implies that $-(\lambda_* + 1/\beta)\eta = \eta_{**}$ (cf.(3.48)), hence $\eta_{**} = 0$, which is impossible. □

We have now the material to write down the Riesz basis property. Recall that $(u_n\ v_n)^T$ are not the functions given exactly by (3.38)–(3.39) but have been suitably normalized to posses the adequate Riesz basis property for the cantilevered beam, (see Theorem 3.3).

THEOREM 3.5. *Let $\alpha = m/\beta$, $\lambda_* = -1/\beta$, λ_n, $n \in \mathbf{Z}$, be the roots of* (3.10). *Assume* (3.42) *and $\eta_* = 2f(\tau_*) \neq 0$. Then $\{F_*,\ F_n,\ n \in \mathbf{Z}\}$ is a Riesz basis for $\mathcal{H}$. Moreover the estimate* (2.25) *is valid with $\delta > 0$ such that*

$$-\delta = \max\{-1/\beta,\ \mathcal{R}e\{\lambda_n\},\ n \in \mathbf{Z}\}, \tag{3.60}$$

which is the optimal rate of decay.

Proof. Let $z = (u\ v\ \eta)^T \in \mathcal{H}$ be given. Since $\eta_* \neq 0$, we can write

$$z = \begin{pmatrix} \tilde{u} \\ \tilde{v} \\ 0 \end{pmatrix} + c_* F_*, \tag{3.61}$$

where

$$c_* = \eta/\eta_*, \quad \tilde{u} = u - c_* u_* \in \mathcal{V}, \quad \tilde{v} = v - c_* v_* \in \mathbf{L}^2(0,1). \tag{3.62}$$

Since $(u_n\ v_n)^T,\ n \in \mathbf{Z}$, is a Riesz basis for $\mathcal{V} \times \mathbf{L}^2(0,1)$, we can write

$$\begin{pmatrix} \tilde{u} \\ \tilde{v} \end{pmatrix} = \sum_{n \in \mathbf{Z}} c_n \begin{pmatrix} u_n \\ v_n \end{pmatrix}, \tag{3.63}$$

where $c_n \in l^2(\mathbf{Z})$, and there exist positive constants C_1, C_2 such that

$$C_1 \sum_{n \in \mathbf{Z}} |c_n|^2 \leq \left\| \begin{pmatrix} \tilde{u} \\ \tilde{v} \end{pmatrix} \right\|^2_{\mathcal{V} \times L^2} \leq C_2 \sum_{n \in \mathbf{Z}} |c_n|^2. \tag{3.64}$$

Note that

$$\left\| \begin{pmatrix} \tilde{u} \\ \tilde{v} \end{pmatrix} \right\|^2_{\mathcal{V} \times L^2} = \int_0^1 (\tilde{u}_{xx}^2 + \tilde{v}^2) dx\ ,\ \|z\|^2_{\mathcal{H}} = \int_0^1 (u_{xx}^2 + v^2) dx + K\eta^2, \tag{3.65}$$

where K is given by (2.9). By using (3.62) we obtain

$$u_{xx}^2 = \tilde{u}_{xx}^2 + 2c_* \tilde{u}_{xx} u_{*xx} + c_*^2 u_{*xx}^2, \tag{3.66}$$

$$v^2 = \tilde{v}^2 + 2c_* \tilde{v} v_* + c_*^2 v_*^2. \tag{3.67}$$

It follows from Young's inequality that

$$\begin{aligned} \left| \int_0^1 2c_* \tilde{u}_{xx} u_{*xx} dx \right| &\leq \sigma \int_0^1 \tilde{u}_{xx}^2 dx + \frac{1}{\sigma} \int_0^1 |\ c_*\ |^2 u_{*xx}^2 dx \\ &\leq \sigma \int_0^1 \tilde{u}_{xx}^2 dx + \frac{M_1}{\sigma} \eta^2, \end{aligned} \tag{3.68}$$

where $\sigma > 0$ is an arbitrary constant and (using (3.62))

$$M_1 = \frac{\int_0^1 u_{*xx}^2}{\eta_*^2}.$$

Similarly we obtain

$$\left| \int_0^1 2c_* \tilde{v} v_* dx \right| \leq \sigma \int_0^1 \tilde{v}^2 dx + \frac{M_2}{\sigma} \eta^2, \tag{3.69}$$

where

$$M_2 = \frac{\int_0^1 v_*^2}{\eta_*^2}.$$

By using (3.66)–(3.69) in (3.65), we obtain

$$\|z\|_{\mathcal{H}}^2 \leq (1+\sigma) \left\| \begin{pmatrix} \tilde{u} \\ \tilde{v} \end{pmatrix} \right\|_{\mathcal{V} \times L^2}^2 + \left(K + M_1 + M_2 + \frac{M_1}{\sigma} + \frac{M_2}{\sigma} \right) \eta^2, \tag{3.70}$$

$$\|z\|_{\mathcal{H}}^2 \geq (1-\sigma) \left\| \begin{pmatrix} \tilde{u} \\ \tilde{v} \end{pmatrix} \right\|_{\mathcal{V} \times L^2}^2 + \left(K + M_1 + M_2 - \frac{M_1}{\sigma} - \frac{M_2}{\sigma} \right) \eta^2 \ . \tag{3.71}$$

From (3.61)–(3.63) it follows that

$$z = \sum_{n \in \mathbf{Z}} c_n F_n + c_* F_*. \tag{3.72}$$

Next we choose $\sigma > 0$ such that

$$\frac{M_1 + M_2}{K + M_1 + M_2} < \sigma < 1,$$

which implies that all coefficients in (3.71) are positive. Since η is proportional to c_* (see (3.62)), it follows from (3.64), (3.70)–(3.71) that there exist positive constants C_3 and C_4 such that the following holds:

$$C_3 \Big(\sum_{n \in \mathbf{Z}} |c_n|^2 + |c_*|^2 \Big) \leq \|z\|_{\mathcal{H}}^2 \leq C_4 \Big(\sum_{n \in \mathbf{Z}} |c_n|^2 + |c_*|^2 \Big). \tag{3.73}$$

It follows from (3.72)–(3.73) that the system $\{F_*, F_n, n \in \mathbf{Z}\}$ is a Riesz basis in $\mathcal{H}$.

Since $F_*,\ F_n,\ n \in \mathbf{Z}$ are all eigenvectors of A, we then have

$$T(t)z = T(t) \left[\sum_{n \in \mathbf{Z}} c_n F_n + c_* F_* \right] = \sum_{n \in \mathbf{Z}} e^{\lambda_n t} c_n F_n + e^{\lambda_* t} c_* F_*. \tag{3.74}$$

That (3.60) determines the optimal decay rate for the semigroup is now an immediate and general consequence of the Riesz basis property in $\mathcal{H}$. □

THEOREM 3.6. *Let $\alpha = m/\beta$, $\lambda_* = -1/\beta$, λ_n, $n \in \mathbf{Z}$, be the roots of* (3.10). *Assume* (3.42) *and $\eta_* = 2f(\tau_*) = 0$. Then $\{F_{**},\ F_n,\ n \in \mathbf{Z}\}$ is a Riesz basis for $\mathcal{H}$. Moreover, for any $\epsilon > 0$, the estimate* (2.25) *is valid for $\delta - \epsilon$, where $-\delta$ is given by* (3.60). *Hence $-\delta$ is again the optimal decay rate.*

Proof. We recall that here $F_* = F_N$ for some $N \in \mathbf{Z}$, F_n being suitably normalized eigenvectors of A. The Riesz basis property can be proven as in Theorem 3.5 by just replacing F_* by F_{**} and using the fact that $\eta_{**} \neq 0$, so that with $c_{**} = \eta/\eta_{**}$ we have

$$z = \begin{pmatrix} u \\ v \\ \eta \end{pmatrix} = \begin{pmatrix} \tilde{u} \\ \tilde{v} \\ 0 \end{pmatrix} + c_{**}F_{**}, \tag{3.75}$$

and for $(\tilde{u}\ \tilde{v})^T$ we use the Riesz basis property of F_n, $n \in \mathbf{Z}$. Then we get

$$z = \sum_{n\in\mathbf{Z}} c_n F_n + c_{**}F_{**}, \tag{3.76}$$

where F_n, $n \in \mathbf{Z}$, are the eigenvectors of A, with $F_N = F_*$, $\lambda_N = \lambda_*$, but $F_{**} \in \mathrm{Ker(A} - \lambda_{\mathrm{N}}\mathrm{I)}^2 \setminus \mathrm{Ker(A} - \lambda_{\mathrm{N}}\mathrm{I)}$. Since F_{**} satisfies $(A - \lambda_* I)F_{**} = F_*$, we get

$$\frac{d}{dt}\Big(e^{\lambda_* t}(tF_* + F_{**})\Big) = e^{\lambda_* t}A(tF_* + F_{**}). \tag{3.77}$$

From (3.76)–(3.77) we get

$$T(t)z = \sum_{n\neq N,\ n\in\mathbf{Z}} e^{\lambda_n t}c_n F_n + e^{\lambda_* t}\Big((c_N + tc_{**})F_* + c_{**}F_{**}\Big). \tag{3.78}$$

Now the fact that the estimate (2.25) holds for $\delta - \epsilon$, for any $\epsilon > 0$, is an immediate and general consequence of the Riesz basis property. Due to the fact that $-\delta$ given by (3.60) may be achieved by λ_*, $\epsilon > 0$ comes from the possible compensation of $e^{2Re\lambda_* t}t^2$ by $e^{(2Re\lambda_* + \epsilon)t}$. If $-\delta > \lambda_* = -1/\beta$, then ϵ is unnecessary. ∎

4. Conclusion. In this paper we studied the stability of a flexible beam with a tip mass. The flexible beam is assumed to be clamped at one end and is free at the other, where a mass is also attached. This model is a variant of the SCOLE model and has been studied before; see, e.g., [1], [9], [13]. To stabilize this hybrid system we apply a boundary control force at the free end of the beam. It is well known that for this model the standard velocity feedback for the control force (e.g., (1.4)), which is widely used in boundary control systems, yields only *asymptotic*, but not *exponential*, stability; see e.g., [9], [13]. In this paper we proposed a (new) control law (see (2.1)), which contains the term $u_{xxxt}(1,t)$ in addition to the standard feedback term $u_t(1,t)$. We then proved that the system is well-posed and that the energy associated with the system decays exponentially to zero if the initial data are in $\mathcal{H}$. We also showed that if the initial data are sufficiently smooth (i.e., in $D(A)$), then the tip mass velocity also decays exponentially to zero. Then we analyzed the spectrum of the system for the special case $m = \alpha\beta$ and proved that the spectrum determines the exponential decay rate for the considered case for almost all $\alpha > 0$.

Appendix A. On the Riesz basis property of eigenvectors of the cantilevered beam with boundary force control. Here our aim is to prove Theorem 3.3. We will consider the set of eigenvectors of the operator B given by (3.15) for the cases $\alpha = 0$ (i.e., uncontrolled cantilevered beam) and $\alpha > 0$ (i.e., controlled cantilevered beam) and show that these two sets are quadratically close. Since the former set of eigenvectors is known to be a Riesz basis in $V \times \mathbf{L}^2(0,1)$, we conclude that the latter set is also a Riesz basis in the same space.

Before we prove the Riesz basis property, first we will show that the number of eigenvalues of the uncontrolled and controlled cantilevered beam are the same, counting multiplicities, in sufficiently large disks. This result will enable us to enumerate the eigenvalues of both systems in a similar way. We recall that the eigenvalues of B for $\alpha \geq 0$ are precisely the roots of (3.10). Since $\lambda = 0$ is not an eigenvalue, equivalently the eigenvalues are the roots of the following function (for $\lambda = i\tau^2$)

$$h(\tau) = \frac{f(\tau)}{\tau^2} = \tau(1 + \cosh\tau\cos\tau) - i\alpha(\sinh\tau\cos\tau - \cosh\tau\sin\tau); \tag{A.1}$$

hence for the uncontrolled case (i.e., $\alpha = 0$), the eigenvalues are the roots of the following function

$$g(\tau) = \tau(1 + \cosh\tau\cos\tau). \tag{A.2}$$

Note that $\tau = 0$ is a simple root of both (A.1) and (A.2) but not an eigenvalue of B for $\alpha \geq 0$. Hence it follows that if $h(\cdot)$ and $g(\cdot)$ have the same number of roots in a large disk, then the same is true for the eigenvalues of the operator B for $\alpha = 0$ and $\alpha > 0$.

LEMMA A.1. *There exists a sequence $R_k \in \mathbf{R}$ such that $R_k \to \infty$ as $k \to \infty$ and the number of roots of* (A.1) *and* (A.2) *are the same, counting multiplicities, in $B(0, R_k)$ where $B(0, R)$ is defined as*

$$B(0, R) = \{\ \tau \in \mathbf{C}\ \ |\ \ |\ \tau\ | \leq R\ \ \}. \tag{A.3}$$

Proof. Let $R > 0$ be given and $\gamma = \{\ \tau \in \mathbf{C}\ \ |\ \ |\ \tau\ | = R\ \ \}$, i.e., a circle of radius R. Since both $h(\cdot)$ and $g(\cdot)$ are analytic in $B(0, R)$, by Rouché's theorem they have the same number of roots, counting multiplicities, if $|\ h(\tau) - g(\tau)\ | < |\ g(\tau)\ |$ for $\tau \in \gamma$. We will show that this is true for some sufficiently large R. For convenience let us define

$$s(\tau) = i\alpha(\sinh\tau\cos\tau - \cosh\tau\sin\tau); \tag{A.4}$$

hence equivalently we need to show the following:

$$\left|\frac{s(\tau)}{g(\tau)}\right| \quad < \quad 1, \qquad \tau \in \gamma. \tag{A.5}$$

Since both $g(\cdot)$ and $s(\cdot)$ are odd functions it is sufficient to consider the upper half plane, and since $\cosh i\tau = \cos\tau$, $\cos i\tau = \cosh\tau$, $\sinh i\tau = i\sin\tau$, $\sin i\tau = i\sinh\tau$, it is sufficient to consider only the first quadrant, i.e., $\tau = Re^{i\theta}$ for $0 \leq \theta \leq \pi/2$.

Let $\tau = Re^{i\theta}$. After straightforward calculations it could be shown that the following holds:

$$|\ s(\tau)\ | \leq \frac{\alpha}{2}(e^{RD} + e^{-RD} + e^{RS} + e^{-RS}), \tag{A.6}$$

$$4\cosh\tau\cos\tau = e^{RD}e^{iRS} + e^{RS}e^{-iRD} + e^{-RS}e^{iRD} + e^{-RD}e^{-iRS}, \tag{A.7}$$

where $D = \cos\theta - \sin\theta$, $S = \cos\theta + \sin\theta$.

For $0 \leq \theta \leq \pi/2$ we have $S \geq 1$ and $S \geq \mid D \mid$; hence $\mid s(\tau) \mid \leq 2\alpha e^{RS}$. For $0 < \theta < \pi/4$ we have $D > 0$; hence for sufficiently large R the following holds:

$$\frac{\mid s(\tau) \mid}{\mid \cosh\tau \cos\tau \mid} \leq \frac{2\alpha}{\mid e^{-RS} \cosh\tau \cos\tau \mid} \leq M \tag{A.8}$$

for some $M > 0$. For $\pi/4 < \theta < \pi/2$ we have $D < 0$, and from (A.6) and (A.7) it easily follows that an estimate similar to (A.8) holds. Hence for $0 < \theta < \pi/2$ and $\theta \neq \pi/4$ we have $\lim_{R\to\infty} \mid \frac{s(\tau)}{g(\tau)} \mid = 0$. For $\theta = 0$ or $\theta = \pi/2$ we have $D = 1$ or $D = -1$, respectively; $S = 1$ and $1 + \cosh\tau\cos\tau = 1 + \cosh R \cos R$ in both cases. Hence if we choose $R = 2n\pi$, we have $\lim_{n\to\infty} \mid \frac{s(\tau)}{g(\tau)} \mid = 0$. We note that this holds if $R \to \infty$ in such a way that $\mid \cos R \mid \geq \delta$ for any $\delta > 0$. For $\theta = \pi/4$ we have $D = 0$, $S > 1$, and $4\cosh\tau\cos\tau = 2\cos RS + 2\cosh RS$; hence $\lim_{R\to\infty} \mid \frac{s(\tau)}{g(\tau)} \mid = 0$. Therefore, for $\tau = Re^{i\theta}$, $R = 2n\pi$, and $0 \leq \theta \leq 2\pi$ we have $\lim_{n\to\infty} \mid \frac{s(\tau)}{g(\tau)} \mid = 0$. Hence there exists a sequence $R_k = 2k\pi$, $k \in \mathbf{N}$, and $k \to \infty$ such that $\lim_{k\to\infty} \mid \frac{s(\tau)}{g(\tau)} \mid < 1$ for $\mid \tau \mid = R_k$. Therefore, by Rouchée's theorem, the number of roots of $g(\cdot)$ and $h(\cdot)$, or equivalently the eigenvalues of the operator B for the cases $\alpha = 0$ and $\alpha > 0$, respectively, are the same in $B(0, R_k)$, counting multiplicities. □

The lemma given above lets us enumerate the eigenvalues of uncontrolled and controlled cantilevered beam in a similar way, at least if they are algebraically simple (see Remark 3 for an extension). In what follows we will give asymptotic formulas for these eigenvalues and then compare the corresponding eigenvectors.

Consider the system and the corresponding eigenvalue problem given by (3.11)–(3.18). From (3.17) it follows that the eigenvalues occur in complex conjugate pairs. Since there are countably many eigenvalues and each eigenvalue is isolated (see Lemma 3.1), the eigenvalues which have positive imaginary part can be numerated by considering the imaginary parts with increasing order. By using asymptotic analysis it can be shown that asymptotically the solutions of (3.17) can be given as ($\lambda = i\tau^2$):

$$\lambda_k = -2\alpha + \mathcal{O}(1/k^2) + i((m\pi)^2 + \alpha\mathcal{O}(1/k)), \tag{A.9}$$

for sufficiently large $k \in \mathbf{N}$, where $m = k + 1/2$; see [12, p. 76]. We note that this estimate can also be obtained by using the wave propagation method (see [2]) for similar estimates. Here the symbol $\mathcal{O}(f(k))$ denotes any function such that $\lim_{k\to\infty} \mathcal{O}(f(k))/f(k)$ exists and is finite.

By using $\lambda_k = i\tau_k^2$, the corresponding τ_k can easily be found as

$$\tau_k = \pm\left[(m\pi + \mathcal{O}(1/k^2)) + i\left(\frac{\alpha}{m\pi} + \mathcal{O}(1/k^3)\right)\right] \tag{A.10}$$

for sufficiently large k. In what follows we will consider (A.10) with + sign; the same conclusions hold with - sign as well (see below). By using (A.10), with + sign, we obtain the following estimates:

$$e^{\tau_k x} = e^{m\pi x}\left((1 + \mathcal{O}(1/k^2)f_1(x)) + i\left(\frac{\alpha x}{m\pi} + \mathcal{O}(1/k^3)f_2(x)\right)\right), \tag{A.11}$$

$$e^{-\tau_k x} = e^{-m\pi x}\left((1 + \mathcal{O}(1/k^2)f_3(x)) - i\left(\frac{\alpha x}{m\pi} + \mathcal{O}(1/k^3)f_4(x)\right)\right), \tag{A.12}$$

$$
\begin{aligned}
e^{i\tau_k x} &= e^{-\frac{\alpha x}{m\pi}}((\cos m\pi x + \cos m\pi x \;\; \mathcal{O}(1/k^3)f_5(x) \\
&\quad - \sin m\pi x \;\; \mathcal{O}(1/k^2)f_6(x)) + i(\sin m\pi x + \sin m\pi x \;\; \mathcal{O}(1/k^3)f_7(x) \\
&\quad + \cos m\pi x \;\; \mathcal{O}(1/k^2)f_8(x))),
\end{aligned} \tag{A.13}
$$

$$
\begin{aligned}
e^{-i\tau_k x} &= e^{\frac{\alpha x}{m\pi}}((\cos m\pi x + \cos m\pi x \;\; \mathcal{O}(1/k^3)f_9(x) \\
&\quad - \sin m\pi x \;\; \mathcal{O}(1/k^2)f_{10}(x)) - i(\sin m\pi x + \sin m\pi x \;\; \mathcal{O}(1/k^3)f_{11}(x) \\
&\quad + \cos m\pi x \;\; \mathcal{O}(1/k^2)f_{12}(x))),
\end{aligned} \tag{A.14}
$$

where the functions $f_i(\cdot)$, $i = 1, \ldots, 12$, are smooth and bounded functions with bounded derivatives. By using (A.11)–(A.14), we obtain the following estimates:

$$
\begin{aligned}
&(\cosh\tau_k + \cos\tau_k)(\sinh\tau_k x - \sin\tau_k x) = \frac{e^{\tau_k}e^{\tau_k x}}{4} + \bigg(- \frac{e^{m\pi}e^{-m\pi x}}{4} \\
&+e^{m\pi}\mathcal{O}(1/k^2)o_1(x) + e^{m\pi x}o_2(x) - \frac{e^{m\pi}}{2}\sin m\pi x + o_3(x)\bigg) \\
&+i(e^{m\pi}\mathcal{O}(1/k)o_4(x) + e^{m\pi x}\mathcal{O}(1/k)o_5(x) + o_6(x)),
\end{aligned} \tag{A.15}
$$

$$
\begin{aligned}
&(\sinh\tau_k + \sin\tau_k)(\cosh\tau_k x - \cos\tau_k x) = \frac{e^{\tau_k}e^{\tau_k x}}{4} + \bigg(\frac{e^{m\pi}e^{-m\pi x}}{4} \\
&+e^{m\pi}\mathcal{O}(1/k^2)o_7(x) + e^{m\pi x}o_8(x) - \frac{e^{m\pi}}{2}\cos m\pi x + o_9(x)\bigg) \\
&+i(e^{m\pi}\mathcal{O}(1/k)o_{10}(x) + e^{m\pi x}\mathcal{O}(1/k)o_{11}(x) + o_{12}(x)),
\end{aligned} \tag{A.16}
$$

where the functions $o_i(\cdot)$, $i = 1, \ldots, 12$, are smooth and bounded functions (as a function of k), and their derivatives are either bounded or satisfy the following:

$$
o_i^{(n)}(x) = (k\pi)^n \hat{o}_i(x), \qquad i = 1, \ldots, 12, \qquad n \in \mathbf{N}, \tag{A.17}
$$

where the functions $\hat{o}_i(\cdot)$ are also smooth and bounded functions. By using (A.15) and (A.16) in (3.18) we obtain

$$
\begin{aligned}
\varphi_1(\tau_k, x) &= \bigg[- \frac{e^{m\pi}e^{-m\pi x}}{2} + e^{m\pi}\mathcal{O}(1/k^2)o_{13}(x) + e^{m\pi x}o_{14}(x) \\
&\quad + \frac{e^{m\pi}}{2}\cos m\pi x - \frac{e^{m\pi}}{2}\sin m\pi x + o_{15}(x)\bigg] \\
&\quad + i[e^{m\pi}\mathcal{O}(1/k)o_{16}(x) + e^{m\pi x}\mathcal{O}(1/k)o_{17}(x) + o_{18}(x)],
\end{aligned} \tag{A.18}
$$

where the functions $o_i(\cdot)$ are of the same form as given in (A.15)–(A.16).

Let $\lambda \in \mathbf{C}$ be an eigenvalue of B (see (3.15)), and let $E \in H = V \times L^2(0,1)$ be the corresponding (unnormalized) eigenvector given by

$$
E = \begin{pmatrix} \varphi_1(\tau, x) \\ i\tau^2\varphi_1(\tau, x) \end{pmatrix}; \tag{A.19}
$$

see (3.18). The norm of E can be found as

$$
\|E\|_H^2 = (\mid \lambda \mid^2 - \lambda^2) \int_0^1 \varphi_1\bar{\varphi}_1 dx - \alpha\lambda \mid \varphi_1(1) \mid^2, \tag{A.20}
$$

where a bar denotes the complex conjugate. Let λ_k and E_k be an eigenvalue, (unnormalized) eigenvector pair. By using (A.18) it easily follows that

$$\int_0^1 (\mathrm{Im}\{\varphi_1\})^2 dx = \mathcal{O}(e^{2k\pi}/(k\pi)^2) \tag{A.21}$$

for k sufficiently large. By using the simple integrals

$$\int_0^1 \cos^2 m\pi x dx = \int_0^1 \sin^2 m\pi x dx = 1/2, \qquad \int_0^1 \sin m\pi x \cos m\pi x dx = \frac{1}{2m\pi}$$

(note that $m = k + 1/2$), it follows from (A.18) that

$$\int_0^1 (Re\{\varphi_1\})^2 dx = C_1 e^{2k\pi} + \mathcal{O}(e^{2k\pi}/(k\pi)) \tag{A.22}$$

for k sufficiently large, where $C_1 > 0$ is a constant. By using (A.9), (A.21), and (A.22) in (A.20) it follows that

$$\|E_k\|_H^2 = C_2 (k\pi)^4 e^{2k\pi} + \mathcal{O}(e^{2k\pi}(k\pi)^3) \tag{A.23}$$

for k sufficiently large, where $C_2 > 0$ is a constant. Hence we define the (approximately) normalized eigenvectors as

$$F_{kr} = \frac{1}{(k\pi)^2 e^{k\pi}} \begin{pmatrix} \varphi_1(\tau_k, x) \\ i\tau_k^2 \varphi_1(\tau_k, x) \end{pmatrix}, \tag{A.24}$$

where τ_k and φ_1 are given by (3.17) and (3.18), respectively.

Now consider the system (3.11)–(3.12) with $\alpha = 0$, i.e., uncontrolled system. By using μ instead of τ, the characteristic equation (3.17) becomes

$$1 + \cosh \mu \cos \mu = 0, \qquad \lambda = i\mu^2, \tag{A.25}$$

whose roots are asymptotically given by

$$\mu_k = m\pi + \mathcal{O}(e^{-m\pi}), \qquad m = k + 1/2 \tag{A.26}$$

for k sufficiently large. It follows that the corresponding function $\varphi_1(\mu_k, x)$ is real. By following the analysis given above, similar to (A.18), we obtain

$$\begin{aligned} \varphi_1(\mu_k, x) &= -\frac{e^{m\pi} e^{-m\pi x}}{2} + e^{m\pi} \mathcal{O}(e^{-m\pi}) o_{19}(x) + e^{m\pi x} o_{20}(x) \\ &\quad + \frac{e^{m\pi}}{2} \cos m\pi x - \frac{e^{m\pi}}{2} \sin m\pi x + o_{21}(x), \end{aligned} \tag{A.27}$$

where the functions $o_i(\cdot)$ are as given in (A.15)–(A.16). Hence, by following the analysis given above, we define the (approximately) normalized eigenvector corresponding to μ_k as

$$G_{kr} = \frac{1}{(k\pi)^2 e^{k\pi}} \begin{pmatrix} \varphi_1(\mu_k, x) \\ i\mu_k^2 \varphi_1(\mu_k, x) \end{pmatrix}. \tag{A.28}$$

THEOREM A.2. *Consider the (approximately) normalized eigenvectors F_k and G_k given by* (A.24) *and* (A.28), *respectively. Then the estimate*

$$\|F_{kr} - G_{kr}\|_H = \mathcal{O}(1/k) \tag{A.29}$$

holds for sufficiently large k.

Proof. From (A.18) and (A.27) it follows that

$$\begin{aligned}\varphi_1(\tau_k,x)-\varphi_1(\mu_k,x) &= e^{m\pi}\mathcal{O}(1/k^2)o_{22}(x)+e^{m\pi x}o_{23}(x)+o_{24}(x)\\ &\quad + i[e^{m\pi}\mathcal{O}(1/k)o_{25}(x)+e^{m\pi x}\mathcal{O}(1/k)o_{26}(x)+o_{27}(x)],\end{aligned} \tag{A.30}$$

where the functions $o_i(\cdot)$ are as given in (A.15). Also note that

$$\begin{aligned}i\tau_k^2\varphi_1(\tau_k,x)-i\mu_k^2\varphi_1(\mu_k,x) &= i\tau_k^2[\varphi_1(\tau_k,x)-\varphi_1(\mu_k,x)]\\ &\quad + i(\tau_k^2-\mu_k^2)\varphi_1(\mu_k,x).\end{aligned} \tag{A.31}$$

From (A.17), (A.30), and (A.31) it follows that

$$\int_0^1 \mid \varphi_{1xx}(\tau_k,x)-\varphi_{1xx}(\mu_k,x) \mid^2 dx = \mathcal{O}(e^{2k\pi}(k\pi)^2), \tag{A.32}$$

$$\int_0^1 \mid \tau_k^2\varphi_1(\tau_k,x)-\mu_k^2\varphi_1(\mu_k,x) \mid^2 dx = \mathcal{O}(e^{2k\pi}(k\pi)^2) \tag{A.33}$$

for k sufficiently large. Hence (A.29) easily follows from (A.32) and (A.33). ∎

Now we consider the algebraic simplicity of the eigenvalues of B for the case $\alpha > 0$ and prove the statement **i** of Theorem 3.3.

LEMMA A.3. *Consider the system given by* (3.11)–(3.12) *for* $\alpha > 0$. *All eigenvalues of B with sufficiently large modulus are algebraically simple.*

Proof. Since the operator B has compact resolvent (see Lemma 3.1), it follows that the spectrum of B consists entirely of isolated points, at most countable, and each eigenvalue has a finite algebraic multiplicity.

Let τ be a root of (3.17), and let $\lambda = i\tau^2$ be the corresponding eigenvalue. From Lemma 3.2 it follows that λ has algebraic multiplicity greater than 1 if and only if $f'(\tau) = 0$; see (3.43).

First note that by using (A.10), (A.13), (A.14), it follows that

$$\cos\tau_k = -(-1)^k\mathcal{O}(1/k^2) - i\left((-1)^k\frac{\alpha}{m\pi}+(-1)^k\mathcal{O}(1/k^3)\right), \tag{A.34}$$

$$\sin\tau_k = (-1)^k + (-1)^k\mathcal{O}(1/k^2) - i((-1)^k\mathcal{O}(1/k^3)). \tag{A.35}$$

By using (A.11), (A.12), (A.34), (A.35) in (3.43) we obtain

$$-i\frac{f'(\tau_k)}{\tau_k^2} = (e^{m\pi}o_1(k)+o_2(k)) + i(-m\pi e^{m\pi}/2 + e^{m\pi}o_3(k)+o_4(k)), \tag{A.36}$$

where $o_i(k)$, $i = 1,\dots,4$ are bounded functions of k. Hence it follows that, for sufficiently large k, we have $f'(\tau_k) \neq 0$, which implies that all eigenvalues with sufficiently large modulus are algebraically simple. ∎

Next we prove that, for almost all $\alpha > 0$, the eigenvalues of B are algebraically simple. Moreover the set of $\alpha > 0$, for which there exists at least one eigenvalue which is not algebraically simple, does not contain a limit point, i.e., any such $\alpha > 0$ is necessarily isolated.

Let $F(\tau)$ be defined as

$$F(\tau) = G(\tau) + i\alpha S(\tau), \tag{A.37}$$

where

$$G(\tau) = -\tau(1 + \cosh\tau\cos\tau), \qquad S(\tau) = \sinh\tau\cos\tau - \cosh\tau\sin\tau. \tag{A.38}$$

We know that for a given $\alpha > 0$, $\lambda = i\tau^2$ is an algebraically simple eigenvalue of B if and only if $F(\tau) = 0,\ F'(\tau) \neq 0$, (see Lemma 3.2). Note that we have

$$F'(\tau) = G'(\tau) + i\alpha S'(\tau). \tag{A.39}$$

Also note that if for some $\alpha > 0$ and $\tau \in \mathbf{C}$ we have $F(\tau) = F'(\tau) = 0$, then by eliminating α in (A.37) and (A.39) we obtain $R(\tau) = 0$, where $R(\tau)$ is given by

$$R(\tau) = G'(\tau)S(\tau) - G(\tau)S'(\tau). \tag{A.40}$$

Note that $G(\tau) = 0$ and $S(\tau) = 0$ cannot be satisfied simultaneously. To see that, assume that for some $\tau \in \mathbf{C}$ we have $G(\tau) = S(\tau) = 0$. Then, since $\tau = 0$ is not an eigenvalue, from (A.38) we obtain $\cos\tau = -1/\cosh\tau$, $\sin\tau = -\sinh\tau/\cosh^2\tau$. Then, by using $\sin^2\tau + \cos^2\tau = 1$, we obtain $\cosh\tau = \pm 1$, and then (A.38) implies $\cos\tau = \mp 1$. It can now easily be shown that such a $\tau \in \mathbf{C}$ does not exist. Hence if $F(\tau) = 0$, then both $G(\tau) \neq 0$ and $S(\tau) \neq 0$ must be true.

LEMMA A.4. *Let, for $a > 0$, the sets $\mathcal{C}_a$ and $\mathcal{C}_\infty$ be defined as*

$$\mathcal{C}_a = \{\alpha \in \mathbf{R},\ 0 < \alpha < a | \exists \tau \in \mathbf{C}, F(\tau) = F'(\tau) = 0\}, \tag{A.41}$$

$$\mathcal{C}_\infty = \{\alpha \in \mathbf{R},\ \alpha > 0 | \exists \tau \in \mathbf{C}, F(\tau) = F'(\tau) = 0\}. \tag{A.42}$$

Then

i. *The set $\mathcal{C}_\infty$, if not empty, is at most countable.*

ii. *The set $\mathcal{C}_a$, if not empty, contains finitely many points.*

Proof. **i.** For some $\alpha > 0$ and $\tau \in \mathbf{C}$ we have $F(\tau) = F'(\tau) = 0$. Then we assume $R(\tau) = 0$, where $R(\tau)$ is given by (A.40). Since $R(\tau)$ is a nonconstant analytic function, it follows that its zero set (i.e., the roots of $R(\tau) = 0$) is at most countable; see, e.g., [16, p. 209, Thm. 10.18]. This also shows that the eigenvalues $\lambda = i\tau^2$ which are not algebraically simple also satisfy $R(\tau) = 0$, and hence are independent of α. From (A.37) we obtain

$$\alpha = i\frac{G(\tau)}{S(\tau)}. \tag{A.43}$$

Since there are countably many $\tau \in \mathbf{C}$ for which the eigenvalues $\lambda = i\tau^2$ are not algebraically simple, and since for these τ (A.43) is satisfied, it follows that there are at most countable many values for $\alpha > 0$ such that there exists at least one eigenvalue with algebraic multiplicity greater that one. Hence the set $\mathcal{C}_\infty$ is countable.

ii. Let $a > 0$ be given and let $0 < \alpha < a$. From Lemma A.3 we know that all eigenvalues with sufficiently large modulus are algebraically simple. Hence there exists a $M > 0$ such that, for all $0 < \alpha < a$ and for all eigenvalues $\lambda = i\tau^2$ which are not algebraically simple, we have $\tau \in B(0, M)$, defined by (A.3). Moreover (A.36) implies that $f'(\tau_k) \neq 0$ for sufficiently large k, uniformly with respect to α, for $0 < \alpha \leq a$.

This fact implies that the constant M is independent of α, for $0 < \alpha \leq a$. However such τ must also satisfy $R(\tau) = 0$, where $R(\tau)$ is given by (A.40). Since $B(0, M)$ is a compact set, the number of roots of $R(\tau) = 0$ in $B(0, M)$ must be finite, for otherwise there will be a limit point of zeros of $R(\tau)$ in $B(0, M)$, which is a contradiction; see, e.g., [16, p. 209, Thm. 10.18]. Since in $B(0, M)$ there are at most finitely many candidates of τ for eigenvalues which are not algebraically simple, it follows from (A.43) that the set $\mathcal{C}_a$ also contains finitely many points. □

The next corollary now proves assertion **ii** of Theorem 3.3.

COROLLARY A.5. **i.** *For almost all $\alpha > 0$ the eigenvalues of the operator B given by* (3.15) *are algebraically simple.*

ii. *If for some $\alpha_0 > 0$ and $\tau_0 \in \mathbf{C}$, $\lambda_0 = i\tau_0^2$ is an eigenvalue which is not algebraically simple, then there exists an open set $U \subset \mathbf{R}$ such that $\alpha_0 \in U$, and for $\alpha \in U$, $\alpha \neq \alpha_0$, the eigenvalues of B are algebraically simple.*

Proof. **i.** This follows easily from Lemma A.4.

ii. Note that the right-hand side of (A.43) is an analytic function around any possible $\tau \in \mathbf{C}$ such that the eigenvalue $\lambda = i\tau^2$ is not algebraically simple. Then the result follows from, e.g., [16, p. 216, Thm. 10.32], and from the fact that all eigenvalues with sufficiently large modulus are algebraically simple. □

To prove that the generalized eigenfunctions of B form a Riesz basis in $\mathcal{H}$, we need the following simple fact.

LEMMA A.6. *Let B be a densely defined closed linear operator in a Hilbert space $\mathcal{H}$. Assume that the spectrum of B consists entirely of, at most countable, isolated points, each of which has a finite algebraic multiplicity. Moreover assume that the eigenvalues are distinct. Then the generalized eigenfunctions are ω-linearly independent (for the definition of ω-independence, see, e.g.,*[7, *p.* 316], *or* [18, *p.* 50]).

Proof. Proof of this fact is essentially the same as given in [7, p. 329] for bounded operators. For closed (unbounded) operators with compact resolvent (discrete in the notation of [5]), we may proceed by using [8, p. 178] or [5, pp. 2292–2293] as follows. Let λ_n and ν_n denote the eigenvalues and their algebraic multiplicity of B, respectively. Let ψ_{ij}, $i = 1, 2, \ldots, n, \ldots$, $j = 1, \ldots, \nu_i$, denote the set of generalized eigenfunctions. Since the spectrum of A does not contain an accumulation point, for each λ_i we can find a constant $r_i > 0$ such that the circle $C_i = \{\lambda \in \mathbf{C} | \mid \lambda - \lambda_i \mid = r_i\}$ does not encircle any eigenvalue other than λ_i. It is well known that the operator

$$P_i = \frac{1}{2\pi i} \int_{C_i} (\lambda I - A)^{-1} d\lambda, \tag{A.44}$$

is well defined and is the projection operator onto the generalized eigenspace corresponding to λ_i; see, e.g., [8, p. 178], [5, pp. 2292–2293]. Now consider the following equation:

$$\sum_{i=1}^{\infty} \sum_{j=1}^{\nu_i} c_{ij} \psi_{ij} = 0. \tag{A.45}$$

By using the projection operator P_i given by (A.44), we obtain

$$P_i \left(\sum_{i=1}^{\infty} \sum_{j=1}^{\nu_i} c_{ij} \psi_{ij} \right) = \sum_{j=1}^{\nu_i} c_{ij} \psi_{ij} = 0. \tag{A.46}$$

Since $\nu_i < \infty$ and the generalized eigenfunctions are linearly independent, it follows from (A.46) that $c_{ij} = 0$, $j = 1, \ldots, \nu_i$. Since this is true for each $i \in \mathbf{N}$, it follows that the generalized eigenfunctions are ω-linearly independent. □

THEOREM A.7. *Let $\alpha > 0$ be given and assume the eigenvalues of the operator B are all algebraically simple (note that this condition holds for almost all $\alpha > 0$; see* Corollary A.5). *Then the set of eigenvectors of B forms a Riesz basis for $\mathcal{H}$.*

Proof. Let F_{kr} and G_{kr} be given by (A.24) and (A.28), respectively. Note that F_{kr} and G_{kr} are the (appropriately) normalized eigenvectors of the operator B, corresponding to given $\alpha > 0$ and $\alpha = 0$, respectively. We note that by Lemma A.1, it is possible to enumerate these eigenvectors similarly, and because of algebraic simplicity we consider only the eigenvectors and not the generalized eigenvectors. This point is important in Theorum 3.5 and Theorum 3.6 in proving the spectrum-determined growth property, which is our main aim.

From Theorum A.2 it follows that for some N we have

$$\sum_{|k|>N} \|F_{kr} - G_{kr}\|_{\mathcal{H}}^2 < \infty; \tag{A.47}$$

see (A.29). Since $N < \infty$, it follows that

$$\sum_{k\in\mathbf{Z}} \|F_{kr} - G_{kr}\|_{\mathcal{H}}^2 < \infty. \tag{A.48}$$

Hence the set of vectors $\{F_{kr}\}$ is quadratically close to the set of vectors $\{G_{kr}\}$. It is well known that the latter set of vectors forms a Riesz basis for $\mathcal{H}$, since for $\alpha = 0$ the operator B becomes a skew adjoint operator. Also by Lemma A.6, the former set of vectors is ω-linearly independent. This implies that the set of vectors $\{F_{kr}\}$ also forms a Riesz basis in $\mathcal{H}$; see, e.g., [18, p. 347, Thm. 11.3]. □

Remark 3. The requirement that the eigenvalues of B for $\alpha > 0$ be algebraically simple is not essential and could be relaxed. Let $\alpha > 0$, and let $\lambda \in \mathbf{C}$ be a root of (A.1), i.e., an eigenvalue of B. It is not known a priori whether the multiplicity of λ as a root of (A.1) and the algebraic multiplicity of λ as an eigenvalue of B are the same. Let us assume that these two multiplicities are the same, and let the set of vectors $\{F_{kr}\}$ include all eigenvectors and the generalized eigenvectors of B. Then by using Lemma A.1, Lemma A.3, Theorem A.2, and Theorem A.7, we conclude that the sets $\{F_{kr}\}$ and $\{G_{kr}\}$ are quadratically close; i.e., (A.48) holds. Hence the set $\{F_{kr}\}$ also forms a Riesz basis in $\mathcal{H}$, and the spectrum-determined growth property stated in Theorum 3.5 and Theorum 3.6 holds. The assumption on the equality of the multiplicities stated above seems to be true; however, the proof of this statement could be rather tedious. If we assume algebraic simplicity, which is generic (i.e., holds for almost all $\alpha > 0$), then these two multiplicities are the same; see Lemma 3.2. This is the basic reason for the assumption on algebraic simplicity.

COROLLARY A.8. *There exists an $a > 0$ such that, for all $0 < \alpha < a$, the set of eigenfunctions of B forms a Riesz basis for $\mathcal{H}$.*

Proof. This fact was proven in [4]. Here we may obtain this result as a corollary by using Lemma A.4, part ii, and Theorem A.7. □

Appendix B. The authors thank the anonymous referees, who suggested many improvements of the paper, and also Bo Peng Rao, who made valuable comments.

REFERENCES

[1] A. V. Balakrishnan, *Compensator design for stability enhancement with collocated controllers*, IEEE Trans. Automat. Control, 36 (1991), pp. 994–1008.

[2] G. Chen and J. Zhou, *The wave propagation method for the analysis of boundary stabilization of vibrating structures*, SIAM J. Appl. Math., 50 (1990), pp. 1254–1283.

[3] G. Chen, S. G. Krantz, D. W. Ma, C. E. Wayne, and H. H. West, *The Euler-Bernoulli beam equation with boundary energy dissipation*, in Operator Methods for Optimal Control Problems, S. J. Lee, ed., Marcell–Dekker, New York, 1987, pp. 67–96.

[4] F. Conrad, *Stabilization of beams by pointwise feedback control*, SIAM J. Control Optim., 28 (1990), pp. 423–438.

[5] N. Dunford and J. T. Schwartz, *Linear Operators*, Vol. 3, Wiley-Interscience, New York, 1971.

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

[7] I. C. Gohberg and M. G. Krein, *Introduction to the Theory of Nonselfadjoint Operators,* Trans. Math. Monogr. 18, AMS, Providence, RI, 1969.

[8] T. Kato, *Perturbation Theory for Linear Operators*, 2nd. ed., Springer-Verlag, New York, 1980.

[9] W. Littman and L. Markus, *Stabilization of a hybrid system of elasticity by feedback boundary damping*, Ann. di Mat. Pura ed Appl., 152 (1988), pp. 281–330.

[10] Ö. Morgül, B. P. Rao, and F. Conrad, *On the stabilization of a cable with a tip mass*, IEEE Trans. Automat. Control, 39 (1994), pp. 2140–2145.

[11] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer–Verlag, New York, 1983.

[12] P. Rideau, *Contrôle d'un assemblage de poutres flexibles par des capteurs-actionneurs ponctuels: étude du spectre du système.* Thèse, Ecole Nationale Supérieriure des Mines de Paris, Sophia-Antipolis, France, 1985.

[13] B. P. Rao, *Stabilization uniforme d'un système hybride en élasticité*, C. R. Acad. Sci. Paris Sér. I Math, 316 (1993), pp. 261–266.

[14] B. P. Rao, *Uniform stabilization of a hybrid system of elasticity*, SIAM J. Control Optim., 33 (1995), pp. 440–454.

[15] B. P. Rao, *Decay estimates of solutions for a hybrid system of flexible structures*, European J. Appl. Math., 4 (1993), pp. 303–319.

[16] W. Rudin, *Real and Complex Analysis*, McGraw-Hill, New York, 1966.

[17] D. L. Russell, *Decay rates for weakly damped systems in Hilbert space obtained with control theoretic methods*, J. Differential Equations, 19 (1975), pp. 344–370.

[18] I. Singer, *Bases in Banach Spaces,* Vol. 1, Springer-Verlag, New York, 1970.

[19] H. Tanabe, *Equations of Evolution*, Pitman, London, 1979.

[20] R. Triggiani, *Lack of uniform stabilization for noncontractive semigroups under compact perturbation*, Proc. Amer. Math. Soc., 105 (1989), pp. 375–383.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 6, pp. 1987–1998, November 1998

008

ON THE VALIDITY OF THE MAXIMUM PRINCIPLE AND OF THE EULER–LAGRANGE EQUATION FOR A MINIMUM PROBLEM DEPENDING ON THE GRADIENT*

ARRIGO CELLINA† AND STEFANIA PERROTTA‡

Abstract. We consider the limiting case $\alpha = \infty$ of the problem of minimizing

$$\int_\Omega (\|\nabla u(x)\|^\alpha + g(u))dx \text{ on } u \in u_0 + W_0^{1,\alpha}(\Omega),$$

where g is differentiable and strictly monotone. If this infimum is finite, it is evidently attained; we show that any minimizing function u satisfies the appropriate form of the Euler–Lagrange equation, i.e., for some function p,

$$\operatorname{div} p(x) = g'(u(x)) \quad \text{for} \quad p(x) \in \partial j_B(\nabla u(x)),$$

where j_B is the indicator function of the closed unit ball in the Euclidean norm of $\mathbb{R}^N$ and ∂ is the subdifferential of the convex function j_B.

Key words. extended valued functions, Euler–Lagrange equations, Hamilton–Jacobi control systems, Pontryagin maximum principle

AMS subject classification. 49K20

PII. S0363012996311319

1. Introduction.

In this paper we consider the problem of minimizing

$$\int_\Omega g(u)dx$$

for $u \in u_0 + W_0^{1,\infty}(\Omega)$, subject to the Hamilton–Jacobi control equation

$$\nabla u(x) = v, \qquad v \in B,$$

where B is the Euclidean unit ball of $\mathbb{R}^N$, i.e., $\{y \in \mathbb{R}^N : \|y\| \le 1\}$. By the convexity and compactness of the control set B, the minimization problem above admits a solution whenever the set of functions u satisfying the control and boundary conditions is nonempty. Under some assumptions on g (that include the linear case), but essentially without assumptions on Ω and on the boundary datum u_0, we show that to a solution u we can associate a map $p \in (L^1(\Omega))^N$ such that, denoting by H the map

$$H(u, p, v) = -g(u) + < p, v >,$$

we have

$$\nabla u(x) = \nabla_p H(x); \quad \operatorname{div} p(x) = -\frac{\partial H}{\partial u},$$

*Received by the editors October 28, 1996; accepted for publication (in revised form) October 30, 1997; published electronically August 31, 1998.
http://www.siam.org/journals/sicon/36-6/31131.html
†Dipartimento di Matematica, Universitá di Milano, Via Saldini 50, I-20133 Milano, Italia (cellina@elanor.mat.unimi.it).
‡Dipartimento di Matematica Pura ed Applicata "G. Vitali," Universitá degli Studi di Modena, Via Campi 213/B, I-41100 Modena, Italia (perrotta@c220.unimo.it).

and almost everywhere (a.e.) $H(u(x), p(x), v(x)) = \max_{w \in B}\{H(u(x), p(x), w)\}$, i.e., the solution satisfies the Pontryagin maximum principle [5].

Equivalently, the problem we consider can be seen as the problem of minimizing the functional

$$F(u) = \int_\Omega (j_B(\nabla u(x)) + g(u))dx$$

for $u \in u_0 + W_0^{1,\infty}(\Omega)$, where j_B is the indicator function of the closed unit ball B. The map $y \to j_B(y)$ is convex, lower semicontinuous, and extended valued. The coercivity requirement for the existence of solutions to the minimum problem is obviously satisfied; hence, when the functional F assumes a finite value for at least one function $u \in u_0 + W_0^{1,\infty}(\Omega)$, the minimization problem admits a solution. Even though the integrand is not differentiable, the convexity of the function $y \to j_B(\|y\|)$ leads one to expect the validity of a Euler–Lagrange inclusion in the form

$$\operatorname{div} p(x) = g'(u(x)) \quad \text{for} \quad p(x) \in \partial j_B(\nabla u(x)).$$

This inclusion can be reduced to the usual language of an equation noticing that

$$\partial j_B(y) = \begin{cases} \{0\}, & \text{if } \|y\| < 1, \\ \{\alpha y \,:\, \alpha \geq 0\}, & \text{if } \|y\| = 1, \\ \emptyset, & \text{if } \|y\| > 1. \end{cases}$$

Hence, establishing the validity of the above differential inclusion for an admissible function u amounts to providing a non-negative function $\alpha \in L^1(\Omega)$, with $\alpha(x) = 0$ when $\|\nabla u(x)\| < 1$, such that

$$\operatorname{div} \alpha(x)\, \nabla u(x) = g'(u(x)).$$

This is what we mean by the Euler–Lagrange equation for this problem; solutions to the above partial differential equation have to be understood in the standard distributional sense. We have that, for this problem, the two formulations of the necessary conditions, namely, the validity of the maximum principle or the validity of the Euler–Lagrange equations, are entirely equivalent. In fact, the following identities hold for $v \in B$:

$$\langle p, v\rangle = \max_{w \in B} \langle p, w\rangle \quad \Leftrightarrow \quad v \in \begin{cases} \{p/\,\|p\|\}, & \text{if } p \neq 0, \\ B, & \text{if } p = 0, \end{cases} \quad \Leftrightarrow \quad p \in \partial j_B(v).$$

It follows from our result, in particular, that whenever the functional F is finite along exactly one function (the boundary function u_0), then u_0 must be a solution to the Euler–Lagrange equation.

Although convex analysis is a guide to write the suitable form of the Euler–Lagrange equation, it is of no help in establishing its validity for this problem. In fact, the basic assumption needed for the applicability of the theory, namely the continuity of the map $\xi \in L^\beta(\Omega) \to \int_\Omega j_B(\|\xi(x)\|)dx$ ([2, Theorem 4.1, p. 59]), is violated in this case, no matter what β is.

In contrast with the usual approach, where regularity of the solution is obtained as a consequence of its being a solution to the Euler–Lagrange equation, in our case we must first prove some regularity of the solution in order to obtain from it the validity of the equation.

Finally, the minimization problem we consider can be seen as a limiting case for $\beta = \infty$ of the problem

$$\text{minimize } \int_\Omega (\|\nabla u(x)\|^\beta + g(u))dx \quad \text{ on } u \in u_0 + W_0^{1,\beta}(\Omega),\ 1 < \beta < \infty.$$

It is known (see [4]) that (under some suitable assumptions) solutions to these problems do exist and satisfy the Euler–Lagrange equation

$$\beta \operatorname{div}(\|\nabla u\|^{\beta-2}\nabla u) = g'(u).$$

Hence, our equation can be seen as a limiting case of the above equation.

Among several other results, a related problem, but with boundary condition identically zero, has been considered by Bhattacharya, Di Benedetto, and Manfredi [1].

2. Main results. We consider the functional

$$F(u) = \int_\Omega (j_B(\|\nabla u(x)\|) + g(u(x)))\, dx$$

and the problem (P) of minimizing $F(u)$ for $u \in u_0 + W_0^{1,\infty}(\Omega)$. We wish to prove the following result.

THEOREM 2.1. *Let Ω be an open bounded subset of $\mathbb{R}^N$, let $g : \mathbb{R} \to \mathbb{R}$ be differentiable and strictly monotonic. Let u_0 in $W^{1,\infty}(\Omega)$ be such that $F(u_0)$ is finite. Then the minimum in problem (P) is attained and any minimizing u is a distributional solution to the Euler–Lagrange inclusion*

$$\operatorname{div} p(x) = g'(u(x)) \quad \text{for} \quad p(x) \in \partial j_{[0,1]}(\|\nabla u(x)\|),$$

i.e., there exists a non-negative function $\alpha \in L^1(\Omega)$, with $\alpha(x) = 0$ when $\|\nabla u(x)\| < 1$ such that

$$\operatorname{div} \alpha(x)\, \nabla u(x) = g'(u(x))$$

in the sense of distributions.

The validity of the theorem is based in the monotonicity of g, needed to establish Lemma 2.3. To prove it we should consider separately the two cases, g increasing and g decreasing. We shall present the proof for the case g increasing. We shall use the notation: for $A \subset \mathbb{R}^N$, $\rho(x, A) = \inf_{y\in A}\{\|x - y\|\}$.

The following lemma is a first regularity result on the solution u.

LEMMA 2.2. *Under the same assumptions as in Theorem 2.1, let u be a solution to problem (P). Then for every x_0 and $r > 0$ such that $B_r(x_0)$ is contained in Ω, we have*

$$\sup\{u(x) - u(x_0) : \|x - x_0\| = r\} = r.$$

Proof. The map u must be, on $B_r(x_0)$, Lipschitzian of Lipschitz constant 1. Hence, the supremum above cannot be larger than r. Assume it is equal to ζr, with $\zeta < 1$. Let η be a Lipschitzian function such that

(i) $\eta(x_0) = -r$,

(ii) $\|\nabla\eta\| = 1$,

(iii) $\eta(x) = 0, x \in \Omega \setminus B_r(x_0)$.

Fix $\lambda \in (\zeta, 1)$ and consider the function

$$\eta_\lambda(x) = \lambda\eta(x) - (u(x) - u(x_0)) + \zeta r.$$

We have that: for $x \in \partial B_r(x_0)$, $\eta_\lambda(x) = \zeta r - (u(x) - u(x_0)) \geq 0$ while $\eta_\lambda(x_0) = -\lambda r + \zeta r < 0$. Call E the connected component of the set $\{\eta_\lambda \leq 0\}$ containing x_0 (the measure of E is positive). The map η_λ^- is then defined to be

$$\eta_\lambda^-(x) = \begin{cases} \eta_\lambda(x), & x \in E, \\ 0, & \text{elsewhere.} \end{cases}$$

We have that $\eta_\lambda^-(x) = 0$ for $x \in \partial B_r(x_0)$ and that $\eta_\lambda^-(x_0) < 0$; moreover,

$$\nabla\eta_\lambda^-(x) = \begin{cases} \lambda\nabla\eta(x) - \nabla u(x), & x \in E, \\ 0, & \text{elsewhere,} \end{cases}$$

so that $\|\nabla\eta_\lambda\| \leq 2$. It is our purpose to show that for parameters $t > 0$ sufficiently small, we have $\|\nabla u + t\nabla\eta_\lambda^-\| \leq 1$.

a) Consider first those $x \in E$ such that $\|\nabla u(x)\| > \frac{1+\lambda}{2}$. We have

$$\langle \nabla u, \nabla\eta_\lambda^- \rangle \leq \lambda\|\nabla u\|\|\nabla\eta\| - \|\nabla u\|^2 = \|\nabla u\|(\lambda - \|\nabla u\|) \leq \|\nabla u\|\frac{\lambda - 1}{2} < 0$$

and

$$|\langle \nabla u, \nabla\eta_\lambda^- \rangle| = -\langle \nabla u, \nabla\eta_\lambda^- \rangle \geq \|\nabla u\|\frac{1-\lambda}{2} \geq \frac{1-\lambda^2}{4}.$$

Hence, for $t \in (0, \frac{1-\lambda^2}{8})$ and a.e. $x \in E$, we have that $2|\langle \nabla u, \nabla\eta_\lambda^- \rangle| > 4t > t\|\nabla\eta_\lambda^-\|^2$. Since

$$\|\nabla u + t\nabla\eta_\lambda^-\|^2 = \|\nabla u\|^2 + t^2\|\nabla\eta_\lambda^-\|^2 + 2t\langle \nabla u, \nabla\eta_\lambda^- \rangle,$$

we obtain

$$\begin{aligned} \|\nabla u + t\nabla\eta_\lambda^-\|^2 &= \|\nabla u\|^2 + t(t\|\nabla\eta_\lambda^-\|^2 + 2\langle \nabla u, \nabla\eta_\lambda^- \rangle) \\ &\leq 1 + t(t\|\nabla\eta_\lambda^-\|^2 - 2|\langle \nabla u, \nabla\eta_\lambda^- \rangle|) < 1. \end{aligned}$$

b) Consider now those $x \in E$ such that $\|\nabla u(x)\| \leq \frac{1+\lambda}{2}$. Then, for t in $(0, \frac{1-\lambda^2}{8})$, we simply have

$$\|\nabla u + t\nabla\eta_\lambda^-\| \leq \|\nabla u(x)\| + t\|\nabla\eta_\lambda^-\| < \frac{1+\lambda}{2} + \frac{1-\lambda}{4}2 = 1.$$

Hence, from the above, the variation η_λ^- is admissible, in the sense that a.e. in Ω, for all t sufficiently small,

$$\|\nabla u + t\nabla\eta_\lambda^-\| \leq 1.$$

For one such t, since: $j_B(\|\nabla u + t\nabla\eta_\lambda^-\|) = 0$, a.e. in Ω; $u + t\eta_\lambda^- \leq u$, a.e. in Ω; $u + t\eta_\lambda^- < u$, a.e. in E, we have

$$F(u + t\eta_\lambda^-) = \int_\Omega g(u(x) + t\eta_\lambda^-(x))\,dx < \int_\Omega g(u(x))\,dx = F(u),$$

a contradiction. The strict monotonicity of g is essential in this step. □

As a simple consequence of the previous lemma, we have the following result.

LEMMA 2.3. *Under the same assumptions as in Lemma* 2.2*, for a.e.* $x \in \Omega$*, we have* $\|\nabla u(x)\| = 1$*; there exist at least one direction* d^x *and a related interval* $[0, b^x)$ *such that for* $\lambda \in [0, b^x), u(x + \lambda d^x) - u(x) = \lambda$ *and* $x + b^x\, d^x \in \partial\Omega$.

Proof. From Lemma 2.2, it follows that to any point $x \in \Omega$ we can associate at least one unit vector (a direction) d^x and (at least) one nonvanishing interval $[0, l)$ such that for $t \in [0, l)$, $u(x + td^x) - u(x) = t$. Given x and d^x, let b^x be such that $[0, b^x)$ is the largest such interval. Let y be on the closure of this segment. When y is in Ω, by the previous lemma we can associate to it at least one direction d^y with the property stated above. This direction d^y must coincide with d^x. Otherwise, choose $x_1 = x + \lambda d^x$ with $\lambda < b^x$ and sufficiently close to it, and choose $y_1 = y + \mu d^y$ with μ positive and sufficiently small, so that the segment from x_1 to y_1 is contained in Ω. Then u is defined on this segment and $u(y_1) - u(x_1) = (u(y_1) - u(y)) + (u(y) - u(x_1)) = \|y_1 - y\| + \|y - x_1\| > \|y_1 - x_1\|$, a contradiction to the fact that u, a solution to the minimum problem, is Lipschitzian with constant 1. In particular, $x + b^x d^x$ must be in $\partial\Omega$ otherwise we would contradict the maximality of b^x. □

Remarks. i) The set of the directions $\{d^x\}$ gives rise to a multivalued map $x \to D(x)$.

ii) From the proof of the above lemma, in particular, we infer that d^y is unique whenever there exist $x \in \Omega$, a direction d^x and t in the interval $(0, b^x)$ such that $y = x + t\, d^x$.

iii) For fixed x and d^x, call (a^x, b^x) the largest open interval such that $u(x + t_1 d^x) - u(x + t_2 d^x) = t_2 - t_1$ for $t_2 > t_1$ and t_1 and t_2 in (a^x, b^x). Whenever x belongs to $S(x) = \{x + td^x : t \in (a^x, b^x)\}$, i.e., when $a^x < 0$, we have that d^x is unique and that a^x and b^x depend only on x, i.e., we can consider the univalent maps $x \to d(x) = d^x, x \to a(x) = a^x$, and $x \to b(x) = b^x$. It will be convenient to set

$$\mathcal{S} = \cup_{\{x \in \Omega\}} S(x).$$

Proof of Theorem 2.1. Let u be a solution to the minimum problem. We have to define a function α with the properties stated in Theorem 2.1 such that for every ϕ in $C_0^\infty(\Omega)$ we have

$$\int_\Sigma \alpha(x) \langle \nabla u(x), \nabla \phi(x) \rangle\, dx + \int_\Sigma g'(u(x)) \phi(x)\, dx = 0.$$

Step a) For k in $\{1, \dots N\}$, let d_k denote the kth component of the vector d. About the properties of the map $x \to d(x)$, we have the following claim, a first regularity result on ∇u.

Claim 2.1. Fix $k \in \{1, \dots, N\}$ and $\varepsilon > 0$. On

$$E_\varepsilon^k = \{x \in \mathcal{S} : (x - \varepsilon d(x), x + \varepsilon d(x)) \subset S(x);\ d_k(x) \geq \frac{1}{\sqrt{N}};\ \rho(x, \partial\Omega) \geq 3\varepsilon\},$$

the map $x \to d(x)$ is Lipschitzian of constant $\frac{2\sqrt{N}}{8}$.

Proof of Claim 2.1. Consider two points P and P' in E_ε^k and set $d = d(P)$, $d' = d(P')$. In the case $\|P - P'\| \geq \frac{\varepsilon}{2\sqrt{n}} \|d - d'\|$, we have

$$\|d(P) - d(P')\| \leq \frac{2\sqrt{N}}{\varepsilon} \|P - P'\|.$$

Hence, we consider the case $\|P-P'\| < \frac{\varepsilon}{2\sqrt{N}}|d-d'|$. Set r to be $\{P+\lambda d : \lambda \in \mathbb{R}\}$ and r' to be $\{P'+\lambda d' : \lambda \in \mathbb{R}\}$. Let $O \in r$ and $O' \in r'$ be the two points of minimal distance for r and r'; then $\langle O'-O, d\rangle = \langle O'-O, d'\rangle = 0$. When $O \neq O'$ we shall refer to the unique three-dimensional space containing r and r' (the case $O = O'$ being similar and simpler). On the plane orthogonal to $O'-O$ and containing r, let r'' be the projection of the line r'. Also let P^* be the nearest point to P on r'', so that $\|P-P'\| \geq \|P-P^*\|$. The point P'' on r'' is defined to be the point having $\|P-O\| = \|P''-O\|$ and lying on the same side (with respect to 0) as P^*. By elementary geometry we have

$$\frac{\|P-P''\|}{\|P-O\|} = \frac{\|d-d'\|}{1}.$$

Consider the triangle O, P, P'' and let H be $\frac{1}{2}P + \frac{1}{2}P''$. We obtain

$$\frac{\|P-P^*\|}{\|P-P''\|} = \frac{\|H-O\|}{\|P''-O\|}.$$

Since, by the definition of E^k_ε, we have that $\frac{\|H-O\|}{\|P''-O\|} \geq \frac{1}{\sqrt{N}}$, we obtain

$$\|P-P''\| = \|P-P^*\|\frac{\|P''-O\|}{\|H-O\|} \leq \sqrt{N}\|P^*-P\|$$

so that

$$\|P-O\| = \frac{\|P-P''\|}{\|d-d'\|} \leq \sqrt{N}\frac{\|P-P^*\|}{\|d-d'\|} \leq \sqrt{N}\frac{\|P-P'\|}{\|d-d'\|} \leq \frac{\varepsilon}{2}.$$

For symmetry reasons, also $\|P'-O'\| \leq \frac{\varepsilon}{2}$. Hence, we have obtained that both O and O' are in Ω, and u is, therefore, defined at O and O'. At this point we are free to assume that $u(O) \geq u(O')$.

Let A and D be the extremes of a segment on $S(P)$ centered on O and of half-length $\frac{\varepsilon}{2}$ and B', C' be the same on $S(P')$ with respect to O'. We have $\|P-A\| \leq \varepsilon$, $\|P-D\| \leq \varepsilon$, and

$$\|P-B'\| \leq \|B'-O'\| + \|O'-P'\| + \|P'-P\| \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} + \frac{\varepsilon}{\sqrt{N}} \leq 2\varepsilon, \quad \|P-C'\| \leq 2\varepsilon.$$

Therefore, all the points A, D, B', and C' lie in the ball $B_{2\varepsilon}(P) \subset \Omega$. On this set, the map u is Lipschitzian of Lipschitz constant 1. Let B and C be the projections of the points B' and C' on the plane orthogonal to $O'-O$ and containing O. We can assume that

$$u(D)-u(A) = \|D-A\| \quad \text{and} \quad u(C')-u(B') = \|C'-B'\| = \|C-B\|.$$

Hence, we have

$$\begin{aligned} \|B'-D\| &\geq u(D)-u(B') = u(D)-u(O)+u(O)-u(O')+u(O')-u(B') \\ &= \|D-O\| + u(O)-u(O') + \|B'-O'\| \geq \|D-O\| + \|B'-O'\|, \end{aligned} \tag{1}$$

while, on the other hand,

$$\begin{aligned} \|B'-D\|^2 &= \|B'-B\|^2 + \|B-D\|^2 \\ &= \|O'-O\|^2 + \|B-O\|^2 + \|O-D\|^2 + 2\langle O-D, B-O\rangle. \end{aligned} \tag{2}$$

By (1) and (2) we obtain

$$\|O'-O\|^2+\|B-O\|^2+\|O-D\|^2+2\langle O-D,B-O\rangle \\ \geq \|D-O\|^2+\|B'-O'\|^2+2\|D-O\|\|B'-O'\|;$$

hence,

$$\|O'-O\|^2 \geq 2\|D-O\|\,\|B-O\|\left(1-\left\langle\frac{O-D}{\|D-O\|},\frac{B-O}{\|B-O\|}\right\rangle\right) \\ =\|D-O\|\,\|B'-O'\|(2-2\langle d,d'\rangle)=\left(\frac{\varepsilon}{2}\right)^2\|d-d'\|^2.$$

It follows then that $\|d-d'\|\leq\frac{2}{\varepsilon}\|O-O'\|\leq\frac{2}{\varepsilon}\|P-P'\|$. Hence, we have

$$\|d(P)-d(P')\|\leq\frac{2\sqrt{N}}{\varepsilon}\|P-P'\|$$

for every P and P' in E^k_ε. This proves the claim. □

Step b) The purpose of this step is to define a countable partition of $\mathcal{S}$ consisting of measurable sets.

Consider the set $\mathcal{P}$ of pairs (p,q), where p and q are integers and q is positive, and let $\sigma:\mathbb{N}\to\mathcal{P}$ be a numbering of this set. Denote by (p_n,q_n) the image $\sigma(n)$. To $n\in\mathbb{N}$ and $k\in 1,\dots,N$, we associate the two disjoint sets

$$E_n^{+,k}=\left\{y\in S(x):x_k=\frac{p_n}{q_n};\ d_k(x)=\sup_{1\leq i\leq N}|d_i(x)|;\ \rho(x,\partial\Omega)\geq\frac{3}{q_n}\right. \\ \left.\text{and } x-\frac{1}{q_n}d(x),x+\frac{1}{q_n}d(x)\in S(x)\right\}$$

and

$$E_n^{-,k}=\left\{y\in S(x):x_k=\frac{p_n}{q_n};\ d_k(x)=-\sup_{1\leq i\leq N}|d_i(x)|;\ \rho(x,\partial\Omega)\geq\frac{3}{q_n}\right. \\ \left.\text{and } x-\frac{1}{q_n}d(x),x+\frac{1}{q_n}d(x)\in S(x)\right\}.$$

In order to obtain a partition of $\mathcal{S}$ we operate in the standard way. Set $\Sigma_1^{+,1}=E_1^{+,1}$ and, in general, $\Sigma_1^{+,k+1}=E_1^{+,k+1}\backslash\{\cup_{i=1,\dots,k}\Sigma_1^{+,i}\}$. Set

$$\sum_{n+1}^{+,1}=E_{n+1}^{+,1}\setminus\left\{\bigcup_{i=1,\dots,N;m=1,\dots,n}\sum_m^{+,i}\right\}$$

and

$$\sum_{n+1}^{+,k+1}=E_{n+1}^{+,k+1}\setminus\left\{\left(\bigcup_{i=1,\dots,N;m=1,\dots,n}\sum_m^{+,i}\right)\bigcup\left(\bigcup_{i=1,\dots,k}\sum_{n+1}^{+,i}\right)\right\}.$$

An analogous procedure is applied to the family $E_n^{-,k}$ to yield the disjoint family $\{\Sigma_n^{-,k}\}$. This second family is defined so as to be disjoint from $\{\Sigma_n^{+,k}\}$ as well.

We have defined a disjoint family. We wish to show that it covers $\mathcal{S}$.

Claim 2.2. $\mathcal{S} = \cup_{k=1,\dots,N;n\in\mathbb{N}}(\Sigma_n^{+,k} \cup \Sigma_n^{-,k})$.

Proof of Claim 2.2. We have only to show that

$$\bigcup_{k=1,\dots,N;n\in\mathbb{N}} \left(\sum_n^{+,k} \bigcup \sum_n^{-,k} \right) \supset \mathcal{S}.$$

Since

$$\bigcup_{k=1,\dots,N;n\in\mathbb{N}} \left(\sum_n^{+,k} \bigcup \sum_n^{-,k} \right) = \bigcup_{k=1,\dots,N;n\in\mathbb{N}} \left(E_n^{+,k} \bigcup E_n^{-,k} \right),$$

we have to show that, for every $x \in \Omega$, $S(x)$ is contained in the set at the right-hand side. Let $x \in S(x')$ for some $x' \in \Omega$. There exists a k such that either $d_k(x) = \sup_{i=1,\dots,N} |d_i(x)|$ or $d_k(x) = -\sup_{i=1,\dots,N} |d_i(x)|$. Let us consider the first case (the other being analogous). Call $\Delta = \rho(x, \partial\Omega)$. Since $S(x')$ is an open interval, there exists δ, $0 < \delta < \frac{\Delta}{2}$, such that

$$\{x + \lambda d(x) : -\delta \le \lambda \le \delta\} \subset S(x').$$

Let q be a positive integer such that $1/q < \delta/(2\sqrt{N})$; there exists p such that $|p/q - x_k| \le 1/2q$. The point $y = x + (\frac{p}{q} - x_k)/d_k(x)d(x)$ has the following properties: its kth component y_k equals p/q; recalling that $d_k(x) \ge 1/\sqrt{N}$, we have that $\|x - y\| = \left|(\frac{p}{q} - x_k)/d_k(x)\right| \le \sqrt{N}/2q < \delta/4$. As a consequence, an interval (on $S(x')$) centered at y and of half-length $1/q$ is contained in $S(y)$ $(= S(x'))$ and contains x. Moreover, $\rho(y, \partial\Omega) \ge \Delta - \delta/4 \ge 7/4\delta \ge 3/q$. Hence, setting $n = \sigma^{-1}(p, q)$, we have $x \in E_n^{+,k}$. This proves Claim 2.2. □

Claim 2.3. The measure of $\Omega\backslash\mathcal{S}$ equals zero.

Proof of Claim 2.3. Since the subset of Ω of those points where u is not differentiable is of measure zero, it is enough to show that the subset of $\Omega\backslash\mathcal{S}$ where u is differentiable is of measure zero. In particular, for x in such a set, we can assume that there exists a unique vector d^x, as defined in Step a); otherwise we would contradict the differentiability at x.

Since $\cup_{k,n}(E_n^{+,k} \cup E_n^{-,k}) = \mathcal{S}$, we shall prove that $m(\Omega\backslash \cup_{k,n} (E_n^{+,k} \cup E_n^{-,k})) = 0$. Assume, on the contrary, that this set is of positive measure and let x_0 be a point of density of it. As it easy to see, the map $x \to D(x)$ as defined in Step a) is upper semicontinuous. In fact, it has a closed graph and its range is contained in the compact set B. Then a well-known criterion for upper semicontinuity applies. It follows then that for every ε there exists δ such that $\|d^x - d(x_0)\| < \varepsilon$ for $\|x - x_0\| < \delta$ and $d^x \in D(x)$. By changing coordinates we shall assume $x_0 = 0$ and $d_N(x_0) = 1$.

Let us consider the (family of) sets $Q_\ell = \{x : 0 \le |x_i| \le \ell,\ i = 1, \dots, N\}$ and let us choose ℓ so small that, for every $x \in Q_\ell$, we have:

i) $\rho(x, \partial\Omega) \ge \ell$,

ii) $d_N(x) \ge \max\{\frac{1}{\sqrt{N}}, \frac{2}{\sqrt{5}}\}$.

Let us consider the subset of Q_ℓ defined by

$$I_\ell = \left\{ x : x_N = 0 \text{ and } |x_i| \le \frac{\ell}{2},\ i = 1, \dots, N-1 \right\}.$$

The $(N-1)$-dimensional measure of I_ℓ is ℓ^{N-1}, while the N-dimensional measure of Q_ℓ is $(2\ell)^N$. Fix t, $\frac{\ell}{3} \le t \le \frac{2}{3}\ell$ and consider, on the hyperplane $\{x_N = t\}$, the set

$$P_t = \left\{ x + t\frac{d(x)}{d_N(x)} : x \in I_\ell \text{ and } d \in D(x) \right\}.$$

Every point y in this set is interior to $S(y)$; $D(y) = d(y)$ so that it is possible to define the map $F_t : P_t \to I_\ell$ defined by

$$F_t(y) = y - t\frac{d(y)}{d_N(y)}.$$

Notice that P_t is contained in E^N_ε as defined in Claim 2.1, with $\varepsilon = \ell/3$; hence the restriction of d to P_t is Lipschitzian with constant $(6\sqrt{N})/\ell$. For y and y' in P_t, we have

$$\begin{aligned}\left\|\frac{d(y)}{d_N(y)} - \frac{d(y')}{d_N(y')}\right\| &\le \frac{\|d(y) - d(y')\|}{d_N(y)} + \frac{\|d(y')\|}{d_N(y)d_N(y')}|d_N(y) - d_N(y')| \\ &\le \sqrt{N}\frac{6\sqrt{N}}{\ell}\|y - y'\| + N\frac{6\sqrt{N}}{\ell}\|y - y'\| \le \frac{12N\sqrt{N}}{\ell}\|y - y'\|.\end{aligned}$$

Hence, the map F_t is Lipschitzian with constant $1 + t\frac{12N\sqrt{N}}{\ell} \le 1 + 8N\sqrt{N}$.

Considering the $(N-1)$-dimensional measure of a subset A of P_t, we have then $m(F_t(A)) \le (1 + 8N\sqrt{N})m(A)$. Hence,

$$m(P_t) \ge \frac{m(F_t(P_t))}{1 + 8N\sqrt{N}} = \frac{m(I_\ell)}{1 + 8N\sqrt{N}} = \frac{\ell^{N-1}}{1 + 8N\sqrt{N}}.$$

The set $\cup_{\ell/3 \le t \le 2\ell/3} P_t$ is contained in $\mathcal{S}$ and, by Fubini's theorem, its N-dimensional measure is at least $\ell^N/(3 + 24N\sqrt{N})$, a fixed fraction of the total measure of Q_ℓ. Hence, x_0 cannot be a point of density. This proves Claim 2.3. □

Claim 2.4. For every $k = 1, \dots, N$, for every n, the sets $\Sigma^{\pm,k}_n$ are measurable.

From now up to Step d) we shall fix a choice of either $+$ or $-$, of k and of n. Hence, for simplicity's sake, we will drop $\pm, n, k$ and simply denote $E^{\pm,k}_n$ by E and $\Sigma^{\pm,k}_n$ by Σ. We shall denote by $\hat{x}$ the $(N-1)$-dimensional vector $(x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_N)$. It is convenient to set $\hat{E}$ to be the subset of $\mathbb{R}^{N-1}$ defined by $\hat{E} = \{\hat{x} : x \in E \cap \{x : x_k = p_n/q_n\}\}$, and analogously for $\hat{\Sigma}$. Consider $(\hat{x}, t)$, $\hat{x} \in \hat{E}, a(\hat{x}) < t < b(\hat{x})$ and define the map

$$\Xi(\hat{x}, t) = x + td(x).$$

This map is uniformly Lipschitz continuous.

Proof of Claim 2.4. As it is easy to see, both the maps $a(\hat{x})$ and $b(\hat{x})$ are lower semicontinuous on $\hat{E}$, and $\hat{E}$ can be described as the intersection of a closed set with the counterimages through a and b of the interval $[-1/q_n, 1/q_n]$; hence it is a measurable set. The subset of $\mathbb{R}^N$ described by $\{(\hat{x}, t) : \hat{x} \in \hat{E}; a(\hat{x}) < t < b(\hat{x})\}$ is measurable and so is E, its image through the Lipschitz continuous map Ξ. Similarly for E. It follows that Σ is measurable. This proves Claim 2.4. □

Step c) We wish to study the properties of the maps $\Xi(\hat{x}, t)$ defined above and of $J\Xi(\hat{x}, t)$.

For a.e. $(\hat{x}, t) \in (\Xi)^{-1}(\Sigma)$, we have that $\nabla\Xi$ exists and a computation shows that it can be obtained as follows. Consider the $N \times N$ matrix $\nabla d(x)$ and form the matrix $I + t\nabla d(x)$. Replace the kth column by the components of $d(x)$, and compute it by setting the kth component of x to be p_n/q_n and the other components to be the components of $\hat{x}$. This is the matrix $\nabla\Xi$. Hence, $J\Xi$ is uniformly bounded on Σ. It is also a.e. different from zero. In fact, differentiating the identity $\|d(x)\| = 1$, we obtain

$(\nabla d(x))\, d(x) = 0$, so that $(I + t\nabla d(x))\, d(x) = d(x)$. By Cramer's rule and the above computation of $\nabla \Xi$, we obtain

$$d_k(x) = \pm \frac{J\Xi(\hat{x}, t)}{det(I + t\nabla d(x))}.$$

Since (on Σ), $|d_k| > \frac{1}{\sqrt{N}}$, we finally have $J\Xi(\hat{x}, t) \neq 0$.

We wish to define a map α and prove it is in $L^1(\Sigma)$. Define first the map β on $(\Xi)^{-1}(\Sigma)$ setting,

$$\beta(\hat{x}, t) = \frac{1}{J\Xi(\hat{x}, t)} \int_{a(\hat{x})}^{t} g'(u(\Xi(\hat{x}, s)))\, J\Xi(\hat{x}, s)\, ds.$$

For $x \in \Sigma$ define α as

$$\alpha(x) = \beta((\Xi)^{-1}(x)).$$

Claim 2.5. $\alpha \in L^1(\Sigma)$.

Proof of Claim 2.5. We recall the change of variables formula ([3, Theorem 2, p. 99]) that states for a function $v \in L^1$ and an invertible and Lipschitzian transformation Ξ, we can write

$$\int v(\Xi(\hat{x}, t)\, J\Xi(\hat{x}, t)\, d(\hat{x}, t) = \int v(x)\, dx.$$

By this formula we obtain

$$\begin{aligned}\int_{\Sigma} g'(u(x))\, dx &= \int_{(\Xi)^{-1}(\Sigma)} g'(u(\Xi(\hat{x}, t)))J\Xi(\hat{x}, t)\, d(\hat{x}, t) \\ &= \int_{\hat{\Sigma}} \left(\int_{a(\hat{x})}^{b(\hat{x})} g'(u(\Xi(\hat{x}, t)))J\Xi(\hat{x}, t)\, dt \right) d\hat{x}.\end{aligned}$$

Similarly, by the change of variables formula and applying the definitions of α and β, we have

$$\begin{aligned}\int_{\Sigma} \alpha(x)\, dx &= \int_{(\Xi)^{-1}(\Sigma)} \beta(\hat{x}, t)J\Xi(\hat{x}, t)\, d(\hat{x}, t) = \int_{\hat{\Sigma}} \left(\int_{a(\hat{x})}^{b(\hat{x})} \beta(\hat{x}, t)J\Xi(\hat{x}, t)\, dt \right) d\hat{x} \\ &= \int_{\hat{\Sigma}} \left(\int_{a(\hat{x})}^{b(\hat{x})} \int_{a(\hat{x})}^{t} g'(u(\Xi(\hat{x}, s)))J\Xi(\hat{x}, s)\, ds\, dt \right) d\hat{x}.\end{aligned}$$

Integrating by parts we obtain that

$$\int_{a(\hat{x})}^{b(\hat{x})} \int_{a(\hat{x})}^{t} g'(u(\Xi(\hat{x}, s)))J\Xi(\hat{x}, s)\, ds\, dt = -\int_{a(\hat{x})}^{b(\hat{x})} (t - b(\hat{x}))g'(u(\Xi(\hat{x}, t)))J\Xi(\hat{x}, t)\, dt.$$

Hence,

$$\begin{aligned}\int_{\Sigma} \alpha(x)\, dx &\leq \int_{\hat{\Sigma}} \operatorname{diam}(\Omega) \left(\int_{a(\hat{x})}^{b(\hat{x})} g'(u(\Xi(\hat{x}, t)))J\Xi(\hat{x}, t)\, dt \right) d\hat{x} \\ &= \operatorname{diam}(\Omega) \int_{\Sigma} g'(u(x))\, dx. \qquad \square\end{aligned}$$

Step d) Since the sets $\Sigma_n^{\pm,k}$ are disjoint and (with the addition of a null set) form a partition of Ω, α is actually defined a.e. on Ω and by adding the previous inequalities over $+$ and $-$ and all k and n, we have that $\alpha \in L^1(\Omega)$.

Setting $p(x) = \alpha(x)\nabla u(x)/(\|\nabla u(x)\|)$, we want to show that the pair $(u(x), p(x))$ is a solution to the Euler–Lagrange equation for the minimization problem (P).

Fix arbitrarily ϕ in $C_0^\infty(\Omega)$ and consider

$$\int_\Omega \alpha(x)\langle \nabla u(x), \nabla\phi(x)\rangle \, dx.$$

Since $\|\langle \nabla u(x), \nabla\phi(x)\rangle\|$ is bounded, the integrand is in $L^1(\Omega)$ and the integral over Ω is the sum of the integrals over $\Sigma_n^{\pm,k}$. We fix one such $\Sigma_n^{\pm,k}$ that we denote by Σ and recall the corresponding notations introduced in Step b). Recall that, by the definition of Ξ and the properties of ∇u,

$$\frac{\partial}{\partial t}\Xi(\hat{x}, t) = \nabla u(\Xi(\hat{x}, t)),$$

independent of t. Hence,

$$\frac{\partial}{\partial t}\phi(\Xi(\hat{x}, t)) = \langle \nabla u(\Xi(\hat{x}, t)), \nabla\phi(\Xi(\hat{x}, t))\rangle.$$

By the change of variables formula and the definition of α, we have

$$\begin{aligned}\int_\Sigma \alpha(x)\langle \nabla u(x), \nabla\phi(x)\rangle \, dx &= \int_{\hat{\Sigma}} \left(\int_{a(\hat{x})}^{b(\hat{x})} \alpha(\Xi(\hat{x}, t)) \frac{\partial}{\partial t}\phi(\Xi(\hat{x}, t)) \, J\Xi(\hat{x}, t) dt \right) d\hat{x} \\ &= \int_{\hat{\Sigma}} \left(\int_{a(\hat{x})}^{b(\hat{x})} \frac{\partial}{\partial t}\phi(\Xi(\hat{x}, t)) \int_{a(\hat{x})}^{t} g'(u(\Xi(\hat{x}, s))) \, J\Xi(\hat{x}, s) \, ds \, dt \right) d\hat{x}.\end{aligned}$$

Integrating by parts we have

$$\begin{aligned}\int_\Sigma \alpha(x)\langle \nabla u(x), \nabla\phi(x)\rangle \, dx = \int_{\hat{\Sigma}} \Bigg(&\left[\phi(\Xi(\hat{x}, t)) \int_{a(\hat{x})}^{t} g'(u(\Xi(\hat{x}, s))) \, J\Xi(\hat{x}, s) \, ds \right]_{a(\hat{x})}^{b(\hat{x})} \\ &- \int_{a(\hat{x})}^{b(\hat{x})} g'(u(\Xi(\hat{x}, t)))\phi(\Xi(\hat{x}, t)) \, J\Xi(\hat{x}, t) \, dt \Bigg) d\hat{x}.\end{aligned}$$

The first term at the right-hand side is zero since $\Xi(\hat{x}, b(\hat{x}))$ belongs to $\partial\Omega$.

In the same way we compute $\int_\Sigma g'(u(x))\phi(x) \, dx$. We have

$$\int_\Sigma g'(u(x))\phi(x) \, dx = \int_{\hat{\Sigma}} \left(\int_{a(\hat{x})}^{b(\hat{x})} g'(u(\Xi(\hat{x}, t)))\phi(\Xi(\hat{x}, t)) \, J\Xi(\hat{x}, t) \, dt \right) d\hat{x}.$$

Hence,

$$\int_\Sigma \alpha(x)\langle \nabla u(x), \nabla\phi(x)\rangle \, dx + \int_\Sigma g'(u(x))\phi(x) \, dx = 0$$

for every $\Sigma_n^{\pm,k}$, hence the same is true on Ω. The pair $(p(x), u(x))$, where $p = \alpha\nabla u/(\|\nabla u\|)$, is a distributional solution to the differential inclusion

$$\operatorname{div} p(x) = g'(u(x)) \quad \text{for} \quad p(x) \in \partial j_{[0,1]}(\|\nabla u(x)\|). \qquad \square$$

REFERENCES

[1] T. Bhattacharya, E. Di Benedetto, and J. Manfredi, *Limits as $t \to \infty$ of $\Delta_p u_p = f$ and related extremal problems. Some topics in linear PDEs*, Rend. Sem. Mat. Univ. Politec. Torino, 1989.

[2] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, North–Holland, Amsterdam, 1976.

[3] L.C. Evans and R.F. Gariepy, *Measure Theory and Fine Properties of Functions*, CRC Press, Boca Raton, FL, 1992.

[4] M. Giaquinta, *Multiple Integrals in the Calculus of Variations and Nonlinear Elliptic Systems*, Princeton University Press, Princeton, NJ, 1983.

[5] L. Pontryagin, V. Boltyanskii, R. Gamkrelidze, and E. Mishchenko, *The Mathematical Theory of Control Processes*, Interscience, New York, 1962.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 6, pp. 1999–2019, November 1998

009

A FEASIBLE DIRECTIONS ALGORITHM FOR OPTIMAL CONTROL PROBLEMS WITH STATE AND CONTROL CONSTRAINTS: CONVERGENCE ANALYSIS*

R. PYTLAK† AND R. B. VINTER†

Abstract. In this paper we describe an optimization algorithm for the computation of solutions to optimal control problems with control, state, and terminal constraints. Inequality and equality constraints are dealt with by means of feasible directions and exact penalty approaches, respectively. We establish a general convergence property of the algorithm which makes no reference to the existence of accumulation points; in this analysis the compactness of the space of relaxed controls is used only to guarantee boundedness of the sequence of penalty parameters. We also demonstrate that relaxed accumulation points of sequences generated by the algorithm satisfy standard first-order necessary conditions of optimality. The algorithm contains a number of computation saving features, including an ε-active strategy for dealing with the "infinite dimensional" inequality constraints. Our convergence analysis provides techniques for studying the convergence properties of related optimization algorithms in which direction-finding subproblems involve the approximation of directional derivatives of the Chebyshev functional associated with state constraints. A companion paper provides details of implementation and numerical examples.

Key words. optimal control, state constrained problems, necessary optimality conditions, numerical algorithms

AMS subject classifications. 49M10, 49J15

PII. S0363012996297649

1. Introduction.

This paper concerns a new first-order, feasible directions algorithm for the solution of the following optimal control problem with pathwise state inequality constraints, labeled **(P)**:

$$\begin{aligned} &\min_u \phi(x(1)) \\ &\text{s.t. } \dot{x}(t) = f(t, x(t), u(t)), \text{ a.e. on } [0,1],\ x(0) = x_0, \\ &u(t) \in \Omega \ \text{ a.e. on } [0,1], \\ &h_i^1(x(1)) = 0\ \forall i \in E, \\ &h_j^2(x(1)) \le 0\ \forall j \in I, \\ &q(t, x(t)) \le 0\ \forall t \in [0,1], \end{aligned} \tag{1.1}$$

expressed in terms of the data: finite sets of index values E, I, functions $f : [0,1] \times \mathcal{R}^n \times \mathcal{R}^m \to \mathcal{R}^n$, $\phi : \mathcal{R}^n \to \mathcal{R}$, $h_i^1 : \mathcal{R}^n \to \mathcal{R}$ for $i \in E$, $h_j^2 : \mathcal{R}^n \to \mathcal{R}$ for $j \in I$ and $q : [0,1] \times \mathcal{R}^n \to \mathcal{R}$, a vector $x_o \in \mathcal{R}^n$, and a set $\Omega \subset \mathcal{R}^m$ of the form

$$\Omega = \{u \in \mathcal{R}^m :\ b_-^i \le u_i \le b_+^i \text{ for } i = 1, 2, \ldots, m\}, \tag{1.2}$$

in which b_-^i, b_+^i, $i = 1, 2, \ldots, m$ are constants. Throughout, T denotes $[0,1]$.

Everything that follows can be adapted to allow for multiple pathwise inequality constraints and also for presence of an integral cost term; we limit ourselves to the above special case for notational simplicity.

*Received by the editors January 24, 1996; accepted for publication (in revised form) October 2, 1997; published electronically August 31, 1998.
http://www.siam.org/journals/sicon/36-6/29764.html
†Centre for Process Systems Engineering, Imperial College, London SW7 2BY, UK (ptk@ps.ic.ac.uk, rbv@ps.ic.ac.uk).

A *control function* $u : T \to \mathcal{R}^m$ is a measurable function which satisfies $u(t) \in \Omega$ a.e. Given any control function, under the hypotheses we shall impose there is a unique absolutely continuous function $x : T \to \mathcal{R}^n$ satisfying $\dot{x}(t) = f(t, x(t), u(t))$ a.e. on T and $x(0) = x_0$. It is denoted by x^u and is referred to as the *state trajectory corresponding to* u.

The control problem **(P)** can be expressed as an optimization problem over the set of control functions

$$\mathcal{U} = \{u : T \to \mathcal{R}^m : \ u \text{ is measurable and } u(t) \in \Omega \text{ a.e. on } T\}$$

with the aid of the functions $\tilde{F}_0 : \mathcal{L}^2_m[T] \to \mathcal{R}$, $\tilde{h}^1_i : \mathcal{L}^2_m[T] \to \mathcal{R}$ for $i \in E$, $\tilde{h}^2_j : \mathcal{L}^2_m[T] \to \mathcal{R}$ for $j \in I$ and $\tilde{q} : \mathcal{L}^2_m[T] \to \mathcal{C}[T]$:

$$\begin{aligned}
\tilde{F}_0(u) &= \phi(x^u(1)),\\
\tilde{h}^1_i(u) &= h^1_i(x^u(1)) \ \forall i \in E,\\
\tilde{h}^2_j(u) &= h^2_j(x^u(1)) \ \forall j \in I,\\
\tilde{q}(u)(t) &= q(t, x^u(t)) \ \forall t \in T.
\end{aligned}$$

The reformulated problem is

$$\min_{u \in \mathcal{U}} \tilde{F}_0(u)$$

s.t.

$$\begin{gathered}
\tilde{h}^1_i(u) = 0 \ \forall i \in E, \ \tilde{h}^2_j(u) \le 0 \ \forall j \in I,\\
\tilde{q}(u)(t) \le 0 \ \forall t \in T.
\end{gathered}$$

The algorithm which we propose has the following features:

a) the algorithm aims to solve a related problem $(\mathbf{P_c})$ in which the equality constraints are replaced by an "exact penalty term" in the cost:

$$\begin{gathered}
\min_{u \in \mathcal{U}} \tilde{F}_c(u)\\
\text{s.t. } \tilde{h}_j(u) \le 0 \ \forall j \in I \text{ and } \tilde{q}(u)(t) \le 0 \ \forall t \in T
\end{gathered}$$

in which

$$\tilde{F}_c(u) = \tilde{F}_0(u)/c + \max_{i \in E} |\tilde{h}^1_i(u)|.$$

(The penalty parameter c is updated according to a simple test, along the lines of that earlier employed by Mayne and Polak [8].)

b) The algorithm generates a sequence of controls whose corresponding state trajectories satisfy the pathwise and endpoint inequality constraints. Search directions are generated by solving a convex control subproblem. The new control is found by conducting an Armijo line search along the direction point obtained from a direction finding subproblem.

Algorithms involving function space iterations for the solution of optimal control problems with pathwise state inequality constraints, with accompanying convergence analysis, have been proposed by Warga [20], Mayne and Polak [9], and Polak, Yang, and Mayne [11]. Both the Warga and Mayne–Polak algorithms involve proximity-type subalgorithms to generate search directions, and their effective implementation is

hampered by the poor performance of proximity-type algorithms applied to the convex sets in infinite dimensional spaces which arise in this context. The algorithm of Warga generates a sequence of relaxed controls, accumulation points of which are shown to satisfy a strong version of the relaxed maximum principle. None of these algorithms involves an ε-active strategy for state constraints, a feature of our algorithm which greatly enhances its efficiency. In the later Polak–Yang–Mayne algorithm, barrier functions are used to eliminate pathwise state constraints from the direction-finding problems. Computation experience of this algorithm is limited, though preliminary findings are promising [11]. The algorithm applies only to problems with no equality constraints.

Machielsen [10] and Alt and Malanowski [1] investigate "function space" second-order methods for solving optimal control problems with state constraints; a local convergence analysis and numerical examples are to be found in [1] and [10], respectively. The fact that the direction-finding subproblems of [10] and [1] are, in general, nonconvex optimal control problems creates difficulties both regarding efficient implementation and global convergence analysis.

A companion paper provides a full discussion of implementational aspects of the algorithm and also numerical examples. The examples include an optimal control problem arising in flight mechanics, concerning optimal control strategies in the presence of windshear, extensively studied by Bulirsh, Montrone, and Pesch [4], [5]. The fact that our feasible directions algorithm provides a solution to the "windshear" problem without recourse to prior information about junction times or control structure (which are required in the method employed in [5]) is evidence of the effectiveness of our algorithm (and indirect, nonlinear programming methods in general) as a computational tool.

What special characteristics of the algorithm provided in this paper promote efficient implementation? One is that search directions generated by the algorithm drive state trajectories into the interior of the state constraint region. This means that satisfaction of the state constraint over the entire time interval T can still be guaranteed, even if we impose the state constraint at only relatively few points in T [16]. Consequently a coarser discretization can be applied to the state constraint than that associated with the parametrization of control functions. This is significant since it is precisely the "dimensionality" of the pathwise constraint which makes it difficult to compute optimal controls for (**P**). Techniques for the approximation of sets on which the state constraint is required to be satisfied were anticipated in an algorithm proposed by Fedorenko [6].

While techniques for deriving conditions on accumulation points generated in both feasible directions and also exact penalty methods in finite dimensional nonlinear programming are now available and well understood, developing a convergence analysis for the optimal control problem (**P**) poses additional difficulties, notably those associated with an inequality constraint function having infinite dimensional range and with the fact that the set $\mathcal{U}$ is not compact. Novel features of the convergence analysis are as follows. We propose the convergence analysis based on the "nonpositive descent function," which along the sequences generated by our algorithm is convergent. Its limit point, equal to zero, is the statement of necessary optimality conditions. A customary result in the literature would be that "relaxed" accumulation points of sequences generated by our algorithm satisfy necessary conditions of optimality in the form of a "relaxed" version of the maximum principle. Our convergence result is stronger in the sense that it is valid for the whole sequence generated by our algo-

rithm. The compactness of the space of relaxed controls is needed only to guarantee boundedness of penalty parameters. The algorithm allows for a computation-saving ε-active strategy in dealing with the "infinite dimensional" inequality constraint. A single, simply stated constraint qualification (hypothesis (**CQ**) below) is invoked both to ensure finite increase of the penalty parameter and to derive properties of accumulation points in place of a pair of constraint qualification–type hypotheses featured, for example, in [8], [9]. We clarify the relationship between standard necessary conditions of optimality and the "nonpositive descent function"–type conditions.

It is to be expected that analytical techniques developed here will also be of benefit in studying the convergence properties of related algorithms for solving optimal control problems, involving Chebyshev-type functional constraints where, owing to the use of a variable stepsize in integration or high order integration procedures, it is either not possible or inconvenient to base the analysis on an a priori discretization of the dynamic equations (see [13]). The algorithm (and its convergence analysis) presented in the paper can easily be adapted to control problems with state constraints whose control functions are defined by piecewise constant (or piecewise polynomial) functions. One such method, a second-order method which exploits the convergence analysis presented here, is described in [13]. It favorably compares with efficient implementations of sequential quadratic programming (SQP) algorithms [7], [22] applied to nonlinear programming problems which are generated by collocation schemes [18].

2. Representation of functional directional derivatives. At each iteration of the feasible directions algorithm, search directions are generated by solving a simplified version of the exact penalty function problem in which the dynamics and cost functional and constraint functionals are replaced by their first-order approximations around the current control function u.

For $d \in \mathcal{U} - u$ we need to consider the first-order approximation $x^u + y^{u,d}$ to x^{u+d} in which the perturbation $y^{u,d}$ to the nominal state trajectory x^u is the unique solution to the linearized equations

$$\begin{aligned} &\dot{y}(t) = f_x(t, x^u(t), u(t))y(t) + f_u(t, x^u(t), u(t))d(t), \\ &y(0) = 0. \end{aligned} \tag{2.1}$$

First-order approximations to the functionals $\tilde{F}_0(u+d) - \tilde{F}_0(u)$, $\tilde{h}^1_i(u+d) - \tilde{h}^1_i(u)$ $(i \in E)$, $\tilde{h}^2_j(u+d) - \tilde{h}^2_j(u)$ $(j \in I)$, and $\tilde{q}(u+d)(t) - \tilde{q}(u)(t)$ $(t \in [0,1])$ can now be defined via $y^{u,d}$ as follows:

$$\begin{aligned} \langle \nabla \tilde{F}_0(u), d\rangle &:= \phi_x(x^u(1))y^{u,d}(1), \\ \langle \nabla \tilde{h}^1_i(u), d\rangle &:= (h^1_i)_x(x^u(1))y^{u,d}(1) \text{ for } i \in E, \\ \langle \nabla \tilde{h}^2_j(u), d\rangle &:= (h^2_j)_x(x^u(1))y^{u,d}(1) \text{ for } j \in I, \\ \langle \nabla \tilde{q}(u)(t), d\rangle &:= q_x(t, x^u(t))y^{u,d}(t) \ \forall\ t \in T. \end{aligned}$$

The notation $\langle \nabla \tilde{F}_0(u), d\rangle$ is intended to convey the suggestion that $\langle \nabla \tilde{F}_0(u), \cdot\rangle$ is a directional derivative associated with some kind of "derivative" $\nabla \tilde{F}_0$ of the functional $\tilde{F}_0$ at u. It is, however, unnecessary to pursue this interpretation (this would require us to specify function spaces and the precise notation of the derivative $\nabla \tilde{F}_0$). As far as describing the feasible directions algorithm and analyzing its convergence properties are concerned, the simplest course is to take the above formulas for $\langle \nabla \tilde{F}_0(u), d\rangle$, etc., as *definitions* of constructs featured in the algorithm whose properties can be

analyzed directly with the help of the results on the relationship between nonlinear control systems and their linear approximations.

The propositions stated below are useful in the analysis of the convergence properties of optimal control algorithms (and our algorithm in particular). Reference will be made in this and the subsequent sections to the following hypotheses.

(H1) $f(t,\cdot,\cdot)$ is continuously differentiable for fixed t, and f, f_x, and f_u are continuous functions. There exists $K<\infty$ such that

$$\|f_x(t,x,u)\| \le K \text{ for all } (t,x,u)\in T\times\mathcal{R}^n\times\Omega. \tag{2.2}$$

(H2) ϕ, h_i^1, $i\in E$, h_j^2, $j\in I$, are continuously differentiable functions. $q(t,\cdot)$ is differentiable for each t, and q, q_x are continuous functions.

Conditions **(H1)** and **(H2)** are regularity hypotheses on the data. Condition (2.2) could be substituted by any condition ensuring the uniform boundedness of state trajectories, for then we can always arrange that (2.2) is satisfied by redefining f for large values of the x variable, values which will never be encountered.

Proofs of the propositions, which are not provided here for the lack of space, are to be found in [15].

PROPOSITION 2.1. *Assume* **(H1)**. *For each* $u\in\mathcal{U}$ *and* $d\in\mathcal{L}_m^2[T]$, (1.1) *and* (2.1) *have unique solutions* (*in the class of absolutely continuous vector valued functions on* $[0,1]$) x^u *and* $y^{u,d}$, *respectively. Furthermore there exist finite constants* c_1, c_2, *and* c_3 *such that*

$$\begin{aligned}\|x^u\|_{\mathcal{L}^\infty} &\le c_1,\\ \|x^u-x^v\|_{\mathcal{L}^\infty} &\le c_2\|u-v\|_{\mathcal{L}^2},\\ \|y^{u,d}\|_{\mathcal{L}^\infty} &\le c_3\|d\|_{\mathcal{L}^2}\end{aligned}$$

$\forall u,\ v\in\mathcal{U},\ d\in\mathcal{L}_m^2[T]$. *In particular,*

$$\|y^{u,d}\|_{\mathcal{L}^\infty} \le c_3\|d\|_{\mathcal{L}^\infty}$$

if $u\in\mathcal{U},\ d\in\mathcal{L}_m^\infty[T]$.

PROPOSITION 2.2. *Assume* **(H1)**. *Then there exists a function* $o_1(\cdot):(0,\infty)\to(0,\infty)$ *such that* $s^{-1}o_1(s)\to 0$ *as* $s\downarrow 0$ *and*

$$\|x^{u+d}-(x^u+y^{u,d})\|_{\mathcal{L}^\infty} \le o_1(\|\,d\|_{\mathcal{L}^\infty})$$

$\forall u\in\mathcal{U}$ *and* $d\in\mathcal{L}_m^\infty[T]$.

PROPOSITION 2.3. *Assume* **(H1)**. *Take a function* $q:T\times\mathcal{R}^n\to\mathcal{R}$ *such that* $q(t,\cdot)$ *is differentiable and* q, q_x *are continuous. Then for any* $\varepsilon>0$ *there exists* $o_2^\varepsilon(\cdot):(0,\infty)\to(0,\infty)$ *such that* $s^{-1}o_2^\varepsilon(s)\to 0$ *as* $s\downarrow 0$ *and*

$$|\max_{t\in R_{\varepsilon,u}}[q(t,x^u(t))+q_x(t,x^u(t))y^{u,d}(t)]-\max_{t\in T}q(t,x^{u+d}(t))| \le o_2^\varepsilon(\|d\|_{\mathcal{L}^\infty}) \tag{2.3}$$

$\forall u\in\mathcal{U}$ *and* $d\in\mathcal{L}_m^\infty[T]$. $R_{\varepsilon,u}$, *the set of times at which the state constraint is* ε*-active, is defined in* (4.1).

PROPOSITION 2.4. *Assume* **(H1)**. *Take a continuously differentiable function* $h:\mathcal{R}^n\to\mathcal{R}$. *Then there exists* $o_3(\cdot):(0,\infty)\to(0,\infty)$ *such that* $s^{-1}o_3(s)\to 0$ *as* $s\downarrow 0$ *and*

$$|h(x^{u+d}(1))-[h(x^u(1))+h_x(x^u(1))y^{u,d}(1)]| \le o_3(\|d\|_{\mathcal{L}^\infty})$$

for all $u \in \mathcal{U}$ *and* $d \in \mathcal{L}^\infty_m[T]$.

PROPOSITION 2.5. *Assume* **(H1)**. *Take continuously differentiable functions* $h_i : \mathcal{R}^n \to \mathcal{R}$, $i \in E$. *Then there exists* $o_4(\cdot) : (0,\infty) \to (0,\infty)$ *such that* $s^{-1}o_4(s) \to 0$ *as* $s \downarrow 0$ *and*

$$|\max_{i\in E}|h_i(x^u(1)) + (h_i)_x(x^u(1))y^{u,d}(1)| - \max_{i\in E}|h_i(x^{u+d}(1))|| \le o_4(\|d\|_{\mathcal{L}^\infty})$$

for all $u \in \mathcal{U}$, $d \in \mathcal{L}^\infty_m[T]$.

3. Relaxed controls. The convergence analysis to follow involves relaxed controls. In this section we briefly review their properties and introduce some notation.

We recall that a *Radon probability measure* ς on the Borel sets of Ω is a regular positive measure such that $\varsigma(\Omega) = 1$. The set of all Radon probability measures is denoted by $rpm(\Omega)$. A *relaxed control*, μ, is a measurable function $\mu : T \to rpm(\Omega)$, where "measurability" is as defined in [19]. The set of relaxed controls is denoted by $\bar{\mathcal{U}}$.

Let $\mathcal{L}^1(T, \mathcal{C}(\Omega))$ denote the space of absolutely integrable functions from T to $\mathcal{C}(\Omega)$. Then the topology imposed on $\bar{\mathcal{U}}$ is the weakest topology such that the mapping

$$\mu \to \int_T \int_\Omega \psi(t,u)\mu(t)(du)dt$$

is continuous for all $\psi \in \mathcal{L}^1(T, \mathcal{C}(\Omega))$.

We recall [19, p. 287] that $\bar{\mathcal{U}}$ is a compact and convex subset of a normed vector space (namely, the dual space of $\mathcal{L}^1(T, \mathcal{C}(\Omega))$ with a "weak" norm whose topology restricted to $\bar{\mathcal{U}}$ coincides with the weak star topology).

Relaxed controls give rise to the relaxed dynamics:

$$\dot{x}(t) = f_r(t, x(t), \mu(t)) := \int_\Omega f(t, x(t), u)\mu(t)(du),\ x(0) = x_0,$$

whose solution we denote by x^μ_r. Extensions to relaxed controls of functions in problem (**P**) are denoted as follows: $\hat{F}_0(\mu) = \phi(x^\mu_r(1))$, $\hat{h}^1_i(\mu) = h^1_i(x^\mu_r(1))$, $\hat{h}^2_j(\mu) = h^2_j(x^\mu_r(1))$, $\hat{q}(\mu)(t) = q(t, x^\mu_r(t))$. We can define then the relaxed problem ($\mathbf{P^r}$) as the problem (**P**) in which functions $\tilde{F}_0$, etc., are substituted by $\hat{F}_0$, etc., ordinary controls u by relaxed controls μ and $\mathcal{U}$ by $\bar{\mathcal{U}}$.

With each ordinary control $u \in \mathcal{U}$ we associate a relaxed control $\mu \in \bar{\mathcal{U}}$ defined by the property $\mu(t)(S) = \delta_{u(t)}(S)$ for all Borel sets $S \subset \Omega$, where $\delta_u(S) = 1$ if $u \in S$ and $\delta_u(S) = 0$ otherwise. We write this control $\mathbf{u}$. Naturally, u and $\mathbf{u}$ give rise to the same state trajectory.

A useful concept, introduced in [3], for studying approximations to relaxed state trajectories is the set of *search directions*:

$$\mathcal{D} := \{d(\cdot,\cdot) : T \times \Omega \to \mathcal{R}^m :\ d(\cdot,u) \text{ is measurable},\ d(t,\cdot) \text{ is continuous},\\ \text{and } u + d(t,u) \in \Omega\ \forall u \in \Omega \text{ a.e. on } T\}.$$

Take $\mu \in \bar{\mathcal{U}}$ and $d \in \mathcal{D}$. $y^{\mu,d}_r$ is the solution to

$$\dot{y}(t) = (f_x)_r(t, x^\mu_r(t), \mu(t))y(t) + \int_\Omega f_u(t, x^\mu_r(t), u)d(t,u)\mu(t)(du),$$
$$y(0) = 0.$$

Notice that if $\mu(t) = \delta_{u(t)}$ and $d(t,u) = v(t) - u$ for some $v \in \mathcal{U}$, then $y_r^{\mu,d} = y^{u,v-u}$.

For $\mu \in \bar{\mathcal{U}}$ and $d \in \mathcal{D}$ we denote

$$\begin{aligned}
\langle \nabla \hat{F}_0(\mu), d\rangle_r &:= \phi_x(x_r^\mu(1)) y_r^{\mu,d}(1),\\
\langle \nabla \hat{h}_i^1(\mu), d\rangle_r &:= (h_i^1)_x(x_r^\mu(1)) y_r^{\mu,d}(1) \text{ for } i \in E,\\
\langle \nabla \hat{h}_j^2(\mu), d\rangle_r &:= (h_j^2)_x(x_r^\mu(1)) y_r^{\mu,d}(1) \text{ for } j \in I,\\
\langle \nabla \hat{q}(\mu)(t), d\rangle_r &:= q_x(t, x_r^\mu(t)) y_r^{\mu,d}(t) \text{ for all } t \in T.
\end{aligned}$$

Under the hypotheses **(H1)** and **(H2)** we deduce from standard properties of relaxed controls [19]. For fixed $d \in \mathcal{D}$, the following mappings are continuous:

$$\begin{gathered}
\mu \to x_r^\mu, \quad \mu \to \langle \nabla \hat{F}_0(\mu), d\rangle_r, \quad \mu \to \langle \nabla \hat{h}_i^1(\mu), d\rangle_r, \ i \in E,\\
\mu \to \langle \nabla \hat{h}_j^2(\mu), d\rangle_r, \ j \in I, \quad \mu \to \langle \nabla \hat{q}(\mu)(\cdot), d\rangle_r,
\end{gathered}$$

where the domain in each case is $\bar{\mathcal{U}}$ and the range spaces $\mathcal{C}(T, \mathcal{R}^n)$, $\mathcal{R}^{|E|}$, $\mathcal{R}^{|I|}$, and $\mathcal{C}(T, \mathcal{R}^n)$, respectively.

4. The algorithm. We begin by describing the direction-finding subproblem and some functions associated with it. First we define an approximation to the active region of the pathwise inequality constraint with reference to the control function

$$R_{\varepsilon,u} := \{t \in T : \ \tilde{q}(u)(t) \geq \max_{\tilde{t} \in T} \tilde{q}(u)(\tilde{t}) - \varepsilon\}. \tag{4.1}$$

Here $\varepsilon > 0$ is a parameter which governs the tightness of the approximation.

For fixed c and u the direction-finding subproblem $\mathbf{P_c}(\mathbf{u})$ for problem $(\mathbf{P_c})$ is

$$\min_{d \in \mathcal{U}-u, \beta \in \mathcal{R}} \beta + 1/(2c)\|d\|^2_{\mathcal{L}^2}$$

s.t.

$$\begin{aligned}
\langle \nabla \tilde{F}_0(u), d\rangle/c + \max_{i \in E} |\tilde{h}_i^1(u) + \langle \nabla \tilde{h}_i^1(u), d\rangle| - \max_{i \in E} |\tilde{h}_i^1(u)| &\leq \beta,\\
\tilde{h}_j^2(u) + \langle \nabla \tilde{h}_j^2(u), d\rangle &\leq \beta \ \forall j \in I,\\
\tilde{q}(u)(t) + \langle \nabla \tilde{q}(u)(t), d\rangle &\leq \beta \ \forall t \in R_{\varepsilon,u}.
\end{aligned}$$

The subproblem can be reformulated as an optimization problem over the space $\mathcal{L}^2_m[T]$ whose objective function is strictly convex and has quadratic growth. It therefore has a unique solution $(\bar{d}, \bar{\beta})$. Since this solution depends on c and u, we may define the *descent function* $\sigma_c(u)$ and the *penalty test function* $t_c(u)$, which will be used to test optimality of a control u and to adjust c, respectively, as

$$\sigma_c(u) = \bar{\beta}$$

and

$$t_c(u) = \bar{\beta} + \max_{i \in E} |\tilde{h}_i^1(u)|/c$$

for given $c > 0$ and $u \in \mathcal{U}$.

A starting point is required which is feasible with respect to the inequality constraints. This point is computed by applying a few iterations of an algorithm which

is a simple variant of the algorithm below (see, e.g., [14]). (Other approaches are possible, such as that described in [12].)

The algorithm is as follows.

ALGORITHM 1. Fix parameters $\varepsilon > 0,\ \gamma,\ \eta \in (0,1),\ c^0 > 0,\ \kappa > 1$.

1. Choose the initial control $u_0 \in \mathcal{U}$ which satisfies $\tilde{h}_j^2(u_0) \leq 0\ \forall j \in I$ and $\tilde{q}(u_0)(t) \leq 0\ \forall t \in T$. Set $k = 0$, $c_{-1} = c^0$.

2. Let c_k be the smallest number chosen from $\{c_{k-1}, \kappa c_{k-1}, \kappa^2 c_{k-1}, \ldots\}$ such that the solution (d_k, β_k) to the direction-finding subproblem $\mathbf{P_{c_k}(u_k)}$ satisfies

$$t_{c_k}(u_k) \leq 0.$$

If $\sigma_{c_k}(u_k) = 0$, then STOP.

3. Let α_k be the largest number chosen from the set $\{1, \eta, \eta^2, \ldots, \}$ such that $u_{k+1} = u_k + \alpha_k d_k$ satisfies the relations

$$\begin{aligned} \tilde{F}_{c_k}(u_{k+1}) - \tilde{F}_{c_k}(u_k) &\leq \gamma \alpha_k \sigma_{c_k}(u_k), \\ \tilde{h}_j^2(u_{k+1}) &\leq 0\ \forall j \in I, \\ \tilde{q}(u_{k+1})(t) &\leq 0\ \forall t \in T. \end{aligned}$$

Increase k by 1. Go to Step 2.

The descent function $\sigma_{c_k}(u_k)$ is nonpositive valued at each iteration. Suppose that subsequences $\{u_k\}_{k\in K},\ \{c_k\}_{k\in K}$ of the sequences of control functions and penalty parameters generated by the algorithm have limit points (in some sense) $\bar{u}$ and $\bar{c}$. We would then expect that $\sigma_{c_k}(u_k) \to \sigma_{\bar{c}}(\bar{u})$ (along the subsequence) and $\sigma_{\bar{c}}(\bar{u}) \geq 0$, a condition which asserts that the direction-finding subproblem for $c_k = \bar{c}$ and $u_k = \bar{u}$ has the solution $(d = 0, \beta = 0)$ and which (together with the feasibility of $\bar{u}$) can be interpreted as a first-order optimality condition satisfied by $\bar{u}$. If $\{c_k\}$ is nondecreasing and bounded (thus convergent), then $\sigma_{c_k}(u_k) \to_{k\to\infty} 0$ would be a stronger result because the existence of a convergent subsequence of $\{u_k\}$ is not requested. Justifying these arguments (under precisely specified hypotheses) is the essence of the convergence analysis to follow.

If u is not a stationary point for the problem $(\mathbf{P})$, then $d \neq 0$ and $\beta < 0$, which implies that

$$\tilde{q}(u)(t) + \langle \nabla \tilde{q}(u)(t), d \rangle < 0\ \forall t \in R_{\varepsilon,u}. \tag{4.2}$$

We can show, under the continuity assumption imposed on $q_x(\cdot,\cdot)$, that $\tilde{q}(u)(t) + \langle \nabla \tilde{q}(u)(t), d \rangle$ is a Lipschitz function with respect to t [16]; therefore there is a finite set of points $A \subset R_{\varepsilon,u}$ such that if

$$\tilde{q}(u)(t) + \langle \nabla \tilde{q}(u)(t), d \rangle < 0\ \forall t \in A,$$

then (4.2) is also satisfied. This property is exploited in the implementable version of the algorithm discussed in [16].

Notice that the only requirement in the directional minimization related to the state constraint is the feasibility of u_{k+1}. Other exact penalty function methods such as that in [9] do not combine the strict descent property (4.2) with the unrestrictive direction minimization procedure regarding the state constraint.

The role of the penalty parameter test function t_c is to ensure that the penalty parameter is large enough in the limit (but finite) to force satisfaction of the equality endpoint constraint. The algorithm ensures that, at the kth iteration,

$$t_{c_k}(u_k) = \sigma_{c_k}(u_k) + \max_{i\in E} |\tilde{h}_i^1(u_k)|/c_k \leq 0.$$

Since under favorable circumstances $\sigma_{c_k}(u_k) \to 0$ and $\{c_k\}$ is bounded, as we will show, we conclude from this inequality that $\max_{i\in E} |\tilde{h}_i^1(\bar{u})| = 0$; i.e., the limiting control satisfies the endpoint equality constraint.

Before concluding this section we need to clarify two points about the algorithm. They are to guarantee that Step 2 and 3 of Algorithm 1 can always be carried out (under the hypotheses of section 2). These gaps are filled by the following proposition in which we invoke a constraint qualification.

(CQ) For each $\mu \in \bar{\mathcal{U}}$ which is feasible w.r.t. the inequality constraints of the problem $(\mathbf{P^r})$ we have $\mathcal{F}(\mu) \neq \emptyset$ and, in the case $E \neq \emptyset$,

$$0 \in \text{interior}[\mathcal{E}(\mu)],$$

where

$$\mathcal{E}(\mu) := \{\{\langle \nabla \hat{h}_i^1(\mu), d\rangle_r\}_{i\in E} \in \mathcal{R}^{|E|} : \ d \in \mathcal{F}(\mu)\}$$

and

$$\begin{aligned}\mathcal{F}(\mu) := \{d \in \mathcal{D} : &\max_{j\in I}[\hat{h}_j^2(\mu) + \langle \nabla \hat{h}_j^2(\mu), d\rangle_r] < 0,\\ &\max_{t\in T}[\hat{q}(\mu)(t) + \langle \nabla \hat{q}(\mu)(t), d\rangle_r] < 0\}.\end{aligned}$$

In the case $I = \emptyset$ (no terminal inequality constraints) we interpret $\max_{j\in I}[\tilde{h}_j^2(\mu) + \cdots] = -\infty$.

(CQ) is related to hypotheses earlier invoked by Mayne and Polak [8]. It is a local controllability condition on values of the (linearized) equality constraint functions with respect to control functions which are strictly feasible w.r.t. the linearized inequality constraints. The role of **(CQ)** is to ensure uniform boundedness of penalty parameter values and that a convergence analysis can be carried out in terms of "normal" extremality conditions (conditions in which the cost multiplier is nonzero) as stated in the next section.

PROPOSITION 4.1.

(i) *Assume that hypotheses* **(H1)**, **(H2)**, *and* **(CQ)** *are satisfied. Then for any* $u \in \mathcal{U}$ *satisfying the endpoint and pathwise inequality constraints of* **(P)** *there exists* $\bar{c} > 0$ *such that for all* $c > \bar{c}$

$$t_c(u) \leq 0.$$

(ii) *Assume that hypotheses* **(H1)** *and* **(H2)** *are satisfied. Then for any* $u \in \mathcal{U}$ *satisfying the endpoint and pathwise inequality constraints and* $c > 0$ *such that* $\sigma_c(u) < 0$, *there exists* $\bar{\alpha} > 0$ *such that if* $\alpha \in [0, \bar{\alpha})$, *then*

$$\begin{aligned}\tilde{F}_c(\tilde{u}) - \tilde{F}_c(u) &\leq \gamma\alpha\sigma_c(u),\\ \tilde{h}_j^1(\tilde{u}) &\leq 0 \ \forall j \in I,\\ \tilde{q}(\tilde{u})(t) &\leq 0 \ \forall t \in T,\end{aligned}$$

where $\tilde{u} = u + \alpha d$ *and* $(d, \beta = \sigma_c(u))$ *is the solution to the direction-finding subproblem corresponding to* c *and* u.

A proof of this proposition is given in section 6.

5. Convergence properties of the algorithm. In this section we show that the descent function $\sigma_c(u)$ converges to zero along the sequences $\{c_k\}$, $\{u_k\}$ generated by Algorithm 1. Furthermore we show that $\{c_k\}$ is bounded and that $\{u_k\}$, regarded as sequences in $\bar{\mathcal{U}}$, satisfy necessary optimality conditions.

Results are given in relation to the necessary conditions (**NC**) in normal form for a control function $\bar{\mu}$, which is feasible for the relaxed problem ($\mathbf{P^r}$), to be a minimizer.

(NC) There exist nonnegative numbers α_j^2, $j \in I$, numbers α_i^1, $i \in E$, an absolutely continuous function $p : [0,1] \to \mathcal{R}^n$ and a nonnegative regular measure ν on the Borel subsets of $[0,1](= T)$ such that

$$
\begin{aligned}
-\dot{p}_r(t) &= (f_x)_r(t, x_r^{\bar{\mu}}(t), \bar{\mu}(t))^T (p_r(t) + \int_{[0,t)} q_x(s, x_r^{\bar{\mu}}(s))\nu(ds)),\\
-(p_r(1) + \int_{[0,1]} q_x(s, x_r^{\bar{\mu}}(s))\nu(ds)) &= \phi_x(x_r^{\bar{\mu}}(1))\\
&\quad + \sum_{i\in E} \alpha_i^1 (h_i^1)_x(x_r^{\bar{\mu}}(1)) + \sum_{j\in I} \alpha_j^2 (h_j^2)_x(x_r^{\bar{\mu}}(1)),\\
\left(p_r(t) + \int_{[0,t)} q_x(s, x_r^{\bar{\mu}}(s))\nu(ds)\right)^T &\int_\Omega f_u(t, x_r^{\bar{\mu}}(t), u) d(t,u) \bar{\mu}(t)(du) \le 0\\
&\qquad \forall d \in \mathcal{D} \text{ a.e. on } [0,1],\\
\operatorname{supp}\{\nu\} &\subset \{t \in T :\ q(t, x_r^{\bar{\mu}}(t)) = 0\} \text{ and } \alpha_j^2 = 0 \text{ if } h_j^2(x_r^{\bar{\mu}}(1)) < 0.
\end{aligned}
\tag{5.1}
$$

Here $\operatorname{supp}\{\nu\}$ denotes the support of the measure ν.

Conditions (**NC**) are standard necessary optimality conditions for a relaxed minimizer (derivable, for example, from [19, Theorem VI.2.3]) valid under hypotheses (**H1**), (**H2**), and (**CQ**) with the exception that (5.1) replaces the customary

$$
\operatorname{supp}\{\bar{\mu}(t)\} \subset \arg\min_{u\in\Omega} r(t)^T f(t, x_r^{\bar{\mu}}(t), u) \text{ a.e. on } T,
\tag{5.2}
$$

where $r(t) = p(t) + \int_{[0,t)} q_x(s, x_r^{\bar{\mu}}(s))\nu(ds)$.

However, (5.2) implies that for all $d \in \mathcal{D}$ and $\varepsilon > 0$

$$
\varepsilon^{-1} \int_\Omega r(t)^T (f(t, x_r^{\bar{\mu}}(t), u + \varepsilon d(t,u)) - f(t, x_r^{\bar{\mu}}(t), u)) \bar{\mu}(t)(du) \le 0 \text{ a.e. on } T.
$$

Passing to the limit as $\varepsilon \downarrow 0$ with the help of the dominated convergence theorem gives (5.1). It follows that (**NC**) are necessary conditions as claimed.

THEOREM 5.1. *Assume that the data for* **(P)** *satisfies hypotheses* **(H1)**, **(H2)**, *and* **(CQ)**. *Let* $\{u_k\}$ *be a sequence of control functions generated by Algorithm* 1, *and let* $\{c_k\}$ *be the corresponding sequence of penalty parameters. Then*

(i) $\{c_k\}$ *is a bounded sequence;*

(ii)

$$
\lim_{k\to\infty} \sigma_{c_k}(u_k) = 0, \quad \lim_{k\to\infty} \max_{i\in E} |\tilde{h}_i^1(u_k)| = 0;
$$

(iii) *if* $\bar{\mu}$ *is any accumulation point of* $\{\mathbf{u}_k\}$ *in* $\bar{\mathcal{U}}$ *(and such an accumulation point always exists), then* $\bar{\mu}$ *is feasible for the relaxed problem* ($\mathbf{P^r}$) *and satisfies necessary conditions* (**NC**).

Notice that if $\{u_k\}$ is a sequence in $\mathcal{U}$ such that $u_k \to \bar{u}$ for some $\bar{u} \in \mathcal{L}^2_m[T]$ with respect to the $\mathcal{L}^2$ norm, then $\mathbf{u}_k \to \bar{\mathbf{u}}$ converges in $\bar{\mathcal{U}}$ ([21]; see also [17]). Part (iii) of Theorem 5.1 may therefore be substituted by the following weaker assertion:

(iii) *if $\bar{u} \in \mathcal{U}$ is any $\mathcal{L}^2$ accumulation point of $\{u_k\}$, then $\bar{u}$ is feasible for* (**P**) *and satisfies* (**NC**).

Notice that the conditions (**NC**), when applied at an ordinary control, reduce to simpler "nonrelaxed" necessary conditions of optimality. We see then that the "relaxed" analysis subsumes the $\mathcal{L}^2$ analysis and improves on it by giving information about asymptotic behavior of the algorithm even when $\mathcal{L}^2$ accumulation points do not exist.

Part (iii) of the theorem implies that if $u \in \mathcal{U}$ is feasible and $\sigma_c(u) = 0$ for some $c > 0$, then u satisfies necessary conditions (**NC**) of optimality. Fix $\varepsilon_{\text{STOP}} > 0$. Part (ii) of the theorem implies that the stopping condition

$$\sigma_{c_k}(u_k) \geq -\varepsilon_{\text{STOP}},\ \max_{i\in E} |\tilde{h}^1_i(u_k)| \leq \varepsilon_{\text{STOP}},\ \max_{j\in I} \tilde{h}^2_j(u_k) \leq 0,\ \max_{t\in T} \tilde{q}(u_k)(t) \leq 0$$

is satisfied after a finite number of iterations. Termination of the algorithm still occurs after a finite number of iterations if the above stopping criterion is supplemented by

$$|\tilde{F}_0(u_{k+1}) - \tilde{F}_0(u_k)| \leq \varepsilon_{\text{STOP}},\quad \|u_{k+1} - u_k\|_{\mathcal{L}^2} \leq \|d_k\|_{\mathcal{L}^2} \leq \varepsilon_{\text{STOP}}.$$

The first inequality follows from (i), (ii), and the first inequality of Proposition 4.1(ii). The second inequality is a consequence of the fact that the optimal value of the subproblem $\mathbf{P_c}(\mathbf{u})$ is nonpositive whence

$$\|d_k\|_{\mathcal{L}^2} \leq -2c_k\sigma_{c_k}(u_k) \ \text{ and (ii)} \ \Rightarrow \lim_{k\to\infty} \|d_k\|_{\mathcal{L}^2} = 0.$$

Notice that (i) and (ii) correspond to the following general convergence result in nonlinear programming related to minimizing a bounded (from below), continuously differentiable function f: $\lim_{k\to\infty} \nabla f(x_k) = 0$. (The condition does not require the existence of accumulation points of $\{x_k\}$ and is relevant, for example, in situations when we seek to minimize $f(x) = e^x$.) We are not aware of convergence results, of this general nature, elsewhere in the constrained nonlinear programming literature.

6. Proof of the convergence theorem, etc. We precede the proof of Proposition 4.1 with a lemma which describes important implications of the (**CQ**).

LEMMA 6.1. *Assume* (**H1**), (**H2**), *and* (**CQ**). *For any relaxed control μ satisfying the inequality constraints for the relaxed problem there exist a neighborhood $\mathcal{O}(\mu)$ of μ, in the relaxed topology, $K_1 > 0$ and $K_2 > 0$ with the following properties: given any $u \in \mathcal{U}$ satisfying the inequality constraints and such that $\mathbf{u} \in \mathcal{O}(\mu)$, there exists $v \in \mathcal{U}$ such that*

$$\max_{i\in E} |\tilde{h}^1_i(u) + \langle \nabla \tilde{h}^1_i(u), v - u\rangle| - \max_{i\in E} |\tilde{h}^1_i(u)| \leq -K_1 \max_{i\in E} |\tilde{h}^1_i(u)|, \tag{6.1}$$

$$\max_{j\in I} [\tilde{h}^2_j(u) + \langle \nabla \tilde{h}^2_j(u), v - u\rangle] \leq -K_1 \max_{i\in E} |\tilde{h}^1_i(u)|, \tag{6.2}$$

$$\max_{t\in T} [\tilde{q}(u)(t) + \langle \nabla \tilde{q}(u)(t), v - u\rangle] \leq -K_1 \max_{i\in E} |\tilde{h}^1_i(u)|, \tag{6.3}$$

$$\text{and} \quad \|v - u\|_{\mathcal{L}^2} \leq \quad K_2 \max_{i\in E} |\tilde{h}^1_i(u)|. \tag{6.4}$$

The lemma is proved in the appendix.

Proof of Proposition 4.1. (i) Fix $u \in \mathcal{U}$ satisfying the inequality constraints. We must find $\hat{c} > 0$ such that if $c > \hat{c}$, then $t_c(u) \leq 0$. If $M(u) := \max_{i \in E} |\tilde{h}_i^1(u)| = 0$, then of course $t_c(u) \leq 0$ for any $c > 0$. If $M(u) > 0$, then according to Lemma 6.1 there exists $\hat{d} \in \mathcal{U} - u$ such that if we set $\varepsilon = K_1 M(u) > 0$ with K_1 as in Lemma 6.1, then

$$\theta(u) < -\varepsilon.$$

Here

$$\begin{aligned}\theta(u) := \max[&\max_{i \in E} |\tilde{h}_i^1(u) + \langle \nabla \tilde{h}_i^1(u), \hat{d}\rangle| - \max_{i \in E} |\tilde{h}_i^1(u)|,\\ &\max_{j \in I}[\tilde{h}_j^2(u) + \langle \nabla \tilde{h}_j^2(u), \hat{d}\rangle],\\ &\max_{t \in T}[\tilde{q}(u)(t) + \langle \nabla \tilde{q}(u)(t), \hat{d}\rangle]].\end{aligned}$$

Because $\sigma_c(u) \leq \langle \nabla \tilde{F}_0(u), \hat{d}\rangle/c + \theta(u)$, from the definition of t_c and Proposition 2.1, we get

$$t_c(u) \leq [W + \max_{i \in E} |\tilde{h}_i^1(u)|]/c + \theta(u),$$

where $W := \max[0, \langle \nabla \tilde{F}_0(u), \hat{d}\rangle]$. It follows that $t_c(u) \leq 0$ for any $c > \hat{c}$ where

$$\hat{c} := \frac{W + M(u)}{-\theta(u)}.$$

(ii) Take $u \in \mathcal{U}$ satisfying the inequality constraints and $c > 0$ such that $\sigma_c(u) < 0$. Let (d, β) be the solution to $\mathbf{P_c(u)}$. Since $\sigma_c(u) \neq 0$, it follows that $d \neq 0$.

We deduce from the differentiability properties of ϕ, h_i^1, h_j^2, and q and Proposition 2.2 that there exists $o : [0, \infty) \to [0, \infty)$ such that $s^{-1} o(s) \to 0$ as $s \downarrow 0$, and the following three inequalities are valid for any $\alpha \in [0, 1]$:

$$\begin{aligned}\tilde{F}_c(u + \alpha d) - \tilde{F}_c(u) \leq\ &\alpha \langle \nabla \tilde{F}_0(u), d\rangle/c + \max_{i \in E} |\tilde{h}_i^1(u) + \alpha \langle \nabla \tilde{h}_i^1(u), d\rangle|\\ &- \max_{i \in E} |\tilde{h}_i^1(u)| + o(\alpha),\end{aligned} \tag{6.5}$$

$$\tilde{h}_j^2(u + \alpha d) \leq \tilde{h}_j^2(u) + \alpha \langle \nabla \tilde{h}_j^2(u), d\rangle + o(\alpha)\ \forall j \in I, \tag{6.6}$$

$$\tilde{q}(u + \alpha d)(t) \leq \tilde{q}(u)(t) + \langle \nabla \tilde{q}(u)(t), d\rangle + o(\alpha)\ \forall t \in T. \tag{6.7}$$

By the convexity of the function $e \to \max_{i \in E} |\tilde{h}_i^1(u) + \langle \nabla \tilde{h}_i^1(u), e\rangle|$,

$$\begin{aligned}&\max_{i \in E} |\tilde{h}_i^1(u) + \alpha \langle \nabla \tilde{h}_i^1(u), d\rangle| - \max_{i \in E} |\tilde{h}_i^1(u)|\\ &\leq \alpha(\max_{i \in E} |\tilde{h}_i^1(u) + \langle \nabla \tilde{h}_i^1(u), d\rangle| - \max_{i \in E} |\tilde{h}_i^1(u)|).\end{aligned}$$

From inequality (6.5), then

$$\begin{aligned}\tilde{F}_c(u + \alpha d) - \tilde{F}_c(u) &\leq \alpha[\langle \nabla \tilde{F}_0(u), d\rangle/c + \max_{i \in E} |\tilde{h}_i^1(u) + \langle \nabla \tilde{h}_i^1(u), d\rangle|\\ &\quad - \max_{i \in E} |\tilde{h}_i^1(u)|] + o(\alpha)\\ &\leq \alpha \sigma_c(u) + o(\alpha).\end{aligned}$$

It follows that

$$\tilde{F}_c(u+\alpha d) - \tilde{F}_c(u) \le \alpha\gamma\sigma_c(u) \ \forall \alpha \in [0,\alpha_1], \tag{6.8}$$

where $\alpha_1 > 0$ is such that $o(\beta) \le \beta(\gamma-1)\sigma_c(u)$ for all $\beta \in [0,\alpha_1]$.

It remains to show that $u+\alpha d$ is feasible w.r.t. the inequality constraints for sufficiently small α. Since $\tilde{h}_j^2(u) \le 0$ and $\alpha \in [0,1]$, (6.6) implies

$$\begin{aligned}\tilde{h}_j^2(u+\alpha d) &\le \alpha[\tilde{h}_j^2(u) + \langle \nabla \tilde{h}_j^2(u), d\rangle] + o(\alpha)\\ &\le \alpha\sigma_c(u) + o(\alpha) \le \alpha\gamma\sigma_c(u) < 0\end{aligned}$$

for all $\alpha \in [0,\alpha_1]$, as required.

We deduce from the differentiability properties of q (Proposition 2.3) that

$$\begin{aligned}\max_{t\in T} \tilde{q}(u+\alpha d)(t) &\le \max_{t\in R_{\varepsilon,u}} [\tilde{q}(u)(t) + \alpha\langle \nabla\tilde{q}(u)(t), d\rangle] + o(\alpha)\\ &\le \alpha\sigma_c(u) + o(\alpha)\\ &\le \alpha\gamma\sigma_c(u) < 0\end{aligned}$$

for $\alpha \in [0,\alpha_1]$, as required. □

Proof of Theorem 5.1. (i) Let $\{u_k\}$ be the sequence generated by Algorithm 1, and let $\{c_k\}$ be the corresponding penalty parameters. Let $\{k_l\}$ be the sequence of index values at which the penalty parameter increases. By extracting a further subsequence (we do not relabel) we can arrange that the sequence $\{\mathbf{u}_{k_l}\}$ has a limit point $\bar{\mu} \in \bar{\mathcal{U}}$ because $\bar{\mathcal{U}}$ is compact. For simplicity of presentation we denote this subsequence by $\{\mathbf{u}_{k_l}\}$. We shall find a number $\hat{c} < \infty$ such that for sufficiently large k_l, $c_{k_l} > \hat{c}$ implies $\sigma_{c_{k_l}}(u_{k_l}) \le -\max_{i\in E} |\tilde{h}_i^1(u_{k_l})|/c_{k_l}$. This contradicts our assumption that the penalty parameter increases along the subsequence. So we may conclude that $\{c_k\}$ is bounded.

Fix k_l such that $\mathbf{u}_{k_l} \in \mathcal{O}(\bar{\mu})$, where $\mathcal{O}(\bar{\mu})$ is in the neighborhood of $\bar{\mu}$ as specified in Lemma 6.1. From the minimizing property of $\sigma_{c_{k_l}}(u_{k_l})$ we deduce

$$\begin{aligned}\sigma_{c_{k_l}}(u_{k_l}) \le{}& 1/(2c_{k_l})\|v_{k_l} - u_{k_l}\|_{\mathcal{L}^2}^2 + \langle \nabla\tilde{F}_0(u_{k_l}), v_{k_l} - u_{k_l}\rangle/c_{k_l}\\ &+ \max\{\max_{i\in E} |\tilde{h}_i^1(u_{k_l}) + \langle \nabla\tilde{h}_i^1(u_{k_l}), v_{k_l} - u_{k_l}\rangle| - \max_{i\in E} |\tilde{h}_i^2(u_{k_l})|,\\ &\quad \max_{j\in I}[\tilde{h}_j^2(u_{k_l}) + \langle \nabla\tilde{h}_j^2(u_{k_l}), v_{k_l} - u_{k_l}\rangle],\\ &\quad \max_{t\in T}[\tilde{q}(u_{k_l})(t) + \langle \nabla\tilde{q}(u_{k_l})(t), v_{k_l} - u_{k_l}\rangle]\}\end{aligned} \tag{6.9}$$

for a control function v_{k_l} satisfying conditions (6.1)–(6.4) of Lemma 6.1 in which v_{k_l}, u_{k_l} replace v, u, respectively. (Notice that Lemma 6.1 may be invoked since $\tilde{h}_j^2(u_{k_l}) \le 0 \ \forall j \in I$ and $\tilde{q}(u_{k_l})(t) \le 0 \ \forall t \in T$.) It follows

$$\begin{aligned}\sigma_{c_{k_l}}(u_{k_l}) \le{}& 1/(2c_{k_l}) \int_0^1 \|v_{k_l}(t) - u_{k_l}(t)\|^2 dt + |\langle \nabla\tilde{F}_0(u_{k_l}), v_{k_l} - u_{k_l}\rangle|/c_{k_l}\\ &- K_1 \max_{i\in E} |\tilde{h}_i^1(u_{k_l})|.\end{aligned}$$

Since the control constraint Ω is bounded and in view of Proposition 2.1, there exists $r > 0$ (independent of k_l) such that

$$\begin{aligned}1/(2c_{k_l}) \int_0^1 \|v_{k_l}(t) - u_{k_l}(t)\|^2 dt + |\langle \nabla\tilde{F}_0(u_{k_l}), v_{k_l} - u_{k_l}\rangle|/c_{k_l}\\ \le (r/c_{k_l})\|v_{k_l} - u_{k_l}\|_{\mathcal{L}^2}.\end{aligned}$$

Hence

$$\sigma_{c_{k_l}}(u_{k_l}) \leq -(K_1 - rK_2/c_{k_l}) \max_{i\in E} |\tilde{h}_i^1(u_{k_l})|.$$

We conclude that

$$\sigma_{c_{k_l}}(u_{k_l}) \leq -\max_{i\in E} |\tilde{h}_i^1(u_{k_l})|/c_{k_l}$$

if $c_{k_l} \geq \hat{c}$, where $\hat{c} = K_1^{-1}(rK_2 + 1)$.

(ii) and (iii) Let $\{u_k\}$ be an infinite sequence generated by the algorithm. We must show that $\lim_{k\to\infty} \sigma_{c_k}(u_k) = 0$ and, if a convergent subsequence of $\{\mathbf{u}_k\}$ has a limit point $\bar{\mu} \in \bar{\mathcal{U}}$, that conditions (**NC**) are satisfied at $\bar{\mu}$.

Stage 1 (convergence analysis). Since the c_k's are bounded and can increase only by multiples of c^0, we must have $c_k = c$ for all $k \geq k_0$, for some k_0 and $c > 0$. In view of the manner in which u_k's are constructed, we have

$$\tilde{F}_c(u_{k+1}) - \tilde{F}_c(u_k) \leq \gamma\alpha_k\sigma_c(u_k)$$

$\forall k \geq k_0$. This means that, $\forall j \geq 1$, $k \geq k_0$

$$\tilde{F}_c(u_{k+j}) - \tilde{F}_c(u_k) \leq \gamma \sum_{i=0}^{j-1} \alpha_{k+i}\sigma_c(u_{k+i}). \tag{6.10}$$

Since $\{\tilde{F}_c(u_k)\}$ is a bounded sequence and $\alpha_k\sigma_c(u_k)$ is nonpositive, we conclude

$$\alpha_k\sigma_c(u_k) \to 0 \quad \text{as } k \to \infty. \tag{6.11}$$

Since $\sigma_c(\mu)$ is bounded as μ ranges over $\bar{\mathcal{U}}$, we can arrange by a subsequence extraction (we do not relabel) that

$$\sigma_c(u_k) \to \beta \text{ for some } \beta \leq 0.$$

We claim that $\beta = 0$. To show this, suppose to the contrary that $\beta < 0$. Then by (6.11) $\alpha_k \to 0$.

We must have

$$\begin{aligned} &\tilde{F}_c(u_k + \eta^{-1}\alpha_k d_k) - \tilde{F}_c(u_k) > \gamma\eta^{-1}\alpha_k\sigma_c(u_k) \\ \text{or} \quad &\max_{j\in I} \tilde{h}_j^2(u_k + \eta^{-1}\alpha_k d_k) > 0 \\ \text{or} \quad &\max_{t\in T} \tilde{q}(u_k + \eta^{-1}\alpha_k d_k)(t) > 0 \end{aligned}$$

for all k sufficiently large.

However, $\|\alpha_k d_k\|_{\mathcal{L}^\infty} \to 0$ as $k \to \infty$. We deduce then from Proposition 2.3 and the continuity of q that

$$\max_{t\in R_{\varepsilon,u_k}} \tilde{q}(u_k + \eta^{-1}\alpha_k d_k)(t) \ = \max_{t\in T} \tilde{q}(u_k + \eta^{-1}\alpha_k d_k)(t)$$

for all k sufficiently large. Since $\sigma_c(u_k) \leq 0$, we conclude that

$$\begin{aligned} \max\{\tilde{F}_c(u_k + \eta^{-1}\alpha_k d_k) - \tilde{F}_c(u_k), &\max_{j\in I} \tilde{h}_j^2(u_k + \eta^{-1}\alpha_k d_k), \\ &\max_{t\in R_{\varepsilon,u_k}} \tilde{q}(u_k + \eta^{-1}\alpha_k d_k)(t)\} \geq \gamma\eta^{-1}\alpha_k\sigma_c(u_k). \end{aligned} \tag{6.12}$$

It follows now from Propositions 2.4–2.5 that there exists a function $o : [0,\infty) \to [0,\infty)$ such that $s^{-1}o(s) \to 0$ as $s \downarrow 0$ and

$$\begin{aligned} \max_{j\in I} \tilde{h}_j^2(u_k + \eta^{-1}\alpha_k d_k) &\le \max_{j\in I} [\tilde{h}_j^2(u_k) + \langle \nabla \tilde{h}_j^2(u_k), \eta^{-1}\alpha_k d_k\rangle] \\ &\quad + o(\eta^{-1}\alpha_k \|d_k\|_{\mathcal{L}^\infty}), \end{aligned} \tag{6.13}$$

$$\begin{aligned} \max_{t\in R_{\varepsilon,u_k}} \tilde{q}(u_k + \eta^{-1}\alpha_k d_k)(t) &\le \max_{t\in R_{\varepsilon,u_k}} [\tilde{q}(u_k)(t) + \langle \nabla \tilde{q}(u_k), \eta^{-1}\alpha_k d_k\rangle] \\ &\quad + o(\eta^{-1}\alpha_k \|d_k\|_{\mathcal{L}^\infty}) \end{aligned} \tag{6.14}$$

and

$$\begin{aligned} \tilde{F}_c(u_k + \eta^{-1}\alpha_k d_k) - \tilde{F}_c(u_k) &\le \langle \nabla \tilde{F}_0(u_k), \eta^{-1}\alpha_k d_k\rangle / c \\ &\quad + \max_{i\in E} |\tilde{h}_i^1(u_k) + \langle \nabla \tilde{h}_i^1(u_k), \eta^{-1}\alpha_k d_k\rangle| \\ &\quad - \max_{i\in E} |\tilde{h}_i^1(u_k)| + o(\eta^{-1}\alpha_k \|d_k\|_{\mathcal{L}^\infty}). \end{aligned} \tag{6.15}$$

Since u_k is feasible w.r.t. the inequality constraints, we have

$$\begin{aligned} \max_{j\in I} [\tilde{h}_j^2(u_k) + \langle \nabla \tilde{h}_j^2(u_k), \eta^{-1}\alpha_k d_k\rangle] &\le \eta^{-1}\alpha_k \max_{j\in I} [\tilde{h}_j^2(u_k) \\ &\quad + \langle \nabla \tilde{h}_j^2(u_k), d_k\rangle] \end{aligned} \tag{6.16}$$

and

$$\begin{aligned} \max_{t\in R_{\varepsilon,u_k}} [\tilde{q}(u_k)(t) + \langle \nabla \tilde{q}(u_k)(t), \eta^{-1}\alpha_k d_k\rangle] &\le \eta^{-1}\alpha_k \max_{t\in R_{\varepsilon,u_k}} [\tilde{q}(u_k)(t) \\ &\quad + \langle \nabla \tilde{q}(u_k)(t), d_k\rangle] \end{aligned} \tag{6.17}$$

for k sufficiently large. Also, by the convexity of $e \to \max_{i\in E} |\tilde{h}_i^1(u_k) + \langle \nabla \tilde{h}_i^1(u_k), e\rangle|$,

$$\begin{aligned} &\max_{i\in E} |\tilde{h}_i^1(u_k) + \langle \nabla \tilde{h}_i^1(u_k), \eta^{-1}\alpha_k d_k\rangle| - \max_{i\in E} |\tilde{h}_i^1(u_k)| \\ &\le \eta^{-1}\alpha_k (\max_{i\in E} |\tilde{h}_i^1(u_k) + \langle \nabla \tilde{h}_i^1(u_k), d_k\rangle| - \max_{i\in E} |\tilde{h}_i^1(u_k)|). \end{aligned} \tag{6.18}$$

Combining inequalities (6.12)–(6.18), noting the definition of $\sigma_c(u_k)$ and the fact that (d_k, β_k) solves $\mathbf{P_c}(\mathbf{u_k})$, and dividing across the resulting inequality by α_k we arrive at

$$\eta^{-1}\sigma_c(u_k) + \alpha_k^{-1} o(\eta^{-1}\alpha_k \|d_k\|_{\mathcal{L}^\infty}) \ge \eta^{-1}\gamma\sigma_c(u_k).$$

We get $\eta^{-1}\beta \ge \eta^{-1}\gamma\beta$ in the limit. But this implies $\gamma \ge 1$, since $\beta < 0$ by assumption. From this contradiction we conclude the validity of $\beta = 0$. Assertion (ii) of the theorem follows from the definition of t_c and part (i).

Let $\{\mathbf{u}_k\}$ be a convergent subsequence with the limit point $\bar{\mu} \in \bar{\mathcal{U}}$. We must show that conditions (**NC**) are satisfied at $\bar{\mu}$. First we establish that $\bar{\mu}$ is feasible for $(\mathbf{P^r})$ and, for some $c > 0$,

$$\begin{aligned} 0 \le \{&\langle \nabla \hat{F}_0(\bar{\mu}), d\rangle_r / c + \max_{i\in E} |\langle \nabla \hat{h}_i^1(\bar{\mu}), d\rangle_r|, \\ &\max_{j\in I} \langle \nabla \hat{h}_j^2(\bar{\mu}), d\rangle_r, \max_{t\in R_{0,\bar{\mu}}} \langle \nabla \hat{q}(\bar{\mu})(t), d\rangle_r\} \end{aligned}$$

for all $d \in \mathcal{D}$.

Since $\mathbf{u}_k \to \bar{\mu}$ we know that $x_r^{\mathbf{u}_k} \to x_r^{\bar{\mu}}$ uniformly. Because u_k is feasible w.r.t. the inequality constraints and the penalty parameter is not updated for k sufficiently large, we have

$$\sigma_c(u_k) \leq -\max_{i\in E} |\tilde{h}_i^1(u_k)|/c,\ \max_{j\in I} \tilde{h}_j^2(u_k) \leq 0,\ \max_{t\in T} \tilde{q}(u_k)(t) \leq 0.$$

But we have shown that $\sigma_c(u_k) \to 0$ as $k \to \infty$. It follows now from the fact that $\mathbf{u}_k \to \bar{\mu} \in \bar{\mathcal{U}}$ that in the limit

$$\max_{i\in E} |\hat{h}_i^1(\bar{\mu})| = 0,\ \max_{j\in I} \hat{h}_j^2(\bar{\mu}) \leq 0,\ \max_{t\in T} \hat{q}(\bar{\mu})(t) \leq 0. \tag{6.19}$$

We have established that $\bar{\mu}$ is feasible for $(\mathbf{P^r})$.

Now choose any sequence $\rho_k \downarrow 0$, $\rho_k \leq 1\ \forall k$ such that

$$\rho_k^{-1}\sigma_c(u_k) \to 0 \ \text{ as } k \to \infty. \tag{6.20}$$

Choose any $d \in \mathcal{D}$. By the convexity of $\mathcal{U}$, $\rho_k d(\cdot, u_k) \in \mathcal{U} - u_k$ for each k.

By definition of σ_c, then

$$\begin{aligned}\sigma_c(u_k) \leq\ & 1/2\rho_k^2 \|d(\cdot,u_k)\|_{\mathcal{L}^2}/c + \max\{\phi_x(x^{u_k}(1))y^{u_k,\rho_k d(\cdot,u_k)}(1)/c \\ & + \max_{i\in E} |h_i^1(x^{u_k}(1)) + (h_i^1)_x(x^{u_k}(1))y^{u_k,\rho_k d(\cdot,u_k)}(1)| - \max_{i\in E} |h_i^1(x^{u_k}(1))|, \\ & \max_{j\in I} [h_j^2(x^{u_k}(1)) + (h_j^2)_x(x^{u_k}(1))y^{u_k,\rho_k d(\cdot,u_k)}(1)], \\ & \max_{t\in T} [q(t,x^{u_k}(t)) + q_x(t,x^{u_k}(t))y^{u_k,\rho_k d(\cdot,u_k)}(t)]\}.\end{aligned} \tag{6.21}$$

Fix $\hat{\varepsilon} > 0$. Since $\rho_k \downarrow 0$ (and consequently $y^{u_k,\rho_k d(\cdot,u_k)} \to 0$ uniformly), and also u_k is feasible w.r.t. the inequality constraints for each k, we have:

$$\max_{j\in I} [h_j^2(x^{u_k}(1)) + (h_j^2)_x(x^{u_k}(1))y^{u_k,\rho_k d(\cdot,u_k)}(1)] \leq \max_{I_{\hat{\varepsilon},\bar{\mu}}} (h_j^2)_x(x^{u_k}(1))y^{u_k,\rho_k d(\cdot,u_k)}(1),$$

$$\max_{t\in T} [q(t,x^{u_k}(t)) + q_x(t,x^{u_k}(t))y^{u_k,\rho_k d(\cdot,u_k)}(t)] \leq \max_{t\in R_{\hat{\varepsilon},\bar{\mu}}} q_x(t,x^{u_k}(t))y^{u_k,\rho_k d(\cdot,u_k)}(t)$$

$\forall k$ sufficiently large. Inserting these inequalities into (6.21), noting that $y^{u_k,\rho_k d(\cdot,u_k)} = \rho_k y^{u_k,d(\cdot,u_k)}$, dividing across by ρ_k, and passing to the limit with the help of (6.20) and continuity of $\hat{F}_0(\cdot)$, $\mu \to \langle \nabla \hat{F}_0(\mu), d\rangle_r$, etc., we obtain

$$\begin{aligned}0 \leq\ & \max\{\phi_x(x_r^{\bar{\mu}}(1))y_r^{\bar{\mu},d}(1)/c + \max_{i\in E} |(h_i^1)_x(x_r^{\bar{\mu}}(1))y_r^{\bar{\mu},d}(1)|, \\ & \max_{j\in I_{\hat{\varepsilon},\bar{\mu}}} (h_j^2)_x(x_r^{\bar{\mu}}(1))y_r^{\bar{\mu},d}(1),\ \max_{t\in R_{\hat{\varepsilon},\bar{\mu}}} q_x(t,x_r^{\bar{\mu}}(t))y_r^{\bar{\mu},d}(t)\}.\end{aligned} \tag{6.22}$$

This inequality is valid for each $\hat{\varepsilon} > 0$ and $d \in \mathcal{D}$.

Again choose arbitrary $d \in \mathcal{D}$ and take $\varepsilon_k \downarrow 0$. For each k let $(h_j^2)_x(x_r^{\bar{\mu}}(1))y_r^{\bar{\mu},d}(1)$ achieve its maximum over $I_{\varepsilon_k,\bar{\mu}}$ at $j = j_k$, and let $q_x(t,x_r^{\bar{\mu}}(t))y_r^{\bar{\mu},d}(t)$ achieve its maximum over $t \in R_{\varepsilon_k,\bar{\mu}}$ at $t = t_k$. Then

$$\begin{aligned}0 \leq\ & \max\{\phi_x(x_r^{\bar{\mu}}(1))y_r^{\bar{\mu},d}(1)/c + \max_{i\in E} |(h_i^1)_x(x_r^{\bar{\mu}}(1))y_r^{\bar{\mu},d}(1)|, \\ & (h_{j_k}^2)_x(x_r^{\bar{\mu}}(1))y_r^{\bar{\mu},d}(1),\ q_x(t_k,x_r^{\bar{\mu}}(t_k))y_r^{\bar{\mu},d}(t_k)\}.\end{aligned} \tag{6.23}$$

Extract a subsequence (we do not relabel) such that $j_k = \bar{j}$ for all k and $t_k \to \bar{t}$ for some index value $\bar{j}$ and some $\bar{t}$. By continuity of the functions involved, $\bar{j} \in I_{0,\bar{\mu}}$, $\bar{t} \in R_{0,\bar{\mu}}$, and (6.23) is valid with $\bar{j}$ and $\bar{t}$ replacing j_k and t_k, respectively.

We have arrived at

$$0 \leq \max\{\phi_x(x_r^{\bar{\mu}}(1))y_r^{\bar{\mu},d}(1)/c + \max_{i\in E} |(h_i^1)_x(x_r^{\bar{\mu}}(1))y_r^{\bar{\mu},d}(1)|,$$
$$\max_{j\in I_{0,\bar{\mu}}} (h_j^2)_x(x_r^{\bar{\mu}}(1))y_r^{\bar{\mu},d}(1), \max_{t\in R_{0,\bar{\mu}}} q_x(t, x_r^{\bar{\mu}}(t))y_r^{\bar{\mu},d}(t)\}.$$

This inequality, which holds for all $d \in \mathcal{D}$, in particular for $(d \equiv 0) \in \mathcal{D}$, is what we set out to prove.

Finally we must attend to the case when Algorithm 1 generates a finite sequence which terminates at a control $\mathbf{u}_{\bar{k}} = \bar{\mu}$, satisfying the stopping criterion. This case is dealt with by applying the preceding arguments to the infinite sequence of controls obtained by "filling in" with repetitions of the following control u_k.

Stage 2 (dualization). The conclusions of Stage 1 can be expressed as

$$\min_{d\in\mathcal{D}} \max_{\gamma\in\mathcal{K}} \Phi(d,\gamma) = 0,$$

where

$$\mathcal{K} := \Big\{\gamma = (\alpha_0, \{\alpha_i^1\}_{i\in E}, \{\alpha_j^2\}_{j\in I}, \nu) \in \mathcal{R}^{1+|E|+|I|} \times \mathcal{C}^\star(T) :$$
$$\alpha_0 \geq 0,\ \alpha_j^2 \geq 0,\ j \in I,\ \sum_{i\in E} |\alpha_i^1| \leq \alpha_0,\ \alpha_0 + \sum_{j\in I} \alpha_j^2 + \int_T \nu(dt) = 1,$$
$$\alpha_j^2 = 0 \ \text{ if } j \notin I_{0,\bar{\mu}},\ \nu \geq 0,\ \operatorname{supp}\{\nu\} \subset R_{0,\bar{\mu}}\Big\}$$

and

$$\Phi(d,\gamma) := \alpha_0\langle\nabla\hat{F}_0(\bar{\mu}), d\rangle_r/c + \sum_{i\in E} \alpha_i^1\langle\nabla\hat{h}_i^1(\bar{\mu}), d\rangle_r$$
$$+ \sum_{j\in I_{0,\bar{\mu}}} \alpha_j^2\langle\nabla\hat{h}_j^2(\bar{\mu}), d\rangle_r + \int_T \langle\nabla\hat{q}(\bar{\mu})(t), d\rangle_r \nu(dt).$$

$\Phi(\cdot,\gamma)$ is a linear function on $\mathcal{L}^1(T, \mathcal{C}(\Omega))$, of which $\mathcal{D}$ is a convex subset. $\Phi(d,\cdot)$ is a bounded linear map and $\mathcal{K}$ is a compact convex set with respect to the product topology of $\mathcal{R}^{1+|E|+|I|} \times \mathcal{C}^\star(T)$, in which the weak star topology is imposed on the last component. It follows from the minimax theorem [2] that there exists some nonzero $\bar{\gamma} \in \mathcal{K}$ such that

$$\min_{d\in\mathcal{D}} \max_{\gamma\in\mathcal{K}} \Phi(d,\gamma) = \min_{d\in\mathcal{D}} \Phi(d,\bar{\gamma}) = 0, \tag{6.24}$$

with $\bar{\gamma} = (\bar{\alpha}_0, \{\bar{\alpha}_i^1\}_{i\in E}, \{\bar{\alpha}_j^2\}_{j\in I}, \bar{\nu})$.

We readily deduce from (**CQ**) that $\bar{\alpha}_0 \neq 0$. By scaling the multipliers we may arrange that $\bar{\alpha}_0/c = 1$.

Now define p to be the solution to the differential equation

$$-\dot{p}_r(t) = (f_x)_r(t, x_r^{\bar{\mu}}(t), \bar{\mu}(t))^T \left(p_r(t) + \int_{[0,t)} q_x(s, x_r^{\bar{\mu}}(s))\bar{\nu}(ds)\right)$$

and

$$-p_r(1) = \int_{[0,1]} q_x(s, x_r^{\bar\mu}(s))\bar\nu(ds) + \phi_x(x_r^{\bar\mu}(1)) + \sum_{i\in E} \bar\alpha_i^1 (h_i^1)_x(x_r^{\bar\mu}(1)) + \sum_{j\in I} \bar\alpha_j^2 (h_j^2)_x(x_r^{\bar\mu}(1)).$$

We have, for any $d \in \mathcal{D}$,

$$\Phi(d,\bar\gamma) + \int_{[0,1]} \left(p_r(t) + \int_{[0,t)} q_x(s, x_r^{\bar\mu}(s))\bar\nu(ds) \right)^T (\dot y_r^{\bar\mu,d}(t) - (f_x)_r(t, x_r^{\bar\mu}(t), \bar\mu(t))^T y_r^{\bar\mu,d}(t) - \int_\Omega f_u(t, x_r^{\bar\mu}(t), u) d(t,u)\bar\mu(t)(du))dt \geq 0.$$

This inequality reduces, via an integration by parts, to

$$\int_{[0,1]} (p_r(t) + \int_{[0,t)} q_x(s, x_r^{\bar\mu}(s))\bar\nu(ds))^T \int_\Omega f_u(t, x_r^{\bar\mu}(t), u) d(t,u)\bar\mu(t)(du)dt \geq 0.$$

These relationships imply that $(x_r^{\bar\mu}, \bar\mu)$ satisfies the stated necessary conditions. ∎

7. Appendix.

Proof of Lemma 6.1. Take any arbitrary $\tilde\mu \in \bar{\mathcal{U}}$ which is feasible w.r.t. the inequality constraints. Let $r > 0$ be a number such that

$$\max_{i\in E} |\hat h_i^1(\mu)| < r$$

$\forall \mu \in \bar{\mathcal{U}}$. We deduce from (**CQ**) that there is a simplex in $\mathcal{E}(\tilde\mu) \subset \mathcal{R}^{n_E}$ ($n_E = |E|$) with vertices $\{e_j\}_{j=0}^{n_E}$ which contains 0 as an interior point. By definition of $\mathcal{E}(\tilde\mu)$, there exist $d_0, \ldots, d_{n_E} \in \mathcal{D}$ and $\delta > 0$ such that for $j = 0, \ldots, n_E$

$$\begin{aligned} \{\langle \nabla \hat h_i^1(\tilde\mu), d_j\rangle_r\}_{i\in E} &= e_j, \\ \max_{i\in I} [\hat h_i^2(\tilde\mu) + \langle \nabla \hat h_i^2(\tilde\mu), d_j\rangle_r] &\leq -\delta, \\ \max_{t\in T} [\hat q(\tilde\mu)(t) + \langle \nabla \hat q(\tilde\mu)(t), d_j\rangle_r] &\leq -\delta. \end{aligned}$$

Let $(\lambda_0, \lambda_1, \ldots, \lambda_{n_E})$ be the barycentric coordinates of 0 w.r.t. the vertices e_j of the simplex; i.e.,

$$0 \left(= \sum_{j=0}^{n_E} \lambda_j e_j \right) = \nabla \hat h^1(\tilde\mu) \circ \sum_{j=0}^{n_E} \lambda_j d_j.$$

Here

$$\nabla \hat h^1(\mu) \circ d := \{\langle \nabla \hat h_i^1(\mu), d\rangle_r\}_{i\in E}.$$

We shall also write

$$\hat h^1(\mu) := \{\hat h_i^1(\mu)\}_{i\in E}.$$

$\nabla \tilde h^1(u) \circ d$ and $\tilde h^1(u)$ are defined analogously.

Since the vertices are in general position and 0 is an interior point, the λ_i's are all positive, and we may find $\delta_1 > 0$ such that

$$\left\{ \lambda_0 - \sum_{j=1}^{n_E} \alpha_j, \lambda_1 + \alpha_1, \ldots, \lambda_{n_E} + \alpha_{n_E} \right\} \in \left\{ \gamma \in \mathcal{R}^{n_E+1} : \ \gamma_j \geq 0 \ \forall j, \ \sum_{j=0}^{n_E} \gamma_j = 1 \right\}$$

whenever $\alpha \in \mathcal{B}(0, \delta_1) \subset \mathcal{R}^{n_E}$. ($\mathcal{B}(0, \delta_1)$ is a ball of radius δ_1.) Furthermore the $n_E \times n_E$ matrix $M(\mu)$ defined by

$$M(\mu)\alpha := \sum_{j=1}^{n_E} \nabla \hat{h}^1(\mu) \circ \alpha_j (d_j - d_0) \tag{7.1}$$

is invertible for $\mu = \tilde{\mu}$ from the definition of $d_j, \ j = 1, \ldots, n_E$.

In consequence of hypothesis (**CQ**) and in view of the continuity properties of the mapping $\mu \to y_r^{\mu,d}$ for fixed d, we may choose a neighborhood $\mathcal{O}(\tilde{\mu})$ of $\tilde{\mu}$ in $\bar{\mathcal{U}}$ and numbers $r > 0$ and $\delta_2 \in (0, r^{-1}]$ such that for any $u \in \mathcal{U}$ satisfying $\mathbf{u} \in \mathcal{O}(\tilde{\mu})$

(i) $$\max_{i \in I} [\tilde{h}_i^2(u) + \langle \nabla \tilde{h}_i^2(u), v_j - u \rangle] \leq -\delta/2 \ \ \forall j,$$

(ii) $$\max_{t \in T} [\tilde{q}(u)(t) + \langle \nabla \tilde{q}(u)(t), v_j - u \rangle] \leq -\delta/2 \ \ \forall j,$$

(iii) $M(u)$ is invertible,

(iv) $$\left\| M(u)^{-1} \nabla \tilde{h}^1(u) \circ \left(\left(\sum_{j=0}^{n_E} \lambda_j v_j \right) - u \right) \right\| \leq \delta_1/2,$$

(v) $$\delta_2 \| M(u)^{-1} \| n_E^{1/2} \leq \delta_1/2.$$

(In (v) the norm is the Euclidean norm and $M(u)$ is defined analogously to $M(\mu)$.)

Here the controls $v_j \in \mathcal{U}$, $j = 0, \ldots, n_E$, are defined to be

$$v_j(t) := u(t) + d_j(t, u(t)).$$

Now suppose that the control u is feasible w.r.t. the inequality constraints and $\tilde{h}^1(u) \neq 0$. Set

$$\alpha = M(u)^{-1} \left[-\nabla \tilde{h}^1(u) \circ \left(\left(\sum_{j=0}^{n_E} \lambda_j v_j \right) - u \right) - \delta_2 \| \tilde{h}^1(u) \|_\infty^{-1} \tilde{h}^1(u) \right], \tag{7.2}$$

in which $\| \tilde{h}^1(u) \|_\infty := \max_{i \in E} |\tilde{h}_i^1(u)|$. Notice that, by properties (iv) and (v), $\|\alpha\| \leq \delta_1$. Also set

$$\hat{v} = v_0 + \sum_{j=1}^{n_E} (\lambda_j + \alpha_j)(v_j - v_0).$$

Because $\|\alpha\| \leq \delta_1$ we have that $\hat{v} \in \mathcal{U}$. Finally we define v to be

$$v = u + (\| \tilde{h}^1(u) \|_\infty / r)(\hat{v} - u).$$

Since $\| \tilde{h}^1(u) \|_\infty / r \leq 1$, it follows that $v \in \mathcal{U}$. We now verify that this control function has the required properties.

Notice first that

$$\|v-u\|_{\mathcal{L}^2} \leq (2d/r)\ \|\ \tilde{h}^1_i(u)\|_\infty, \tag{7.3}$$

where d is a bound on the $\mathcal{L}^2_m[T]$ norms of elements in $\mathcal{U}$.

We have from (7.1) and (7.2) that

$$\begin{aligned} M(u)\alpha &= \nabla\tilde{h}^1(u)\circ\sum_{j=1}^{n_E}\alpha_j(v_j-v_0) \\ &= -\nabla\tilde{h}^1(u)\circ\left(\sum_{j=1}^{n_E}\lambda_j(v_j-v_0)+v_0-u\right)-\delta_2\|\tilde{h}^1(u)\|_\infty^{-1}\tilde{h}^1(u). \end{aligned}$$

By definition of $\hat{v}$,

$$\nabla\tilde{h}^1(u)\circ(\hat{v}-u) = -\delta_2\|\tilde{h}^1(u)\|_\infty^{-1}\tilde{h}^1(u).$$

But then

$$\nabla\tilde{h}^1(u)(v-u) = -(\delta_2/r)\tilde{h}^1(u).$$

Since $\delta_2/r \leq 1$, it follows that

$$\begin{aligned} &\max_{i\in E}|\tilde{h}^1_i(u)+\langle\nabla\tilde{h}^1_i(u),v-u\rangle| - \max_{i\in E}|\tilde{h}^1_i(u)| \\ &\qquad\leq -(\delta_2/r)\|\tilde{h}^1(u)\|_\infty. \end{aligned} \tag{7.4}$$

We deduce from property (i) that

$$\left\langle\nabla\tilde{h}^2_j(u), v_0+\sum_{i=1}^{n_E}(\lambda_i+\alpha_i)(v_i-v_0)-u\right\rangle \leq -\tilde{h}^2_j(u)-\delta/2\ \forall j\in I.$$

It follows that

$$\langle\nabla\tilde{h}^2_j(u),v-u\rangle \leq (\|\tilde{h}^1(u)\|_\infty/r)(-\tilde{h}^2_j(u)-\delta/2)\ \forall j\in I.$$

Since $\|\tilde{h}^1(u)\|_\infty/r \leq 1$ and $\tilde{h}^2_j(u)\leq 0$ for all $j\in I$, we deduce that

$$\tilde{h}^2_j(u)+\langle\nabla\tilde{h}^2_j(u),v-u\rangle \leq -(\delta/(2r))\|\tilde{h}^1(u)\|_\infty\ \forall j\in I. \tag{7.5}$$

Likewise we deduce from property (ii) that

$$\tilde{q}(u)(t)+\langle\nabla\tilde{q}(u)(t),v-u\rangle \leq -(\delta/(2r))\|\tilde{h}^1(u)\|_\infty\ \forall t\in T. \tag{7.6}$$

Surveying inequalities (7.3)–(7.6), we see that v satisfies all relevant conditions for completion of the proof when we set $K_1=\min\{\delta_2,\delta/(2r)\}$ and $K_2=2d/r$, numbers whose magnitudes do not depend on our choice of u. ∎

REFERENCES

[1] W. Alt and K. Malanowski, *The Lagrange–Newton method for state-constrained optimal control problems*, Comput. Optim. Appl., 4 (1995), pp. 217–239.

[2] J.P. Aubin and I. Ekeland, *Applied Nonlinear Analysis*, Wiley Interscience, New York, 1984.

[3] T.E. Baker and E. Polak, *On the optimal control of systems described by evolution equations*, SIAM J. Control Optim., 32 (1994), pp. 224–260.

[4] R. Bulirsch, F. Montrone, and H.J. Pesch, *Abort landing in the presence of windshear as a minimax optimal control problem, part* 1: *Necessary conditions*, J. Optim. Theory Appl., 70 (1991), pp. 1–23.

[5] R. Bulirsch, F. Montrone, and H.J. Pesch, *Abort landing in the presence of windshear as a minimax optimal control problem, part* 2: *Multiple shooting and homotopy*, J. Optim. Theory Appl., 70 (1991), pp. 223–254.

[6] R.P. Fedorenko, *Priblizhyonnoye reshenyie zadach optimalnovo upravlenya*, Nauka, Moscow, 1978.

[7] A.S. Drud, *A GRG code for large sparse dynamic nonlinear optimization problems*, Math. Programming, 31 (1985), pp. 153–191.

[8] D.Q. Mayne and E. Polak, *An exact penalty function algorithm for optimal control problems with control and terminal equality constraints, part* 1 *and part* 2, J. Optim. Theory Appl., 32 (1980), pp. 211–246; pp. 345–364.

[9] D.Q. Mayne and E. Polak, *An exact penalty function algorithm for control problems with state and control constraints*, IEEE Trans. Automat. Control, AC-32 (1987), pp. 380–387.

[10] K.C.P. Machielsen, *Numerical Solution of Optimal Control Problems with State Constraints by Sequential Quadratic Programming in Function Space*, CWI Tract 53, Centrum Wisk. Inform., Amsterdam, 1988.

[11] E. Polak, T.H. Yang, and D.Q. Mayne, *A method of centers based on barrier functions for solving optimal control problems with continuum state and control constraints*, SIAM J. Control Optim., 31 (1993), pp. 159–179.

[12] E. Polak and L. He, *Unified steerable phase*I*–phase*II *method of feasible directions for semi-infinite optimization*, J. Optim. Theory Appl., 69 (1991), pp. 83–107.

[13] R. Pytlak, *Runge–Kutta based procedure for optimal control of differential-algebraic equations*, J. Optim. Theory Appl., 97 (1998), pp. 675–705.

[14] R. Pytlak and R.B. Vinter, *PH*2*SOL Solver: An* $O(N)$ *Implementation of an Optimization Algorithm for a General Optimal Control Problem*, Research Report C93–36, Centre for Process Systems Engineering, Imperial College, 1993.

[15] R. Pytlak and R.B. Vinter, *A Feasible Directions Type Algorithm for Optimal Control Problems with State and Control Constraints: Convergence Analysis*, Research Report C96–24, Centre for Process Systems Engineering, Imperial College, 1996.

[16] R. Pytlak and R.B. Vinter, *A Feasible Directions Algorithm for Optimal Control Problems with State and Control Constraints: Implementation*, J. Optim. Theory Appl., submitted.

[17] G.I. Stassinopoulos and R.B. Vinter, *Conditions for convergence in the computation of optimal controls*, J. Inst. Math. Appl., 22 (1978), pp. 1–14.

[18] P. Tanartkit and L.T. Biegler, *Stable decomposition for dynamic optimization*, Indust. Engrg. Chem. Res., 34 (1995), pp. 1253–1266.

[19] J. Warga, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.

[20] J. Warga, *Iterative procedure for constrained and unilateral optimization problems*, SIAM J. Control Optim., 20 (1982), pp. 360–376.

[21] L.J. Williamson and E. Polak, *Relaxed controls and the convergence of optimal control algorithms*, SIAM J. Control Optim., 14 (1976), pp. 737–757.

[22] S.J. Wright, *Structured interior point methods for optimal control*, in Proc. 30th CDC, Brighton, UK, 1991, pp. 1711–1716.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 6, pp. 2020–2039, November 1998

010

NEWTON'S LAW AND INTEGRABILITY OF NONHOLONOMIC SYSTEMS*

A. M. BLOCH† AND P. E. CROUCH‡

Abstract. In this paper we consider nonholonomic control systems on Riemannian manifolds. Such systems evolve on subbundles of tangent bundles, defined by the nonholonomic constraints. This paper promotes the view of such systems as the restriction to the nonholonomic subbundle of "Newton law"-type problems on the entire tangent bundle, defined by, in general, non-Riemannian connections. These connections should be related to specific geometric properties of the nonholonomic system. We introduce a particular class of connections and demonstrate the richness of the class through four examples—the rolling ball, the constrained particle, the rolling penny, and the generalized rolling ball. This class of connections is strongly related to questions of integrability of the original nonholonomic system. This, in turn, provides additional insight into the relation between nonholonomic control systems formulated as kinematic equations and those that are formulated as the full dynamic equations.

Key words. Newton's law, nonholonomic, integrability, connections

AMS subject classifications. 93C15, 53C05, 53C20, 58F07, 93B17

PII. S0363012995291634

1. Introduction.

In this paper we consider nonholonomic control systems on Riemannian manifolds, as a general framework in which to consider control problems associated with classical nonholonomic mechanical systems. There has been much work on nonholonomic mechanics and associated control systems, including work by the authors [1], [2], [3], and others including Vershik and Gershkovich [4], Vershik and Fadeev [5], and Bloch, Krishnaprasad, Marsden, and Murray [9]. An abstract notion of mechanics may be cast in terms of second-order equations on manifolds through the extra structure of a connection and from which a notion of "Newton law" systems may be formed. Systems with constraints limited to only the configuration variables, denoted holonomic systems, can be identified with Lagrangian systems. Constraints on such systems which include the phase variables, not simply the configuration variables, yield nonholonomic systems through application of d'Alembert's principle. These systems are not naturally identified as Lagrangian systems. If one treats the constraints as in a constrained variational problem, rather than invoking d'Alembert's principle, one obtains a class of systems called vakonomic systems [4]–[5]. Nonholonomic and vakonomic systems are contrasted in Vershik and Fadeev [5] (see also Bloch and Crouch [2]).

Since nonholonomic systems are not naturally Lagrangian, one can ask about other natural structures identified with nonholonomic systems. Vershik and Gershkovich [4] identify a new connection on a reduced phase space on which nonholonomic systems are given as a "Newton law" system. Other approaches include Bates and

*Received by the editors September 13, 1995, accepted for publication (in revised form) July 7, 1997; published electronically August 31, 1998.
http://www.siam.org/journals/sicon/36-6/29163.html

†Department of Mathematics, University of Michigan, Ann Arbor, MI 48109 (abloch@math.lsa.umich.edu). The research of this author was partially supported by the National Science Foundation PYI grant DMS-91-57556, AFOSR grant F49620-96-1-0100, a Guggenheim Fellowship, and the Institute for Advanced Study.

‡Center for Systems Science and Engineering, Arizona State University, Tempe, AZ 85287-5506 (peter.crouch@asu.edu). The research of this author was partially supported by NATO grant CRG 910926.

Sniatycki [6], Bloch, Krishnaprasad, Marsden, and Murray [9]. In this regard it is interesting to note the fundamental contributions by Cartan [7] and Vershik and Fadeev [5] (see also Arnold [10]).

In this work we expand upon the previous note [8] by the authors and take a different point of view from the work described above. The fundamental perspective is that nonholonomic systems may be viewed as "Newton law" systems defined by connections on the entire phase space which are related to the system geometry.

Although there are many choices for these connections, we concentrate on one particular class which has a very appealing structure. This sheds some light on the ability to generate nonholonomic systems from "Newton law" systems defined by metric connections. The structure also sheds light on the kinematic and dynamic formulations on nonholonomic systems and the associated integrability problems.

To be more specific, we suppose that $M^n, \langle \cdot, \cdot \rangle$ is a Riemannian manifold with symmetric Riemannian connection ∇ and covariant derivative $D/\partial t$. We sometimes denote the metric by G. We may describe a class of nonholonomic control systems on M, driven by external forces, by following the works of Bloch and Crouch [1], [2], [3]. These are generally defined by systems of equations:

$$\frac{D^2 q}{\partial t^2} = \sum_{i=1}^{m} \lambda_i W_i + F; \quad q \in M, \quad \omega_i(\dot{q}) \equiv 0, \quad 1 \leq i \leq m, \tag{1}$$

where ω_i, $1 \leq i \leq m$, are m independent constraint forms on M satisfying

$$\omega_i(X) = \langle W_i, X \rangle; \qquad X \in \Gamma(TM),$$

where $\Gamma(V)$ denotes the space of sections of a vector bundle V, $W_i \in \Gamma(TM)$, $1 \leq i \leq m$, and $F \in \Gamma(TM)$ is an arbitrary external force field.

System (1) is nonholonomic precisely when the distribution $N \subset TM$ defined by

$$N_p = \{X_p : X \in \Gamma(TM), \omega_i(X) \equiv 0, 1 \leq i \leq m\}, \quad p \in M, \tag{2}$$

is not integrable. Let $a_{ij} = \omega_i(W_j)$, $1 \leq i, j \leq m$. The independence of the forms ω_i implies that the matrix $[a_{ij}]_{1 \leq i,j \leq m}$ is invertible on the whole of M. By differentiating the constraints we obtain from (1)

$$\frac{D\omega_i}{\partial t}(\dot{q}) + \omega_i\left(\sum_j \lambda_j W_j + F \right) \equiv 0, \quad 1 \leq i \leq m,$$

and so we may solve for the multipliers λ_i as

$$\lambda_k = -\sum_j a_{kj}^{-1} \left(\frac{D\omega_j}{\partial t}(\dot{q}) + \omega_j(F) \right), \quad 1 \leq k \leq m.$$

We may, therefore, obtain an equivalent formulation of the system described by (1) in the form

$$\frac{D^2 q}{\partial t^2} + \sum_{k,j} W_k a_{kj}^{-1} \frac{D\omega_j}{\partial t}(\dot{q}) = F - \sum_{k,j} W_k a_{k,j}^{-1} \omega_j(F), \tag{3}$$

$$\omega_i(\dot{q}) = 0, \quad 1 \leq i \leq m.$$

Much effort has gone into understanding or rationalizing this system of equations. We briefly review the approach of Vershik and Gershkovich [4] and Fadeev and Vershik [5].

For any $X \in \Gamma(TM)$ we set

$$\pi_N(X) = X - \sum_{ki} W_k a_{ki}^{-1} \omega_i(X).$$

It is evident that $\pi_N(X) \in \Gamma(N)$ and $\pi_N(\pi_N(X)) = \pi_N(X)$. Thus π_N is the projection onto the subbundle $N \subset TM$. By differentiating the constraints we find that

$$\begin{aligned}\pi_N\left(\frac{D^2q}{\partial t^2}\right) &= \frac{D^2q}{\partial t^2} - \sum_{i,k} W_k a_{ki}^{-1} \omega_i\left(\frac{D^2q}{\partial t^2}\right) \\ &= \frac{D^2q}{\partial t^2} + \sum_{i,k} W_k a_{ki}^{-1} \frac{D\omega_i}{\partial t}(\dot{q}).\end{aligned}$$

We may, therefore, write (3) in the form

$$\pi_N\left(\frac{D^2q}{\partial t^2}\right) = \pi_N(F), \quad \dot{q} \in N. \tag{4}$$

Definition 1.1. *In general, if $\bar{\nabla}$ is any connection on M, with corresponding covariant derivative $\bar{D}/\partial t$, then we define the following second-order system*

$$\frac{\bar{D}^2q}{\partial t^2} = \bar{F}, \quad q \in M \tag{5}$$

to be a "Newton law" system on M, with external forces modeled by the vector field $\bar{F}$ on M.

Thus, the nonholonomic system (1) may be viewed as simply the projection (4) of the Newton law system, $D^2q/\partial t^2 = F$, onto the subbundle N. However, even more is true. We may define a new connection ∇' on M by setting

$$\nabla'_X Y = \nabla_X Y + \sum_{k,i=1}^{m} W_i a_{ik}^{-1} (\nabla_X \omega_k)(Y); \quad X, Y \in \Gamma(TM), \tag{6}$$

which, in turn, defines a covariant derivative $D'/\partial t$. Moreover, from (6), the connection ∇' has the property

$$\omega_i(\nabla'_X Y) = \omega_i(\nabla_X Y) + (\nabla_X \omega_i)(Y) = X(\omega_i(Y)).$$

Thus, if $X, Y \in \Gamma(N)$, $\nabla'_X Y \in \Gamma(N)$, and so $\nabla'|_N$ defines a connection on the subbundle N. Thus, we may write the nonholonomic system (1) in the form

$$\frac{D'^2q}{\partial t^2} = \pi_N(F); \dot{q} \in N, \tag{7}$$

and view the system as a Newton law system on the bundle N. These observations were made by Fadeev and Vershik [5]. We take a different perspective in this paper and begin by noting that system (7) defines a perfectly good Newton law system on all of TM:

$$\frac{D'^2q}{\partial t^2} = \pi_N(F), \quad q \in M. \tag{8}$$

DEFINITION 1.2. *Consider a Newton law system on M, defined by a connection $\bar{\nabla}$ and external force $\bar{F}$ of the form*

$$\frac{\bar{D}^2 q}{\partial t^2} = \bar{F}, \quad q \in M.$$

We say that the system has the restriction property if it restricts to the subbundle N, and on N it coincides with the nonholonomic system (1).

It follows that system (8) has the restriction property. However, the choice of connection ∇' seems to be an arbitrary choice, and motivated by conversations with Vershik, we ask if there are not more natural choices, related to the geometry of the nonholonomic system? To begin an analysis of this question in this paper, we consider another n dimensional bundle over M, denoted V, with connection ∇^V, and a vector bundle isomorphism $A : TM \to V$, and introduce a class of connections on TM, parameterized by A, and a symmetric two tensor $S : TM \otimes TM \to \mathbb{R}$ on M, denoted $\nabla^{(A,S)}$. Corresponding to $\nabla^{(A,S)}$, we may construct a covariant derivative $D^{(A,S)}/\partial t$, and pose a similar question. When does the system

$$\frac{D^{(A,S)^2} q}{\partial t} = \pi_N(F), \quad q \in M \tag{9}$$

have the restriction property? We answer this question by making various assumptions on the nonholonomic system and choices for V, ∇^V, and A. The defining property of the connection $\nabla^{(A,S)}$ is that when the system (9) is restricted to N it may be rewritten as

$$\frac{D^V v}{\partial t} = A_q(\pi_N(F)); \quad \dot{q} = A_q^{-1}(v) \in N.$$

The particular choices of V, ∇^V, and A of course impact the form and properties of these transformed equations. We investigate the results for the particular choices made in answering the question above, especially the question of integrability.

We may also ask another question. Is there another metric g on M, (which is different from G), generating a Riemannian connection ∇^g on M, such that the Newton law system

$$\frac{D^{g^2} q}{\partial t^2} = \pi_N(F), \quad q \in M$$

has the restriction property? We provide a partial answer to this question by providing sets of conditions on g, A, and S, so that $\nabla^g = \nabla^{(A,S)}$. Examples we have examined thus far do not seem to arise from such a Riemannian connection.

We now outline the specific agenda for the paper. In section 2 we define the general class of connections $\nabla^{(A,S)}$ on M and prove some properties of these connections and the associated Newton law systems. In section 3 we further develop our understanding of the general class of connections by introducing a specific subclass of nonholonomic systems (in which M must be parallelizable and including a large class of systems of interest) and examining the proposed class of connections for the specific class of nonholonomic systems. In section 4 we study the question of existence of Newton law systems generated by Riemannian connections which have the restriction property. In section 5 we discuss the relationship between the class of connections $\nabla^{(A,S)}$ and integrability of the corresponding Newton law, or nonholonomic systems. Finally,

in section 6 we give four examples within the subclass identified in section 3 and demonstrate the results of the paper in each case. These examples illustrate that the connections introduced in section 2 are of real interest to the understanding of nonholonomic systems in general.

2. Connections defined by bundle maps. We introduce another vector bundle V over M; isomorphic to TM, $\pi : V \to M$, with $V_q = \pi^{-1}(q)$ and $\dim(V_q) = n$ for $q \in M$. We assume that V comes equipped with a connection ∇^V, which defines a covariant derivative $D^V/\partial t$ on V. We let $A : TM \to V$ be a vector bundle isomorphism so that

$$A_q : T_qM \to V_q$$

is a nonsingular vector space isomorphism for each $q \in M$, and hence $A^{-1} : V \to TM$ is well defined. Assume also that we are given a Newton law system on M,

$$\frac{\bar{D}^2 q}{\partial t^2} = \bar{F}$$

defined by a connection $\bar{\nabla}$ on M. We are interested in how this system behaves under the bundle map

$$v = A_q(\dot{q}). \tag{10}$$

To differentiate this expression we must introduce a connection on the bundle $L(TM; V)$ over M, of all bundle maps from TM to V. The connections $\bar{\nabla}$ and ∇^V allow us to define the induced connection $\bar{\nabla}$ on $L(TM, V)$ by setting

$$(\bar{\nabla}_X A)(Y) \stackrel{\Delta}{\to} = \nabla^V_X(AY) - A(\bar{\nabla}_X Y), A \in \Gamma(L(TM; V)), X, Y \in \Gamma(TM). \tag{11}$$

The implication for the notation is that the connection ∇^V is fixed, whereas the connection $\bar{\nabla}$ on M is allowed to vary. We may now differentiate equation (10) to obtain

$$\frac{D^V v}{\partial t} = \frac{\bar{D} A_q}{\partial t}(\dot{q}) + A_q\left(\frac{\bar{D}^2 q}{\partial t^2}\right). \tag{12}$$

Substituting our Newton law system (5), we obtain the following system on V:

$$\begin{aligned} \frac{D^V v}{\partial t} &= \frac{\bar{D} A_q}{\partial t}(A_q^{-1} v) + A_q(\bar{F}), \\ \dot{q} &= A_q^{-1}(v). \end{aligned} \tag{13}$$

Thus, by "changing coordinates" as in (10), we have complicated our Newton law system (5) by the introduction of the term

$$\frac{\bar{D} A_q}{\partial t}(A_q^{-1} v) = \frac{\bar{D} A_q}{\partial t}(\dot{q}). \tag{14}$$

We can, therefore, ask the natural question "Can we choose the connection $\bar{\nabla}$ so that the term (14) vanishes identically?" We may recast this question by introducing another definition.

DEFINITION 2.1. *$A \in \Gamma(L(TM;V))$ is Killing with respect to $\bar{\nabla}$ if*

$$(\bar{\nabla}_X A)(Y) + (\bar{\nabla}_Y A)(X) \equiv 0, \quad X, Y \in \Gamma(TM), \tag{15}$$

or

$$(\bar{\nabla}_X A)(X) \equiv 0, \quad X \in \Gamma(TM).$$

Clearly, the last question may now be rephrased as "Is it possible to choose $\bar{\nabla}$ on M so that A is Killing?" Our choice of terminology is clearly reminiscent of the definition in the case of vector fields $X \in \Gamma(TM)$.

Recall that the Riemannian connection ∇, defined by the metric G, is uniquely determined by the properties

(i) $X\langle Y, Z\rangle = \langle \nabla_X Y, Z\rangle + \langle Y, \nabla_X Z\rangle$, (i.e., $\nabla G \equiv 0$);

(ii) $T(X,Y) \overset{\Delta}{=} \nabla_X Y - \nabla_Y X - [X,Y] = 0$.

T is known as the torsion tensor corresponding to ∇. In general, we say any connection $\bar{\nabla}$ on M is symmetric if the torsion tensor corresponding to $\bar{\nabla}$ is identically zero.

THEOREM 2.2. *The unique symmetric connection ∇^A on M, so that $A \in \Gamma(L(TM;V))$ is Killing with respect to ∇^A is given by*

$$\nabla^A_X Y = \nabla_X Y + \frac{1}{2} A^{-1}((\nabla_X A)(Y) + (\nabla_Y A)(X)); \quad X, Y \in \Gamma(TM). \tag{16}$$

Proof. We first show that A is indeed Killing with respect to ∇^A defined by (16). From the definition (11), we find that

$$\begin{aligned}(\nabla^A_X A)(Y) &= \nabla^V_X(AY) - A(\nabla^A_X Y) \\ &= \nabla^V_X(AY) - A(\nabla_X Y) - \frac{1}{2}((\nabla_X A)(Y) + (\nabla_Y A)(X)).\end{aligned}$$

Thus

$$\begin{aligned}(\nabla^A_X A)(Y) + (\nabla^A_Y A)(X) &= \nabla^V_X(AY) + \nabla^V_Y(AX) - A(\nabla_X Y + \nabla_Y X) \\ &\quad -((\nabla_X A)(Y) + (\nabla_Y A)(X)) \\ &= 0.\end{aligned}$$

Since ∇^A is a symmetric connection, if $\bar{\nabla}$ is any other symmetric connection, we may write

$$\bar{\nabla}_X Y = \nabla^A_X Y + S(X,Y),$$

where S is a symmetric two tensor. Thus

$$\begin{aligned}(\bar{\nabla}_X A)(X) &= \nabla^V_X(AX) - A(\bar{\nabla}_X X) = \nabla^V_X(AX) - A(\nabla^A_X X) - AS(X,X) \\ &= (\nabla^A_X A)(X) - AS(X,X).\end{aligned}$$

But A is Killing with respect to ∇^A, so

$$(\bar{\nabla}_X A)(X) = -AS(X,X).$$

Since A is full rank, if A is to be Killing with respect to $\bar{\nabla}$, we must have $S \equiv 0$. Hence, $\bar{\nabla} = \nabla^A$ and ∇^A is unique. ∎

Since we are interested primarily in nonholonomic systems (1), it is interesting to study this result in the context of Newton law systems (5) which have the restriction property. In this case we can weaken the requirement in (13) that the term $\bar{D}A/\partial t(\dot{q})$ vanish identically, and simply require that the term vanish for $\dot{q} \in N$. In this case the transformation (10) maps the nonholonomic system (1) into a system of the form

$$\frac{D^V v}{\partial t} = A_q(\pi_N(F)), \quad \dot{q} = A_q^{-1}(v), \quad \dot{q} \in N. \tag{17}$$

DEFINITION 2.3. *$A \in \Gamma(L(TM;V))$ is Killing on N with respect to $\bar{\nabla}$ if*

$$(\bar{\nabla}_X A)(Y) + (\bar{\nabla}_Y A)(X) = 0 \qquad X, Y \in \Gamma(N), \tag{18}$$

or

$$(\bar{\nabla}_X A)(X) = 0 \qquad X \in \Gamma(N).$$

THEOREM 2.4. *Let $\nabla^{(A,S)}$ be a symmetric connection on M for which $A \in \Gamma(L(TM;V))$ is Killing on N, with respect to $\nabla^{(A,S)}$. Then*

$$\nabla_X^{(A,S)} Y = \nabla_X Y + \frac{1}{2} A^{-1}((\nabla_X A)(Y) + (\nabla_Y A)(X)) + S(X,Y), \quad X, Y \in \Gamma(TM), \tag{19}$$

for some symmetric two tensor S such that $S|_N \equiv 0$.

Proof. Since $S(X,Y) = 0$, for $X, Y \in \Gamma(N)$, the proof of the previous theorem demonstrates that A is indeed Killing on N with respect to $\nabla^{(A,S)}$, as defined in (19). The same proof demonstrates that the additional term S, such that $S|_N \equiv 0$, is the only flexibility in the definition of such a connection. □

Corresponding to the connection $\nabla^{(A,S)}$ on M, defined by (19), there exists a covariant differentiation $D^{(A,S)}/\partial t$. From (19) we have

$$\frac{D^{(A,S)^2} q}{\partial t^2} = \frac{D^2 q}{\partial t^2} + A^{-1}\left(\frac{DA}{\partial t}\right)(\dot{q}) + S(\dot{q}, \dot{q}). \tag{20}$$

We may now ask another question. When does the Newton law system

$$\frac{D^{(A,S)^2} q}{\partial t^2} = \pi_N(F), \quad q \in M$$

have the restriction property? Since $S|_N \equiv 0$, by comparing equations (3) and (20), we obtain the necessary and sufficient condition

$$A^{-1}\frac{DA}{\partial t}(\dot{q}) - \sum_{k,j} W_k a_{kj}^{-1} \frac{D\omega_j}{\partial t}(\dot{q})\Big|_N \equiv 0. \tag{21}$$

Now we set $A(W_k) \stackrel{\Delta}{\to} = \hat{W}_k$, $1 \le k \le m$, and

$$B \stackrel{\Delta}{\to} = A - \sum_{j,k=1}^{m} \hat{W}_k a_{kj}^{-1} \omega_j = A \circ \pi_N. \tag{22}$$

It follows that $B|_{N^\perp} = 0$ and

$$(\nabla_X B)(Y) = (\nabla_X A)(Y) - \sum_{j,k=1}^{m} \nabla_X(\hat{W}_k a_{kj}^{-1})\omega_j(Y) - \sum_{j,k=1}^{m} \hat{W}_k a_{kj}^{-1}(\nabla_X \omega_j)(Y),$$

so

$$A^{-1}(\nabla_X B)(X) = A^{-1}(\nabla_X A)(X) - \sum_{j,k=1}^{m} W_k a_{kj}^{-1}(\nabla_X \omega_j)(X); X \in \Gamma(N).$$

Since A is full rank, condition (21) is, therefore, equivalent to the fact that B is Killing on N with respect to ∇.

THEOREM 2.5. *Given a vector bundle V, connection ∇^V, symmetric tensor S, $S|_N = 0$, then a necessary and sufficient condition for the Newton law system on M*

$$\frac{D^{(A,S)^2} q}{\partial t} = \pi_N(F), \quad q \in M$$

to have the restriction property is that $A \circ \pi_N$ is Killing on N with respect to ∇.

3. A special class of nonholonomic systems. Given a vector bundle V, connection ∇^V, we wish to find examples of bundle maps $A : TM \to V$ so that $A \circ \pi_N$ is Killing on N with respect to ∇. To work toward this goal we restrict our attention to a special subclass of nonholonomic systems which we describe here.

DEFINITION 3.1. *A one form ν on M is said to be a Killing form with respect to a connection $\bar{\nabla}$ on M if*

$$(\bar{\nabla}_X \nu)(X) \equiv 0, \qquad X \in \Gamma(TM). \tag{23}$$

A one form ν on M is said to be a Killing form on N with respect to a connection $\bar{\nabla}$ on M if

$$(\bar{\nabla}_X \nu)(X) \equiv 0, \qquad X \in \Gamma(N). \tag{24}$$

We note that if $\nu(X) = \langle V, X\rangle$, $X \in \Gamma(TM)$, then

$$\begin{aligned} X(\nu(Y)) &= \langle \nabla_X V, Y\rangle + \langle V, \nabla_X Y\rangle \\ &= (\nabla_X \nu)(Y) + \nu(\nabla_X Y). \end{aligned}$$

Thus $(\nabla_X \nu)(Y) = \langle \nabla_X V, Y\rangle$, and, in particular, $(\nabla_X \nu)(X) = \langle \nabla_X V, X\rangle$. Hence, ν is Killing with respect to ∇ if and only if V is Killing in the classical sense. We now describe the importance of Killing forms for Newton law and nonholonomic systems.

LEMMA 3.2. *If ν is a Killing form with respect to $\bar{\nabla}$, then $\nu(\dot{q})$ is a constant of motion for the Newton law system $\bar{D}^2 q/\partial t^2 = 0$.*

Proof.

$$\frac{d}{dt}\nu(\dot{q}) = \left(\frac{\bar{D}\nu}{\partial t}\right)(\dot{q}) + \nu\left(\frac{\bar{D}^2 q}{\partial t^2}\right) = 0. \qquad \square$$

LEMMA 3.3 (Arnold [10]). *If ν is a Killing form on N with respect to ∇ such that $\nu(N^\perp) = 0$, then $\nu(\dot{q})$ is a constant of motion for the nonholonomic system* (3) *with $F \equiv 0$.*

Proof.

$$\frac{d}{dt}\nu(\dot{q}) = \frac{D\nu}{\partial t}(\dot{q}) + \nu\left(\frac{D^2q}{\partial t^2}\right)$$
$$= \frac{D\nu}{\partial t}(\dot{q}) + \sum_{k,j} \nu(W_k)a_{kj}^{-1}\frac{D\omega_j}{\partial t}(\dot{q}) = 0,$$

since for $\dot{q} \in N$, $\frac{D\nu}{\partial t}(\dot{q}) = 0$, and $W_k \in N^{\perp}$. □

Clearly, the existence of Killing one forms as in Lemmas 3.2 and 3.3 is linked to the integrability of the systems. We discuss this further in section 5.

Assumption 3.1.

(i) M is parallelizable;

(ii) In the nonholonomic system (1) we may complete the set $\{\omega_1, \dots, \omega_m\}$ to a basis of $\Gamma(T^*M)$ by the set $\{\nu_1, \dots, \nu_{n-m}\}$ of one forms ν_k that are Killing on N with respect to ∇ and $\nu_k(N^{\perp}) = 0$, $1 \le k \le n-m$.

Let V_k, $1 \le k \le n-m$ be the vector fields defined by $\nu_k(X) = \langle V_k, X\rangle$, $X \in \Gamma(TM)$, and set $A(V_k) = \hat{V}_k$, $1 \le k \le n-m$, $b_{kj} = \nu_k(V_j)$, $1 \le k, j \le n-m$. Assumption 3.1 ensures that the matrix $[b_{kj}]_{1\le k,j\le n-m}$ is invertible on all of M. It follows that

$$B = A \circ \pi_N = \sum_{j,k=1}^{n-m} \hat{V}_j b_{jk}^{-1}\nu_k,$$

and so

$$A = \sum_{j,k=1}^{m} \hat{W}_j a_{jk}^{-1}\omega_k + \sum_{j,k=1}^{n-m} \hat{V}_j b_{jk}^{-1}\nu_k. \tag{25}$$

LEMMA 3.4. *$B = A\circ\pi_N$ is Killing on N with respect to ∇, under Assumption* 3.1*, if and only if*

$$\nabla_X^V\left(\sum_{j=1}^{n-m} \hat{V}_j b_{jk}^{-1}\right) \equiv 0, \qquad 1 \le k \le n-m, \quad X \in \Gamma(N). \tag{26}$$

Proof.

$$(\nabla_X B)(X) = \sum_k \nabla_X^V\left(\sum_j \hat{V}_j b_{jk}^{-1}\right)\nu_k(X) + \sum_{jk} \hat{V}_j b_{jk}^{-1}(\nabla_X\nu_k)(X).$$

Since ν_k is Killing on N with respect to ∇, $(\nabla_X B)(X) \equiv 0$, for $X \in \Gamma(N)$ if and only if condition (26) is satisfied. □

The condition expressed by (26) is not particularly attractive. We, therefore, make some choices for V, ∇^V, and A, namely, the following assumption.

Assumption 3.2. Under Assumption 3.1 set $V = TM$, $\hat{V}_k = \sum_m V_m b_{mk}$, $\hat{W}_k = \sum_m W_m a_{mk}$ so that

$$A = \sum_m V_m\nu_m + \sum_m W_m\omega_m \tag{27}$$

and

$$\nabla^V V_k = 0, \quad 1 \le k \le n-m, \quad \nabla^V W_k = 0, \quad 1 \le k \le m.$$

COROLLARY 3.5. *Under Assumptions* 3.1 *and* 3.2, $B = A \circ \pi_N$ *is always Killing on* N *with respect to* ∇.

COROLLARY 3.6. *Under Assumptions* 3.1 *and* 3.2, *then for any symmetric tensor* S, $S|_N = 0$, *the Newton law system on* M

$$\frac{D^{(A,S)^2} q}{\partial t^2} = \pi_N(F), \quad q \in M$$

has the restriction property. The corresponding nonholonomic system may be rewritten in the form

$$\frac{D^V v}{\partial t} = \sum_{i=1}^{n-m} V_i \nu_i(F), \quad \dot{q} = A_q^{-1}(v), \quad v, \dot{q} \in N. \tag{28}$$

Clearly, there may be other means of satisfying the property that $A \circ \pi_N$ be Killing on N with respect to ∇ in addition to assumptions 3.1 and 3.2. In particular, Theorem 2.5 does not preclude cases in which M is not parallelizable and more interesting choices of V and ∇^V are appropriate.

4. Riemannian connections compatible with bundle maps. In this section we consider the question, "Does there exist a metric g on M such that $\nabla^g = \nabla^{(A,S)}$ for some bundle map $A : TM \to V$ and symmetric tensor S." We first examine this question in general, and then under Assumptions 3.1 and 3.2, making a natural choice for g. We begin by giving a useful characterization of ∇^g.

LEMMA 4.1. *The unique Riemannian connection* ∇^g *on* M *corresponding to the metric* g *on* M *is given by*

$$\begin{aligned} g(Z, \nabla^g_X Y) = g(Z, \nabla_X Y) + \frac{1}{2}\{&(\nabla_Y g)(X,Z) \\ &+ (\nabla_X g)(Y,Z) - (\nabla_Z g)(X,Y)\}, \quad X, Y, Z \in \Gamma(TM). \end{aligned} \tag{29}$$

Proof. Note that ∇ is, of course, the unique Riemannian connection corresponding to the metric G. Clearly ∇^g, defined by (29), is a symmetric connection. Hence, if ∇^g, defined by (29), also satisifes $\nabla^g g \equiv 0$, it must be the unique Riemannian connection corresponding to the metric g. It is, therefore, sufficient to prove that $\nabla^g g \equiv 0$, or

$$Zg(X,Y) = g(\nabla^g_Z X, Y) + g(X, \nabla^g_Z Y)$$

or

$$(\nabla_Z g)(X,Y) + g(\nabla_Z X, Y) + g(X, \nabla_Z Y) = g(\nabla^g_Z X, Y) + g(X, \nabla^g_Z Y). \tag{30}$$

Now, from (29) we have

$$\begin{aligned} g(\nabla^g_Z X, Y) + g(X, \nabla^g_Z Y) = g(\nabla_Z X, Y) + g(X, \nabla_Z Y) \\ + \frac{1}{2}[(\nabla_Z g)(X,Y) + (\nabla_X g)(Z,Y) \\ - (\nabla_Y g)(Z,X) + (\nabla_Z g)(Y,X) \\ + (\nabla_Y g)(Z,X) - (\nabla_X g)(Z,Y)]. \end{aligned}$$

But this is just the left-hand side of (30). □

We may now compare ∇^g, given by (29), and ∇^A, defined by (16). In particular, we see that

$$g(Z, \nabla^A_X Y) = g(Z, \nabla_X Y) + \frac{1}{2} g(Z, A^{-1}((\nabla_X A)(Y) + (\nabla_Y A)(X))).$$

Comparing this expression with that of ∇^g we see that $\nabla^g = \nabla^A$ if and only if

$$g(Z, A^{-1}((\nabla_X A)(Y) + (\nabla_Y A)(X))) = (\nabla_Y g)(X, Z) + (\nabla_X g)(Y, Z) - (\nabla_Z g)(X, Y).$$

Since this expression is symmetric in X and Y, we may simplify this condition, as in the following lemma.

LEMMA 4.2. *$\nabla^g = \nabla^A$ if and only if*

$$g(Z, A^{-1}(\nabla_X A)(X)) = (\nabla_X g)(X, Z) - \frac{1}{2}(\nabla_Z g)(X, X), \quad X, Z \in \Gamma(TM). \tag{31}$$

Ascertaining solutions A and g of (31) is a hard problem in general, but there is an obvious candidate for g, namely,

$$g(X, Y) = h(AX, AY), \quad X, Y \in \Gamma(TM), \tag{32}$$

where h is a metric on V with $\nabla^V h \equiv 0$. We may substitute this special case into the condition (31) to obtain a condition on the map A alone. Since $(\nabla_Z g)(X, Y) = Z(g(X, Y)) - g(\nabla_Z X, Y) - g(X, \nabla_Z Y)$ for the special case of (32) we have

$$\begin{aligned}(\nabla_Z g)(X, Y) &= (\nabla^V_Z h)(AX, AY) + h(\nabla^V_Z (AX), AY) + h(AX, \nabla^V_Z (AY)) \\ &\quad - h(A\nabla_Z X, Y) - h(AX, A\nabla_z Y) \\ &= h((\nabla_Z A)(X), AY) + h(AX, (\nabla_Z A)(Y)).\end{aligned}$$

Thus from (29) and (16) we have

$$\begin{aligned}h(AZ, A\nabla^g_X Y) = h(AZ, A\nabla^A_X Y) &+ \frac{1}{2} h(AX, (\nabla_Y A)(Z) - (\nabla_Z A)(Y)) \\ &+ \frac{1}{2} h(AY, \nabla_X A)(Z) - (\nabla_Z A)(X)).\end{aligned}$$

We may summarize these observations in the following result.

THEOREM 4.3. *For the metric $g(X, Y) = h(AX, AY)$, $\nabla^V h \equiv 0$, we have*

$$\nabla^g = \nabla^{(A,S)},$$

where S is the symmetric tensor defined by

$$\begin{aligned}h(AZ, AS(X, Y)) = \frac{1}{2} h(AX, (\nabla_Y A)(Z) - (\nabla_Z A)(Y)) \\ + \frac{1}{2} h(AY, (\nabla_X A)(Z) - (\nabla_Z A)(X)).\end{aligned} \tag{33}$$

We may apply this result in the situation of Theorem 2.5 to obtain the following corollary.

COROLLARY 4.4. *The Newton law system on M, with g given by* (32),

$$\frac{D^{g^2}q}{\partial t^2} = \frac{D^{(A,S)^2}q}{\partial t^2} = \pi_N(F), \quad q \in M$$

has the restriction property if and only if

(34) (i) $A \circ \pi_N$ is Killing on N with respect to ∇;
(ii) $S|_N \equiv 0$, S defined in (33).

We now examine condition (34) (ii) in the case where Assumptions 3.1 and 3.2 hold. We set

(35) $$\hat{\omega}_k(W_m) = \delta_{k,m}, \qquad \hat{\omega}_k(V_m) \equiv 0,$$
$$\hat{\nu}_k(V_m) = \delta_{k,m}, \qquad \hat{\nu}_k(W_m) \equiv 0,$$

so that $\{\hat{\omega}_1, \dots, \hat{\omega}_m, \hat{\nu}_1, \dots, \hat{\nu}_{n-m}\}$ is a dual frame for $\{W_1, \dots, W_m, V_1, \dots, V_{n-m}\}$ and choose the metric h to be

(36) $$h = \frac{1}{2}\sum_{k=1}^{m} \hat{\omega}_k \otimes \hat{\omega}_k + \frac{1}{2}\sum_{k=1}^{n-m} \hat{\nu}_k \otimes \hat{\nu}_k.$$

We must now check that indeed $\nabla^V h \equiv 0$, where $V = TM$ and ∇^V is defined in Assumption 3.2. Since

$$\nabla^V_Z X = \sum_{k=1}^{m} Z(\hat{\omega}_k(X))W_k + \sum_{k=1}^{n-m} Z(\hat{\nu}_k(X))V_k,$$

it is clear that

$$(\nabla^V_Z h)(X,Y) = Z(h(X,Y)) - h(\nabla^V_Z X, Y) - h(X, \nabla^V_Z Y) \equiv 0.$$

Now we may simplify the definition of S in (33), using the expression (36) for h and (27) for A. We see that

$$(\nabla_X A)(Y) = \sum_m \nabla^V_X V_m \nu_m(Y) + \sum_m \nabla^V_X W_m \omega_m(Y)$$
$$+ \sum_m V_m(\nabla_X \nu_m)(Y) + \sum_m W_m(\nabla_X \omega_m)(Y).$$

Now we have the identity

$$d\nu(X,Y) = (\nabla_X \nu)(Y) - (\nabla_Y \nu)(X), \quad X, Y \in \Gamma(TM), \quad \nu \in \Gamma(T^*M).$$

Thus, from the definition of ∇^V, we see that

$$(\nabla_X A)(Y) - (\nabla_Y A)(X) = \sum_m V_m d\nu_m(X,Y) + \sum_m W_m d\omega_m(X,Y),$$

and hence we have from (33)

$$\sum_m \nu_m(Z)\nu_m(S(X,Y)) + \sum_m \omega_m(Z)\omega_m(S(X,Y))$$
$$= \frac{1}{2}\sum_m (\nu_m(X)\, d\nu_m(Y,Z) + \omega_m(X)\, d\omega_m(Y,Z))$$
$$+ \frac{1}{2}\sum_m (\nu_m(Y)\, d\nu_m(X,Z) + \omega_m(Y)\, d\omega_m(X,Z)).$$

We deduce the following result.

LEMMA 4.5. *If Assumptions* 3.1 *and* 3.2 *hold, and if* S *is the tensor defined in* (33)*, then* $S|_N \equiv 0$ *if and only if*

$$d\nu_k(X,\cdot) \equiv 0, \quad 1 \le k \le n-m, \quad X \in \Gamma(N). \tag{37}$$

It turns out that in all of the examples in section 6, this condition is never satisfied. This negative result does not, of course, exclude the possibility that we can find solutions to the equations

$$\nabla^g = \nabla^{(A,S)}$$

for some metric g on M, and pairs (A,S) satisfying condition (34), which yields a Newton law system with the restriction property, as in Corollary 4.4.

5. Integrability of nonholonomic systems. In this section we consider the implications for the preceding analysis for the integrability of nonholonomic systems. We recall that if we are given a differential equation on $\mathbb{R}^N$,

$$\dot{x} = f(x,t), \quad x \in \mathbb{R}^N,$$

then the system is said to be integrable, if through coordinate transforms, it may be reduced to a system of algebraic relations and quadratures. Of course, this property may also be a local one, defined only in a particular open domain of $\mathbb{R}^N$. In the case of differential equations defined on a Riemannian manifold, the situation is complicated by the geometry in the case of a global property, but locally the properties are identical. Our analysis and assumptions in this paper are particularly concerned with the global integrability properties of a nonholonomic system defined on a Riemannian manifold, even though our examples are such that the manifold is parallelizable, if not Euclidean.

First we consider the situation studied in section 2, where we are given V, ∇^V, the map A, and corresponding Newton law system on M,

$$\frac{D^{A^2}q}{\partial t^2} = \bar{F}, \quad q \in M. \tag{38}$$

Since ∇^A is the unique connection so that A is Killing with respect to ∇^A, we have that if $v = A_q\dot{q}$ we may rewrite this system in the form

$$\frac{D^V v}{\partial t} = A_q(\bar{F}), \quad \dot{q} = A_q^{-1}(v), \quad (q,v) \in V. \tag{39}$$

In general, this global transformation of coordinates does nothing to simplify the integration of the system of equations. This depends upon the choice of vector bundle V and connection ∇^V. In particular, if V is parallelizable, with frame $\{Z_1,\dots,Z_n\}$, and dual frame $\{z_1,\dots,z_n\}$ with $\nabla^V Z_k \equiv 0$, $1 \le k \le n$, then independent of the original connection ∇ on M we are able to simplify system (39) as follows. Setting $v_i = z_i(v)$, $1 \le i \le n$, we have $v = \sum_{i=1}^n v_i z_i(q)$, so now equations (39) become

$$\begin{aligned} \dot{v}_i &= z_i(\bar{F}), \quad 1 \le i \le n, \\ \dot{q} &= \sum_{i=1}^n A_q^{-1}(Z_i)v_i, \quad q \in M. \end{aligned} \tag{40}$$

Thus the "dynamics" of (38), i.e., the first equation in (39), has been reduced to quadratures, and one is essentially left with the "kinematics," i.e., the second equation in (39). In the case of classical dynamical systems, (with $\bar{F} \equiv 0$), the v_i are indeed constants and we would say that they are n integrals of the motion, through which we can reduce the dynamics to a system evolving on a phase space of dimension n only and not $2n$.

In the case of a nonholonomic system (1), Theorem (2.5) provides a means of reviewing it as a restriction to N of the Newton law system

$$\frac{D^{(A,S)^2}q}{\partial t^2} = \pi_N(F), \quad q \in M,$$

assuming, of course, that $S|_N \equiv 0$ and $A \circ \pi_N$ is Killing with respect to ∇. We may apply the same analysis as above to the restriction to N and transform the nonholonomic system to the form (39). If we now also insist that Assumptions 3.1 and 3.2 hold, then we may set

$$\{Z_1, \dots, Z_n\} = \{W_1, \dots, W_m, V_1, \dots, V_{n-m}\},$$
$$\{z_1, \dots, z_n\} = \{\hat{\omega}_1, \dots, \hat{\omega}_m, \hat{\nu}_1, \dots, \hat{\nu}_{n-m}\},$$

$$v = \sum_{i=1}^{n-m} v_i V_i(q) + \sum_{i=1}^{m} w_i W_i(q),$$
$$A = \sum_{m} V_m \nu_m + \sum_{m} W_m \omega_m,$$
$$A^{-1} = \sum_{j,k} V_m b_{jk}^{-1} \hat{\nu}_k + \sum_{j,k} W_m a_{jk}^{-1} \hat{\omega}_k.$$

The nonholonomic system (1) is then reduced to system (40), which in this case takes the form

$$\dot{v}_i = \nu_i(F), \quad 1 \le i \le n-m, \quad \dot{w}_i = 0, \quad 1 \le i \le m,$$
$$\dot{q} = \sum_{k,j=1}^{n-m} V_j b_{jk}^{-1} v_k.$$

Indeed, under Assumptions 3.1 and 3.2 we may now make precise the relationship between nonholonomic control systems formulated as kinematic or dynamic systems. If we set $u_i = v_i(F)$, $1 \le i \le n-m$, then the nonholonomic control system modelled with dynamics is

$$\dot{v}_i = u_i, \quad \dot{w}_i = 0, \quad \dot{q} = \sum_{j,k=1}^{m} V_j b_{jk}^{-1} v_k,$$

while the control system modeled on kinematics alone is simply

$$\dot{q} = \sum_{j,k=1}^{m} V_j b_{jk}^{-1} u_k.$$

The reader should compare this discussion with the less succinct discussion in Bloch and Crouch [2]. Note that the process of reducing the full nonholonomic dynamics to the kinematics is a different reduction to the procedures detailed in Bloch and Crouch [1], and Bloch, Krishnaprasad, Marsden, and Murray [9].

In the special case illustrated above, where we are able to reduce the system to quadratures and the kinematics, the issue of complete integrability remains. In general we cannot expect a nonholonomic system to satisfy Assumption 3.1, even though the four examples we give in section 6 do satisfy this assumption. For example, the chain of trailers, analyzed in Crouch and Jakubczyk [11], does not satisfy Assumption 3.1. In some sense Assumption 3.1 provides $n-m$ integrals, while the nonholonomic constraints provide another m integrals, and these integrals are all linear in the velocities. Indeed, Lemmas 3.2 and 3.3 demonstrate this fact for the classical Newton law system and unforced nonholonomic motion. In general, however, we cannot expect linearity in the velocities, even if we are able to find integrals. For material on the integrability of nonholonomic systems, see Zenkov [12], Hermans [13], and Arnold [10]. Arnold does demonstrate that the rolling ball is indeed integrable, although he makes use of an abstract result, applicable to all systems admitting an invariant measure. We make some further comments about this example, which is also one of the examples we treat in the next section.

6. Examples. In these examples we refer the reader to the cited references for details of the notation employed.

Example 6.1 (rolling penny, Bloch, and Crouch [2]).

$$\begin{pmatrix} \ddot{x} \\ \ddot{y} \\ \ddot{\theta} \\ \ddot{\phi} \end{pmatrix} = \lambda_1 \begin{pmatrix} 1 \\ 0 \\ -\cos\phi \\ 0 \end{pmatrix} + \lambda_2 \begin{pmatrix} 0 \\ 1 \\ -\sin\phi \\ 0 \end{pmatrix} + u_1 \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} + u_2 \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \tag{41}$$

with the nonholonomic constraints: $\dot{x} = \cos\phi\dot{\theta}$, $\dot{y} = \sin\phi\dot{\theta}$, and Euclidean metric structure.

We make the definitions

$$\begin{aligned} \omega_1 &= dx - \cos\phi\, d\theta, \qquad \omega_2 = dy - \sin\phi\, d\theta, \\ W_1 &= \frac{\partial}{\partial x} - \cos\phi\frac{\partial}{\partial\theta}, \qquad W_2 = \frac{\partial}{\partial y} - \sin\phi\frac{\partial}{\partial\theta}, \\ \nu_2 &= d\phi, \qquad \nu_1 = d\theta + \cos\phi\, dx + \sin\phi\, dy, \\ V_2 &= \frac{\partial}{\partial\phi}, \qquad V_1 = \frac{\partial}{\partial\theta} + \cos\phi\frac{\partial}{\partial x} + \sin\phi\frac{\partial}{\partial y}. \end{aligned}$$

Although W_1, W_2, V_1, and V_2 are not mutually orthogonal, W_1 and W_2 are orthogonal to V_1 and V_2. Clearly, ν_2 is Killing. We calculate $(D\nu_1/\partial t)(\dot{q})$.

$$\begin{aligned} \frac{D\nu_1}{\partial t}(\dot{q}) &= -\sin\phi\dot{\phi}\dot{x} + \cos\phi\dot{\phi}\dot{y} = \dot{\phi}(\dot{y}\cos\phi - \dot{x}\sin\phi) \\ &= \cos\phi\nu_2(\dot{q})\omega_2(\dot{q}) - \sin\phi\nu_2(\dot{q})\omega_1(\dot{q}). \end{aligned}$$

Thus ν_1 is not Killing, but it is Killing when restricted to N. Thus, this example does satisfy Assumption 3.1, and we may impose Assumption 3.2, in which case we may view system (41) as the restriction of a Newton law system defined by a connection $\nabla^{(A,S)}$ described in Corollary 3.6.

We note that if $\eta_1 = \sin\phi\, d\phi$, $\eta_2 = \cos\phi\, d\phi$, then

$$(\nabla_X\nu_1)(Y) = \eta_2(X)\omega_2(Y) - \eta_1(X)\omega_1(Y)$$

and

$$d\nu_1(X,Y) = (\nabla_X \nu_1)(Y) - (\nabla_Y \nu_1)(X) = (\eta_2 \wedge \omega_2 - \eta_1 \wedge \omega_1)(X,Y).$$

In particular, even though $d\nu_1|_N \equiv 0$, condition (37) is not satisfied and we are unable to cast $\nabla^{(A,S)}$ as a metric connection.

Example 6.2 (Bates–Sniatycki example of a constrained particle [6]).

$$\begin{pmatrix} \ddot{x} \\ \ddot{y} \\ \ddot{z} \end{pmatrix} = \lambda \begin{pmatrix} -y \\ 0 \\ 1 \end{pmatrix} + \frac{u_1}{\sqrt{1+y^2}} \begin{pmatrix} 1 \\ 0 \\ y \end{pmatrix} + u_2 \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \tag{42}$$

with the nonholonomic constraint $\dot{z} = y\dot{x}$ and Euclidean metric structure.

We make the following definitions:

$$\begin{aligned} &\omega = dz - ydx, \qquad W = \frac{\partial}{\partial z} - y\frac{\partial}{\partial x}, \\ &\nu_1 = \frac{1}{\sqrt{1+y^2}}(dx + ydz), \qquad V_1 = \frac{1}{\sqrt{1+y^2}}\left(\frac{\partial}{\partial x} + y\frac{\partial}{\partial z}\right), \\ &\nu_2 = dy, \qquad V_2 = \frac{\partial}{\partial y}. \end{aligned}$$

Thus, W, V_1, and V_2 form an orthogonal set. Clearly, ν_2 is Killing. We calculate $(D\nu_1/\partial t)(\dot{q})$.

$$\begin{aligned} \frac{D\nu_1}{\partial t}(\dot{q}) &= \frac{d}{dt}\left(\frac{1}{\sqrt{1+y^2}}\right)\dot{x} + \frac{d}{dt}\left(\frac{y}{\sqrt{1+y^2}}\right)\dot{z} \\ &= \frac{\nu_2(\dot{q})\omega(\dot{q})}{(1+y^2)^{3/2}}. \end{aligned}$$

Thus ν_1 is not Killing, but it is Killing when restricted to N. Hence, the example does satisfy Assumption 3.1, and we may impose Assumption 3.2, in which case we may view system (42) as the restriction of a Newton law system defined by a connection $\nabla^{(A,S)}$ as described in Corollary 3.6.

We note that if we set $\eta = (dy/(1+y^2)^{3/2})$, then

$$(\nabla_X \nu_1)(Y) = \eta(X)\omega(Y)$$

and

$$d\nu_1(X,Y) = (\eta \wedge \omega)(X,Y).$$

In particular, even though $d\nu_1|_N \equiv 0$, condition (37) is not satisfied, and we are unable to cast $\nabla^{(A,S)}$ as a metric connection.

Example 6.3 (rolling ball, Bloch, and Crouch [2]).

$$\begin{aligned} J\dot{\nu} &= S(\nu)J\nu + \lambda_1 Pe_1 + \lambda_2 Pe_2, \qquad \nu \in \mathbb{R}^3, \\ \dot{P} &= S(\nu)P, \qquad P \in SO(3), \\ m\ddot{x} &= \lambda_2 + u_1, \qquad x, y \in \mathbb{R}, \\ m\ddot{y} &= -\lambda_1 + u_2, \end{aligned} \tag{43}$$

with the nonholonomic constraints

$$e_2^T P^T \nu + \dot{x} = 0, \qquad e_1^T P^T \nu - \dot{y} = 0,$$

and metric structure

$$\langle (\nu_A, \dot{x}_A, \dot{y}_A), (\nu_B, \dot{x}_B, \dot{y}_B) \rangle = \frac{1}{2} \nu_A^T J \nu_B + \frac{1}{2} m (\dot{x}_A \dot{x}_B + \dot{y}_A \dot{y}_B).$$

Here ν is the angular velocity, measured in axes fixed in the rotating ball; P is the angular position, represented by a special orthogonal matrix; x and y are the coordinates of the center, and center of mass of the ball relative to inertial axes. J is the inertia tensor of the ball and $S(a)$ is the skew-symmetric matrix representation of the three vector a, which is uniquely defined by the identity

$$S(a)b = b \times a, \quad a, b \in \mathbb{R}^3,$$

where "$\times$" is the vector cross product in $\mathbb{R}^3$.

This system is 10-dimensional, evolving on $T(\mathbb{R}^2 \times SO(3))$. For simplicity we denote by $a^T \partial/\partial\nu$ ($a^T\, d\nu$) the right invariant vector field (one form) on $SO(3)$ with generator a. We may obtain

$$\begin{aligned}
W_1 &= (J^{-1} P e_1)^T \frac{\partial}{\partial \nu} - \frac{1}{m} \frac{\partial}{\partial y}, & \omega_1 &= (P e_1)^T\, d\nu - dy, \\
W_2 &= (J^{-1} P e_2)^T \frac{\partial}{\partial \nu} + \frac{1}{m} \frac{\partial}{\partial x}, & \omega_2 &= (P e_2)^T\, d\nu + dx, \\
V_1 &= (P e_1)^T \frac{\partial}{\partial \nu} + \frac{\partial}{\partial y}, & \nu_1 &= (JP e_1)^T\, d\nu + m\, dy, \\
V_2 &= (P e_2)^T \frac{\partial}{\partial \nu} - \frac{\partial}{\partial x}, & \nu_2 &= (JP e_2)^T\, d\nu - m\, dx, \\
V_3 &= (P e_3)^T \frac{\partial}{\partial \nu}, & \nu_3 &= (JP e_3)^T\, d\nu.
\end{aligned}$$

V_3 corresponds to the fact that along the motion $e_3^T P^T J \nu$ is a constant. In general, we could insert another torque, u, exerted about Pe_3, by adding a term uPe_3 to the first equation. Although W_1, W_2, V_1, V_2, V_3 is not an orthogonal set relative to the metric, $\{W_1, W_2\}$ is orthogonal to $\{V_1, V_2, V_3\}$, from which it is easy to see that they form a spanning set for the tangent spaces $T_q(\mathbb{R}^2 \times SO(3))$.

It was demonstrated in Bloch and Crouch [1], that $(Pe_k)^T(\partial/\partial\nu)$ are Killing vector fields, relative to the metric structure so V_1, V_2, V_3 are indeed Killing, even without restricting to N. It is also interesting to integrate the system equations using the easily verified identities

$$\begin{aligned}
&\frac{d}{dt} \nu_1(\dot{q}) = u_2, \qquad \frac{d}{dt} \nu_2(\dot{q}) = -u_1, \\
&\frac{d}{dt} \nu_3(\dot{q}) = \omega_2(\dot{q}) = \omega_1(\dot{q}) \equiv 0.
\end{aligned}$$

Setting $\mathbb{J} = J + m(Pe_1 e_1^T P^T + Pe_2 e_2^T P^T)$ and $e_3^T P^T J \nu = d (= \text{const})$, we obtain equations (40) in the form

$$\begin{aligned}
\dot{P} &= S(\mathbb{J}^{-1} P(e_1 a_2 + e_2(-a_1) + e_3 d)) P, \\
\dot{x} &= -e_2^T P^T \mathbb{J}^{-1} P(e_1 a_2 + e_2(-a_1) + e_3 d), \\
\dot{y} &= e_1^T P^T \mathbb{J}^{-1} P(e_1 a_2 + e_1(-a_1) + e_3 d), \\
\dot{a}_1 &= u_1, \qquad \dot{a}_2 = u_2.
\end{aligned} \tag{44}$$

It is useful to rewrite equations (43) in terms of $M = \mathbb{J}\nu$. Note that

$$Pe_1e_1^TP + Pe_2e_2^TP = I - Pe_3e_3^TP = -S(Pe_3)S(Pe_3),$$

so $M = \mathbb{J}\nu = J\nu + mS(Pe_3)S(\nu)Pe_3$. From (44) we see that

$$\nu = \mathbb{J}^{-1}P(e_1a_2 + e_2(-a_1) + e_3d),$$

so

$$\dot{M} = \frac{d}{dt}\mathbb{J}\nu = S(\nu)M + P(e_1u_2 - e_2u_1).$$

Thus these equations, together with the kinematics (44) are equivalent to the original system equations (43). It turns out that the rolling ball system (with $u_1 \equiv u_2 \equiv 0$) is completely integrable, as demonstrated by Arnold [10]. This is not obvious from the reduced order kinematic equations (44), and indeed Arnold demonstrates integrability in another way—by applying an abstract result to the subsystem in M and P:

$$\dot{M} = S(\nu)M, \quad \dot{P} = S(\nu)P, \quad \nu = \mathbb{J}^{-1}M. \tag{45}$$

As is clear from (44), once these equations are integrated, the remaining states may be obtained by quadratures. The main ingredient in the proof of Arnold's result is noting that system (45) admits an invariant measure,

$$(m - (Pe_3)^T(J + mI)^{-1}Pe_3)^{1/2}$$

and four integrals, M^TM, M^TPe_3, $(Pe_3)^TPe_3$, and $M^T\nu$. It is clear that the first three expressions are integrals from the structure of equations (45). To establish that $M^T\nu$ is an integral we proceed as follows. Since, for the rolling ball system (43), the total kinetic energy of the system is conserved for $u_1 \equiv u_2 \equiv 0$ (see [2]), we have that

$$r = \nu^TJ\nu + m\dot{x}^2 + m\dot{y}^2$$

is an integral of the motion (43). Using the nonholonomic constraints we see that

$$r = \nu^TJ\nu + m[(\nu^TPe_2)^2 + (\nu^TPe_1)^2].$$

But $(\nu^TPe_2)^2 + (\nu^TPe_1)^2 = ||Pe_3 \times \nu||^2 = ||S(\nu)Pe_3||^2$. Thus, $M^T\nu = \nu^TJ\nu + m||S(\nu)Pe_3||^2$ is indeed an integral for motion (43) and reduced motion (45).

The "angular" momentum M is, in fact, the nonholonomic momentum map discussed in Bloch, Krishnaprasad, Marsden, and Murray [9]. As a final remark, it is very interesting to contrast the work described here with that of Cartan [7]. Indeed, many of the constructions in Cartan [7] for the rolling ball example use exactly the frame discussed in our Example 6.3. Clearly there is much further structure underlying many of the observations made in this paper.

Example 6.4 (generalized rolling ball, Bloch, and Crouch [1], [2]). We assume that G is a compact, semisimple Lie group of dimension L, with Lie algebra $\mathfrak{d}$. We assume that $\langle\cdot,\cdot\rangle_G$ is a right invariant metric on G, with

$$\langle \dot{g}_A, \dot{g}_B\rangle_G = \mathcal{K}(\dot{g}_A, J_g\dot{g}_B),$$

where $\mathcal{K}$ is a bi-invariant metric on G and $J_g : T_gG \to T_gG$ is a positive definite operator for each $g \in G$. The generalized rolling ball system described in Bloch and Crouch [1], [2], is defined by the system of equations

$$\begin{aligned} \frac{D^2 g}{\partial t^2} &= \sum_{k=1}^{N} \lambda_k J_g^{-1} X_k^l, \quad g \in G, \quad L > N, \\ m\ddot{x}_k &= -\lambda_k + u_k, \quad 1 \le k \le N, \\ \dot{x}_k &= \langle \dot{g}, J_g^{-1} X_k^l \rangle_G, \quad 1 \le k \le N \quad \text{(nonholonomic constraints)}, \end{aligned} \tag{46}$$

where $D/\partial t$ is the covariant derivative corresponding to the metric connection determined by $\langle \cdot, \cdot \rangle$, and X^l is the left invariant vector field corresponding to an element $X \in \mathfrak{d}$. We assume that $X_1, \ldots, X_N, X_{N+1}, \ldots, X_L$ is a basis for $\mathfrak{d}$. There is a natural metric on the configuration space $M = G \times \mathbb{R}^N$ determined by

$$\langle (\dot{g}_A, \dot{x}_A), (\dot{g}_B, \dot{X}_B) \rangle_M = \frac{1}{2} \langle \dot{g}_A, \dot{g}_B \rangle_G + \frac{m}{2} \dot{x}_A^T \dot{x}_B. \tag{47}$$

It was proved in Bloch and Crouch [1] that X^l is a Killing vector field for $X \in \mathfrak{d}$. We define one forms on $G \times \mathbb{R}^N$ by

$$\begin{aligned} \omega_k(\dot{g}, \dot{x}) &= \langle \dot{g}, J_g^{-1} X_k^l \rangle_G - \dot{x}_k, \quad 1 \le k \le N, \\ \nu_k(\dot{g}, \dot{x}) &= \langle \dot{g}, X_k^l \rangle_G + \dot{x}_k m, \quad 1 \le k \le N, \\ \nu_k(\dot{g}, \dot{x}) &= \langle \dot{g}, X_k^l \rangle_G, \quad N+1 \le k \le L. \end{aligned}$$

The corresponding vector fields are

$$\begin{aligned} W_k &= J_g^{-1} X_k^l - \frac{1}{m} \partial/\partial x_k, \quad 1 \le k \le N, \\ V_k &= X_k^l + \partial/\partial x_k, \quad 1 \le k \le N, \\ V_k &= X_k^l, \quad N+1 \le k \le L. \end{aligned} \tag{48}$$

Clearly, V_k and ν_k are Killing with respect to the metric connection determined by metric (47). We also compute, for $0 \le k, j \le N$,

$$\begin{aligned} \langle W_k, V_j \rangle_M &= \frac{1}{2} \langle J_g^{-1} X_k^l, X_j^l \rangle_G + \frac{m}{2} \left(- \frac{1}{m} \right) \left\langle \frac{\partial}{\partial x_k}, \frac{\partial}{\partial x_j} \right\rangle_{\mathbb{R}^n} \\ &= \frac{1}{2} \mathcal{K}(X_k^l, X_j^l) - \frac{1}{2} \delta_{kj}. \end{aligned}$$

Thus if we assume that $X_1, \ldots, X_L$ is an orthonormal basis of $\mathfrak{d}$ with respect to $\mathcal{K}$, we see that $\{W_1, \ldots, W_N\}$ is orthogonal to $\{V_1, \ldots, V_L\}$. It follows that the system (46) and choice of vector fields (48) satisfy our Assumption (3.1). Hence, system (46) may be viewed as the restriction of a Newton law system defined by a connection described in Corollary 3.6. Note that in this example, just as in Example 6.3, we do not have to restrict to the subbundle nonholonomic constraints.

Setting

$$\begin{aligned} v_k &= \nu_k(\dot{g}, \dot{x}), \quad 1 \le k \le N, \\ d_k &= \nu_k(\dot{g}, \dot{x}), \quad N+1 \le k \le L, \end{aligned}$$

we can show (using analysis in [1], [2]) that $\dot{d}_k = 0$, $\dot{v}_k = u_k$. By the independence of the vector fields W_k, V_k, we may solve the $L + N$ equations

$$0 = \omega_k(\dot{g}, \dot{x}), 1 \leq k \leq N, \quad v_k = \nu_k(\dot{g}, \dot{x}), 1 \leq k \leq N, \quad d_k = \nu_k(\dot{g}, \dot{x}), \quad N{+}1 \leq k \leq L,$$

for $(\dot{g}, \dot{x})$ in terms of v_k and d_k. Hence, as in section (5), we may integrate the $2(L + N)$-dimensional system (46), to obtain the corresponding $(L + N)$-dimensional kinematic system.

REFERENCES

[1] A. M. Bloch and P. E. Crouch, *Nonholonomic control systems on Riemannian manifolds*, SIAM J. Control Optim., 33 (1995), pp. 126–148.

[2] A. M. Bloch and P. E. Crouch, *Nonholonomic and Vakonomic control systems on Riemannian Manifolds*, Fields Institute Communications, Vol. 1, M. J. Enos, ed., 1993, pp. 25–52.

[3] A. M. Bloch and P. E. Crouch, *Controllability of nonholonomic systems on Riemannian Manifolds*, in Proc. IEEE Conf. on Decision and Control, Tucson, AZ, 1992, pp. 1594–1596.

[4] A. M. Vershik and V. Ya. Gershkovich, *Nonholonomic problems and the theory of distributions*, Acta Appl. Math., 12 (1988), pp. 181–209.

[5] A. M. Vershik and L. D. Fadeev, *Lagrange mechanics in an invariant setting*, Selecta Math. Soviet., 1 (1981), pp. 339–350.

[6] L. Bates and J. Sniatycki, *Nonholonomic reduction*, Rep. Math Phys., 32 (1993), pp. 99–115.

[7] E. Cartan, *Sur La Représentation Géométriques Des Systèms Matériels Nonholonomes*, Collected Works, Oevres Complètes, Gauthier–Villars, Paris, 1952.

[8] A. M. Bloch and P. E. Crouch, *Another view of nonholonomic mechanical control systems*, in Proc. 1995 IEEE Conf. on Decision and Control, New Orleans, LA, 1995, pp. 1066–1071.

[9] A. M. Bloch, P. S. Krishnaprasad, J. E. Marsden, and R. M. Murray, *Nonholonomic mechanical systems with symmetry*, Arch. Rationál Mech. Anal., 136 (1996), pp. 21–99.

[10] V. Arnold, *Dynamical Systems* III, Springer-Verlag, New York, 1988.

[11] P. E. Crouch and B. Jakubczyk, *Dynamic Transformations and Chains of Mechanical Systems*, in Proc. IEEE Conf. on Decision and Control, Orlando, FL, 1994.

[12] D. V. Zenkov, *The geometry of the routh problem*, J. Nonlinear Sci., 5 (1996), pp. 503–519.

[13] J. Hermans, *A symmetric sphere rolling in a surface*, Nonlinearity, 8 (1995) pp. 493–515.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 6, pp. 2040–2060, November 1998

011

SINGULAR PERTURBATION OF A FINITE HORIZON PROBLEM WITH STATE-SPACE CONSTRAINTS*

FABIO BAGAGIOLO† AND MARTINO BARDI‡

Abstract. We study the singular perturbation of optimal control problems for nonlinear systems with constraints on the fast state variables and a cost functional either of Bolza type or involving the exit time of the system from a given domain. Under a controllability assumption on the fast variables, we show that these variables become controls in the limit problem. Our method consists of passing to the limit in the associated Hamilton–Jacobi–Bellman (HJB) equations by means of some tools in the theory of viscosity solutions.

Key words. singular perturbations, optimal control, nonlinear systems, viscosity solutions, Hamilton–Jacobi equations, Bolza problem, exit time problems, state constraints

AMS subject classifications. 49L25, 49L20, 93C73, 35F25, 35F30

PII. S0363012996314476

Introduction. In this paper we study a class of singular perturbation problems for nonlinear systems of the form

$$\begin{cases} y'(t) = f(y(t), z(t), \alpha(t)), & \alpha(t) \in A, \qquad t > 0, \\ z'(t) = \dfrac{1}{\varepsilon} g(y(t), z(t), \beta(t)), & \beta(t) \in B, \qquad t > 0, \\ (y(0), z(0)) = (x, \zeta), \end{cases} \tag{0.1}$$

where $\varepsilon > 0$, A and B are compact, f and g are Lipschitzean in the state variables and continuous, with Bolza cost functional

$$J_\varepsilon(x, \zeta, t, \alpha, \beta) := \int_0^t e^{-\lambda s} l(y(s), z(s), \alpha(s)) ds + e^{-\lambda t} h(y(t), z(t)), \tag{0.2}$$

where l and h are bounded and continuous, $\lambda \geq 0$, and the fast state variables $z(\cdot)$ satisfy the constraint

$$z(t) \in \overline{\Omega} \qquad \forall t > 0, \tag{0.3}$$

where Ω is an open connected subset of $\mathbb{R}^M$ with Lipschitz boundary.

Problems of this kind have been extensively studied by many authors; see, e.g., the survey paper by Kokotović [19], the books by Kokotović, Khalil, and O'Reilly [20] and Bensoussan [9], and a recent article [23]. However, we will make here rather different assumptions and use different methods. Typically, in fact, one assumes some solvability with respect to ζ of the algebraic equation

$$g(x, \zeta, b) = 0, \tag{0.4}$$

*Received by the editors December 30, 1996; accepted for publication (in revised form) December 9, 1997; published electronically August 31, 1998. This research was partially supported by the M.U.R.S.T. project "Problemi non lineari nell'analisi e nelle applicazioni fisiche, chimiche, biologiche."

http://www.siam.org/journals/sicon/36-6/31447.html

†Dipartimento di Matematica, Università di Trento, Via Sommarive 14, I-38050 Povo-Trento, Italy (bagagiol@science.unitn.it).

‡Dipartimento di Matematica, Università di Padova, Via Belzoni 7, I-35131 Padova, Italy (bardi@math.unipd.it).

and expects that the limit optimal control problem involve the system

$$\begin{cases} y'(t) = f(y(t), z(t), \alpha(t)), & \alpha(t) \in A, \quad t > 0, \\ 0 = g(y(t), z(t), \beta(t)), & \beta(t) \in B, \quad t > 0, \\ y(0) = x, \end{cases} \tag{0.5}$$

which is obtained by setting $\varepsilon = 0$ in (0.1).

In this paper, instead of the solvability of (0.4), we will assume a controllability condition on the fast variables $z(\cdot)$ of the following type:

$$\overline{\text{co}}\, g(y, z, B) \supseteq \overline{B}_{\mathbb{R}^M}(0, 1) \quad \forall\ y \in \mathbb{R}^N,\ z \in \overline{\Omega}, \tag{0.6}$$

where $\overline{\text{co}}$ denotes the closed convex hull and the set on the right-hand side is the closed unit ball centered at 0 in $\mathbb{R}^M$ (see section 4 for more general assumptions). The optimal control problem we obtain in the limit as $\varepsilon \to 0$ is the minimization of the functional

$$J(x, t, z, \alpha) := \int_0^t e^{-\lambda s} l(y(s), z(s), \alpha(s)) ds + e^{-\lambda t} \inf_{\zeta \in \overline{\Omega}} h(y(t), \zeta) \tag{0.7}$$

for the system

$$\begin{cases} y'(t) = f(y(t), z(t), \alpha(t)), & z(t) \in \overline{\Omega}, \quad t > 0, \\ y(0) = x, \end{cases} \tag{0.8}$$

instead of (0.5). Our main result states that the value function $V(x, t)$ of this problem is the limit (uniform on compact sets), as $\varepsilon \to 0$, of the value functions $V_\varepsilon(x, \zeta, t)$ of the problems corresponding to $\varepsilon > 0$.

The heuristic explanation of this result is the following. The controllability condition (0.6) allows the system to reach any point $\xi \in \overline{\Omega}$ from any starting point ζ in a lap of time of order $\varepsilon|\zeta - \xi|$. Therefore, the value function V_ε is less and less sensitive to ζ as $\varepsilon \to 0$. Moreover, since the control β does not appear in the equations for the slow variables $y(\cdot)$, nor in the cost functional J_ε, its role reduces to driving optimally the z variable within $\overline{\Omega}$. Since this can be done arbitrarily fast as $\varepsilon \to 0$, β disappears in the limit problem, and z takes the role of a control varying in $\overline{\Omega}$.

Our method is also quite different from usual. In fact we consider the viscosity solution of the problem

$$\begin{cases} u_t + \lambda u + H\left(y, z, \nabla_y u, \dfrac{1}{\varepsilon} \nabla_z u\right) = 0, & \text{in } \mathbb{R}^N \times \overline{\Omega} \times]0, +\infty[, \\ u(x, \zeta, 0) = h(x, \zeta), & \text{in } \mathbb{R}^N \times \overline{\Omega}, \end{cases} \tag{0.9}$$

and prove that it converges to the viscosity solution of

$$\begin{cases} u_t + \lambda u + \sup_{z \in \overline{\Omega}} H(y, z, \nabla_y u, 0) = 0, & \text{in } \mathbb{R}^N \times]0, +\infty[, \\ u(x, 0) = \inf_{z \in \overline{\Omega}} h(x, z), & \text{in } \mathbb{R}^N. \end{cases} \tag{0.10}$$

Then we get the desired result because the value functions V_ε and V are, respectively, the viscosity solution of (0.9) and (0.10) with Bellman's Hamiltonian

$$\begin{aligned} H(x, \zeta, p, q) := & \sup_{a \in A} \Big\{ - f(x, \zeta, a) \cdot p - l((x, \zeta, a) \Big\} \\ & + \sup_{b \in B} \Big\{ - g(x, \zeta, b) \cdot q \Big\}, \quad \forall (x, \zeta) \in \mathbb{R}^N \times \overline{\Omega},\ \forall (p, q) \in \mathbb{R}^N \times \mathbb{R}^M. \end{aligned} \tag{0.11}$$

Note that the Hamilton–Jacobi equation in (0.9) has to be solved in a set which is not open, thus solutions must be interpreted in the sense of constrained viscosity solutions introduced by Soner [24]; see also Capuzzo Dolcetta and Lions [10].

Our convergence result for the solutions of the Cauchy problem (0.9) is more general than we need for the specific application, because it holds for any Hamiltonian satisfying a comparison principle and the conditions

$$\begin{aligned} &\lim_{|q|\to+\infty} H(y,z,p,q) = +\infty, \qquad \text{uniformly for bounded } y,z,p; \\ &\varepsilon_1 \le \varepsilon_2 \Rightarrow H\left(y,z,p,\frac{q}{\varepsilon_1}\right) \ge H\left(y,z,p,\frac{q}{\varepsilon_2}\right), \end{aligned} \tag{0.12}$$

but not necessarily convex in the variables p and q as is Bellman's Hamiltonian defined by (0.11). In [21] Lions outlined a proof of a similar result for the case $\Omega = \mathbb{R}^M$, $\lambda = 0$, and under hypotheses of uniform continuity of the solutions and of the initial data h, but with a weaker assumption on the Hamiltonian, namely, $H(y,z,p,q) \ge H(y,z,p,0)$ for all (y,z,p,q), instead of the second of (0.12). His proof is based on approximating (0.9) with the addition of a small viscosity term, obtaining uniform gradient bounds, independent of ε, for the solutions of these parabolic approximations, and using the Ascoli–Arzelà theorem to extract a convergent sequence of V_ε. The hard estimates involved in this procedure are based on earlier work of Jensen and Lions [18] on partial differential equations (PDE) methods for singular perturbation problems in the optimal control of diffusion processes. Here, instead, we use the method of weak limits in viscosity sense, or "relaxed half-limits," introduced by Barles and Perthame [6], [7], and a comparison theorem for merely semicontinuous sub- and supersolutions of the limit HJB equation, which allow us to pass to the limit in the equations with only L^∞-estimates on V_ε.

To show the flexibility of our method we also apply it to optimal control problems where the cost functional is computed until the exit time of the system from a given domain, either open or closed. In this case the value functions satisfy a Dirichlet boundary value problem for a stationary HJB equation, and in the limit procedure described previously we use the notion of boundary condition in viscosity sense and a comparison theorem due to Ishii [16] (see also [7], [5], and [4]).

A simple application of the main result is given in section 3. We consider control problems for the system (0.8) with controls $z(\cdot)$ restricted to Lipschitzean functions with Lipschitz constant less than $1/\varepsilon$ and initial point $\zeta \in \overline{\Omega}$. This corresponds to the system (0.1) with $g(x,\zeta,b) = b \in \overline{B}_{\mathbb{R}^M}(0,1)$. From the main result we obtain that, as $\varepsilon \to 0$, these value functions converge (uniformly on compact sets) to the value function of the limit problem with measurable controls $z(\cdot)$. This problem is studied in Chapter V, Section 6, of the book [9] in the special case $\Omega = \mathbb{R}^M$ by different methods. The issue of Lipschitz controls was also studied by Barron, Evans, and Jensen [8] in the more general framework of differential games (see also the references therein) in the case that the set $\overline{\Omega}$ where the controls are constrained is the unit cube and the dynamics and the running cost are 1-periodic in the corresponding variables. They use the theory of viscosity solutions and gradient estimates on approximating parabolic equations.

We believe that our method is general enough to apply to several other singular perturbation problems, in particular for differential games. A similar analysis has been done for the infinite horizon problem (which is simpler with our PDE approach) in [2], [3], [4], and recently extended in [1] to the case when the slow state variables

$y(\cdot)$ are constrained by means of a comparison theorem of Ishii and Koike [17]. We mention also that similar problems have been studied recently by Subbotina [27] within Subbotin's theory of minimax solutions [26], and that Soner studied singular perturbations for some stochastic control problems by viscosity solutions methods in [25] .

We refer to the books [20], [9] and the references therein for the motivations and applications of singular perturbation problems in optimal control. We recall that the pioneering work on the theory of viscosity solutions is due to Crandall and Lions [13], Crandall, Evans, and Lions [12], and Lions [22]. For a comprehensive account of the theory of viscosity solutions and its applications to optimal control, with special attention to discontinuous solutions and the weak limit technique employed here, we refer to the books by Barles [5], and Bardi and Capuzzo Dolcetta [4] for first-order equations, and by Fleming and Soner [15] for second-order equations.

The paper is organized as follows. In section 1 we list the precise assumptions and give some preliminary results. In section 2 we prove the main theorem, and in section 3 we apply it to the problem of Lipschitz controls. In section 4 we relax the controllability assumption (0.6) and illustrate the convergence result for singular perturbations of exit time problems. An Appendix contains the proof of the continuity of the value function for problems with state constraints.

1. Statement of the problem and basic results. Let Ω be an open connected subset of $\mathbb{R}^M$, A, B be two compact sets and $f : \mathbb{R}^N \times \mathbb{R}^M \times A \to \mathbb{R}^N$, $g : \mathbb{R}^N \times \mathbb{R}^M \times B \to \mathbb{R}^M$ be two continuous functions. For every $(x, \zeta) \in \mathbb{R}^N \times \overline{\Omega}$ and for every $\varepsilon > 0$, let us consider the controlled dynamical system (0.1), where the controls $\alpha(\cdot)$ and $\beta(\cdot)$ are measurable functions defined on $[0, +\infty[$ and taking values in A and B, respectively. The dynamics f and g are supposed to be Lipschitz functions on $(y, z) \in \mathbb{R}^N \times \mathbb{R}^M$ uniformly on $a \in A$ and $b \in B$, respectively; moreover, we suppose that f has linear growth with respect to $y \in \mathbb{R}^N$, i.e., there exists $K > 0$ such that

$$|f(y, z, a)| \le K(1 + |y|) \qquad \forall\ (y, z, a) \in \mathbb{R}^N \times \overline{\Omega} \times A.$$

Under these assumptions on the dynamics, the system (0.1) has a unique continuous solution (trajectory), which continuously depends on the initial datum (x, ζ). When no ambiguity arises about the dependence on ε, on (x, ζ), and on the choice of the controls α and β, we shall denote this trajectory by $(y(t), z(t))$.

We want the trajectories of the system (0.1) to respect the *state-space constraint* (0.3). Hence, we consider the possibly empty set of admissible controls for the system (0.1), which depends on ε and on the starting point (x, ζ):

$$(\mathcal{A} \times \mathcal{B})^{\varepsilon}_{(x,\zeta)} := \big\{(\alpha, \beta) : [0, +\infty[\to A \times B \text{ measurable} \mid z(t) \in \overline{\Omega} \quad \forall\ t \ge 0\big\}. \tag{1.1}$$

Now we consider a finite horizon optimal control problem. Let $l : \mathbb{R}^N \times \mathbb{R}^M \times A \to \mathbb{R}$ be a bounded continuous function, uniformly continuous with respect to $(y, z) \in \mathbb{R}^N \times \mathbb{R}^M$ uniformly in $a \in A$; let $h : \mathbb{R}^N \times \mathbb{R}^M \to \mathbb{R}$ be a bounded continuous function, which is uniformly continuous in the bounded sets of $\mathbb{R}^N$ uniformly in $z \in \overline{\Omega}$, i.e., for every $R > 0$ there exists a continuous increasing function $\omega_R : [0, +\infty[\to \mathbb{R}$ such that $\omega_R(0) = 0$ and

$$|h(y_1, z) - h(y_2, z)| \le \omega_R(|y_1 - y_2|) \qquad \forall z \in \overline{\Omega},\ \forall y_1, y_2 \in \mathbb{R}^N,\ |y_1|, |y_2| \le R. \tag{1.2}$$

Let us consider the *cost functional* J_ε defined in (0.2), where $(\alpha, \beta) \in (\mathcal{A} \times \mathcal{B})^{\varepsilon}_{(x,\zeta)}$, $t \in [0, +\infty[$, $(x, \zeta) \in \mathbb{R}^N \times \overline{\Omega}$ and the subscript ε means that the trajectories are

relative to system (0.1) with (x,ζ) as initial point and the admissible controls α and β. The optimal control problem is to minimize the cost J_ε over the set of admissible controls, therefore respecting the state-space constraint $z(t) \in \overline{\Omega}$. Hence, we define the *value function*:

$$V_\varepsilon(x,\zeta,t) := \inf_{(\alpha,\beta)\in(\mathcal{A}\times\mathcal{B})^\varepsilon_{(x,\zeta)}} J_\varepsilon(x,\zeta,t,\alpha,\beta). \tag{1.3}$$

If the set of admissible controls is not empty for every initial point, then, by the hypotheses on l and h, V_ε is a real-valued function.

Now we define another control problem, which will turn out to be the limit problem as ε goes to zero. Let us consider the controlled dynamical system (0.8), where $x \in \mathbb{R}^N$, $\alpha \in \mathcal{A} := \{\alpha : [0,+\infty[\to A \text{ measurable}\}$ and $z \in \mathcal{Z} := \{z : [0,+\infty[\to \overline{\Omega} \text{ measurable}\}$.

The trajectories of system (0.8) will be denoted by $y(\cdot)$. Then we define the following *final cost*:

$$\tilde{h}(y) := \inf_{z\in\overline{\Omega}} h(y,z), \tag{1.4}$$

which is bounded and continuous, and consider the cost functional J defined in (0.7). We have the following value function:

$$V(x,t) := \inf_{(z,\alpha)\in\mathcal{Z}\times\mathcal{A}} J(x,t,z,\alpha), \tag{1.5}$$

which is a real-valued function. Hence, we have a control problem without state-space constraint and with α and z as controls.

Before stating our main result, we need some other hypotheses. The first is a regularity assumption on the domain Ω, whose boundary should be piecewise of class $C^{1,1}$ and should have a uniform *interior cone property*, i.e., there exist two positive constants h, r, a bounded uniformly continuous function $\eta : \overline{\Omega} \to \mathbb{R}^M$, and functions $g_i \in C^{1,1}(\mathbb{R}^M)$, $i = 1,\dots,q$, $q \in \mathbb{N}$, such that

$$\begin{aligned} &\big\{\xi \in \mathbb{R}^M \big| |\xi - (z + t\eta(z))| \le rt\big\} \subseteq \Omega \qquad \forall z \in \overline{\Omega}, \forall t \in [0,h], \\ &\overline{\Omega} = \big\{z \in \mathbb{R}^M : g_i(z) \le 0 \quad \forall i = 1,\dots,q\big\}, \\ &\big|\nabla g_i(z)\big| > 0 \qquad \forall z \text{ such that } g_i(z) = 0. \end{aligned} \tag{1.6}$$

The next hypothesis states the controllability in the z-variables, i.e.,

$$\overline{\text{co}}\, g(y,z,B) \supseteq \big\{\xi \in \mathbb{R}^M \big| |\xi| \le 1\big\} \quad \forall\, y \in \mathbb{R}^N,\ z \in \overline{\Omega}, \tag{1.7}$$

where $\overline{\text{co}}$ means the closed convex hull of the set. Observe that (1.7) implies the existence of an "inward field" on the boundary of Ω, i.e.,

$$\forall z \in \partial\Omega\ \forall y \in \mathbb{R}^N\ \exists b \in B \text{ such that } g_i(z) = 0 \Rightarrow g(y,z,b)\cdot\nabla g_i(z) < 0. \tag{1.8}$$

THEOREM 1.1. *Under all the hypotheses stated before, the sequence of the value functions defined in* (1.3) *uniformly converges to the value function defined in* (1.5)*, over the compact sets of* $\mathbb{R}^N \times \overline{\Omega} \times]0,+\infty[$ *as* $\varepsilon \to 0$*. Moreover, if the final cost* h *does not depend on the variable* $z \in \overline{\Omega}$*, then the uniform convergence is over the compact sets of* $\mathbb{R}^N \times \overline{\Omega} \times [0,+\infty[$.

Next, we give some basic results about optimal control and viscosity solutions of Hamilton–Jacobi–Bellman equations. Let us consider the optimal control problem with state-space constraint given by system (0.1), the constraint (0.3), the cost functional (0.2), and the value function (1.3). Under all the hypotheses stated before, the following proposition holds.

PROPOSITION 1.2. *For every $\varepsilon > 0$, for every $(x,\zeta) \in \mathbb{R}^N \times \overline{\Omega}$, the set of admissible controls is not empty. Moreover, the value function V_ε is continuous and bounded over the sets $\mathbb{R}^N \times \overline{\Omega} \times [0,T]$, for every $T \in]0,+\infty[$.*

This proposition is a special case of Theorem A.1 in the Appendix.

Now we consider an open subset $\mathcal{O} \subseteq \mathbb{R}^m$, and a continuous real-valued function H defined on $\overline{\mathcal{O}} \times \mathbb{R}^m$. Consider the Cauchy problem

$$(P)\begin{cases} u_t(x,t) + \lambda u(x,t) + H(x, \nabla u(x,t)) = 0, & \text{in } \overline{\mathcal{O}} \times]0,T[, \\ u(x,0) = h(x), & \text{in } \overline{\mathcal{O}}, \end{cases} \tag{1.9}$$

where ∇ means the gradient with respect to the spatial variables and u_t the time derivative.

DEFINITION 1.3. *We say that a bounded and continuous function $u : \overline{\mathcal{O}} \times [0,T[\to \mathbb{R}$ is a constrained viscosity solution of (P) if u satisfies the initial condition and for every test function $\varphi \in C^1(\overline{\mathcal{O}} \times]0,T[)$, the following holds:*

(1.10)
if $(x_0,t_0) \in \mathcal{O} \times]0,T[$ is a local maximum point for $u - \varphi$ in $\mathcal{O} \times]0,T[$, then
$$\varphi_t(x_0,t_0) + \lambda u(x_0,t_0) + H(x_0, \nabla\varphi(x_0,t_0)) \le 0;$$

(1.11)
if $(x_0,t_0) \in \overline{\mathcal{O}} \times]0,T[$ is a local minimum point for $u - \varphi$ in $\overline{\mathcal{O}} \times]0,T[$, then
$$\varphi_t(x_0,t_0) + \lambda u(x_0,t_0) + H(x_0, \nabla\varphi(x_0,t_0)) \ge 0.$$

If u satisfies (1.10) *only, then we say that u is a viscosity subsolution in $\mathcal{O} \times]0,T[$ of* (1.9)*; if u satisfies* (1.11) *only, then we say that u is a viscosity supersolution in $\overline{\mathcal{O}} \times]0,T[$ of* (1.9).

Next, let us consider the following Cauchy problem, in $\mathbb{R}^N \times \overline{\Omega} \times [0,T[$ for every $T > 0$:

$$(P_\varepsilon)\begin{cases} (V_\varepsilon)_t(x,\zeta,t) + \lambda V_\varepsilon(x,\zeta,t) + H\left(x,\zeta,\nabla_y V_\varepsilon(x,\zeta,t), \dfrac{1}{\varepsilon}\nabla_z V_\varepsilon(x,\zeta,t)\right) = 0 \\ V_\varepsilon(x,\zeta,0) = h(x,\zeta), \end{cases} \tag{1.12}$$

where ∇_y and ∇_z, respectively, mean the gradient with respect to the variable $y \in \mathbb{R}^N$ and $z \in \overline{\Omega} \subset \mathbb{R}^M$ and the Hamiltonian H is defined in (0.11). Note that (P_ε) is a particular case of (P).

THEOREM 1.4. *The value function V_ε defined in* (1.3) *is the unique constrained viscosity solution of* (P_ε)*, among the bounded and continuous functions in $\mathbb{R}^N \times \overline{\Omega} \times [0,T]$.*

Proof. Using standard techniques, the proof of the fact that V_ε is a constrained viscosity solution is easy (see, for instance, [24] or [4]). The uniqueness follows from a comparison result, which we state next. □

THEOREM 1.5. *Let $u, v : \mathbb{R}^N \times \overline{\Omega} \times [0,T] \to \mathbb{R}$ be two bounded and continuous functions, such that u is a subsolution in $\mathbb{R}^N \times \Omega \times]0,T[$ of* (1.12)*, v is a supersolution in $\mathbb{R}^N \times \overline{\Omega} \times]0,T[$ of* (1.12)*, and $u(x,\zeta,0) \le v(x,\zeta,0)$ for every $(x,\zeta) \in \mathbb{R}^N \times \overline{\Omega}$. Then $u \le v$ in $\mathbb{R}^N \times \overline{\Omega} \times [0,T]$.*

Proof. This result is obtained using standard techniques. In particular, we use test functions with penalization terms as in [24] (see also [10], [4]). The detailed proof can be found in [2]. □

Finally, let us consider the following Cauchy problem in $\mathbb{R}^N \times [0,T[$ without state-space constraint:

$$(\mathrm{P}_0)\begin{cases} V_t(x,t) + \lambda V(x,t) + \sup_{z\in\overline{\Omega}} H(x,z,\nabla V(x,t),0) = 0, \\ V(x,0) = \tilde{h}(x), \end{cases} \tag{1.13}$$

where ∇ means the gradient with respect to the variable $y \in \mathbb{R}^N$ and the initial datum $\tilde{h}$ is defined in (1.4). Note that (P_0) is a particular case of (P).

THEOREM 1.6. *The value function V defined in* (1.5) *is continuous and bounded in $\mathbb{R}^N \times [0,T]$ and it is the only viscosity solution of* (P_0) *in $\mathbb{R}^N \times]0,T[$ among the continuous and bounded functions.*

For the proof, see, for instance, [4]. In particular, the uniqueness follows from a comparison result.

THEOREM 1.7. *Let $u, v : \mathbb{R}^N \times [0,T] \to \mathbb{R}$ be bounded. If u is an upper semicontinuous subsolution and v is a lower semicontinuous supersolution in $\mathbb{R}^N \times]0,T[$ of* (1.13) *and $u(x,0) \le v(x,0)$, for every $x \in \mathbb{R}^N$, then $u \le v$ in $\mathbb{R}^N \times [0,T]$.*

2. The singular perturbation problem. This section is devoted to the proof of Theorem 1.1. Hence, we put ourselves in the framework of section 1. Since we need some technical results, we shall break the proof into several lemmas.

The value functions V_ε are equibounded in $\mathbb{R}^N \times \overline{\Omega} \times [0,T]$ for every $T > 0$. Hence, for every $(x,\zeta,t) \in \mathbb{R}^N \times \overline{\Omega} \times [0,+\infty[$, we can define the following real-valued *weak limits*:

$$\begin{aligned} \underline{V}(x,\zeta,t) &:= \liminf_{(y,z,\tau,\varepsilon)\to(x,\zeta,t,0)} V_\varepsilon(y,z,\tau), \\ \overline{V}(x,\zeta,t) &:= \limsup_{(y,z,\tau,\varepsilon)\to(x,\zeta,t,0)} V_\varepsilon(y,z,\tau), \end{aligned} \tag{2.1}$$

where the limits are taken in $(y,z,\tau,\varepsilon) \in \mathbb{R}^N \times \overline{\Omega} \times [0,+\infty[\times]0,+\infty[$. Note that $\underline{V}$ is lower semicontinuous and $\overline{V}$ is upper semicontinuous.

LEMMA 2.1. *For every $(x,\zeta,t) \in \mathbb{R}^N \times \overline{\Omega} \times [0,+\infty[$, the following inequality holds:*

$$V(x,t) \le \underline{V}(x,\zeta,t).$$

Proof. Let us define the following function:

$$\tilde{V}(x,\zeta,t) := V(x,t) \qquad \forall (x,\zeta,t) \in \mathbb{R}^N \times \overline{\Omega} \times [0,+\infty[. \tag{2.2}$$

It is easy to check that $\tilde{V}$ is a subsolution of (1.12) for all $\varepsilon > 0$.

On the other hand, we have, for every $\varepsilon > 0$,

$$\tilde{V}(x,\zeta,0) = V(x,0) = \tilde{h}(x) \le h(x,\zeta) = V_\varepsilon(x,\zeta,0).$$

Hence, by Theorem 1.5, we get for every $\varepsilon > 0$ and for every $(x,\zeta,t) \in \mathbb{R}^N \times \overline{\Omega} \times [0,+\infty[$: $V(x,t) = \tilde{V}(x,\zeta,t) \le V_\varepsilon(x,\zeta,t)$. By (2.1) and the continuity of V, we obtain

$$V(x,t) = \liminf_{(y,\tau)\to(x,t)} V(y,\tau) \le \liminf_{(y,z,\tau,\varepsilon)\to(x,\zeta,t,0)} V_\varepsilon(y,z,\tau) = \underline{V}(x,\zeta,t). \qquad \square$$

LEMMA 2.2. *For every* $(x,\zeta)\in\mathbb{R}^N\times\overline{\Omega}$,

$$\overline{V}(x,\zeta,0)=h(x,\zeta). \tag{2.3}$$

Proof. Let us note that the Hamiltonian in (0.9) can be written as

$$H\left(x,\zeta,p,\frac{1}{\varepsilon}q\right)=\sup_{a\in A}\Big\{-f(x,\zeta,a)\cdot p-l(x,\zeta,a)\Big\}-\frac{1}{\varepsilon}\inf_{b\in B}\Big\{g(x,\zeta,b)\cdot q\Big\}.$$

By (1.7), the last term of the right-hand side is nonnegative. Hence, if $\varepsilon_1\le\varepsilon_2$, then V_{ε_2} is a supersolution of (1.12) with $\varepsilon=\varepsilon_1$. So, by Theorem 1.5

$$0<\varepsilon_1\le\varepsilon_2\Rightarrow V_{\varepsilon_1}\le V_{\varepsilon_2}\qquad\text{in }\mathbb{R}^N\times\overline{\Omega}\times[0,+\infty[. \tag{2.4}$$

Using (2.4), the definition of $\overline{V}$ (2.1) and the fact that the V_ε are continuous and $V_\varepsilon(x,\zeta,0)=h(x,\zeta)$, it is not hard to get (2.3). ∎

LEMMA 2.3. *Let* $\mathcal{O}\subseteq\mathbb{R}^m$ *be open and connected and* $u:\mathcal{O}\to\mathbb{R}$ *be a bounded upper semicontinuous function solving in the viscosity sense*

$$|\nabla u|\le 0\quad \textit{in } \mathcal{O}.$$

Then u *is constant.*

Proof. We shall prove that u is locally Lipschitz. Hence, u solves the equation almost everywhere and we get the conclusion.

Let M be a lower bound for u in $\mathcal{O}$. Let us take $x_0\in\mathcal{O}$ and consider $r_0>0$ such that the open ball B with radius r_0 and centre x_0 is contained in $\mathcal{O}$. Denoting by B' the open ball with radius $r_0/2$ and centre x_0, let us take $x'\in B'$ and consider the function

$$\phi(y)=u(y)-C|x'-y|^2\quad\forall y\in\mathcal{O},$$

where the constant $C>0$ will be fixed later. The function ϕ is upper semicontinuous and, hence, it reaches its maximum in the closure of B. Let y' be a point of maximum. We claim that y' belongs to B'. In fact, $\phi(x')=u(x')\ge M$ and, for a suitable choice of the constant C, $\phi(y)\le M$ for all $y\in B\setminus B'$. Hence, by definition of subsolution,

$$2C|x'-y'|\le 0,$$

which implies $y'=x'$ and, hence, we get

$$u(y)-u(x')\le C|y-x'|\quad\forall y\in B.$$

By the arbitrariness of $x'\in B'$, u is Lipschitz in B'. ∎

LEMMA 2.4. *The upper weak limit* $\overline{V}$ *is constant with respect to the variable* $\zeta\in\Omega$, *for every* $(x,t)\in\mathbb{R}^N\times]0,+\infty[$.

Proof. First, we prove that $\overline{V}$ solves

$$|\nabla_z\overline{V}|\le 0\qquad\text{in }\mathbb{R}^N\times\Omega\times]0,+\infty[, \tag{2.5}$$

where ∇_z denotes the gradient of $\overline{V}$ with respect to ζ. Let us take a test function φ and a strict local maximum point (x_0,ζ_0,t_0) for $\overline{V}-\varphi$ in $\mathbb{R}^N\times\Omega\times]0,+\infty[$. Hence, there exists a sequence $(x_\varepsilon,\zeta_\varepsilon,t_\varepsilon)$ of local maxima in $\mathbb{R}^N\times\Omega\times]0,+\infty[$ for $V_\varepsilon-\varphi$, such

that, for $\varepsilon \to 0$, the sequence converges to (x_0, ζ_0, y_0) and $V_\varepsilon(x_\varepsilon, \zeta_\varepsilon, t_\varepsilon)$ converges to $\overline{V}(x_0, \zeta_0, t_0)$. Consequently, at $(x_\varepsilon, \zeta_\varepsilon, t_\varepsilon)$, we have

$$\varphi_t + \lambda V_\varepsilon + \sup_A \big\{ - f \cdot \nabla_y \varphi - l \big\} \le \frac{1}{\varepsilon} \inf_B \big\{ g \cdot \nabla_z \varphi \big\}. \tag{2.6}$$

Using the controllability condition (1.7), we have the following inequality:

$$\frac{1}{\varepsilon} \inf_{b \in B} \big\{ g(x_\varepsilon, \zeta_\varepsilon, b) \cdot \nabla_z \varphi(x_\varepsilon, \zeta_\varepsilon, t_\varepsilon) \big\} \le -\frac{1}{\varepsilon} \big| \nabla_z \varphi(x_\varepsilon, \zeta_\varepsilon, t_\varepsilon) \big|.$$

Since in a neighborhood of (x_0, ζ_0, t_0) the left-hand side of (2.6) is bounded; we then get $\nabla_z \varphi(x_0, \zeta_0, t_0) = 0$ and hence (2.5).

Next, applying Lemma 2.3, we get the conclusion. □

Now, our purpose is to compare the function $\overline{V}$ with the function V, in particular, we want $\overline{V}(x, \zeta, t) \le V(x, t)$. This will be done using the comparison result for the limit problem (P_0), Theorem 1.7. Unfortunately, if h is not constant with respect to $\zeta \in \overline{\Omega}$, then the inequality $\overline{V}(x, \zeta, 0) \le V(x, 0)$ does not hold. Note that, by Lemma 2.4, the function

$$\mathcal{V}(x, t) := \inf_{\zeta \in \Omega} \overline{V}(x, \zeta, t) \qquad \forall (x, t) \in \mathbb{R}^N \times [0, +\infty[.$$

satisfies

$$\mathcal{V}(x, t) = \overline{V}(x, \zeta, t) \qquad \forall (x, \zeta, t) \in \mathbb{R}^N \times \Omega \times]0, +\infty[. \tag{2.7}$$

LEMMA 2.5. *For every* $(x, t) \in \mathbb{R}^N \times [0, +\infty[$

$$\mathcal{V}(x, t) \le V(x, t). \tag{2.8}$$

Proof. First, we prove that $\mathcal{V}$ is a subsolution in $\mathbb{R}^N \times]0, +\infty[$ of (1.13). Let us consider a test function $\varphi \in C^1(\mathbb{R}^N \times]0, +\infty[)$ and a point (x_0, t_0) of strict local maximum for $\mathcal{V} - \varphi$ in $\mathbb{R}^N \times]0, +\infty[$. Let us take an arbitrary point $\zeta_0 \in \Omega$ and consider the function $\phi(x, \zeta, t) := \varphi(x, t) + |\zeta - \zeta_0|^2$. Using Lemma 2.4 and (2.7), we can say that (x_0, ζ_0, t_0) is a strict local maximum point for $\overline{V} - \phi$. Hence, there exists a sequence $(x_\varepsilon, \zeta_\varepsilon, t_\varepsilon)$ of maximum point for $V_\varepsilon - \phi$, converging to (x_0, ζ_0, t_0) and, moreover, $V_\varepsilon(x_\varepsilon, \zeta_\varepsilon, t_\varepsilon)$ converges to $\overline{V}(x_0, \zeta_0, t_0)$. Then, by (1.12), we get

$$\begin{aligned} &\varphi_t(x_\varepsilon, t_\varepsilon) + \lambda V_\varepsilon(x_\varepsilon, \zeta_\varepsilon, t_\varepsilon) \\ &+ \sup_{a \in A} \big\{ - f(x_\varepsilon, \zeta_\varepsilon, a) \cdot \nabla_y \varphi(x_\varepsilon, t_\varepsilon) - l(x_\varepsilon, \zeta_\varepsilon, a) \big\} \\ &\le \frac{1}{\varepsilon} \inf_{b \in B} \big\{ g(x_\varepsilon, \zeta_\varepsilon, b) \cdot 2(\zeta_\varepsilon - \zeta_0) \big\} \le 0, \end{aligned}$$

where the last inequality holds by virtue of the controllability condition (1.7). Hence, passing to the limit for $\varepsilon \to 0$, recalling (2.7), using the arbitrariness of $\zeta_0 \in \Omega$ and the continuity of f and l, we obtain

$$\varphi_t + \lambda \mathcal{V} + \sup_{\overline{\Omega} \times A} \big\{ - f \cdot \nabla \varphi - l \big\} \le 0 \qquad \text{at } (x_0, t_0).$$

On the other hand, by Lemma 2.2, the limit problem (P_0) and the definition of $\tilde{h}$, it turns out that $\mathcal{V}(x, 0) = V(x, 0)$, for every $x \in \mathbb{R}^N$. Hence, since V is a supersolution of (1.13), Theorem 1.7 gives the conclusion. □

PROPOSITION 2.6. *The sequence of value functions V_ε uniformly converges to V, over the compact sets of $\mathbb{R}^N \times \Omega\times]0,+\infty[$.*

Proof. Using (2.7), Lemma 2.1, Lemma 2.5, and the definitions of the weak limits (2.1), we obtain

$$V(x,t) \le \underline{V}(x,\zeta,t) \le \overline{V}(x,\zeta,t) \le V(x,t) \qquad \forall (x,\zeta,t) \in \mathbb{R}^N \times \Omega\times]0,+\infty[,$$

and the conclusion easily follows. ∎

Proposition 2.6 is close to the first statement of Theorem 1.1. The only difference is that we have not yet proved the convergence up to the boundary of Ω. To complete the proof, we consider a stronger controllability condition than (1.7), i.e.,

$$g(y,z,B) \supseteq \left\{\xi \in \mathbb{R}^M \middle| |\xi| \le 1\right\} \qquad \forall (y,z) \in \mathbb{R}^N \times \overline{\Omega}, \tag{2.9}$$

which, however, is not restrictive. Indeed, the next lemma shows that, for g and B satisfying (1.7), there exist $\tilde{g}$ and $\tilde{B}$ satisfying (2.9), such that the corresponding Cauchy problems (P_ε) are the same. Hence, the corresponding value functions V_ε and $\tilde{V}_\varepsilon$ coincide by the uniqueness statement of Theorem 1.4.

LEMMA 2.7. *There exist a compact set $\tilde{B}$ and a function $\tilde{g} : \mathbb{R}^N \times \mathbb{R}^M \times \tilde{B} \to \mathbb{R}^M$ with the same regularity properties as g, satisfying in addition* (2.9)*, such that equations* (1.12) *are the same.*

Proof. Let us consider the set

$$\Lambda := \left\{ (\lambda_1,\dots,\lambda_{M+1}) \in \mathbb{R}^{M+1} \middle| \lambda_i \ge 0 \ \forall i = 1,\dots,M+1, \ \sum_{i=1}^{M+1} \lambda_i = 1 \right\}$$

and define $\tilde{B} := B^{M+1} \times \Lambda$. We denote the elements $(b_1,\dots,b_{M+1},\lambda_1,\dots,\lambda_{M+1})$ of $\tilde{B}$ by $\tilde{b}$ and we define

$$\tilde{g}(y,z,\tilde{b}) := \sum_{i=1}^{M+1} \lambda_i g(y,z,b_i) \qquad \forall (y,z,\tilde{b}) \in \mathbb{R}^N \times \mathbb{R}^M \times \tilde{B}.$$

It is easy to prove that $\tilde{B}$ is compact and that $\tilde{g}$ has the same regularity properties as g. Moreover, by Carathéodory's theorem (see, for instance, [11])

$$\operatorname{cog}(y,z,B) = \tilde{g}(y,z,\tilde{B}) \qquad \forall (y,z) \in \mathbb{R}^N \times \mathbb{R}^M. \tag{2.10}$$

Since $\tilde{g}(y,z,\tilde{B})$ is closed, from (1.7) we see that $\tilde{g}$ has the property (2.9).

Now, let us note that the last term of the Hamiltonian is equal to

$$-\frac{1}{\varepsilon} \inf_{\xi \in \operatorname{cog}(x,\zeta,B)} \left\{ \xi \cdot \nabla_z V_\varepsilon(x,\zeta,t) \right\}.$$

Hence, by (2.10), the corresponding equations (1.12) are the same. ∎

In the sequel, we shall suppose that (2.9) holds for g and B.

LEMMA 2.8. *For every $(x,\zeta,t) \in \mathbb{R}^N \times \partial\Omega\times]0,+\infty[$, there exist $0 < \bar{t} < t$, a continuous path $\gamma : [0,\bar{t}] \to \mathbb{R}^N \times \overline{\Omega}$, $\gamma(\tau) = (y_\tau, z_\tau)$, and a constant $C > 0$, such that*

$$\begin{cases} (y_0,z_0) = (x,\zeta), & \\ (y_\tau,z_\tau) \in \mathbb{R}^N \times \Omega & \forall \tau \in]0,\bar{t}], \\ V_\varepsilon(x,\zeta,t) - V_\varepsilon(y_\tau,z_\tau,t-\tau) \le C\tau & \forall \tau \in [0,\bar{t}], \forall \varepsilon \in]0,1]. \end{cases} \tag{2.11}$$

Proof. Let $(x,\zeta,t) \in \mathbb{R}^N \times \partial\Omega \times]0,+\infty[$ be fixed and $\eta \in \mathbb{R}^M$, $0 < \bar{t} < t$ be such that

$$\tau \in]0,\bar{t}] \Rightarrow \zeta + \tau\eta \in \Omega.$$

Such a η exists by virtue of (1.6). It is not restrictive to suppose $|\eta| \leq 1$. We take an arbitrary control $\alpha \in \mathcal{A}$ and consider the continuous trajectory in $\mathbb{R}^N$, starting from x and defined by

$$\overline{y}(\tau) = x + \int_0^\tau f(\overline{y}(s), \zeta + s\eta, \alpha(s))ds.$$

Next, for every $0 < \varepsilon \leq 1$, let us consider the set

$$D(\varepsilon) := \big\{(\tau,b) \in [0,\bar{t}] \times B \big| g(\overline{y}(\tau), \zeta + \tau\eta, b) = \varepsilon\eta\big\}.$$

The set $D(\varepsilon)$ is compact and, by virtue of (2.9), it is not empty. Hence, by a selection lemma, see, for instance, [14], for every such ε, there exists a measurable function $\beta^\varepsilon : [0,\bar{t}] \to B$ such that

$$g(\overline{y}(\tau), \zeta + \tau\eta, \beta^\varepsilon(\tau)) = \varepsilon\eta, \qquad \text{a.e. } \tau \in [0,\bar{t}].$$

Then, for every $0 < \varepsilon \leq 1$, the following dynamical system

$$\begin{cases} y'(\tau) = f(y(\tau), z(\tau), \alpha(\tau)), & \tau \in]0,\bar{t}[, \\ z'(\tau) = \dfrac{1}{\varepsilon} g(y(\tau), z(\tau), \beta^\varepsilon(\tau)), & \tau \in]0,\bar{t}], \\ (y(0), z(0)) = (x,\zeta) \end{cases}$$

has the unique solution: $\tau \to (\overline{y}(\tau), \zeta + \tau\eta) \in \mathbb{R}^N \times \Omega$. If we define $\gamma(\tau) = (y_\tau, z_\tau) := (\overline{y}(\tau), \zeta + \tau\eta)$, then γ has the first two properties listed (2.11). Next, we prove the third one. Take any $\tau \in]0,\bar{t}]$, $\mu > 0$ and $0 < \varepsilon \leq 1$. Consider a control $(\alpha',\beta') \in \mathcal{A}^\varepsilon_{(y_\tau,z_\tau)}$ such that

$$V_\varepsilon(y_\tau, z_\tau, t-\tau) + \mu \geq J_\varepsilon(x,\zeta,t-\tau,\alpha',\beta'). \tag{2.12}$$

Then we define the control

$$(\alpha''(s),\beta''(s)) = \begin{cases} (\alpha(s), \beta^\varepsilon(s)), & \text{if } 0 \leq s < \tau, \\ (\alpha'(s-\tau), \beta'(s-\tau)), & \text{if } s \geq \tau, \end{cases}$$

which is admissible for (x,ζ). Hence, using (2.12) and the controls (α',β') and (α'',β''), we obtain the following estimate:

$$\begin{aligned} &V_\varepsilon(x,\zeta,t) - V_\varepsilon(y_\tau, z_\tau, t-\tau) \\ &\qquad \leq J_\varepsilon(x,\zeta,t,\alpha'',\beta'') - J_\varepsilon(y_\tau, z_\tau, t-\tau, \alpha',\beta') + \mu. \end{aligned} \tag{2.13}$$

From this, it is easy to deduce the last inequality in (2.11). □

LEMMA 2.9. *The sequence of value functions V_ε converges pointwise to the function V in the set $\mathbb{R}^N \times \overline{\Omega} \times]0,+\infty[$.*

Proof. The pointwise convergence in $\mathbb{R}^N \times \Omega \times]0,+\infty[$ immediately follows from Proposition 2.6. Hence, let us consider $(x,\zeta,t) \in \mathbb{R}^N \times \partial\Omega \times]0,+\infty[$. Using the notations of Lemma 2.8, let us fix $\delta > 0$ and take $\tau \in]0,\bar{t}[$ and $0 < \bar{\varepsilon} \leq 1$ such that

$$\begin{aligned} & C\tau \leq \frac{\delta}{3}, \\ & \left|V_\varepsilon(y_\tau, z_\tau, t-\tau) - V(y_\tau, t-\tau)\right| \leq \frac{\delta}{3} \qquad \forall 0 < \varepsilon \leq \bar{\varepsilon}, \\ & \left|V(y_\tau, t-\tau) - V(x,t)\right| \leq \frac{\delta}{3}, \end{aligned} \tag{2.14}$$

where the second inequality comes from Lemma 2.8 and Proposition 2.6 and the third one by the continuity of V. Hence, using Lemma 2.8 again, we obtain

$$V_\varepsilon(x,\zeta,t) - V(x,t) \leq \delta$$

and we get the conclusion because $V_\varepsilon \geq V$ (see Lemma 2.1). ∎

Proof of Theorem 1.1. Recalling (2.4), we know that the sequence of value functions V_ε is monotone in $\mathbb{R}^N \times \overline{\Omega} \times [0,+\infty[$. By Lemma 2.9, in the set $\mathbb{R}^N \times \overline{\Omega} \times]0,+\infty[$, the V_ε are continuous and pointwise converge to the continuous function V. Hence, the sequence uniformly converges to V, over any compact sets. Moreover, if the final cost h does not depend on $z \in \overline{\Omega}$, then by (1.4) both V_ε and V satisfy the same initial condition. Hence, by (2.3) and Lemma 2.4, we can conclude that $\overline{V}$ is constant in the variable $\zeta \in \Omega$, for every $(x,t) \in \mathbb{R}^N \times [0,+\infty[$. From this, proceeding as before, we obtain the uniform convergence over any compact set of $\mathbb{R}^N \times \overline{\Omega} \times [0,+\infty[$. ∎

3. The case of Lipschitz controls. In this section we give a simple application of Theorem 1.1. We consider $\Omega \subseteq \mathbb{R}^M$, $\zeta \in \overline{\Omega}$, $\varepsilon > 0$, $\lambda \geq 0$, $f : \mathbb{R}^N \times \mathbb{R}^M \to \mathbb{R}^N$, and $l, h : \mathbb{R}^N \times \mathbb{R}^M \to \mathbb{R}$. All the regularity hypotheses on Ω, f, l, and h for applying the results of the previous sections, are supposed to hold. Let us consider the following controlled dynamical system in $\mathbb{R}^N$:

$$\begin{cases} y'(t) = f(y(t), z(t)), & t > 0, \\ y(0) = x, \end{cases} \tag{3.1}$$

where the controls z belong to the set

$$\mathcal{L}^\varepsilon_\zeta := \left\{ z : [0,+\infty[\to \overline{\Omega} \;\middle|\; |z(t_1) - z(t_2)| \leq \frac{1}{\varepsilon}|t_1 - t_2|,\ z(0) = \zeta \right\}.$$

Note that $\mathcal{L}^\varepsilon_\zeta$ is the set of Lipschitz functions with Lipschitz constant less than $1/\varepsilon$ and with initial point ζ.

We want to minimize the cost functional

$$J(x,t,z) := \int_0^t e^{-\lambda s} l(y_x(s;z), z(s)) ds + e^{-\lambda t} h(y_x(t;z), z(t)). \tag{3.2}$$

Hence, we define the value function

$$V_\varepsilon(x,t,\zeta) := \inf_{z \in \mathcal{L}^\varepsilon_\zeta} J(x,t,z),$$

which depends also on the initial point ζ for the control z.

Next, we consider the optimal control problem defined by system (3.1), measurable controls $z \in \mathcal{Z}$ taking value in $\overline{\Omega}$ and value function

$$V(x,t) := \inf_{z\in\mathcal{Z}} \left(\int_0^t e^{-\lambda s} l(y_x(s;z), z(s))ds + e^{-\lambda t} \inf_{\zeta\in\overline{\Omega}} h(y_x(t;z),\zeta) \right).$$

PROPOSITION 3.1. *The value functions V_ε uniformly converge to the value function V, over any compact set of $\mathbb{R}^N \times \overline{\Omega} \times]0,+\infty[$. Moreover, if h does not depend on $\zeta \in \overline{\Omega}$, then the convergence is uniform over any compact set of $\mathbb{R}^N \times \overline{\Omega} \times [0,+\infty[$.*

Proof. We set $\mathcal{B} := \{\beta : [0,+\infty[\to \{\xi \in \mathbb{R}^M, |\xi| \le 1\} \text{ measurable}\}$ and consider the system

$$\begin{cases} y'(t) = f(y(t), z(t)), & t > 0, \\ z'(t) = \dfrac{1}{\varepsilon}\beta(t), & t > 0, \\ (y(0), z(0)) = (x, \zeta). \end{cases} \tag{3.3}$$

This is a special case of system (0.1) (where A is a singleton) and the hypotheses of Theorem 1.1 are satisfied. Then, we consider the set $\mathcal{B}^\varepsilon_{(x,\zeta)}$ of admissible controls with respect to the state-space constraint: $z(t) \in \overline{\Omega}$. By Theorem 1.1, if we take J defined in (3.2) and the value function

$$V^\varepsilon(x,\zeta,t) := \inf_{\beta\in\mathcal{B}^\varepsilon_{(x,\zeta)}} J(x,t,z),$$

they uniformly converge to V over the compact sets, as desired. On the other hand, a function z is Lipschitz, with Lipschitz constant less than $1/\varepsilon$, if and only if it satisfies the equation $z' = (1/\varepsilon)\beta$ almost everywhere, with $\beta \in \mathcal{B}$. Therefore, it is easy to see that $V_\varepsilon = V^\varepsilon$ and the proof is complete. □

Now, we consider the set $\mathcal{L}$ of Lipschitz controls $z : [0,+\infty[\to \overline{\Omega}$, with arbitrary Lipschitz constant and without fixed initial point. Again, we consider the corresponding value function

$$V_{\mathcal{L}}(x,t) := \inf_{z\in\mathcal{L}} J(x,t,z).$$

Since piecing together two elements of $\mathcal{L}$ does not necessary give an element of $\mathcal{L}$, the dynamic programming principle (DPP) does not obviously hold. Hence, the standard proof that the function $V_{\mathcal{L}}$ satisfies the Hamilton–Jacobi equation does not work. However, $V_{\mathcal{L}}$ turns out to coincide with V by Proposition 3.1 because

$$V(x,t) \le V_{\mathcal{L}}(x,t) \le V_\varepsilon(x,t,\zeta) \qquad \forall \varepsilon > 0,\ \forall (x,\zeta,t) \in \mathbb{R}^N \times \overline{\Omega} \times [0,+\infty[.$$

4. Remarks and extensions. In this section, we give some ideas about possible extensions and applications of our method to other control problems.

First of all, we consider the controllability hypothesis (0.6). If we replace (0.6) with the following assumption:

$$\forall (y,z) \in \mathbb{R}^N \times \overline{\Omega}\ \exists r > 0 \text{ such that } \overline{\text{co}}\, g(y,z,B) \supseteq \overline{B}(0,r), \tag{4.1}$$

then our method still works. The difference between (0.6) and (4.1) is that in (4.1) the radius of the ball depends on the point (y,z). Actually, we can further enlarge

the class of perturbation problems as follows. We replace the equation for the fast variables z in (0.1) with

$$z'(t) = g_1(y(t), z(t), \beta(t)) + \frac{1}{\varepsilon} g(y(t), z(t), \beta(t)), \tag{4.2}$$

with g and g_1 satisfying the standard uniform Lipschitz continuity assumption and g verifying (4.1). We note that the function $\varepsilon g_1 + g$ does not necessarily satisfy (0.6); however, since g_1 is locally bounded, for every (y, z) we can take suitably small ε, such that the function $\varepsilon g_1 + g$ satisfies condition (4.1) in a neighborhood of (y, z). For instance, a linear system

$$z' = M_1 y + M_2 z + \frac{1}{\varepsilon} M_3 b,$$

where M_1, M_2 and M_3 are matrices, satisfies (4.1) if $M_3 B \supseteq \overline{B}(0, r)$ for some r, but the vector field $\varepsilon(M_1 y + M_2 z) + M_3 b$ satisfies (0.6) only if M_1 and M_2 are both null.

THEOREM 4.1. *For every $\varepsilon > 0$, let V_ε be the value function of the control problem where the fast variables are driven by (4.2). Then, the sequence $(V_\varepsilon)_{\varepsilon>0}$ uniformly converges in any compact set of $\mathbb{R}^N \times \overline{\Omega} \times]0, +\infty[$, as $\varepsilon \to 0$, to the value function V of the same limit problem as in section 2.*

Sketch of the proof. Using the Lipschitz continuity of g and g_1 in (y, z) uniformly with respect to b, it is not hard to prove that the function

$$r(y, z, \varepsilon) := \max \left\{ r > 0 \Big| \overline{\text{co}} \Big[g_1(y, z, B) + \frac{1}{\varepsilon} g(y, z, B) \Big] \supseteq \overline{B}(0, r) \right\}$$

is well defined and continuous in $(y, z, \varepsilon) \in \mathbb{R}^N \times \overline{\Omega} \times]0, +\infty[$. Moreover, for every (y, z), $r(y, z, \varepsilon) > 0$ for suitable small ε, by (4.1); hence for every compact set of $\mathbb{R}^N \times \overline{\Omega}$ and small ε, r has a strictly positive minimum. Thus, modifying appropriately the proofs of Lemmas 2.2, 2.4, 2.5, and 2.8 to the new system (4.2), (4.1), Theorem 4.1 can be easily proved. □

Remark 4.1. We can consider the system (0.1) with the second equation

$$z'(t) = \frac{1}{\varepsilon} \left(g_1(y(t), z(t), \beta(t)) + g(y(t), z(t), \beta(t)) \right), \tag{4.3}$$

and make the proofs of Lemmas 2.2, 2.4, and 2.5 work again, provided that, for every $\varepsilon > 0$, the value function V_ε is a subsolution of

$$\nabla_z V_\varepsilon \cdot g_1 \leq 0. \tag{4.4}$$

In this case, the sign of the uncontrollable terms in the proofs of the lemmas is the correct one. Hence, we get the usual convergence result, at least in any compact set of $\mathbb{R}^N \times \Omega \times]0, +\infty[$. □

Next we explain how our method applies to exit time problems. With the notions and hypotheses of the Introduction and section 1, we take an open bounded set $\mathcal{O} \subset \mathbb{R}^N$ with smooth boundary, in particular with the interior cone property (i.e., the obvious analogue of the first property in (1.6)). For every trajectory $(y(\cdot), z(\cdot))$ of the system (0.1) we define the exit time from $\mathcal{O}$, which depends on the starting point (x, ζ), on the admissible control (α, β), and on $\varepsilon > 0$:

$$\tau^\varepsilon = \tau(x, \zeta, \alpha, \beta, \varepsilon) := \min \left\{ t \geq 0 \big| y(t) \in \partial\mathcal{O} \right\}. \tag{4.5}$$

In the same way, we define the exit time from $\mathcal{O}$ for the limit system (0.8), which depends on the starting point x and on the measurable controls z and α:

$$\tau = \tau(x, z, \alpha) := \min\left\{t \geq 0 \middle| y(t) \in \partial\mathcal{O}\right\}. \tag{4.6}$$

We consider the following cost functionals:

$$J_\varepsilon(x, \zeta, \alpha, \beta) := \int_0^{\tau^\varepsilon} e^{-\lambda s} l(y(s), z(s), \alpha(s)) ds + e^{-\lambda \tau^\varepsilon} h(y(\tau^\varepsilon)), \tag{4.7}$$

$$J(x, z, \alpha) := \int_0^{\tau} e^{-\lambda s} l(y(s), z(s), \alpha(s)) ds + e^{-\lambda \tau} h(y(\tau)), \tag{4.8}$$

where $\lambda > 0$ and the terminal cost h is a continuous function in a neighborhood $B(\partial\mathcal{O}, \delta)$ of $\partial\mathcal{O}$. Next we define the value functions for the perturbed problems and for the limit problem as, respectively,

$$V_\varepsilon(x, \zeta) := \inf_{(\alpha,\beta)\in(\mathcal{A}\times\mathcal{B})^\varepsilon_{(x,\zeta)}} J_\varepsilon(x, \zeta, \alpha, \beta) \tag{4.9}$$

and

$$V(x) := \inf_{(z,\alpha)\in\mathcal{Z}\times\mathcal{A}} J(x, z, \alpha). \tag{4.10}$$

Let us now suppose that the terminal cost h satisfies the compatibility condition

$$h(x) \leq J(x, z, \alpha) \qquad \forall x \in B(\partial\mathcal{O}, \delta),\ \forall (z, \alpha) \in \mathcal{Z} \times \mathcal{A}, \tag{4.11}$$

and the limit system has the controllability property

$$\sup_{(\zeta,a)\in\overline{\Omega}\times A} f(x, \zeta, a) \cdot \nu(x) > 0 \quad \forall x \in \partial\mathcal{O}, \tag{4.12}$$

where $\nu(x)$ is the exterior normal to $\mathcal{O}$. These assumptions ensure the continuity of the value function V by the results in Chapter 4 of [4] (see also [22] for similar results; note that (4.11) implies $h(x) \leq V(x)$ in $B(\partial\mathcal{O}, \delta)$; thus V is lower semicontinuous at all $x \in \partial\mathcal{O}$, whereas (4.12) implies the upper semicontinuity of V at boundary points). The value functions V_ε, however, may be discontinuous.

Under the previous assumptions, the following theorem holds.

THEOREM 4.2. *When ε tends to 0, the sequence of value functions V_ε converges to the value function V uniformly over the compact sets of $\mathcal{O} \times \Omega$.*

Sketch of the proof. By the assumptions (4.11) and (4.12), $V \in BUC(\overline{\mathcal{O}})$ and it is the solution of the boundary value problem

$$\begin{cases} \lambda V + \sup_{z\in\overline{\Omega}} H(y, z, \nabla V, 0) = 0 & \text{in } \mathcal{O}, \\ V = h & \text{on } \partial\mathcal{O}, \end{cases} \tag{4.13}$$

where the Hamiltonian H is defined in (0.11).

Now, for every $\varepsilon > 0$, we denote by V_ε^* the upper semicontinuous envelope of V_ε. By the results of Ishii [16], V_ε^* is subsolution of

$$\begin{cases} \lambda V_\varepsilon^* + H(y, z, \nabla_y V_\varepsilon^*, \frac{1}{\varepsilon}\nabla_z V_\varepsilon^*) \leq 0 & \text{in } \mathcal{O} \times \Omega, \\ V_\varepsilon^* \leq h \text{ or } \lambda V_\varepsilon^* + H(y, z, \nabla_y V_\varepsilon^*, \frac{1}{\varepsilon}\nabla_z V_\varepsilon^*) \leq 0 & \text{on } \partial\mathcal{O} \times \Omega, \end{cases} \tag{4.14}$$

where the boundary condition is also interpreted in the viscosity sense.

Then, as in (2.1), we define the lower weak limit $\underline{V}$ and the upper weak limit $\overline{V}$. Also in this case, for every strict local maximum point (x_0, ζ_0) for $\overline{V} - \varphi$ there exists a sequence $(x_\varepsilon, \zeta_\varepsilon) \in \overline{\mathcal{O}} \times \Omega$ of maximum points for $V_\varepsilon^* - \varphi$ converging to (x_0, ζ_0), such that $V_\varepsilon^*(x_\varepsilon, \zeta_\varepsilon)$ converges to $\overline{V}(x_0, \zeta_0)$. Hence, as in Lemma 2.4, we can say that $\overline{V}$ is constant with respect $\zeta \in \Omega$ in $\mathcal{O} \times \Omega$. Then, if we define the function

$$\mathcal{V}(x) := \inf_{\zeta \in \Omega} \overline{V}(x, \zeta) \qquad \forall x \in \overline{\mathcal{O}}, \tag{4.15}$$

we obtain that $\mathcal{V} = \overline{V}$ in $\mathcal{O}$. As in Lemma 2.5, we can prove that $\mathcal{V}$ is an upper semicontinuous subsolution of

$$\begin{cases} \lambda \mathcal{V} + \sup_{z \in \overline{\Omega}} H(y, z, \nabla \mathcal{V}, 0) = 0 & \text{in } \mathcal{O}, \\ \mathcal{V} = h \text{ or } \lambda \mathcal{V} + \sup_{z \in \overline{\Omega}} H(y, z, \nabla \mathcal{V}, 0) = 0 & \text{on } \partial \mathcal{O}. \end{cases} \tag{4.16}$$

The only difference is the boundary condition, and we observe that if $\mathcal{V}(x_0) > h(x_0)$ for some $x_0 \in \partial \mathcal{O}$, then, along the above sequence $(x_\varepsilon, \zeta_\varepsilon)$ converging to (x_0, ζ_0) with arbitrary $\zeta_0 \in \Omega$, we have $V_\varepsilon^*(x_\varepsilon, \zeta_\varepsilon) > h(x_\varepsilon)$, at least for small ε. Therefore, the Hamilton–Jacobi equation holds at such points by (4.14), and by letting $\varepsilon \to 0$ we obtain that $\mathcal{V}$ satisfies the Hamilton–Jacobi equation at the boundary point x_0.

Now we apply the comparison result in Ishii [16] (see also [4]) and get

$$\overline{V}(x, \zeta) \le V(x) \qquad \forall (x, \zeta) \in \mathcal{O} \times \Omega.$$

Since the inequality $V \le \underline{V}$ follows from the definitions of V, V_ε, and $\underline{V}$, we easily get the conclusion. □

Remark 4.2. The same convergence result as in Theorem 4.2 holds if we consider the control problem with exit time from the closure of $\mathcal{O}$. In this case we define

$$\hat{\tau}^\varepsilon = \hat{\tau}(x, \zeta, \alpha, \beta, \varepsilon) := \inf \left\{ t \ge 0 \middle| y(t) \notin \overline{\mathcal{O}} \right\}, \tag{4.17}$$

$$\hat{\tau} = \hat{\tau}(x, z, \alpha) := \inf \left\{ t \ge 0 \middle| y(t) \notin \overline{\mathcal{O}} \right\}. \tag{4.18}$$

Let $\hat{V}_\varepsilon$ and $\hat{V}$ be the value functions defined as in (4.9) and (4.10) with cost functionals J_ε and J as in (4.7) and (4.8) with τ^ε and τ replaced by $\hat{\tau}^\varepsilon$ and $\hat{\tau}$, respectively.

Then, under the hypotheses (4.11) and (4.12), $\hat{V}_\varepsilon \to \hat{V} = V$ when $\varepsilon \to 0$, where V is the value function defined in (4.10). In fact, $\hat{V} = V$ because the upper semicontinuous and lower semicontinuous envelopes $\hat{V}^*$ and $\hat{V}_*$ are, respectively, sub- and supersolution of (4.16), hence, by a comparison theorem, $\hat{V}^* \le V \le \hat{V}_*$ (see [16], [4]). On the other hand, the functions $\hat{V}_\varepsilon$ do not necessarily coincide with V_ε defined in (4.9), but $\hat{V}_\varepsilon^*$ satisfies (4.14) for every ε. Thus the upper weak limit of $\hat{V}_\varepsilon$ is a subsolution of (4.16) and the conclusion follows as in the previous proof. □

Appendix: Continuity of the value function. In this Appendix we shall use the notation $B(x, r)$ for the open ball with radius r and centre x.

We consider a general controlled dynamical system in $\mathbb{R}^m$:

$$\begin{cases} y'(t) = f(y(t), \alpha(t)), & t > 0, \\ y(0) = x, \end{cases} \tag{A.1}$$

with the constraint on the trajectories

$$y(t) := y_x(t; \alpha) \in \overline{\Omega} \qquad \forall t > 0. \tag{A.2}$$

We have the measurable controls $\alpha \in \mathcal{A}$, which takes values in a compact set A. Hence, we consider the cost functional

$$J(x,t,\alpha) := \int_0^t e^{-\lambda s} l(y(s),\alpha(s))ds + e^{-\lambda t} h(y(t)) \tag{A.3}$$

and the value function

$$V(x,t) := \inf_{\alpha\in\mathcal{A}_x} J(x,t,\alpha), \tag{A.4}$$

where $\mathcal{A}_x$ is the set of admissible controls. The hypotheses are the following:

$$\begin{aligned}
&\Omega \subseteq \mathbb{R}^m \text{ is open and connected and satisfies (1.6), with } M \text{ replaced by } m;\\
&f : \mathbb{R}^m \times A \to \mathbb{R}^m \text{ is continuous;}\\
&|f(x,a) - f(y,a)| \le L|x-y| \qquad \forall x,y \in \mathbb{R}^m, \forall a \in A;\\
&|f(x,a)| \le K(1+|x|) \qquad \forall x,y \in \mathbb{R}^m, \forall a \in A;\\
&l : \mathbb{R}^m \times A \to \mathbb{R} \text{ is continuous;}\\
&|l(x,a) - l(y,a)| \le \omega(|x-y|),\ |l(x,a)| \le K \qquad \forall x,y \in \mathbb{R}^m, \forall a \in A;\\
&h : \mathbb{R}^m \to \mathbb{R} \text{ is continuous and bounded,}
\end{aligned} \tag{A.5}$$

where $\omega(0) = 0$ and ω is continuous and increasing. Note, that the fourth inequality in (A.5), follows from the continuity of f and the compactness of A. Moreover, we will use the following controllability assumption at points of $\partial\Omega$:

$$\forall z \in \partial\Omega\ \exists a \in A \text{ such that } g_i(z) = 0 \Rightarrow f(z,a)\cdot\nabla g_i(z) < 0. \tag{A.6}$$

It is easy to see that, for every $\varepsilon > 0$, the problem described in the previous section by (0.1), (0.2), (0.3), and (1.3) is a particular case of the last one, with $m = N + M$.

THEOREM A.1. *Under the assumptions above, $\mathcal{A}_x$ is not empty and V is bounded and continuous in $\overline{\Omega}\times[0,T]$, for every $T > 0$.*

We need a lemma. First, let us define

$$\begin{aligned}
I_t(x,\alpha) &:= \int_0^t e^{-\lambda s} l(y_x(s;\alpha),\alpha(s))ds \qquad \forall x \in \overline{\Omega},\ \forall \alpha \in \mathcal{A},\ \forall t > 0,\\
G_i &:= \{x \in \mathbb{R}^N : g_i(x) \le 0\}, \qquad i = 1,\dots,q.
\end{aligned} \tag{A.7}$$

LEMMA A.2. *For every $x_0 \in \overline{\Omega}$ there exist constants $r_0 > 0$, $t^* > 0$, $C > 0$ such that, for every $x \in B(x_0,r_0)\cap\overline{\Omega}$ and for every $\alpha \in \mathcal{A}$, there exists $\overline{\alpha} \in \mathcal{A}$ such that*

$$y_x(t;\overline{\alpha}) \in \overline{\Omega} \qquad \forall t \in [0,t^*], \tag{A.8}$$

$$\left|I_{t^*}(x,\overline{\alpha}) - I_{t^*}(x,\alpha)\right| \le C \max\left\{\operatorname{dist}(y_x(t,\alpha),G_i) : t \in [0,t^*], i = 1,\dots,q\right\}. \tag{A.9}$$

Proof. This proof is a modification of the Soner's one [24]. If $x_0 \in \Omega$, then we are done, since, by the third and the fourth of (A.5) and Gronwall's lemma, we get

$$\begin{cases} |y_x(t;\alpha) - y_z(t;\alpha)| \le e^{Lt}|x-z| & \forall x,z \in \mathbb{R}^m,\ \forall \alpha \in \mathcal{A},\ \forall t \ge 0;\\ |y_x(t;\alpha) - x| \le Kt(1+|x|)e^{Kt} & \forall x \in \mathbb{R}^m,\ \forall \alpha \in \mathcal{A},\ \forall t \ge 0, \end{cases} \tag{A.10}$$

and so, for a suitable small time interval independent on the control, all the trajectories starting from the same point do not go out of a fixed compact set. Hence, let us suppose $x_0 \in \partial\Omega$. By (A.5) and (A.6), there exist a nonempty set $I \subseteq \{1,\dots,q\}$, a constant control $\overline{a} \in A$, and two positive numbers δ_0, ξ_0, such that

$$(A.11)\qquad \begin{cases} g_i(x_0) = 0 & \forall i \in I,\ g_j(x_0) < 0,\ \forall j \notin I, \\ B(x_0,\delta_0) \subseteq \overset{\circ}{G}_j & \forall j \notin \mathrm{I}, \\ f(x,\overline{a})\cdot\nabla g_i(x) < -\xi_0 < 0 & \forall x \in B(x_0,\delta_0)\cap\overline{\Omega},\ \forall i \in I. \end{cases}$$

Now, let us take $r_0 < \delta_0$, a point $x \in B(x_0,r_0)\cap\overline{\Omega}$ and fix $t' > 0$ sufficiently small such that

$$(A.12)\qquad \begin{cases} y_x(t;\alpha) \in B(x_0,\delta_0) & \forall t \in [0,t'],\ \forall \alpha \in \mathcal{A}, \\ f(y_x(t;\alpha),\overline{a})\cdot\nabla g_i(y_x(t,\alpha)) \le -\xi_0 < 0 & \forall t \in [0,t'],\ \forall i \in I, \end{cases}$$

which is possible by (A.11). So we have

$$(A.13)\qquad y_z(t;\alpha) \in G_j \qquad \forall t \in [0,t'],\ \forall z \in B(x_0,r_0)\cap\overline{\Omega},\ \forall \alpha \in \mathcal{A},\ \forall j \notin I.$$

Then, let us take a measurable control $\alpha \in \mathcal{A}$ and define

$$(A.14)\qquad \begin{aligned} &t_0 := \inf\big\{t \in]0,t'] : y_x(t;\alpha) \in \partial\Omega\big\}, \qquad t_0 := t' \text{ if } y_x(t;\alpha) \notin \partial\Omega\ \forall t \in [0,t'], \\ &\mu := \max\big\{\mathrm{dist}(y_x(t;\alpha),G_s) : t \in [0,t'],\ s = 1,\dots,q\big\}, \\ &\overline{\alpha}(t) := \alpha(t)\chi_{[0,t_0[}(t) + \overline{a}\chi_{[t_0,t_0+h\mu]}(t) + \alpha(t-h\mu)\chi_{]t_0+h\mu,+\infty[}(t), \end{aligned}$$

where $h > 0$ will be fixed later and χ_S denotes the characteristic function of the set S.

By (A.12), for proving (A.8), we have only to prove that $y_x(t;\overline{\alpha}) \in G_i$, that is $g_i(y_x(t;\overline{\alpha})) \le 0$, for every fixed $t \in [0,t']$ and every $i \in I$ (note that only for the moment t' is our main candidate as t^*).

If $0 \le t \le t_0$ then, by virtue of definition (A.14), there is nothing to prove. If $t_0 < t \le t_0 + h\mu$ and $t \le t'$, then using (A.12), (A.13), and the fact that $y_x(t;\overline{\alpha})$ is a solution of (A.1), we get for every $i \in I$:

$$g_i(y_x(t;\overline{\alpha})) = g_i(y_x(t_0;\overline{\alpha})) + \int_{t_0}^{t} \nabla g_i(y_x(s;\overline{\alpha}))\cdot f(y_x(s;\overline{\alpha}),\overline{\alpha}(s))ds \le -\xi_0(t-t_0) < 0,$$

and, hence, we get $y_x(t;\overline{\alpha}) \in \overline{\Omega}$ for every $t \le t'$,$t \in [t_0,t_0+h\mu]$. If $t \le t'$ and $t > t_0 + h\mu$, then for a suitable $\tau > t_0$ it is $t = \tau + h\mu$. Then, using the definition of $\overline{\alpha}$, we have

$$(A.15)\qquad \begin{aligned} &g_i(y_x(\tau+h\mu;\overline{\alpha})) = g_i(y_x(t_0;\overline{\alpha})) + \int_{t_0}^{t_0+h\mu} \nabla g_i(y_x(s;\overline{\alpha}))\cdot f(y_x(s,\overline{\alpha}),\overline{a})ds \\ &+ \int_{t_0+h\mu}^{\tau+h\mu} \nabla g_i(y_x(s;\overline{\alpha}))\cdot f(y_x(s;\overline{\alpha}),\alpha(s-h\mu))ds \\ &\le g_i(y_x(t_0;\overline{\alpha})) - \xi_0 h\mu + \int_{t_0}^{\tau} \nabla g_i(y_x(s+h\mu;\overline{\alpha}))\cdot f(y_x(s+h\mu;\overline{\alpha}),\alpha(s))ds. \end{aligned}$$

Now, using (A.5) and the fact that g_i is $C^{1,1}$ and, hence, bounded and Lipschitz, with its first derivatives over compact sets and using (A.10), we obtain, after adding and subtracting suitable terms

$$(A.16)\qquad g_i(y_x(\tau+h\mu;\overline{\alpha})) \le -\xi_0 h\mu + Ch\mu\int_{t_0}^{\tau}\Big[e^{L(s-t_0)}ds + (g_i\circ y_x(\cdot,\alpha))'(s)\Big]ds,$$

where the constant $C > 0$ depends on x_0 only. Next, we suppose that the radius r_0 is small enough, such that, in the ball $B(x_0, r_0)$, the following inequality holds: $g_i(z) \leq \mathrm{dist}(z, G_i)$ for every $z \in B(x_0, r_0)$. The last is not restrictive; it is sufficient to divide the function g_i by its Lipschitz constant in the ball $B(x_0, 2\delta_0)$ and to note that the set G_i remains unchanged after this operation. Hence, if we take $0 < t^* \leq t'$ such that

$$t^* \leq \frac{1}{L} \log\left(1 + \frac{L\xi_0}{2C}\right),$$

then we get

$$g_i(y_x(\tau + h\mu; \overline{\alpha})) \leq \mu\left(1 - \frac{\xi_0}{2}h\right).$$

Hence, if we take $h = 2/\xi_0$, (A.8) is proved.

Now, let us note that it is not restrictive to suppose the running cost l to be a Lipschitz function on $x \in \mathbb{R}^M$ uniformly in $a \in A$. In fact, we are interested in the continuity of the value function and hence, if the value function with Lipschitz running cost is continuous, then, taking a uniformly approximating sequence of Lipschitz running cost, we have a uniformly approximating sequence of continuous value functions. So, for Lipschitz l, using the construction of $\overline{\alpha}$ and standard estimates, it is easy to prove that (A.9) holds. □

Remark A.1. It is easy to convince oneself that, for x in a compact subset of $\overline{\Omega}$, the constants r_0, t^* and C of Lemma A.2, and also the quantity h, can be uniformly chosen. □

Proof of Theorem A.1. Let us prove that $\mathcal{A}_x \neq \emptyset$ for every $x \in \overline{\Omega}$. Let us take $x_0 \in \overline{\Omega}$ and let t_0^* be the supremum of the values for t^*. If $t_0^* = +\infty$, then we are done. If it is not the case, let us take $\alpha_0 \in \mathcal{A}$ such that $y_{x_0}(t; \alpha_0) \in \overline{\Omega}$ for every $0 \leq t \leq t_0^*$. In a similar way, we define t_1^* with respect to $x_1 := y_{x_0}(t_0^*; \alpha_0)$. If $t_1^* = +\infty$, we are done. If it is not the case, let us take $\alpha_1 \in \mathcal{A}$ such that $y_{x_1}(t; \alpha_1) \in \overline{\Omega}$ for every $0 \leq t \leq t_1^*$ and take t_2^* with respect to $x_2 := y_{x_1}(t; \alpha_1)$. Then we consider the measurable control $\overline{\alpha}$, constructed by gluing the controls α_n; it belongs to $\mathcal{A}_{x_0}$ if and only if $\sum_{n=0}^{+\infty} t_n^* = +\infty$. But if this is not true, then, by (A.10), the trajectory $y_{x_0}(\cdot; \overline{\alpha})$ does not exit from a compact subset of $\overline{\Omega}$. Hence, by Remark A.1, the values t_n^* are bounded away from zero, which is a contradiction to $\sum_{n=0}^{+\infty} t_n^* < +\infty$.

The boundedness of the value function is obvious; then we prove the continuity in every set $\overline{\Omega} \times [0, T]$ with $T > 0$. Let us take $T > 0$, $x_0 \in \overline{\Omega}$ and consider a ball B with x_0 as centre. We shall show that, for a suitable ball B' inside B and centered in x_0, the value function is uniformly continuous in $B' \cap \overline{\Omega} \times [0, T]$. Hence, the continuity in $\overline{\Omega} \times [0, T]$ will be proved. By Remark A.3, we can take the constants t^*, r_0, and C uniformly in B. Using (A.10), we take B' such that, $y_x(t; \alpha) \in B$ for every $t \in [0, T]$, for every $\alpha \in \mathcal{A}$ and for every $x \in B'$. Hence, using Lemma A.2 and proceeding as in Soner [24], we get the following estimates, which hold for every $x, z \in B' \cap \overline{\Omega}$, for every α, for every $\overline{\alpha}$ as in Lemma A.2 with respect to x and α, and for every $t \in [0, t^*]$:

$$\begin{aligned} &\big|y_x(t; \overline{\alpha}) - y_z(t; \alpha)\big| \leq C\big|x - z\big|, \\ &\big|I_t(x, \overline{\alpha}) - I_t(z, \alpha)\big| \leq C\big|x - z\big|, \end{aligned} \tag{A.17}$$

where $C > 0$ and I_t is defined as in (A.6). Moreover, starting from two points $x, z \in B'$, the same estimates hold even for every $t \in [0, T]$, if $T > t^*$. In fact, if

$T \le t^*$, during the time-interval $[0,T]$, for every fixed $\alpha \in \mathcal{A}$, the trajectories starting from a point of B' do not exit from B. Hence, we can repeat the same argument, used for getting the estimate, with respect to the starting points $y_x(t^*;\overline{\alpha})$, $y_z(t^*;\alpha)$ and to the same value t^*. Let us iterate this argument. At every nth step, we construct, as in Lemma A.2, a control $\overline{\alpha}_n$ depending on the starting point $y_x(nt^*;\overline{\alpha}_{n-1})$ and on the control α. Hence, for every $x \in B' \cap \overline{\Omega}$ and for every $\alpha \in \mathcal{A}$ there exists a control $\overline{\alpha}$ such that $y_x(t;\overline{\alpha}) \in \overline{\Omega}$ and all the previous estimates hold for every $t \in [0,T]$.

Next, let us fix $\delta > 0$. Then, by definition of V (A.4), for a suitable $\alpha \in \mathcal{A}_z$ we have

$$
\begin{aligned}
&V(x,t) - V(z,s) \\
&\qquad \le I_t(x,\overline{\alpha}) - I_s(z,\alpha) + e^{-\lambda t} h(y_x(t;\overline{\alpha})) - e^{-\lambda s} h(y_z(s;\alpha)) + \delta.
\end{aligned} \tag{A.18}
$$

Hence, using the estimates (A.17), the definition of I_t, denoting by ω_h a modulus of continuity of h in B, and adding and subtracting suitable terms, we get

$$
V(x,t) - V(z,s) \le C_1 \big|t-s\big| + C\big|x-z\big| + \omega_h\big(C\big|x-z\big|\big) + \delta,
$$

for a suitable positive constant C_1 and for C as in (A.17). Hence, we reach the conclusion by the arbitrariness of $\delta > 0$. □

REFERENCES

[1] F. Antoniali, *Problemi di controllo ottimo con variabili veloci e vincoli di stato: equazioni di Bellman e perturbazione singolare*, thesis, Università di Padova, Italy, July 1996.

[2] F. Bagagiolo, *Soluzioni di viscosità vincolate di equazioni di Bellman e perturbazione singolare in problemi di controllo ottimo*, thesis, Università di Padova, Italy, July 1993.

[3] F. Bagagiolo, M. Bardi, and I. Capuzzo Dolcetta, *A viscosity solutions approach to some asymptotic problems in optimal control*, in Partial Differential Equations Methods in Control and Shape Analysis, G. Da Prato and J.-P. Zolesio, eds., Marcel Dekker, New York, 1997, pp. 27–37.

[4] M. Bardi and I. Capuzzo Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton–Jacobi-Bellman Equations*, Birkhäuser, Boston, 1997.

[5] G. Barles, *Solutions de Viscosité des Équations de Hamilton–Jacobi*, Mathématiques et Applications 17, Springer-Verlag, Paris, 1994.

[6] G. Barles and B. Perthame, *Discontinuous solutions of deterministic optimal stopping time problems*, RAIRO Modél. Math. Anal. Numér., 21 (1987), pp. 557–579.

[7] G. Barles and B. Perthame, *Exit time problems in optimal control and vanishing viscosity method*, SIAM J. Control Optim., 26 (1988), pp. 1133–1148.

[8] E.N. Barron, L.C. Evans, and R. Jensen, *Viscosity solutions of Isaac's equations and differential games with Lipschitz controls*, J. Differential Equations, 53 (1984), pp. 213–233.

[9] A. Bensoussan, *Perturbation Methods in Optimal Control*, Wiley/Gauthier–Villars, Chichester, 1988.

[10] I. Capuzzo Dolcetta and P.L. Lions, *Hamilton–Jacobi equations with state constraints*, Trans. Amer. Math. Soc., 318 (1990), pp. 643–683.

[11] L. Cesari, *Optimization–Theory and Applications*, Springer-Verlag, New York, 1983.

[12] M.G. Crandall, L.C. Evans, and P.L. Lions, *Some properties of viscosity solutions of Hamilton–Jacobi equations*, Trans. Amer. Math. Soc., 282 (1984), pp. 487–502.

[13] M.G. Crandall and P.L. Lions, *Viscosity solutions of Hamilton–Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.

[14] W.H. Fleming and R. Rishel, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, Berlin, 1975.

[15] W.H. Fleming and H.M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.

[16] H. Ishii, *A boundary value problem of the Dirichlet type for Hamilton–Jacobi equations*, Ann. Scuola Norm. Pisa Cl. Sci. (4), 16 (1989), pp. 105–135.

[17] H. Ishii and S. Koike, *A new formulation of state constraints problems for first order PDE's*, SIAM J. Control Optim., 36 (1996), pp. 554–571.

[18] R. Jensen and P.L. Lions, *Some asymptotic problems in fully nonlinear elliptic equations and stochastic control*, Ann. Scuola Norm. Pisa Cl. Sci. (4), 11 (1989), pp. 129–176.
[19] P.V. Kokotović, *Applications of singular perturbation techniques to control problems*, SIAM Rev., 26 (1984), pp. 501–550.
[20] P.V. Kokotović, H.K. Khalil, and J. O'Reilly, *Singular Perturbation Methods in Control: Analysis and Design*, Academic Press, London 1986.
[21] P.L. Lions, *Equations de Hamilton–Jacobi et solutions de viscosité*, in Ennio De Giorgi Colloquium, Paris, 1983, Res. Notes Math. 125, Pitman, Boston, 1985, pp. 83–97.
[22] P.L. Lions, *Generalized Solutions of Hamilton–Jacobi Equations*, Pitman, Boston, 1982.
[23] M. Quincampoix and H. Zhang, *Singular perturbations in non-linear optimal control systems*, Differential Integral Equations, 8 (1995), pp. 931–944.
[24] H.M. Soner, *Optimal control problems with state space constraint* I, SIAM J. Control Optim., 24 (1986), pp. 551–561.
[25] H.M. Soner, *Singular perturbations in manufacturing*, SIAM J. Control Optim., 31 (1993), pp. 132–146.
[26] A.I. Subbotin, *Generalized Solutions of First Order PDEs: The Dynamic Optimization Perspective*, Birkhäuser, Boston, 1995.
[27] N.N. Subbotina, *Asymptotic properties of minimax solutions of Isaacs–Bellman equations in differential games with fast and slow motions*, J. Appl. Math. Mech., 60 (1996), pp. 883–890.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 6, pp. 2061–2081, November 1998

012

DISSIPATIVITY AND THE LUR'E PROBLEM FOR PARABOLIC BOUNDARY CONTROL SYSTEMS*

L. PANDOLFI†

Abstract. Dissipativity is studied for a class of boundary control systems whose free evolution is described by a holomorphic semigroup. It is proved that the system is dissipative if and only if there exists a (bounded selfadjoint) solution of the corresponding linear operator inequality. In addition, the Lur'e problem is solved under particular assumptions.

Key words. singular regulator, distributed control, Popov function, holomorphic semigroups, Lur'e problem

AMS subject classifications. 43N10, 93C20

PII. S0363012997315439

1. Introduction.

The quadratic regulator problem with stability is a complex body of results; see [24]. Extensions to distributed systems were given in [25] for uniformly continuous semigroup systems and in [14] for C_0-semigroup systems. A partial extension to a class of boundary control systems is in [13].

In this paper we present a further extension to a class of boundary control problems, a result on the spectral factorization of the related Popov function, and a proof of existence of solutions to the corresponding Lur'e problem.

Now we describe the problems that we are going to study. We denote X and U to be complex Hilbert spaces, and we consider a continuous quadratic form (QF) on $X \times U$,

$$F(x,u) = \langle x, Qx\rangle + \langle x, Su\rangle + \langle Su, x\rangle + \langle u, Ru\rangle, \tag{1}$$

where "$\langle\cdot,\cdot\rangle$" denotes both the inner product in the Hilbert space X and that in the Hilbert space U. We assume that Q, S, R are linear bounded operators in the proper spaces and that $Q = Q^*$ and $R = R^*$, but we are not assuming either that $Q \geq 0$ or that $R \geq 0$ (this last condition will be a consequence of Property P1 described below).

We consider the boundary control system (S),

$$\dot{x} = A(x - Du), \qquad x(0) = x_0 \tag{2}$$

($x \in X$, $u \in U$). Systems of the form (2) have been studied in many papers in order to describe boundary control systems.

The pair of the quadratic functional (QF) and of the system (S) is called a *Popov pair* ((S),(QF)).

We must give a meaning to the solutions of the equation that describes system (S). Formally, it is given by

$$x(t) = x(t; x_0, u) = e^{At}x_0 - A\int_0^t e^{A(t-s)}Du(s)\,\mathrm{d}\,s\,. \tag{3}$$

*Received by the editors January 24, 1997; accepted for publication (in revised form) January 20, 1998; published electronically August 31, 1998. This research was partially supported by the Italian Ministero della Ricerca Scientifica e Tecnologica within the program of GNAFA–CNR and by NATO CRG program SA.5–2–05 (CRG940161).

http://www.siam.org/journals/sicon/36-6/31543.html

†Politecnico di Torino, Dipartimento di Matematica, Corso Duca degli Abruzzi, 24, 10129 Torino, Italy (lucipan@polito.it).

Of course, this equation does not make any sense in X unless we introduce some restrictions. Two main classes under which this expression defines an element of the space X are singled out in [9]. In this paper we study systems of the first class: holomorphic semigroup systems. More precisely, we assume the following.

(i) Let α be any number larger than the exponential order $\omega(A)$ of the semigroup. We define $X_\gamma = \mathrm{dom}(\alpha I - A)^\gamma$, a space which is independent of α as long as $\alpha > \omega(A)$. Analogously, $X_\gamma^{(*)} = \mathrm{dom}(\alpha I - A^*)^\gamma$. We assume that there exists a number $\tilde{\gamma} \in (0,1)$ such that $\mathrm{Im}D \subseteq X_{1-\tilde{\gamma}}$. Hence $B = -AD = (\alpha I - A)^{\tilde{\gamma}}(\alpha I - A)^{1-\tilde{\gamma}}D - \alpha D$ belongs to $\mathcal{L}(U, (X_\gamma^{(*)})')$.

We note that here and in the following we use the same symbol A in order to denote the original operator A and its extension $A^{*'} \in \mathcal{L}(X, (\mathrm{dom}A^*)')$.

(ii) The pair (A, D) is stabilizable by a bounded feedback operator. This means that there exists an operator $K \in \mathcal{L}(X, U)$ with the following property: we consider the operator A_K defined by

$$A_K x = A(I - DK)x, \qquad \mathrm{dom}A_K = \{x \in X, (I - DK)x \in \mathrm{dom}A\}\,.$$

We require that the operator A_K generate an exponentially stable semigroup (which turns out to be a holomorphic semigroup; see [10, part 2, section 4].

Under these assumptions, it is known that the function $x(\cdot)$ is almost everywhere (a.e.) X-valued and locally square integrable on $(0, +\infty)$, but in general it is not continuous; see [10] and references therein.

In this paper we call "trajectory" of the control system (S) a pair $(x(\cdot), u(\cdot))$, which satisfies (3) for a given $x_0 \in X$. The function $x(\cdot)$ given in (3) will be called the "state function," which corresponds to the "initial datum" x_0 and to the "input" or "control function" $u(\cdot)$.

We consider the following quadratic functional computed along the trajectories of the control systems (S):

$$J(x_0; u) = \int_0^{+\infty} F(x(t), u(t))\, \mathrm{d}\, t\,, \tag{4}$$

where $x(t) = x(t; x_0, u)$.

The first problem that we consider is the "regulator problem with stability." This means that we want to give equivalent conditions for the following property.

Property P1. There exists a number $\eta \in \mathbb{R}$ such that for each $x_0 \in X$ we have

$$\inf\{J(x_0; u)\,,\ u(\cdot) \in L^2(0, +\infty; U) \text{ such that } x(\cdot) \in L^2(0, +\infty; X)\} \geq \eta \|x_0\|^2\,.$$

Here $x(\cdot) = x(\cdot; x_0, u)$.

We stress that we are not assuming simply integrability of $F(x(\cdot), u(\cdot))$, but we require that possible trajectories which are not square integrable be discarded from consideration. Hence we consider only the controls $u(\cdot)$ which belong to the set

$$\mathcal{M}_{x_0} = \{u(\cdot) \in L^2(0, +\infty; U) \text{ such that } x(\cdot) \in L^2(0, +\infty; X)\}\,.$$

The stabilizability assumption guarantees that for each x_0 the set $\mathcal{M}_{x_0}$ is nonempty but, in general, $\mathcal{M}_{x_0}$ will depend upon x_0. If it happens that the semigroup e^{At} is exponentially stable, then $\mathcal{M}_{x_0} = L^2(0, +\infty; U)$.

When Property P1 is satisfied, the pair of system (S) and of the quadratic functional (QF) is called *dissipative* (or, in special applications, *passive* or *positive.*)

We already noted that the previous problem was studied and solved in [14] in the case in which $\mathrm{Im}D \subseteq \mathrm{dom}A$ (distributed control action). The extensions of the results in [14] to boundary control systems may follow two routes, the first one being a direct extension of the chain of lemmas in [14] to our system. Instead, following an idea which we introduced already in the previous papers [15, 16], we prefer a different route. We associate an "augmented system" to system (2). This is a distributed control system. Then we can directly apply the results in [14] to the augmented system in order to obtain the corresponding results for the boundary control system.

The methods used for the analysis of Property P1 were already used in the papers [11, 12, 17, 18].

A second problem that we are going to study is the Lur'e problem. A precise description, with assumptions and references, will be given in section 2, after the description of the main results which concern Property P1. Roughly speaking, the Lur'e problem is the following problem: we consider the case that the semigroup is exponentially stable. A necessary condition which is satisfied when Property P1 holds is that a certain operator valued function defined on the imaginary axis is nonnegative; see the last statement of Theorem 1. This function is called the Popov function of the system. This condition is necessary *but not sufficient* for Property P1, and it is required that we find *additional conditions* under which Property P1 holds. This problem is known as the Lur'e problem. It is an important and classical problem in control theory which has not yet been completely clarified even for finite dimensional systems. See [2, 5] for recent finite dimensional results.

The plan of the paper is as follows: in the next section we present our main results; the augmented system is derived in section 3 and it is studied in section 3.1, where Theorem 1 is proved under the further assumption that the semigroup e^{At} is exponentially stable. The stability assumption is removed in section 4.

If the system is stable, the Popov function (defined in the next section) has an analytic extension (with singularities) to the right half plane, and it makes sense to study the factorization properties of this function; see section 5 where the Lur'e problem is solved.

2. Main results and description of the Lur'e problem. If P_1 and P_2 are continuous selfadjoint operators on X, we write $P_1 \geq P_2$ if $P_1 - P_2 \geq 0$.

The first result that we are going to prove is the following one.

THEOREM 1. *We have that Property P*1 *holds if and only if there exists a linear bounded operator P on X, $P = P^*$, such that the following holds for each $x \in \mathrm{dom}A$, $u \in U$:*

$$\langle Ax, P(x + Du)\rangle + \langle P(x + Du), Ax\rangle + F(x + Du, u) \geq 0. \tag{5}$$

In this case, $\inf_{u \in \mathcal{M}_{x_0}} J(x_0; u) = \langle x_0, \hat{P}x_0\rangle$, *where the operator $\hat{P}$ is the maximal solution to the inequality* (5).

*If Property P*1 *holds, then*

$$\begin{cases} F(x + Du, u) \geq 0 \text{ for each } \ x \in \mathrm{dom}A\,, u \in U \text{ such that} \\ \\ \text{there exists a real } \omega \text{ for which } i\omega(x + Du) = Ax\,, \end{cases} \tag{6}$$

and $R \geq 0$.

Inequality (5) is known as the *dissipation inequality* or *linear operator inequality* (LOI).

We note that if it happens that $i\omega$ belongs to the resolvent set of the operator A, then the last condition can be written in the form

$$\Pi(i\omega) = F(-i\omega(i\omega I - A)^{-1}Du + Du, u) = F((i\omega I - A)^{-1}Bu, u) \geq 0\,. \tag{7}$$

The function $\omega \to F((i\omega I - A)^{-1}Bu, u)$ is called the *Popov function* of system (S).

A case in which the Popov function is defined for each real ω is the case that the semigroup e^{At} is exponentially stable. In this case $(i\omega I - A)^{-1}Bu = (i\omega I - A)^{-1}A^{\tilde{\gamma}}A^{1-\tilde{\gamma}}Du \to 0$ for $\omega \to \infty$ (we use the fact that the semigroup is holomorphic here). Hence, if Property P1 holds and if the semigroup is exponentially stable, then

$$0 \leq \lim_{\omega\to\infty} F((i\omega I - A)^{-1}Bu, u) = \langle u, Ru\rangle;$$

i.e., R is positive semidefinite.

We shall prove that the condition $R \geq 0$ is implied by Property P1 even if the semigroup is not exponentially stable.

Remark 1. The assumption that the semigroup is *holomorphic* has a crucial role in the proof of the positivity of R. In particular, for hyperbolic control systems, the condition $\Pi(i\omega) \geq 0$ for every real ω *does not imply* $R \geq 0$. See [20] for a counterexample.

Now we can describe the Lur'e problem and the results that we are going to prove concerning this problem.

The Lur'e problem is an important and difficult problem, which can be described as follows: we assume that the semigroup is *exponentially stable* so that the Popov function is defined on the imaginary axis. Let us assume that the Popov function $\Pi(i\omega)$ is nonnegative, $\langle u, \Pi(i\omega)u\rangle \geq 0$ for each u. The problem is to find additional conditions under which there exists a solution P to (LOI). The oldest finite dimensional results are due to [8, 23]. In those papers the additional condition was *complete controllability.* Recent papers show that a much weaker condition than controllability is sufficient; see [2, 5]. See [4, 19] for recent proofs which hold for distributed systems under the assumption that e^{At} is a C_0-*group,* still under a controllability assumption. These proofs essentially reproduce the proof given in [25], which was based on the assumption that e^{At} is a *uniformly continuous group.* Instead, the arguments in the present paper are closer to the paper [1], where the Lur'e problem is studied for *semigroup* systems with *distributed* control under quite demanding regularity assumptions on the input and output operators. In fact, we study a problem analogous to the one studied in [1] under the additional condition that the semigroup is holomorphic but with *control acting through the boundary* and less stringent assumptions on the output operator.

The Lur'e problem is difficult even for finite dimensional systems, and we are able to attack it only under particular conditions.

We describe the assumptions under which we study the Lur'e problem.

1. We assume that the input u is *scalar* so that $S \in X$ and $D \in X$ and the Popov function is a scalar function.
2. We assume that the semigroup is holomorphic and *exponentially stable*, and moreover, we assume $\sigma(A) = \sigma_p(A) = \{z_n\}$; each z_n is a simple eigenvalue and the sequence $\{v_n\}$ of the normalized eigenvectors is a Riesz basis of X. The eigenvalues of A are denoted $z_n = -x_n + iy_n$ (it will be assumed that $x_n > 0$ in section 5).

3. There exists a number μ such that

$$\left|\frac{y_k}{x_k}\right| < \mu\,.$$

4. There exists an exponent $\delta < 1$ and a *positive* number H such that, for $|\omega|$ large enough,

$$|\omega|^{\delta}|\Pi(i\omega)| > H\,. \tag{8}$$

Now we state our result and then we discuss briefly the previous assumptions.

The result that we prove is the following.

THEOREM 2. *Let assumptions* 1–4 *hold. If* $\Pi(i\omega) \geq 0$ *for every real* ω*, then there exists a solution* $P = P^* \in \mathcal{L}(X)$ *of the (LOI)* (5).

The proof of this result rests on the same ideas as [1]; see section 5.

Now we discuss the assumptions. The assumption that the spectrum of the generator is point spectrum with simple eigenvalues and that the eigenvectors are a Riesz basis is satisfied in many cases of practical interest.

As to condition 3, we note that the spectrum of an analytic semigroup is contained in a sector. In principle it could be a sector of amplitude π. We are assuming that the amplitude is less then π. This assumption is not too restrictive. In particular, in the very important case that the operator A is selfadjoint, the eigenvalues are real (and in this case the proof of Theorem 2 can be simplified).

If we compare the previous assumptions with the existing finite dimensional theory, we see that the weakest condition under which the (finite dimensional) Lur'e problem was solved is that there exists at least one point ω_0 such that $P(i\omega_0) \neq 0$ (if the control is scalar). We expect that this condition *is not* sufficient in the case of distributed systems since for finite dimensional systems it implies that there are only finitely many zeros of $P(i\omega)$ and a polynomial rate of decrease at ∞. Both these conditions are not true for distributed systems.

This observation explains assumption 4, which is the most demanding (also from the point of view of practical verification). We note that this assumption is extremely strong for systems with *distributed control action:* the simple Popov function is $\Re e\, 1/[1+i\omega] = \mathrm{O}(1/\omega^2)$ for $\omega \to +\infty$. In fact, assumption 4 can be relaxed if the operator S is more regular. We note also that assumption 4 is not used if the system has only finitely many eigenvalues. Hence, our proof *can* be applied to finite dimensional systems, even if assumption 4 does not hold when $R = 0$.

Finally, we show an example of a system which satisfies assumptions 1–4. We consider the heat equation in one space dimension with scalar control and observation

$$x_t = x_{ss}, \quad t > 0\,, \quad 0 < s < 1\,; \qquad x(t,0) = u(t)\,, \quad x_s(t,1) = 0\,.$$

It is well known that assumptions 2 and 3 are satisfied.

Let $Q = 0$, $R = 0$, and

$$S^*(x(\cdot)) = \int_0^1 x(s)\,\mathrm{d}\,s,$$

and let us consider assumption 4. A simple calculation shows that the transfer function is

$$S^*A(i\omega I - A)^{-1}D = -\frac{1}{\sigma}\frac{\sinh\sigma}{\cosh\sigma}, \quad \sigma = e^{i\pi/4}\sqrt{\omega}$$

(when $\omega > 0$; otherwise $\sigma = e^{-i\pi/4}\sqrt{\omega}$). It is a long but elementary computation to see that there exist numbers $0 < m < M$ such that

$$\frac{m}{\sqrt{\omega}} \leq \Pi(i\omega) \leq \frac{M}{\sqrt{\omega}}\,.$$

Hence, assumption 4 is satisfied, too.

3. Preliminaries: Stable semigroup systems. In this section we assume that the operator A generates an exponentially stable semigroup, and we show that it is possible to introduce an *augmented* Popov pair, whose analysis directly provides the proof of Theorem 1. The special feature of this new Popov pair is that it is described by a *distributed* (i.e., not boundary) control system.

A preliminary observation is the following one:

$$\inf_{u\in L^2(0,+\infty;U)} J(x_0;u) \leq J(x_0,0) \leq M\|x_0\|^2$$

for some number M. We note that $\inf_u J(x_0;u)$ is computed with respect to any $u \in L^2(0,+\infty;U)$ since we are assuming that system (S) is exponentially stable.

The state function of an exponentially stable system depends continuously, in the $L^2(0,+\infty;X)$ norm, on the input function $u(\cdot) \in L^2(0,+\infty;U)$ and also on the initial datum $x_0 \in X$ (see [10]). This observation is used in the following proof.

LEMMA 3. *Let e^{At} be an exponentially stable semigroup. We have the following.*

1. *Let H be a dense subspace of X and let us assume that*

$$\inf_{u\in L^2(0,+\infty;U)} J(x_0;u) \geq \eta\|x_0\|^2 \tag{9}$$

for each initial condition $x_0 \in H$. Then inequality (9) *holds for each initial condition $x_0 \in X$.*

2. *Let $\mathcal{H}$ be a dense subspace of $L^2(0,+\infty;U)$, and let us define the two numbers*

$$\alpha(x_0) = \inf_{u\in L^2(0,+\infty;U)} J(x_0;u), \tag{10}$$

$$\beta(x_0) = \inf_{u\in\mathcal{H}} J(x_0;u)\,. \tag{11}$$

Then $\alpha(x_0) = \beta(x_0)$.

Proof. First, we prove item 1. By contradiction, let us assume that we can find $\tilde{x} \in X$ and a control $\tilde{u}(\cdot) \in L^2(0,+\infty;U)$ such that $J(\tilde{x},\tilde{u}) < (\eta-\epsilon)\|\tilde{x}\|^2$ for some $\epsilon > 0$.

Let us choose a sequence $\{x_n\}$ $x_n \to \tilde{x}$, $x_n \in H$. Then, $x(\cdot;x_n,\tilde{u}) \to x(\cdot;\tilde{x},\tilde{u})$ in $L^2(0,+\infty;X)$ so that we have $J(x_n;\tilde{u}) < (\eta-\epsilon)\|\tilde{x}\|^2$ if n is large enough. As $(\eta-\epsilon)\|x_n\|^2 \to (\eta-\epsilon)\|\tilde{x}\|^2$ we see that $(\eta-\epsilon)\|\tilde{x}\|^2 < (\eta-\epsilon/2)\|x_n\|^2$ for large n. This contradicts condition (9), which is assumed on H.

Next, we prove item 2. It is clear that $\alpha(x_0) \leq \beta(x_0)$. By contradiction, let us assume that $\alpha(x_0) < \beta(x_0) - \epsilon$ so that we can find $u_0 \in L^2(0,+\infty;U)$ such that $\alpha(x_0) < J(x_0;u_0) < \beta(x_0) - \epsilon$. We choose a sequence $\{u_n\}$ in $\mathcal{H}$ which converges to u_0 in the norm of $L^2(0,+\infty;U)$. Stability of e^{At} implies that $J(x_0;u_n) \to J(x_0;u)$ so that, for large n, $J(x_0;u_n) < \beta(x_0) - \epsilon$. This is a contradiction. □

This lemma shows that we can confine ourselves to study the boundedness from below of the quadratic functional $J(x_0;u)$ by using regular controls, in particular by using C^1 controls.

If the control $u(\cdot)$ is of class C^1, then we have

$$x(t) = x(t; x_0, u) = e^{At}\{x_0 - Du(0)\} + Du(t) - \int_0^t e^{A(t-s)} D\dot{u}(s)\, \mathrm{d}\, s,$$

i.e.,

$$x(t) = e^{At}\{x_0 - Du(0)\} + Du(t) - \int_0^t e^{A(t-s)} Dv(s)\, \mathrm{d}\, s, \tag{12}$$

$$u(t) = u(0) + \int_0^t v(s)\, \mathrm{d}\, s\,. \tag{13}$$

This observation is the essence of the method that we shall use: the functions $\xi(\cdot)$, $u(\cdot)$, $v(\cdot)$ in (12), (13) represent the trajectories of the *augmented system* (AS),

$$\left\{ \begin{array}{lcllcl} \dot{\xi} & = & A\xi - Dv, & \xi(0) & = & x_0 - Du_0, \\ \dot{u} & = & v, & u(0) & = & u_0\,. \end{array} \right. \tag{14}$$

We shall see that it is possible to study a suitable quadratic functional along the trajectories of the augmented system in order to get informations on the original problem.

We note that the block operators which describe the augmented system are the operators $\mathcal{A}_0$, $\mathcal{D}$:

$$\mathcal{A}_0 = \left[\begin{array}{cc} A & 0 \\ 0 & 0 \end{array} \right], \qquad \mathcal{D} = \left[\begin{array}{c} -D \\ I \end{array} \right].$$

Remark 2. Also, the subspace $\{u(\cdot) \in L^2(0, +\infty; U) \cap C^1(0, +\infty; U)\,,\ u(0) = 0\}$ is dense in $L^2(0, +\infty; U)$ so that we could even consider $u_0 = 0$ from the beginning. It seems more interesting to show that the vector $u(0) = u_0$ has no influence on the results, even if it is nonzero (see the next section). Moreover, the applications of the results in [14] are more direct if we do not assume $u(0) = 0$.

3.1. The quadratic regulator problem with stability. The previous arguments suggest the following considerations. We define the quadratic functional (QFS)

$$\Phi(\xi, u, v) = F(\xi + Du, u)\,. \tag{15}$$

We note that this functional does not depend on v. We introduce the quadratic cost

$$J(\xi_0, u_0; v) = \int_0^{+\infty} \Phi(\xi(t), u(t), v(t))\, \mathrm{d}\, t$$

computed along the trajectories of the augmented system (AS) described by (12), (13). This cost *does not involve the input v explicitly.* If $u(\cdot)$ is a C^1 input function, then

$$J(x_0 - Du(0), u(0); v)_{|_{v=\dot{u}}} = J(x_0; u)\,.$$

Hence, Lemma 3 shows that Property P1 holds if and only if for every x_0, for every $u_0 \in U$, and for every $v(\cdot) \in L^2(0, +\infty; U)$ we have

$$J(x_0 - Du_0, u_0; v) \geq \eta \|x_0\|^2 = \eta \|(x_0 - Du_0) + Du_0\|^2\,,$$

provided that the corresponding *trajectory* $(\xi(\cdot), u(\cdot), v(\cdot))$ of the augmented system (AS) is square integrable. Hence, we shall denote by $\mathcal{M}_{\xi_0,u_0}$ the set

$$\mathcal{M}_{\xi_0,u_0} = \{v(\cdot) \in L^2(0,+\infty;U) \text{ which produces a square integrable trajectory of system (AS) whose initial datum is } (\xi_0,u_0)\},$$

and we consider the following property.

Property P2. There exists a real number η such that

$$\inf_{v\in\mathcal{M}_{\xi_0,u_0}} J(\xi_0,u_0;v) > \eta\|\xi_0 + Du_0\|^2 .$$

The previous considerations show that Property P1 and Property P2 are equivalent.

Now we observe that the augmented system (AS) *is not exponentially stable but is stabilizable.* In fact, $v=-u$ is a stabilizing feedback, because e^{At} is exponentially stable.

A result from [14] shows that for stabilizable systems with distributed controls we can study, instead of Property P2, the following apparently weaker condition.

Property P3. For each (ξ_0,u_0) we have $\inf_{v\in\mathcal{M}_{\xi_0,u_0}} J(\xi_0,u_0;v) > -\infty$.

Property P3 is characterized in Theorem 3 in [14] as follows. Property P3 is equivalent to the existence of an operator $W = W^* \in \mathcal{L}(X\times U)$ such that

$$\begin{aligned} &\langle \mathcal{A}_0\mathcal{X} + \mathcal{D}v, W\mathcal{X}\rangle + \langle W\mathcal{X}, \mathcal{A}_0\mathcal{X} + \mathcal{D}v\rangle + \Phi(\xi,u,v) \geq 0 \\ &\forall \mathcal{X} = \operatorname{col}[\xi,u] \in \operatorname{dom}\mathcal{A}_0,\ \forall v \in U . \end{aligned} \tag{16}$$

The operators $\mathcal{A}_0$, $\mathcal{D}$ are the operators of the augmented system (AS).

Inequality (16) is the inequality (LOI) written for the augmented system. We stress that Property P3 and (16) are equivalent conditions provided that the pair $(\mathcal{A},\mathcal{D})$ is stabilizable, as in our case.

From [14], Property P3 implies that

$$\inf_{v\in\mathcal{M}_{\xi_0,u_0}} J(\xi_0,u_0;v) = \left\langle \begin{bmatrix} \xi_0 \\ u_0 \end{bmatrix} \hat{W}, \begin{bmatrix} \xi_0 \\ u_0 \end{bmatrix} \right\rangle,$$

where $\hat{W}$ is the *maximal* solution to (LOI). Hence there exists $M \leq \inf\sigma(\hat{W})$ such that

$$\inf_{v\in\mathcal{M}_{\xi_0,u_0}} J(\xi_0,u_0;v) > M\left\| \begin{bmatrix} \xi_0 \\ u_0 \end{bmatrix} \right\|^2 .$$

This proves the equivalence of Property P2 and Property P3. Furthermore, if Property P3 holds, then for every $(\omega,\xi,u,v) \in \mathbb{R}\times X\times U\times U$ such that $(i\omega I - \mathcal{A})\mathcal{X} = \mathcal{D}v$ ($\mathcal{X}$ =col$[\xi,u]$) we have that $\Phi(\xi,u,v)\geq 0$ (the frequency domain inequality).

These results are proved in [14, Theorem 3]. Now we shall elaborate on these properties in order to obtain conditions for Property P1. A crucial observation is that the left-hand side of (16) does not contain a quadratic term of v. Consequently inequality (16) implies that the coefficient of v is zero, i.e., that $W\mathcal{D}=0$. In other terms, this shows that the operator W solves the Lur'e problem for the augmented system, with the functional $\Phi(\xi,u,v)$ (constant with respect to v).

From $W\mathcal{D}=0$ we see that any operator W which satisfies (16) has the following block form:

$$W = \begin{bmatrix} P & PD \\ D^*P & D^*PD \end{bmatrix}, \qquad P = P^* \in \mathcal{L}(X). \tag{17}$$

We shall write $\hat{P}$ for the $(1,1)$ block of the maximal solution $\hat{W}$. In fact, if $W_{i\,j}$, $1 \le i,\, j \le 2$, are the blocks of W, then the equation $W\mathcal{D} = 0$ shows that $-W_{1\,1}D + W_{1\,2} = 0$, $-W_{2\,1}D + W_{2\,2} = 0$, and $W_{2\,1} = W_{1\,2}^*$, since W is selfadjoint.

Now we prove the following lemma.

LEMMA 4. *Let us define*

$$\alpha = \inf_{u \in L^2(0,+\infty;U)} J(x_0;u) \qquad \gamma = \inf_{u_0} \inf_{v \in \mathcal{M}_{x_0+Du_0,u_0}} J(x_0 + Du_0, u_0; v)\,.$$

Then $\alpha = \gamma$.

Proof. The inequality $\alpha \le \gamma$ is clear. We assume that $\alpha + \epsilon < \gamma - \epsilon$ $(\epsilon > 0)$ and we show a contradiction.

We know that

$$\gamma - \epsilon < \gamma \le J(x_0 + Du_0, u_0; v) \tag{18}$$

for each $u_0 \in U$ and $v(\cdot) \in \mathcal{M}_{x_0+Du_0,u_0}$.

We know that (AS) is stabilizable. Let $\mathcal{K}$ be a stabilizing feedback and denote by $(AS)_{\mathcal{K}}$ the closed loop. Let $J_{\mathcal{K}}(\xi,u;v)$ be the cost obtained from the quadratic functional $\Phi_{\mathcal{K}}(\xi,u,w) = \Phi(\xi,u,\mathcal{K}X + w) = \Phi(\xi,u)$ (since $v = \mathcal{K}X + w$ does not appear explicitly in the quadratic functional Φ), $X = \operatorname{col}[\xi,u]$, i.e., the original cost computed along the trajectories of system $(AS)_{\mathcal{K}}$. Then,

$$\gamma = \inf_{u_0} \inf_{w(\cdot) \in L^2(0,+\infty;U)} J_{\mathcal{K}}(x_0 + Du_0, u_0; w),$$

i.e.,

$$\gamma - \epsilon < \gamma \le J_{\mathcal{K}}(x_0 + Du_0, u_0; w),$$

for each $u_0 \in U$ and $w(\cdot) \in L^2(0,+\infty;U)$ since we noted that

$$\{J(x_0 + Du_0; u_0; v)\ v \in \mathcal{M}_{x_0+Du_0,u_0}\} = \{J_{\mathcal{K}}(x_0 + Du_0; u_0; w)\ w \in L^2(0,+\infty;U)\}\,.$$

Stability of $(AS)_{\mathcal{K}}$ implies that the transformation

$$(x_0,u_0,w) \to J_{\mathcal{K}}(x_0 + Du_0, u_0; w), \qquad X \times U \times L^2(0,+\infty;U) \to \mathbb{R}$$

is continuous. Hence we can find $\sigma > 0$ such that

$$\gamma - \epsilon < J_{\mathcal{K}}(\tilde{x} + D\tilde{u}, \tilde{u}; \tilde{w}) \tag{19}$$

provided that $\|\tilde{x} - x_0\|_X < \sigma$ and also $\|\tilde{u} - u_0\|_U < \sigma$, $\|\tilde{w}(\cdot) - w(\cdot)\|_{L^2} < \sigma$. The elements u_0 and $w(\cdot)$ are arbitrary in U and in $L^2(0,+\infty;U)$. Hence, (19) holds provided that $\|\tilde{x} - x_0\|_X < \sigma$ and each $\tilde{u} \in U$, $\tilde{w}(\cdot) \in L^2(0,+\infty;U)$.

Now let $\hat{u}$ be such that

$$\alpha \le J(x_0;\hat{u}) < \alpha + \epsilon\,.$$

We can assume that $\hat{u}$ belongs to $C^1(0,+\infty;U) \cap L^2(0,+\infty;U)$, and we can also assume that $\|u(0) - u_0\|_U \le \sigma$, since $w = \dot{u}$ does not appear explicitly in $J_K(\xi_0,u_0;w)$. Hence,

$$\alpha \le J(x_0 + D\hat{u}(0), \hat{u}(0); \hat{u}'(\cdot)) < \alpha + \epsilon < \gamma - \epsilon\,.$$

This contradicts (19). ∎

The previous result implies that

$$\inf_{u\in\mathcal{M}_{x_0}} J(x_0;u) = \inf_{u_0} \left\langle \begin{bmatrix} x_0 - Du_0 \\ u_0 \end{bmatrix}, \begin{bmatrix} \hat{P} & \hat{P}D \\ D^*\hat{P} & D^*\hat{P}D \end{bmatrix} \begin{bmatrix} x_0 - Du_0 \\ u_0 \end{bmatrix} \right\rangle = \langle x_0, \hat{P}x_0\rangle,$$

and we have the following.

THEOREM 5. *A necessary and sufficient condition for Property P*1 *is the existence of an operator* $\hat{P} = \hat{P}^* \in \mathcal{L}(X)$ *such that*

$$\inf_{u(\cdot)\in L^2(0,+\infty;U)} J(x_0;u) = \langle x_0, \hat{P}x_0\rangle\,.$$

We show that $\hat{P}$ solves a suitable dissipation inequality.

Let W of the form (17) be a solution of (16). Its $(1,1)$ block P solves

$$\langle A\xi, P(\xi + Du)\rangle + \langle P(\xi + Du), A\xi\rangle + F(\xi + Du, u) \geq 0 \tag{20}$$

for every $\xi \in \mathrm{dom}A$, $u \in U$, as wanted.

The previous observation applies in particular to the maximal solution $\hat{W}$, hence to $\hat{P}$. Now we show that, if a solution P to (20) exists, this implies Property P1. This is easy since, if P satisfies (20), then we construct a solution W to (16) (see (17)), and this implies Property P3 from [14]. Hence, Property P1 holds.

This proves the first part of Theorem 1 under the further assumption that the semigroup e^{At} is exponentially stable. Now we consider the frequency domain condition. We note that the equality $(i\omega I - \mathcal{A})\mathcal{X} = \mathcal{D}v$ can be written as

$$(i\omega I - A)\xi = -Dv, \qquad i\omega u = v \tag{21}$$

so that, if $i\omega(\xi + Du) = A\xi$, then we have the required inequality

$$F(\xi + Du, u) \geq 0\,. \tag{22}$$

Equivalently, we proved that $F(x,u) \geq 0$ when $x - Du \in \mathrm{dom}A$ (i.e., x satisfies the prescribed boundary conditions) and $i\omega x = A(x - Du)$ for some real ω.

Remark 3. The crucial assumption in this section is that e^{At} is an exponentially stable semigroup. The assumption that the semigroup is holomorphic *is not used* in the previous arguments. It is used, instead, in the proof that $R \geq 0$; see section 2. In fact, Property P1 does not imply $R \geq 0$ in general; see [20] for a counterexample.

In the next section we examine the case that system (S) is stabilizable, while in section 5, we return to stable systems in order to study the Lur'e problem. For this reason we consider now some additional material for stable systems, which will be used in section 5.

We noted already that if system (S) is stable, then the augmented system can be stabilized by using the simple feedback $v = -u$. We apply this special feedback and we get a special stabilized augmented system (SAS). The operators of system (SAS) are

$$\mathcal{A}_S = \begin{bmatrix} A & D \\ 0 & -I \end{bmatrix}, \qquad \mathcal{D} = \begin{bmatrix} -D \\ I \end{bmatrix}.$$

In fact, the input matrix is not affected by the feedbacks. Also, the quadratic functional is not affected since it does not depend upon v.

We compare the Popov function $\Pi(i\omega)$ of the original Popov pair ((S),(QF)) with the Popov function $P(i\omega)$ of the Popov pair of the stabilized augmented system ((SAS),(QFS)).

The Popov function of ((S),(QF)) is the function $\omega \to \Pi(\omega)$ in (7).

We note that

$$\begin{aligned}\Phi(\xi,u,v) &= \langle \xi + Du, Q[\xi + Du]\rangle + \langle \xi + Du, Su\rangle + \langle Su, \xi + Du\rangle + \langle u, Ru\rangle \\ &= \left\langle \begin{bmatrix} \xi \\ u \end{bmatrix}, \begin{bmatrix} Q & S+QD \\ D^*Q + S^* & R + D^*QD \end{bmatrix} \begin{bmatrix} \xi \\ u \end{bmatrix} \right\rangle = \left\langle \begin{bmatrix} \xi \\ u \end{bmatrix}, \mathcal{Q} \begin{bmatrix} \xi \\ u \end{bmatrix} \right\rangle\end{aligned}$$

and

$$\begin{aligned}&\langle w, P(i\omega)w\rangle = \langle (i\omega I - \mathcal{A}_{\mathcal{S}})^{-1}\mathcal{D}w, \mathcal{Q}(i\omega I - \mathcal{A}_{\mathcal{S}})^{-1}\mathcal{D}w\rangle \\ &= -\langle \mathcal{D}^*(i\omega I + {\mathcal{A}_{\mathcal{S}}}^*)^{-1}\mathcal{Q}(i\omega I - \mathcal{A}_{\mathcal{S}})^{-1}\mathcal{D}w, w\rangle \\ &= \frac{1}{1+\omega^2} F(Dw - i\omega(i\omega I - A)^{-1}Dw, w) = \frac{1}{1+\omega^2}\langle w, \Pi(i\omega)w\rangle\end{aligned}$$

admits the holomorphic extension

$$\langle w, P(z)w\rangle = -\langle \mathcal{D}^*(zI + {\mathcal{A}_{\mathcal{S}}}^*)^{-1}\mathcal{Q}(zI - \mathcal{A}_{\mathcal{S}})^{-1}\mathcal{D}w, w\rangle. \tag{23}$$

This extension is bounded in a strip $|\Re e\, z| < \epsilon$, since $\mathcal{A}_{\mathcal{S}}$ generates an exponentially stable semigroup.

4. Stabilizable systems. In this section we show that the results in section 3 can be extended to stabilizable systems.

We stabilize system (S) first by applying a stabilizing feedback $u = Kx + v$. We get a new system, the *stabilized system,* described by

$$\dot{x} = (A + BK)x + Bv(t), \tag{24}$$

where $A_K = A + BK$ generates a stable holomorphic semigroup; see [10, section 4]. Moreover,

$$(zI - A_K)^{-1} = [I + (zI - A)^{-1}ADK]^{-1}(zI - A)^{-1}$$

for every z in a sector contained in $\rho(A) \cap \rho(A_K)$, $|z|$ large enough (see [10, section 4]). Hence,

$$\begin{aligned}&\|(zI - A_K)^{-1}B\| \\ &= \|[I + (zI - A)^{-1}ADK]^{-1}(zI - A)^{-1}[(\alpha I - A)^{\tilde{\gamma}}(\alpha I - A)^{1-\tilde{\gamma}}D - \alpha D]\|\end{aligned}$$

is of the order of $1/|z|^{1-\tilde{\gamma}}$. In particular, it is a bounded operator and the first resolvent equation shows that $D_K = A_K^{-1}AD \in X$. This shows that system (24) can be written as

$$\dot{x} = A_K[x - D_K v], \tag{25}$$

and we can apply the results of the previous section to this stabilized system.

We denote by (SS) the system described by (25) (the *stabilized system*).

The cost that we should associate with system (SS) is

$$J_K(x_0; v) = \int_0^{+\infty} F(x(t), u(t))\, \mathrm{d}\, t = \int_0^{+\infty} F(x(t), v(t) + Kx(t))\, \mathrm{d}\, t$$

and, of course,

$$\inf_{v\in L^2(0,+\infty;U)} J_K(x_0;v) = \inf_{u\in\mathcal{M}_{x_0}} J(x_0;u)$$

(we used already an argument like this in the proof of Lemma 4). Consequently we can check Property P1 with respect to system (SS), and system (SS) produces a *stabilizable* augmented system, the pair (A,D) being replaced by the pair (A_K,D_K). For maximum clarity let us call (ASS) the augmented system which is obtained starting from system (SS). Hence system (ASS) is the following system, where $\xi = x - D_K v$:

$$\dot\xi = A_K\xi - D_K w, \qquad \dot v = w\,.$$

We associate the following cost with system (ASS):

$$J_K(\xi_0,v_0;w) = \int_0^{+\infty} F_K(\xi,v,w)\,\mathrm{d}\,t\,, \quad F_K(\xi,v,w) = F(\xi + D_K v, v + K(\xi + D_K v))\,.$$

Moreover, the set $\mathcal{M}_K(\xi_0,v_0)$ is defined as $\mathcal{M}(\xi_0,u_0)$ but with respect to (ASS).

The same arguments as those used in the proofs of Lemmas 3 and 4 show the following.

LEMMA 6. *We have that*

$$\inf_{u\in\mathcal{M}_{x_0}} J(x_0,u) = \inf_{v\in L^2(0,+\infty;U)} J_K(x_0,v) > -\infty \qquad \forall x_0\in X$$

if and only if

$$\inf_{w\in\mathcal{M}_K(\xi_0,v_0)} J_K(\xi_0,v_0;w) > -\infty \qquad \forall(\xi_0,u_0)\in X\times U\,.$$

Lemma 6, Theorem 4, and following considerations show that Property P1 holds if and only if the dissipation inequality written with respect to system (ASS) and to the functional $J_K(\xi_0,v_0;w)$ admits a solution, i.e., from the arguments of the previous section, if and only if there exists $P = P^* \in \mathcal{L}(X)$ such that the quadratic form

$$\langle A_K\xi, P(\xi + D_K v)\rangle + \langle P(\xi + D_K v), A_K\xi\rangle + F(\xi + D_K v, v + K(\xi + D_K v)) \tag{26}$$

is nonnegative for each $\xi \in \mathrm{dom}A_K$ and for each $v\in U$.

We make a formal computation now, which is justified below.

Let the vectors ξ and v in (26) be fixed. We introduce the vectors ζ, u, x defined by

$$\xi = \zeta - D_K v, \qquad u = v + K\zeta, \qquad x = \zeta - Du\,. \tag{27}$$

In terms of ζ, v we have

$$\begin{aligned}&\langle A(I-DK)[\zeta - D_K v], P\zeta\rangle = \langle A_K\zeta - ADv, P\zeta\rangle,\\ &F(\xi + D_K v, v + K(\xi + D_K v)) = F(\zeta, v + K\zeta)\,.\end{aligned}$$

If we introduce the vectors x, u we get

$$\begin{aligned}&\langle A_K\xi, P\zeta\rangle = \langle Ax, P(x + Du)\rangle,\\ &F(\zeta, v + K\zeta) = F(x + Du, u),\end{aligned} \tag{28}$$

and this shows that the dissipation inequality (20) is equivalent to the dissipation inequality for the augmented system, i.e., that Theorem 1 holds.

The following lemma justifies the previous assertions.

LEMMA 7. *Let $\xi \in \mathrm{dom}A_K$, v be given, and x be defined as in* (27). *We have that $x \in \mathrm{dom}A$ and $A_K\xi = Ax$ so that equality* (28) *holds.*

Proof. We introduce the operator $A^{*'} \in \mathcal{L}(X, (\mathrm{dom}A^*)')$. We know that $A^{*'}$ is an extension to X of the operator A in the following sense: let i be the (continuous and dense) injection of $\mathrm{dom}A^*$ into X so that its adjoint $i' \in \mathcal{L}(X, (\mathrm{dom}A^*)')$ is the continuous and dense injection of X in $(\mathrm{dom}A^*)'$ (the operator i' is the restriction to $\mathrm{dom}A^*$ of linear continuous functionals on X; see [22, section 5.1]). Then, $x \in \mathrm{dom}A$ if and only if $A^{*'}x = i'\tilde{x} \in i'X$ and in this case $Ax = \tilde{x}$. So we prove the lemma as follows: we show that $A^{*'}x = i'A_K\xi$. This means that we show that the two functionals $A^{*'}x$ and $i'A_K\xi$ act in the same way on $\mathrm{dom}A^*$.

Let $\langle\langle\cdot,\cdot\rangle\rangle$ be the pairing of $\mathrm{dom}A^*$ and its dual. We prove that for each $y \in \mathrm{dom}A^* \subseteq \mathrm{dom}A_K^*$ we have

$$\langle\langle i'A_K\xi, y\rangle\rangle = \langle\langle A^{*'}x, y\rangle\rangle.$$

We have

$$\langle\langle i'A_K\xi, y\rangle\rangle = \langle A_K\xi, y\rangle = \langle \xi, A_K^*y\rangle, \qquad \langle\langle A^{*'}x, y\rangle\rangle = \langle x, A^*y\rangle\,.$$

We shall prove that

$$\langle D_Kv, A_K^*y\rangle = \langle A_K^{-1}ADv, A_K^*y\rangle = \langle Dv, A^*y\rangle\,. \tag{29}$$

This equality being granted, we have

$$\begin{aligned}\langle \xi, A_K^*y\rangle &= \langle \zeta - D_Kv, A_K^*y\rangle = \langle \zeta, A_K^*y\rangle - \langle D_Kv, A_K^*y\rangle\\ &= \langle (I-DK)\zeta, A^*y\rangle - \langle Dv, A^*y\rangle = \langle \zeta - D[K\zeta + v], A^*y\rangle\\ &= \langle \zeta - Du, A^*y\rangle = \langle x, A^*y\rangle\end{aligned}$$

as wanted. It remains to be proven that (29) holds if $y \in \mathrm{dom}A^*$.

We noted that the vector $A_K^{-1}ADv = (A_K^{*'})^{-1}A^{*'}Dv$ belongs to X. This means that the operator $A^{*'}Dv$, which is an element of $(\mathrm{dom}A^*)'$, is extensible to $(\mathrm{dom}A_K^*)$ and $\langle A_K^{-1}ADv, A_K^*y\rangle$ is the action of $ADv = A^{*'}Dv$ as an element of $(\mathrm{dom}A_K^*)'$ on $y \in \mathrm{dom}A^* \subseteq \mathrm{dom}A_K^*$. The value taken by $A^{*'}Dv$ on y is equal to $\langle Dv, A^*y\rangle$ as wanted. This completes the proof. ∎

The proved equality justifies (28).

Now we consider the frequency domain inequality (FDC) for (ASS), i.e.,

$$F(\xi + D_Kv, v + K(\xi + D_Kv)) \geq 0 \quad \text{if } i\omega(\xi + D_Kv) = A_K\xi\,. \tag{30}$$

More explicitly, with ζ, x, u defined in (27),

$$F(\xi + D_Kv, v + K(\xi + D_Kv)) = F(\zeta, v + K\zeta) = F(x + Du, u) \geq 0$$

if $Ax = A_K\xi = i\omega(\xi + D_Kv) = i\omega(x + Du)$.

An important consequence of the (FDC) is easily seen from (30). We fix v and we read (30) with

$$\xi = \xi(\omega) = -(i\omega I - A_K)^{-1}i\omega D_Kv\,.$$

Hence, $\lim_{|\omega|\to+\infty}\xi(\omega) = -D_K v$ so that

$$F(\xi(\omega) + D_K v,\, v + K[\xi(\omega) + D_K v]) \to F(0, v) = \langle v, Rv\rangle,$$

and we get the following theorem.

THEOREM 8. *The condition $R \geq 0$ is implied by Property P*1.

5. Scalar systems: Spectral factorization and the Lur'e problem. In this section we solve the Lur'e problem. Hence, we assume that the conditions in Theorem 2 are satisfied. The proof makes use of the same ideas as in [1] so that the first step is the construction of the spectral factorization of the positive Popov function. This part of the proof follows ideas which are completely different from those in [1]. The demanding assumption 4 is not used in this part of the proof. But, the assumption that the semigroup is holomorphic is explicitly used.

5.1. Spectral factorization of $P(i\omega)$. In this section we put ourselves in the following case, which is in particular the case of the augmented system (SAS) when the control is scalar.

(i) The control u enters as a distributed control and takes scalar values.
(ii) The semigroup generated by A is holomorphic and stable.
(iii) The Popov function is not identically zero.

We prove the following fact.

THEOREM 9. *Under the previous assumptions, and if $P(i\omega) \geq 0$ for each real ω, there exists a function $M_0(i\omega)$ which is holomorphic and bounded in $\Re e\, z > 0$ and such that*

$$P(i\omega) = M_0^*(i\omega) M_0(i\omega)\,.$$

We use a penalization method for the proof. We consider the sequence of Popov functions

$$P_n(z) = P(z) + \frac{1}{n}$$

which correspond to the "penalized" cost obtained by the addition of $\frac{1}{n}\|u\|^2$ to (QFS). We said already that the extension $P(z)$ is holomorphic (and bounded) in a strip $|\Re e\, z| < \epsilon$. In this strip $|P_n(z)|$ is bounded by $\max|P(z)| + 1$. In particular, it is uniformly bounded with respect to n.

We can exhibit explicitly a spectral factorization of the function $P_n(z)$. This function satisfies the conditions in Theorem 2 of [14] so that there exists a stabilizing solution W_n of the corresponding dissipation inequality. We use this solution in order to construct the factor

$$M_n(z) = \frac{1}{\sqrt{n}} + \sqrt{n}\mathcal{D}^* W_n [zI - \mathcal{A}_{\mathcal{S}}]^{-1}\mathcal{D}\,.$$

The function $M_n(z)$ is holomorphic and bounded in $\Re e\, z > 0$. It is easily seen, as for finite dimensional systems, that

$$P_n(i\omega) = |M_n(i\omega)|^2\,. \tag{31}$$

In fact, it happens that W_n solves the Riccati equation

$$\langle \mathcal{A}_{\mathcal{S}} X, W_n Y\rangle + \langle W_n X, \mathcal{A}_{\mathcal{S}} Y\rangle + \langle X, \mathcal{Q} Y\rangle = \langle X, W_n \mathcal{D}^* n \mathcal{D} W_n Y\rangle \qquad \forall X, Y \in \mathrm{dom}\mathcal{A}_{\mathcal{S}}$$

so that we have also

$$(i\omega I + \mathcal{A}_{\mathcal{S}}{}^*)^{-1} W_n \mathcal{D}^* n \mathcal{D} W_n (i\omega I - \mathcal{A}_{\mathcal{S}})^{-1}$$
$$= W_n (i\omega I - \mathcal{A}_{\mathcal{S}})^{-1} - (i\omega I + \mathcal{A}_{\mathcal{S}}{}^*)^{-1} W_n + (i\omega I + \mathcal{A}_{\mathcal{S}}{}^*)^{-1} \mathcal{Q} (i\omega I - \mathcal{A}_{\mathcal{S}})^{-1}.$$

Once this formula is known, a direct computation shows (31). Furthermore, $M_n^{-1}(z)$ is holomorphic and bounded in the right half plane:

$$M_n^{-1}(z) = \sqrt{n} - n^{3/2} \mathcal{D}^* W_n (zI - \mathcal{A}_{\mathcal{S}} - \sqrt{n} \mathcal{D} \mathcal{D}^* W_n)^{-1} \mathcal{D}.$$

The proof that the previous function is the inverse of $M_n(z)$ is based on the second resolvent identity, [7, p. 197, Theorem 5.103].

The boundedness of $M_n^{-1}(z)$ shows that the function $M_n(z)$ is an *outer* function.

The fact that $M_n(z)$ is holomorphic and bounded in $\Re e\, z > 0$ shows that for each z we have

$$\sup_{\Re e\, z>0} \|M_n(z)\|^2 = \sup_{\omega \in \mathbb{R}} |M_n(i\omega)|^2 = \sup_{\omega} \|P(i\omega) + 1/n\| \leq 1 + \|P(\cdot)\|_\infty . \tag{32}$$

Montel's theorem [3, p. 153] shows that $M_n(z)$ has a subsequence, still denoted $\{M_n(z)\}$, which converges to an H^∞ function $M_0(z)$ *uniformly on compact sets of* $\Re e\, z > 0$. We noted that boundedness holds in a strip around the imaginary axis, too, so that $\{M_n(z)\}$ converges to $M_0(z)$ for $\Re e\, z > -\epsilon$, $\epsilon > 0$, uniformly on compact sets.

It follows that $M_0(i\omega)$ is *continuous on the imaginary axis* and $|P(i\omega)| = |M_0(i\omega)|^2$.

We noted that the functions $M_n(z)$ are *outer* functions. We prove that $M_0(z)$ is an outer function, too.

We invoke a characterization of outer functions from [6, Theorem 4.6]. A function $L(z)$ is outer if and only if there exists a particular $z = x + iy$ with $x > 0$ such that the following equality holds:

$$\log |L(z)| = \frac{1}{\pi} \int_{-\infty}^{+\infty} \log |L(it)| \frac{x}{x^2 + (y-t)^2} \, \mathrm{d}\, t \,. \tag{33}$$

In fact, if the previous equality holds for one z it holds for every z such that $\Re e\, z > 0$.

The function $M_0(z)$ is not zero since we are assuming that the Popov function is not identically zero. Hence there exists a point x_0 of the real axis in which $M_0(x_0) \neq 0$. Also, the functions $M_n(\cdot)$ are not zero in this point since an outer function is never zero in the right half plane. Moreover, (33) holds for every $M_n(\cdot)$ and $M_n(x_0) \to M(x_0)$, so that $\log |M_n(it)| \to \log |M_0(it)|$ for each t. We prove that

$$\int_{-\infty}^{+\infty} \log |M_n(it)| \frac{x_0}{x_0^2 + t^2} \, \mathrm{d}\, t \longrightarrow \int_{-\infty}^{+\infty} \log |M_0(it)| \frac{x_0}{x_0^2 + t^2} \, \mathrm{d}\, t \,.$$

For this, we introduce the functions $\log^+ a = \max\{0, \log a\}$ and $\log^- a = \min\{0, \log a\}$. We note that $\log^+ |M_n(it)|$ is nonnegative and bounded above, since $\{M_n(it)\}$ is bounded. The *dominated convergence theorem* shows that

$$\int_{-\infty}^{+\infty} \log^+ |M_n(it)| \frac{x_0}{x_0^2 + t^2} \, \mathrm{d}\, t \to \int_{-\infty}^{+\infty} \log^+ |M_0(it)| \frac{x_0}{x_0^2 + t^2} \, \mathrm{d}\, t \,.$$

In order to show the corresponding convergence for the integrals with $\log^-$ we invoke the *monotone convergence theorem.*

By assumption, $P(it) \geq 0$ so that

$$|M_n(it)| = \left[P(it) + \frac{1}{n}\right]^{1/2} \geq \left[P(it) + \frac{1}{n+1}\right]^{1/2} = |M_{n+1}(it)|\,.$$

Hence, $\{|M_n(it)|\}$ is a *decreasing* sequence of functions. Consequently, $\{\log^- |M_n(it)|\}$ is a *decreasing sequence* of *negative* functions. Convergence of the integrals follows from monotone convergence theorem.

Consequently we proved the following.

THEOREM 10. *The function $M_0(i\omega)$ is an outer factor of $P(i\omega)$ and $P(i\omega) = |M_0(i\omega)|^2$.*

COROLLARY 11. *The factor $M_0(z)$ does not have zeros in $\Re e\, z > 0$. Moreover, the function $M_0(z)$ belongs to H^2 of the right half plane.*

Proof. The first statement follows since an outer function does not have unstable zeros. The second statement follows since when $R = 0$ we have from (23) that

$$|M_0(i\omega)|^2 = |P(i\omega)| \leq \frac{K}{|\omega|^2}$$

and $M_0(i\omega)$ is continuous; in particular it is continuous for $\omega = 0$.

This shows that $M_0(z)$ is H^∞ and square integrable on the imaginary axis; hence it is H^2. □

COROLLARY 12. *The function $M_0(z)$ is the Laplace–Fourier transformation of a function $\check{M}_0(t)$ whose support is in $t \geq 0$ and which is square integrable.*

Of course, we are not asserting that the factor $M(z)$ has an inverse in H^∞.

5.2. The Lur'e problem. In this section we prove Theorem 2. As we said, we use the same method as in [1]; hence we make use of the factorization of $\Pi(i\omega)$ found in the previous section, with the properties in Corollaries 11 and 12. Moreover, we prefer again to work with the stabilized augmented system so that we show that there exists an operator W which solves the Lur'e equation (LuE),

$$2\Re e\, \langle \mathcal{A}_{\mathcal{S}}\Xi, W\Xi\rangle + \langle \Xi, \mathcal{Q}\Xi\rangle \geq 0 \quad \forall \Xi \in \mathrm{dom}\mathcal{A}_{\mathcal{S}}\,, \qquad W\mathcal{D} = 0\,. \tag{LuE}$$

Here $\mathcal{Q}$ is the operator matrix of the functional (QFS) (see (23)) and $\mathcal{D}$ is the input operator which is now a vector in $X \times \mathbb{C}$ since we are considering scalar controls.

We prove the following lemmas first.

LEMMA 13. *Let the assumptions of Theorem 2 hold. Then there exist a number $c > 0$ and a number $\gamma < 1/2$ such that $|\bar{z}_k|^{1+\gamma}|M_0(-\bar{z}_k)| > c$.*

Proof. We assume first that $\Pi(i\omega)$ does not have $i\omega$-axis zeros. Later we remove this assumption.

We recall the assumptions that

$$\left|\frac{y_k}{x_k}\right| < \mu, \qquad \Pi(i\omega) > \frac{\mathrm{const}}{\omega^\beta}\,,\ \beta < 1\,.$$

Moreover, we recall that $P(i\omega) = \Pi(i\omega)/(1+\omega^2)$ so that there exists $s < 3$ such that

$$P(i\omega)|\omega|^s > k > 0. \tag{34}$$

We can assume that the previous estimate holds for each ω since we consider the case in which the Popov function does not have zeros on the imaginary axis.

Let us recall that $-\bar{z}_k = x_k + iy_k$, $x_k > 0$.

We know that $M_0(z)$ is an outer function so that from [6] we know that

$$\log|M_0(-\bar{z}_k)| = \frac{1}{\pi}\int_{-\infty}^{+\infty} \log P(it)^{1/2} \frac{x_k}{x_k^2 + (y_k - t)^2}\, \mathrm{d}\, t,$$

and also

$$\log\{|\bar{z}_k|^{1+\gamma}|M_0(-\bar{z}_k)|\} = \frac{1}{\pi}\int_{-\infty}^{+\infty} \log|\bar{z}_k|^{1+\gamma} P(it)^{1/2} \frac{x_k}{x_k^2 + (y_k - t)^2}\, \mathrm{d}\, t$$

$$= \frac{1}{\pi}\int_{-\infty}^{+\infty} \log|\bar{z}_k|^{\gamma+1} P(i[y_k - tx_k])^{1/2} \frac{1}{1+t^2}\, \mathrm{d}\, t$$

$$= \frac{1}{2\pi}\int_{-\infty}^{+\infty} \log \frac{[x_k^2 + y_k^2]^{\gamma+1}}{|y_k - tx_k|^s} \left[|y_k - tx_k|^s P(i[y_k - tx_k])\right] \frac{1}{1+t^2}\, \mathrm{d}\, t \tag{35}$$

$$= \frac{1}{2\pi}\int_{-\infty}^{+\infty} \log \frac{[x_k^2 + y_k^2]^{\gamma+1}}{|y_k - tx_k|^s} \frac{1}{1+t^2}\, \mathrm{d}\, t \tag{36}$$

$$+\frac{1}{2\pi}\int_{-\infty}^{+\infty} \log \left[|y_k - tx_k|^s P(i[y_k - tx_k])\right] \frac{1}{1+t^2}\, \mathrm{d}\, t. \tag{37}$$

The integral in (37) is bounded from below by $(\log k)/2$; see (34).

We write the integral in (36) as follows:

$$\int_{-\infty}^{+\infty} \log \frac{[x_k^2 + y_k^2]^{\gamma+1}}{|y_k - tx_k|^s} \frac{1}{1+t^2}\, \mathrm{d}\, t$$

$$= \pi \log x_k^{2(\gamma+1)-s} + \pi \log \left[1 + \left(\frac{y_k}{x_k}\right)^2\right]^{\gamma+1} - s\int_{-\infty}^{+\infty} \log\left|\frac{y_k}{x_k} - t\right| \frac{1}{1+t^2}\, \mathrm{d}\, t\,.$$

It is now sufficient to notice that the last integral is bounded from above (so that minus the integral is bounded from below). This is clear since

$$\log\left|\frac{y_k}{x_k} - t\right| \le \log(\mu + t)\,.$$

The second addendum is nonnegative and $\log x_k^{2(\gamma+1)-s}$ is bounded from below if $2(\gamma+1) - s \ge 0$. We know that $s = 3 - \epsilon$ so that γ must satisfy $\gamma > \frac{1}{2} - \epsilon/2$. In particular there exists $\gamma < 1/2$ with this property.

Now we remove the assumption that $\Pi(i\omega)$ does not have imaginary axis zeros. By assumption, the zeros on the imaginary axis must be finite in number since we assumed inequality (34) for large $|\omega|$. Hence, it is sufficient to show how one of them, say it_0, can be handled. Moreover, we can confine ourselves to consider only those z_k with $|z_k|$ large. Hence we consider z_k such that $|t_0/z_k| < 1/2$.

A zero t_0 of $P(i\omega)$ must have even multiplicity $2n$. We remove the zero by multiplying and dividing the term $P(i\omega)$ by the factor

$$\left[\frac{1 + i(t - t_0)}{t - t_0}\right]^{2n}.$$

It is still true that

$$|t|^s \left[\frac{|1 + i(t - t_0)|}{|t - t_0|}\right]^{2n} P(it) > m > 0\,.$$

We replace the integrand in (35) with

$$\log\left\{\frac{[x_k^2+y_k^2]^{\gamma+1}}{|y_k-tx_k|^s}\frac{|y_k-tx_k-t_0|^{2n}}{|1+i(y_k-tx_k-t_0)|^{2n}}\right.$$
$$\left.\cdot\frac{|1+i(y_k-tx_k-t_0)|^{2n}}{|y_k-tx_k-t_0|^{2n}}|y_k-tx_k|^s P(i[y_k-tx_k])\right\}.$$

This operation replaces the integral (37) with

$$\int_{-\infty}^{+\infty}\log|y_k-tx_k|^s\left\{\frac{|1+i[y_k-tx_k]-it_0|}{|y_k-tx_k-t_0|}\right\}^{2n}P(i[y_k-tx_k])\frac{1}{1+t^2}\,\mathrm{d}\,t,$$

which is larger than $\frac{1}{2}\log m$; this operation produces a new addendum,

$$\frac{1}{2\pi}\int_{-\infty}^{+\infty}\log\left|\frac{y_k-tx_k-t_0}{1+i[y_k-tx_k]-it_0}\right|^{2n}\frac{1}{1+t^2}\,\mathrm{d}\,t$$
$$=\frac{n}{\pi}\int_{-\infty}^{+\infty}\log\left|\frac{i(t-t_0)}{1+i(t-t_0)}\right|\frac{x_k}{x_k+(t-y_k)^2}\,\mathrm{d}\,t=\log\left|\frac{z_k-it_0}{1+z_k-it_0}\right|,$$

since the function $\frac{z-it_0}{1+z-it_0}$ is H^∞ and outer. Boundedness from below follows since

$$\left|\frac{z_k-it_0}{z_k+1-it_0}\right|\geq\frac{1}{2}\frac{1}{1+|1-it_0|/|\tilde{z}|},$$

where $\tilde{z}$ is one of the points $-\bar{z}_k$ with minimal norm among those for which $|\frac{t_0}{z_k}|<1/2$.

This completes the proof. □

Now we prove the following.

LEMMA 14. *There exists a vector $q\in(\mathrm{dom}(-\mathcal{A}_{\mathcal{S}})^\alpha)'$, $0\leq\alpha<1/2$, which satisfies*

$$\int_0^{+\infty}e^{\mathcal{A}_{\mathcal{S}}{}^*t}q\check{M}_0(t)\;dt=\int_0^{+\infty}e^{\mathcal{A}_{\mathcal{S}}{}^*t}\mathcal{Q}e^{\mathcal{A}_{\mathcal{S}}t}\mathcal{D}\;dt\,. \tag{38}$$

This vector q also satisfies

$$\check{M}_0(t)=\begin{cases}\langle\mathcal{D},e^{\mathcal{A}_{\mathcal{S}}{}^*t}q\rangle & \text{if}\quad t>0,\\ 0 & \text{if}\quad t<0\,.\end{cases}$$

Proof. We note that the integral on the left of (38) makes sense for every $q\in(\mathrm{dom}(-\mathcal{A}_{\mathcal{S}}{}^*)^\alpha)'$ if $\alpha<1/2$, since we know that $\check{M}_0(t)$ is square integrable. In fact, it is known that

$$\|(-\mathcal{A}_{\mathcal{S}}{}^*)^\alpha e^{\mathcal{A}_{\mathcal{S}}{}^*t}\|\leq\frac{\mathrm{const}}{t^\alpha}e^{-\sigma t}$$

for a positive number σ; see [21]. For this same reason the integral on the right side defines a vector in $\mathrm{dom}(-\mathcal{A}_{\mathcal{S}}{}^*)^{1-\epsilon}$ for each $\epsilon>0$; see [10, part 1, Theorem 2.8]. Consequently, we must solve an equation of the form

$$\int_0^{+\infty}e^{\mathcal{A}_{\mathcal{S}}{}^*t}q\check{M}_0(t)\,\mathrm{d}\,t=(-\mathcal{A}_{\mathcal{S}}{}^*)^{-(1-\epsilon)}l \tag{39}$$

for a given vector $l\in X$.

We introduce the basis v_k of the eigenvectors of $\mathcal{A}_\mathcal{S}$. We multiply scalarly both sides of (39) by v_k. We put $l_k = \langle v_k, l\rangle$ and $q_k = \langle v_k, (-\mathcal{A}_\mathcal{S}^*)^{-\alpha} q\rangle = \langle v_k, \tilde{q}\rangle$, $\tilde{q} \in X$. We see that q is a solution if and only if

$$\frac{1}{(-z_k)^{1-\epsilon}} l_k = \int_0^{+\infty} \langle (-\mathcal{A}_\mathcal{S})^\alpha e^{\mathcal{A}_\mathcal{S} t} v_k, \tilde{q}\rangle \overline{\check{M}_0(t)}\, \mathrm{d}\, t \tag{40}$$

$$= (-z_k)^\alpha \overline{M}_0(-z_k) q_k\,. \tag{41}$$

This shows that q_k is well defined:

$$q_k = \frac{1}{(-z_k)^{\alpha+1-\epsilon}\overline{M_0(-\bar{z}_k)}} l_k\,.$$

In fact, $M_0(z)$ does not have zeros in the right half plane, since it is an outer function. In order to prove the existence of q we must ascertain that the sequence $\{q_k\}$ belongs to l^2, i.e., that the sequence $\{M_0(-\bar{z}_k)(-z_k)^{\alpha+1-\epsilon}\}$ is bounded away from zero, since we know that $\{l_k\}$ belongs to l^2. This follows from the previous lemma with $\alpha = \gamma + \epsilon$. We know that it is possible to choose $\gamma < \frac{1}{2}$ so that α can be chosen less then $\frac{1}{2}$ too, since the positive number ϵ is arbitrary.

Now we prove the formula for $\check{M}_0(t)$. We introduce the functions

$$K(t) = \begin{cases} e^{\mathcal{A}_\mathcal{S} t}\mathcal{D} & \text{if} \quad t > 0, \\ 0 & \text{if} \quad t < 0\,, \end{cases} \qquad H(t) = \begin{cases} \mathcal{Q} e^{\mathcal{A}_\mathcal{S} t}\mathcal{D} & \text{if} \quad t > 0, \\ 0 & \text{if} \quad t < 0\,, \end{cases}$$

and we consider the Fourier transformation of

$$t \to \int_0^{+\infty} \langle K(s), H(t+s)\rangle\, \mathrm{d}\, s, \tag{42}$$

which is $P(-i\omega)$, i.e., $M^*(-i\omega)M(-i\omega)$. This is also the Fourier transformation of

$$\int_0^{+\infty} \check{M}_0(s)\overline{\check{M}_0(t+s)}\, \mathrm{d}\, s\,. \tag{43}$$

We replace $K(\cdot)$ and $H(\cdot)$ with their expressions and we use (38). We see that the difference of the integrals in (42) and (43) is

$$\int_0^{+\infty} \{\check{M}_0(t+s) - \langle \mathcal{D}, e^{\mathcal{A}_\mathcal{S}^*(t+s)} q\rangle\} \overline{\check{M}_0(s)}\, \mathrm{d}\, s$$

for each t, and it is identically zero since both integrals are equal to $P(-i\omega)$.

The function $M_0(z)$ has a set of zeros of null measure, since it is a nonzero H^∞ function, and this implies that the brace is zero a.e., as wanted. □

Now we define W:

$$\langle \Xi', W\Xi\rangle = \int_0^{+\infty} \langle \Xi', e^{\mathcal{A}_\mathcal{S}^* t}\mathcal{Q} e^{\mathcal{A}_\mathcal{S} t}\Xi\rangle\, \mathrm{d}\, t - \int_0^{+\infty} \langle \Xi', e^{\mathcal{A}_\mathcal{S}^* t} q\rangle \cdot \overline{\langle \Xi, e^{\mathcal{A}_\mathcal{S}^* t} q\rangle}\, \mathrm{d}\, t\,.$$

The last integral makes sense since each factor is square integrable.

For each $\Xi \in \mathrm{dom}\mathcal{A}_\mathcal{S}$, we have

$$2\Re e\, \langle \mathcal{A}_\mathcal{S}\Xi, W\Xi\rangle = -\langle \Xi, \mathcal{Q}\Xi +\rangle + |\langle\langle \Xi, q\rangle\rangle|^2 \geq -\langle \Xi, \mathcal{Q}\Xi\rangle$$

and

$$\langle \Xi, W\mathcal{D}\rangle = \int_0^{+\infty} \langle e^{\mathcal{A}_S t}\Xi, \mathcal{Q}e^{\mathcal{A}_S t}\mathcal{D}\rangle \,\mathrm{d}\,t - \int_0^{+\infty} \langle \Xi, e^{\mathcal{A}_S{}^* t}q\rangle\overline{\langle e^{\mathcal{A}_S{}^* t}q, \mathcal{D}\rangle} \,\mathrm{d}\,t$$
$$= \int_0^{+\infty} \langle e^{\mathcal{A}_S t}\Xi, \mathcal{Q}e^{\mathcal{A}_S t}\mathcal{D}\rangle \,\mathrm{d}\,t - \int_0^{+\infty} \langle \Xi, e^{\mathcal{A}_S{}^* t}q\rangle\overline{\check{M}_0(t)} \,\mathrm{d}\,t = 0\,,$$

as wanted. This ends the proof of the existence of solutions to the Lur'e problem (LuE).

Acknowledgment. The author thanks the referee for useful comments, which improved the style of the paper.

REFERENCES

[1] A. V. Balakrishnan, *On a generalization of the Kalman-Yakubovich Lemma*, Appl. Math. Optim., 31 (1995), pp. 177–187.

[2] A. N. Churilov, *On the solvability of matrix inequalities*, Mat. Zametki, 36 (1984), pp. 725–732.

[3] J. B. Conway, *Functions of One Complex Variable*, Springer-Verlag, Berlin, 1973.

[4] R. F. Curtain, *The Kalman–Yakubovich–Popov lemma for Pritchard–Salamon systems*, Systems Control Lett., 27 (1996), pp. 67–72; *Correction to the Kalman–Yakubovich–Popov lemma for Pritchard–Salamon systems*, Systems Control Lett., 28 (1996), pp. 237–238.

[5] L. E. Faibusovich, *Matrix Riccati inequality: existence of solutions*, Systems Control Lett., 9 (1987), pp. 59–64.

[6] J. B. Garnett, *Bounded Analytic Functions*, Academic Press, New York, 1981.

[7] E. Hille and R.S. Phillips, *Functional Analysis and Semigroups*, AMS, Providence, RI, 1957.

[8] R. E. Kalman, *Lyapunov functions for the problem of Lur'e in automatic control*, Proc. Nat. Acad. Sci. U.S.A., 49 (1963), pp. 201–205.

[9] I. Lasiecka and R. Triggiani, *Differential and Algebraic Riccati Equations with Applications to Boundary/Point Control Problems*: *Continuous Theory and Approximation Theory*, Lecture Notes in Control and Inform. Sci. 164, Springer–Verlag, Berlin, 1991.

[10] I. Lasiecka and R. Triggiani, *The regulator problem for parabolic equations with Diriclet boundary control. Part* 1: *Riccati's feedback synthesis and regularity of optimal solutions*, Appl. Math. Optim., 16 (1987), pp. 147–168; *Part* 2: *Galerkin approximation*, Appl. Math. Optim., 16 (1987), pp. 187–216.

[11] I. Lasiecka, D. Lukes, and L. Pandolfi, *Input dynamics and nonstandard Riccati equations with applications to boundary control of damped wave and plate equations*, J. Optim. Theory Appl., 84 (1995), pp. 549–574.

[12] I. Lasiecka, L. Pandolfi, and R. Triggiani, *A singular control approach to highly damped second–order abstract equations and applications*, Appl. Math. Optim., 36 (1997), pp. 67–107.

[13] A. L. Likhtarnikov and V. A. Yakubovich, *The frequency theorem for equation of evolutionary type*, Siberian Math. J., 17 (1976), pp. 1069–1085.

[14] J.-Cl. Louis and D. Wexler, *The Hilbert space regulator problem and operator Riccati equation under stabilizability*, Ann. Soc. Sci. Bruxelles Sér. I, 105 (1991), pp. 137–165.

[15] L. Pandolfi, *Some properties of the frequency domain description of boundary control systems*, J. Math. Anal. Appl., 142 (1989), pp. 219–241.

[16] L. Pandolfi, *Generalized control systems, boundary control systems, and delayed control systems*, Math. Control Signals Systems, 3 (1990), pp. 165–181.

[17] L. Pandolfi, *From singular to regular control systems*, in Control of Partial Differential Equations, G. Da Prato and M. Tubaro, eds., Marcel Dekker, New York, 1994, pp. 153–165.

[18] L. Pandolfi, *The standard regulator problem for systems with input delays: An approach through singular control theory*, Appl. Math. Optim., 31 (1995), pp. 119–136.

[19] L. Pandolfi, *The Kalman–Yakubovich–Popov theorem*: *An overview and new results for hyperbolic control systems*, Nonlinear Anal., 30 (1997), pp. 735–745.

[20] L. Pandolfi, *The Kalman–Yakubovich–Popov theorem for stabilizable hyperbolic boundary control systems*, Integral Equations Operator Theory, to appear.

[21] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.

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

[23] V. A. Yakubovich, *Solution of certain matrix inequalities occurring in control theory*, Dokl. Akad. Nauk USSR, 143 (1962), pp. 1304–1307.

[24] V.A. Yakubovich, *The frequency theorem in control theory*, Siberian Math. J., 14 (1973), pp. 384–419.

[25] V. A. Yakubovich, *A frequency theorem for the case in which the state and control spaces are Hilbert spaces with an applications to some problems in the synthesis of optimal controls. Part* I, Siberian Math. J., 15 (1974), pp. 457-476; *Part* II, Siberian Math. J., 16 (1975), pp. 828–845.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 6, pp. 2082–2102, November 1998

013

VALUATION OF INVESTMENTS IN REAL ASSETS WITH IMPLICATIONS FOR THE STOCK PRICES*

THOMAS S. KNUDSEN†, BERNHARD MEISTER‡, AND MIHAIL ZERVOS§

Abstract. A general model for the valuation of natural resource investments is formulated and analyzed within a stochastic control theoretic framework. Using dynamic programming, the value of such an investment with a general payoff function is determined under the assumption that the commodity price process is given by a stochastic differential equation. The analysis results in closed form analytic solutions which can easily be computed and exhibits qualitatively different optimal behaviors, depending on parameter values. Implications for stocks and options are also considered.

Key words. stochastic control, optimal stopping, real assets, options

AMS subject classifications. 93E20, 93E03, 90A09, 90A11

PII. 0363012997315816

1. Introduction.

The area of *real options* has recently attracted considerable interest (see Dixit and Pindyck [12] for a review). This approach to contractual claims on real assets concentrates on their optionlike characteristics and uses option theory to evaluate them. Techniques which are similar to the ones developed in finance, especially to the seminal results of Black and Scholes, offer a new perspective and turn out to be very useful in the valuation of investment decisions in industry. To some extent, the same is true for the dynamic programming approach. Unlike the traditional but, in some ways, still orthodox net present value approach, the real options approach, as well as the dynamic programming approach can incorporate some of the uncertainty, irreversibility, and timing on which investment decisions depend.

In this paper, dynamic programming is used to study investments in industry and, in particular, in the natural resource industry. More specifically, we consider the problem of evaluating an investment in industry under the assumptions that it produces a single commodity and its value depends on the commodity price as well as on the way in which production is scheduled. Our model generalizes the one studied in section 6.3 of Dixit and Pindyck [12], notably in the directions of adding an abandonment option and of considering a much more general running payoff function. With reference to the natural resource industry, our model is closely related to the model studied by Brennan and Schwartz [7] using the contingent claim approach and which was further analyzed by Paddock, Siegel, and Smith [20]. Also, other related models have been studied by McDonald and Siegel [19], Pindyck [21], Dixit [11], Cortazar and Schwartz [9], Brekke and Øksendal [6], and Shirakawa [23]. At this point, it is important to emphasize that despite adopting the dynamic programming

*Received by the editors January 29, 1997; accepted for publication (in revised form) November 24, 1997; published electronically August 31, 1998.
http://www.siam.org/journals/sicon/36-6/31581.html

†Graduate School of Systems Management, The University of Tsukuba, Tokyo, 3-29-1 Otsuka, Bunkyo-ku, Tokyo 112, Japan (knudsen@gssm.otsuka.tsukuba.ac.jp). The work of this author was supported by the Japanese Society for the Promotion of Science.

‡Physics Department, Tokyo Institute of Technology, 2-12-2 O-okayama, Meguro-ku, Tokyo 152, Japan (meister@th.phys.titech.ac.jp). The work of this author was supported by the Japanese Society for the Promotion of Science.

§Department of Statistics, School of Mathematics and Statistics, University of Newcastle, Newcastle upon Tyne NE1 7RU, UK (Zervos.Mihail@ncl.ac.uk).

approach, we have implicitly solved the problem for the contingent claim approach as well, provided the convenience yield of the commodity is constant. Further information regarding this point can be found in Paddock, Siegel, and Smith [20] and the references therein, as well as in Dixit and Pindyck [12]; following either of the two approaches, the value of the project/firm is shown to satisfy equivalent nonlinear differential equations which are connected by a simple change of variables which involves only trivial algebra.

Apart from giving a price for an investment, our analysis also addresses the question of how production should be optimally scheduled. Our model is formulated as a stochastic control problem in which one has to decide on the production rate (i.e., the production per time unit) and the project's abandonment time. With regard to the production rate level, we make the assumption that this can be changed instantly and without cost to any value within a given set of admissible values. Also, with reference to the natural resource industry, we assume that the investment/firm under consideration has access to an infinite amount of the resource. Undoubtedly, this assumption is unrealistic from the perspective of a specific investment. However, it provides a certain approximation of reality which is further supported by the fact that we obtain easily computable results in a closed analytic form (see also the discussion at the end of section 2).

The first step of our analysis is to establish the dynamic programming equation, which is a variational inequality of the form encountered in the theory of optimal stopping. Optimal stopping problems have been addressed by many authors in numerous papers. Notable contributions to the solution of the general problem with a probabilistic approach include Fakeev [14], Bismut and Skalli [5], El Karoui [13] and a number of references therein. In a Markovian setting, Bensoussan and Lions [3] and Krylov [18] have studied optimal stopping problems and have proved under very general conditions that the corresponding value functions satisfy appropriate variational inequalities. In this paper, we will adopt an approach which is classical in the theory of stochastic optimal control (see, for example, Fleming and Soner [15, section IV.3] and which consists of finding a solution of the dynamic programming differential equation which satisfies the assumptions of a "verification theorem" which identifies this solution with the control problem's value function. In particular, we prove an appropriate "verification theorem," we explicitly solve the dynamic programming differential equation, and we derive an optimal strategy. An important feature of our results is that the optimal strategy can take qualitatively different forms, depending on parameter values. It is worth mentioning that similar analyses of related problems have been made by Brekke and Øksendal [6], who study a very general model of optimal switching related to investment decisions, and Davis and Zervos [10], who study a problem of combined singular stochastic control and optimal stopping.

Our analysis has a further implication. By deriving a value which is dependent solely on the commodity price, a connection between asset valuation and equity prices is established. In order to fix ideas, consider a company which produces a single commodity and whose total asset value uncertainty depends only on the uncertainty linked with the commodity price (e.g., the firm does not have debt). By assigning an asset value $v(x)$ to such a company, given that the commodity price is x, we obtain an expression for the company's stock price in terms of the commodity price. Furthermore, under the assumption that the commodity price follows a geometric Brownian motion, we show that the company's asset value, and therefore its stock price, is *not* a geometric Brownian motion (note that in Bensoussan, Crouhy, and Galai [1], [2], a similar observation is made for the stock price of a firm for which the total asset value is debt and equity, and the total asset value follows a geometric

Brownian motion). However, we show that, for the simplest case considered here, the Black & Scholes formula can be used to calculate an upper bound for the value of a European option on the stock of the company, as well as an approximate value for such options which is valid for short times to maturity and high commodity prices. At this point, it is worth noting that the idea of establishing a connection between a firm's total asset value and the firm's stock price under various conditions is not novel. For example, in [1], [2] an extensive analysis in this direction is carried out. However, a central drawback of these analyses is the simplifying assumption that the company's asset value always follows a geometric Brownian motion.

The paper is organized as follows. In section 2, a stochastic control problem which models the decisions on how to optimally schedule production and on which the foundations of our analysis are laid is formulated and discussed. Section 3 is concerned with establishing the Hamilton–Jacobi–Bellman (HJB) equation, which takes the form of a variational inequality, and proving a general existence result, whereas in section 4, an associated ODE is studied. The HJB equation is explicitly solved and the optimal strategy is derived in section 5 under certain additional hypotheses, whereas in section 6, the solution is further developed for a special case which arises in the comparison of our model to the model developed by Brennan and Schwartz [7] and which has special significance for the natural resource industry. In section 7, a European option written on the firm's stock is analyzed. Finally, section 8 contains a summary of our results as well as a description of certain possible extensions of our research.

2. Formulation of the control problem. Let $(\Omega, \mathcal{F}, P)$ be a probability space equipped with a filtration $(\mathcal{F}_t)$ satisfying the usual conditions of right continuity and augmentation by P-negligible sets and carrying a standard one-dimensional $(\mathcal{F}_t)$-Brownian motion W. We will denote by $\mathcal{T}$ the set of all $(\mathcal{F}_t)$-stopping times and by $\mathcal{C}$ the set of all progressively measurable processes U with values in a compact subset of the real line $\mathcal{U}$.

We model the *commodity price* by the solution of the SDE

$$dX_t = b(X_t)dt + \sigma(X_t)dW_t \ , \qquad X_0 = x > 0 \ , \tag{2.1}$$

where the following assumption holds.

Assumption A1. $b, \sigma : \mathbb{R}_+ \to \mathbb{R}_+$ are given functions such that (2.1) has a unique strong solution with values in $]0, \infty[$, P-almost surely (a.s.).

Specific assumptions under which (2.1) has a unique solution can be found in any book treating the subject of SDEs (e.g., see Corollary 5.5.16 and Proposition 5.5.17 in Karatzas and Shreve [16], or Theorem IX.3.5 in Revuz and Yor [22]), whereas Feller's Test for Explosions (see [16, Theorem 5.5.29]) yields necessary and sufficient conditions for the solution of (2.1) to have values in $]0, \infty[$, P-a.s.

A *production rate process* will be any process $U \in \mathcal{C}$, whereas an *abandonment time* will be any stopping time $\tau \in \mathcal{T}$. The set of *admissible strategies* Π will be the family of all pairs (U, τ) such that $U \in \mathcal{C}$ and $\tau \in \mathcal{T}$.

With any admissible strategy $(U, \tau) \in \Pi$, we associate the payoff

$$J_x(U, \tau) = E\left\{ \int_0^\tau e^{-rt} \overline{h}(X_t, U_t)dt + \mathrm{I}_{\{\tau < \infty\}} e^{-r\tau} g(X_\tau) \right\} \ , \tag{2.2}$$

where $\overline{h} :]0, \infty[\times \mathcal{U} \to \mathbb{R}$ and $g :]0, \infty[\to \mathbb{R}$ are given functions; given $(x, u) \in]0, \infty[\times \mathcal{U}$, $\overline{h}(x, u)$ represents the *running payoff* resulting if the commodity price is x and the production rate is u, whereas $-g(x)$ represents the project's *abandonment*

cost. Note that, from a financial point of view, the abandonment cost should not depend on the commodity price, and so, as far as modeling is concerned, g should be a constant. However, for the part of our analysis that we shall consider nonconstant g, such a generality adds no extra cost. The objective of the control problem is to maximize $J_x(U,\tau)$ over Π. Accordingly, we define the value function v by

$$v(x) = \sup_{(U,\tau)\in\Pi} J_x(U,\tau) \ . \tag{2.3}$$

The following assumptions on $\overline{h}$ and g will ensure that the optimization problem is well posed in the sense that there are no policies with infinite payoff. (Obviously, any project in the real world complies with such a restriction.)

Assumption A2. The running payoff function $\overline{h}$ is upper semicontinuous and if $h :]0,\infty[\to\mathbb{R}$ is the function defined by

$$h(x) := \max_{u\in\mathcal{U}} \overline{h}(x,u) \ , \tag{2.4}$$

then

$$E\int_0^\infty e^{-rt}|h(X_t)|dt < \infty \tag{2.5}$$

for every initial condition $x > 0$. Also, given any $x > 0$, the abandonment payoff g satisfies

$$E \operatorname*{ess\,sup}_{\tau\in\mathcal{T}} \left\{ \mathrm{I}_{\{\tau<\infty\}} e^{-r\tau} |g(X_\tau)| \right\} < \infty \ . \tag{2.6}$$

Note that since $\overline{h}$ is upper semicontinuous, h is upper semicontinuous as well (see Bertsekas and Shreve [4, Proposition 7.32]). Moreover (see [4, Proposition 7.33]), there exists a Borel measurable $u :]0,\infty[\to\mathcal{U}$ such that

$$h(x) = \overline{h}(x,u(x)) \ . \tag{2.7}$$

The very general Assumptions A1 and A2 will be used to prove part of the results of section 3. In the same section, we will obtain a general existence result under the following additional assumption.

Assumption A3. The functions b, σ, h are twice continuously differentiable,

$$\sigma(x) > 0 \qquad \forall x \in]0,\infty[\ ,$$

and there exist constants C, k such that $\forall x, y \in]0,\infty[$

$$|b(x) - b(y)| + |\sigma(x) - \sigma(y)| \le C|x - y| \ ,$$

$$|b'(x)| + |b''(x)| + |\sigma'(x)| + |\sigma''(x)| \le C(1 + x^k) \ ,$$

$$|h(x)| + |h'(x)| + |h''(x)| + |g(x)| + g'(x)| + |g''(x)| \le C(1 + x^k) \ .$$

The analysis of sections 4–7 will assume that the commodity price follows a geometric Brownian motion, that g is constant, and that $\overline{h}$ satisfies some different conditions. More specifically, we will impose the following assumptions.

Assumption A4.a. The functions b, σ are given by $b(x) = bx$, $\sigma(x) = \sqrt{2}\sigma x$, for some constants b, σ.

Assumption A4.b. The function $\overline{h}$ is upper semicontinuous, and if h is defined by (2.4), then h is increasing and satisfies (2.5) as well as $\lim_{x\to\infty} h(x) = \infty$. Also, $g \equiv K$ for some constant $K \in \mathbb{R}$.

The additional hypotheses on $\overline{h}$ are satisfied in all economically sensible cases since the running payoff function should be an increasing function of the commodity price and should tend to infinity when the price explodes.

With reference to the natural resource industry, the assumption that an investment/firm has access to an infinite amount of the resource can intuitively be viewed as reasonable as long as either production in the distant future has little effect on the present value due to large discounting or a replacement cost for produced resource has been incorporated into the model through appropriate choice of $\overline{h}$. In the latter case, every unit produced is replaced by a new one which is "added" to the reserves at a given cost (such a cost can account for the exploration and development of new reserves); in effect, the firm would then have an unlimited supply.

3. Existence of an optimal strategy. Consider the control problem described in the previous section. With reference to Assumption A2, since $h(x)$ is the best rate of return given that the commodity price is x, it is intuitively clear that the optimal production process should be indistinguishable from $u \circ X$. In this way, the problem reduces to optimally choosing the abandonment time τ, i.e., to an optimal stopping problem. As a consequence, with reference to standard results of the theory of optimal stopping, we should expect that the value function satisfies the following variational inequality

$$\max\left\{\frac{1}{2}\sigma^2 w'' + bw' - rw + h\ ,\ g - w\right\} = 0\ . \tag{3.1}$$

In general, (3.1) does not admit a unique solution even within the space of infinitely differentiable functions.

Example 3.1. For $\sigma(x) = \sqrt{2}x$, $b(x) = x$, $r = 4$, $h(x) = x$, $g(x) = 0$, it is straightforward to verify that each of the functions defined by

$$w(x) = Ax^2 + Bx^{-2} + \frac{1}{3}x\ ,$$

where $A, B \geq 0$, satisfies (3.1). □

On the other hand, with reference to the theory of optimal stopping, and as we will see in subsequent sections, we should expect that the value function is *not* twice continuously differentiable. For this reason, we consider solutions of the HJB equation (3.1) which belong to a Sobolev space $W^{2,p}_{loc}(]0,\infty[)$, $1 \leq p \leq \infty$. Recall (see, for example, Brezis [8, Chapter VIII]) that a function $w \in W^{2,p}_{loc}(]0,\infty[)$, $1 \leq p \leq \infty$ admits a representative which is continuous with continuous classical first derivative; moreover, it has a second derivative in the sense of distributions which is a function $w'' \in L^p_{loc}(]0,\infty[)$. Therefore, whenever we write $w \in W^{2,p}_{loc}(]0,\infty[)$ we will refer to this continuous representative and we will denote by w' its continuous classical first derivative and by w'' any function which identifies with its distributional second derivative. Also, when we say that $w \in W^{2,p}_{loc}(]0,\infty[)$ satisfies the HJB equation (3.1), we mean that equality holds in (3.1) for Lebesgue almost all x. Of course, since we consider the solvability of (3.1) in $W^{2,p}_{loc}(]0,\infty[)$, we have to *assume* that $g \in W^{2,p}_{loc}(]0,\infty[)$, as long as abandonment can ever be optimal.

Note that identifying the value function of the control problem with a solution of (3.1) in $W^{2,p}_{loc}(]0,\infty[)$ incorporates the so-called smooth pasting condition of optimal stopping which, in the context of the problem studied here, requires that $v'(x) = g'(x)\ \forall x \in \mathcal{S} := \{x \in]0,\infty[:\ v(x) = g(x)\}$ and, in particular, on the boundary of this set. The "smooth pasting condition" is a necessary condition for a wide class of optimal stopping problems; related results can be found in Shiryayev [24, section 3.8] and Krylov [18, Corollary 4.7.9].

We now prove a "verification theorem" that we will use in subsequent sections and which relates the value function of the control problem with a solution of the HJB equation (3.1).

THEOREM 3.2. *Consider the control problem described in section* 2*, and assume that* A1 *and* A2 *hold. Suppose that the HJB equation* (3.1) *has a solution* $w \in W^{2,p}_{loc}(]0,\infty[)$ *for some* $p \in [1,\infty]$ *such that if* M *is the process defined by*

$$M_t = \int_0^t e^{-rs}\sigma(X_s)w'(X_s)dW_s\ , \qquad t \geq 0\ , \tag{3.2}$$

then, for every constant $T > 0$*, the stopped process* M^T *is a martingale. Given any initial condition* $x > 0$*,*

a) $v(x) \leq w(x)$*, and*

b) *if*

$$\liminf_{t\to\infty} e^{-rt}E|w(X_t)| = 0\ , \tag{3.3}$$

then $v(x) = w(x)$ *and the optimal strategy is given by*

$$\tilde{U}_t = u(X_t)\ , \qquad \tilde{\tau} = \inf\{t \geq 0:\ X_t \in \mathcal{S}\}\ , \tag{3.4}$$

where u *satisfies* (2.7) *and* $\mathcal{S} := \{x \in]0,\infty[:\ w(x) = g(x)\}$.

Proof. a) Fix an arbitrary admissible strategy $(U,\tau) \in \Pi$. An application of Itô–Tanaka's formula (see Theorem IV.1.5, Corollary IV.1.6, and the remarks thereafter in Revuz and Yor [22]) yields

$$\begin{aligned} e^{-rt}w(X_t) = w(x) &+ \int_0^t e^{-rs}\left[\frac{1}{2}\sigma^2(X_s)w''(X_s) + b(X_s)w'(X_s) - rw(X_s)\right]ds \\ &+ \int_0^t e^{-rs}\sigma(X_s)w'(X_s)dW_s\ . \end{aligned}$$

This implies that

$$\begin{aligned} &\int_0^{\tau\wedge t} e^{-rs}h(X_s)ds + e^{-r(\tau\wedge t)}g(X_{\tau\wedge t}) \\ &\quad = w(x) + e^{-r(\tau\wedge t)}\left[g(X_{\tau\wedge t}) - w(X_{\tau\wedge t})\right] + \int_0^{\tau\wedge t} e^{-rs}\sigma(X_s)w'(X_s)dW_s \\ &\qquad + \int_0^{\tau\wedge t} e^{-rs}\left[\frac{1}{2}\sigma^2(X_s)w''(X_s) + b(X_s)w'(X_s) - rw(X_s) + h(X_s)\right]ds\ . \end{aligned} \tag{3.5}$$

Since w satisfies (3.1), we obtain

$$\int_0^{\tau\wedge t} e^{-rs}h(X_s)ds + e^{-r(\tau\wedge t)}g(X_{\tau\wedge t}) \leq w(x) + \int_0^{\tau\wedge t} e^{-rs}\sigma(X_s)w'(X_s)dW_s\ .$$

Taking expectations, we find that

$$E\left\{\int_0^{\tau\wedge t} e^{-rs}h(X_s)ds + e^{-r(\tau\wedge t)}g(X_{\tau\wedge t})\right\} \leq w(x)\ .$$

Letting $t \to \infty$, we obtain (by (2.5), (2.6), and the dominated convergence theorem)

$$E\left\{\int_0^{\tau} e^{-rt}h(X_t)dt + \mathrm{I}_{\{\tau<\infty\}}e^{-r\tau}g(X_\tau)\right\} \leq w(x)\ .$$

However, this and the fact that (because of (2.2), (2.7), and (2.4))

$$J_x(U,\tau) \leq E\left\{\int_0^{\tau} e^{-rt}h(X_t)dt + \mathrm{I}_{\{\tau<\infty\}}e^{-r\tau}g(X_\tau)\right\} \tag{3.6}$$

imply that $v(x) \leq w(x)$.

b) If $\tilde{U}$ and $\tilde{\tau}$ are as in (3.4), then (3.5) and (3.1) imply that

$$\begin{aligned}\int_0^{\tilde{\tau}\wedge t} e^{-rs}\overline{h}(X_s,\tilde{U}_s)ds + \mathrm{I}_{\{\tilde{\tau}\leq t\}}e^{-r\tilde{\tau}}g(X_{\tilde{\tau}}) &\\ = \int_0^{\tilde{\tau}\wedge t} e^{-rs}h(X_s)ds + \mathrm{I}_{\{\tilde{\tau}\leq t\}}e^{-r\tilde{\tau}}g(X_{\tilde{\tau}}) &\\ = w(x) - \mathrm{I}_{\{\tilde{\tau}>t\}}e^{-rt}w(X_t) + \int_0^{\tilde{\tau}\wedge t} e^{-rs}\sigma(X_s)w'(X_s)dW_s\ .&\end{aligned}$$

Taking expectations, we obtain

$$E\left\{\int_0^{\tilde{\tau}\wedge t} e^{-rs}\overline{h}(X_s,\tilde{U}_s)ds + \mathrm{I}_{\{\tilde{\tau}\leq t\}}e^{-r\tilde{\tau}}g(X_{\tilde{\tau}})\right\} = w(x) - e^{-rt}E\left\{\mathrm{I}_{\{\tilde{\tau}>t\}}w(X_t)\right\}\ .$$

In view of (3.3), we can pass to the limit $t \to \infty$ through an appropriate sequence to obtain $J_x(\tilde{U},\tilde{\tau}) = w(x)$, which, combined with part a) of the theorem, implies that $v(x) = w(x)$. □

Note that, among other things, the preceding theorem asserts that if the value function of the control problem satisfies the HJB equation (3.1), then it is "minimal" in the set of all solutions of (3.1), which may be uncountably many.

We now have the following existence result.

THEOREM 3.3. *Consider the control problem described in section* 2 *and assume that* A1–A3 *hold. The value function* v *belongs to* $W^{2,\infty}_{loc}(]0,\infty[)$ *and satisfies* (3.1)*, whereas the optimal strategy is given by* (3.4)*.*

Proof. Consider the optimal stopping problem defined by

$$\hat{v}(x) = \sup_{\tau\in\mathcal{T}} E\left\{\int_0^{\tau} e^{-rt}h(X_t)dt + \mathrm{I}_{\{\tau<\infty\}}e^{-r\tau}g(X_\tau)\right\}\ .$$

The value function $\hat{v}$ belongs to $W^{2,\infty}_{loc}(]0,\infty[)$, satisfies (3.1), and the stopping time $\tilde{\tau}$ defined by (3.4) is optimal (see Krylov [18, Theorem 6.4.14]). Now, in view of (3.6) and the fact that (3.6) holds with equality for $U = \tilde{U}$, it is clear that $v = \hat{v}$, and the proof is complete. □

It is well known that if $|w'|$ is bounded by a polynomial, then the process M defined by (3.2) is a square integrable martingale if stopped at any constant time. The following lemma provides a similar condition which will be of use in section 5.

LEMMA 3.4. *Consider the SDE* (2.1)*, assume that* b*,* σ *satisfy* A4.a*, and let* M *be the process defined by* (3.2)*. If there exist constants* C*,* k *such that*

$$|w'(x)| \leq C(x^{-k} + x^k) \quad \forall x > 0 \ ,$$

then, for every $T > 0$*, the stopped process* M^T *is a square integrable martingale.*

Proof. In the case that we consider here, the unique strong solution of the SDE (2.1) is given by (see Karatzas and Shreve [16, section 5.6.C])

$$X_t = x \exp\left\{(b - \sigma^2)t + \sqrt{2}\sigma W_t\right\} \ , \quad t \geq 0 \ . \tag{3.7}$$

Therefore, given any reals κ, λ, and any $t \geq 0$,

$$\begin{aligned} e^{\kappa t} E X_t^\lambda &= x^\lambda \exp\left\{\left[\sigma^2\lambda^2 + (b - \sigma^2)\lambda + \kappa\right] t\right\} \, E \exp\left\{-\sigma^2\lambda^2 t + \sqrt{2}\sigma\lambda W_t\right\} \\ &= x^\lambda \exp\left\{\left[\sigma^2\lambda^2 + (b - \sigma^2)\lambda + \kappa\right] t\right\} \ , \end{aligned} \tag{3.8}$$

and so,

$$E \int_0^T e^{\kappa t} X_t^\lambda dt = \int_0^T e^{\kappa t} E[X_t^\lambda] dt < \infty \quad \forall T > 0 \ .$$

As a consequence, given any $T > 0$,

$$\begin{aligned} E M_T^2 &= 2\sigma^2 E \int_0^T e^{-2rt} X_t^2 |w'(X_t)|^2 dt \\ &\leq 2\sigma^2 C^2 E \int_0^T e^{-2rt} \left(X_t^{-2k+2} + 2X_t^2 + X_t^{2k+2}\right) dt \\ &< \infty \ , \end{aligned}$$

which implies that the stopped local martingale M^T has integrable quadratic variation and therefore is a square integrable martingale (see Revuz and Yor [22, Proposition IV.1.23]. □

4. Study of a fundamental ODE. In the following section, we will explicitly solve the control problem which arises when the drift and dispersion of the SDE (2.1) satisfy Assumption A4.a by finding a solution of the HJB equation which satisfies the requirements of the Verification Theorem 3.2. In this case, the HJB equation takes the form

$$\max\left\{\sigma^2 x^2 w''(x) + bxw'(x) - rw(x) + h(x) \ , \ g(x) - w(x)\right\} = 0 \ . \tag{4.1}$$

It is well known that the general solution of the ODE

$$\sigma^2 x^2 w''(x) + bxw'(x) - rw(x) + h(x) = 0 \tag{4.2}$$

which is associated with (4.1) is given by

$$w_g(x) = Ax^m + Bx^n \ , \tag{4.3}$$

where $A, B \in \mathbb{R}$ and m, n are given by

$$m = \frac{1}{2\sigma^2}\left[\sigma^2 - b - \sqrt{(\sigma^2 - b)^2 + 4\sigma^2 r}\right] \tag{4.4}$$

and

$$n = \frac{1}{2\sigma^2}\left[\sigma^2 - b + \sqrt{(\sigma^2 - b)^2 + 4\sigma^2 r}\right]. \tag{4.5}$$

The following proposition is concerned with the construction and certain properties of a special solution of (4.2). Note that part of these results are similar to results presented in section 2 of Kobila [17]; in particular, conclusions b) and c) are the same as Propositions 2.2 and 2.4 in this reference, respectively.

PROPOSITION 4.1. *Consider a measurable function* $h :]0,\infty[\to \mathbb{R}$*, and let* X *be the solution of the SDE* (2.1) *under* A4.a. *The following statements are equivalent:*

i) $E\int_0^\infty e^{-rt}|h(X_t)|dt < \infty$ *for every initial condition* $x > 0$.

ii) *There exists an initial condition* $x > 0$ *such that* $E\int_0^\infty e^{-rt}|h(X_t)|dt < \infty$.

iii) $x \to x^{-m-1}h(x) \in L^1(]0, y[)$ *and* $x \to x^{-n-1}h(x) \in L^1(]y,\infty[)$ $\forall y > 0$.

iv) *There exists a* $y > 0$ *such that* $x \to x^{-m-1}h(x) \in L^1(]0, y[)$ *and* $x \to x^{-n-1}h(x) \in L^1(]y,\infty[)$.

If i)–iv) *are true, then*

a) $\liminf_{x\to\infty} x^{-n}|h(x)| = \liminf_{x\downarrow 0} x^{-m}|h(x)| = 0$

and

$$w_p(x) := \frac{1}{\sigma^2(n-m)}\left[x^m\int_0^x s^{-m-1}h(s)ds + x^n\int_x^\infty s^{-n-1}h(s)ds\right], \quad x \in]0,\infty[\tag{4.6}$$

defines a real valued function such that

b) w_p *is twice differentiable in the classical sense and is a special solution of the ODE* (4.2),

c) *there exists a constant* C *such that*

$$|w_p'(x)| \le C(x^{m-1} + x^{n-1}) \quad \forall x > 0\ ,$$

d) w_p *satisfies*

$$w_p(x) = E\int_0^\infty e^{-rt}h(X_t)dt \quad \forall x > 0\ , \tag{4.7}$$

and

e) $\lim_{t\to\infty} e^{-rt}E|w_p(X_t)| = 0$.

Proof. iii) $\Leftrightarrow$ iv): Assume that there exists a $y \in]0,\infty[$ such that

$$C_1 := \int_0^y s^{-m-1}|h(s)|ds < \infty \quad \text{and} \quad C_2 := \int_y^\infty s^{-n-1}|h(s)|ds < \infty\ .$$

Given any $x \in]0, y[$, it is clear that

$$\int_0^x s^{-m-1}|h(s)|ds \le C_1\ . \tag{4.8}$$

On the other hand, since $m - n < 0$,

$$\begin{aligned}\int_x^\infty s^{-n-1}|h(s)|ds &= \int_x^y s^{m-n}\cdot s^{-m-1}|h(s)|ds + C_2\\ &\le x^{m-n}\int_x^y s^{-m-1}|h(s)|ds + C_2\\ &\le C_1 x^{m-n} + C_2\ .\end{aligned} \tag{4.9}$$

Similarly, given any $x \in]y, \infty[$,

$$\int_0^x s^{-m-1}|h(s)|ds \leq C_1 + C_2 x^{n-m} \tag{4.10}$$

and

$$\int_x^\infty s^{-n-1}|h(s)|ds \leq C_2 \ . \tag{4.11}$$

However, these bounds prove that iv) $\Rightarrow$ iii). The reverse implication is obvious.

Proof of b). If iii) is satisfied, then w_p is well defined, in which case it is trivial to verify b).

Proof of c). The bounds (4.8)–(4.11) imply that

$$x^m \int_0^x s^{-m-1}|h(s)|ds, \quad x^n \int_x^\infty s^{-n-1}|h(s)|ds \quad \leq \quad C_1 x^m + C_2 x^n \ . \tag{4.12}$$

Using these, we calculate

$$\begin{aligned} |w_p'(x)| &= \frac{1}{\sigma^2(n-m)} \left| m x^{m-1} \int_0^x s^{-m-1} h(s) ds + n x^{n-1} \int_x^\infty s^{-n-1} h(s) ds \right| \\ &\leq \frac{1}{\sigma^2(n-m)} \left(|m| x^{m-1} \int_0^x s^{-m-1} |h(s)| ds + n x^{n-1} \int_x^\infty s^{-n-1} |h(s)| ds \right) \\ &\leq \frac{1}{\sigma^2} \big(C_1 x^{m-1} + C_2 x^{n-1} \big), \end{aligned}$$

which proves c).

i) $\Leftrightarrow$ iii) and ii) $\Leftrightarrow$ iv): Assume first that h is positive and bounded. In this case, it is clear that all of the statements i)–iv) are true. Also, it is easy to check that both w_p and $x \to x w_p'(x)$ are bounded. Applying Itô–Tanaka's formula and the occupation times formula, and using the fact that w_p satisfies (4.2) (because iii), and therefore b), is true), we obtain

$$\int_0^t e^{-rs} h(X_s) ds = w_p(x) - e^{-rt} w_p(X_t) + \sqrt{2}\sigma \int_0^t e^{-rs} X_s w_p'(X_s) dW_s \ . \tag{4.13}$$

Taking expectations and passing to the limit $t \to \infty$, we obtain (4.7).

Now assume that h is an arbitrary positive function, and consider the sequences of functions (h_k) and $(w_{p,k})$, where $h_k(x) = h(x) \wedge k$ and $w_{p,k}$ is defined by (4.6) with h_k in place of h. Since (h_k) converges pointwise to h and (4.7) holds with $w_{p,k}$, h_k in place of w_p, h, respectively, for every k, the monotone convergence theorem implies (4.7), where both sides may be equal to ∞.

In particular, we have just proved that for every measurable h,

$$E \int_0^\infty e^{-rt} |h(X_t)| dt = \frac{1}{\sigma^2(n-m)} \left[x^m \int_0^x s^{-m-1} |h(s)| ds + x^n \int_x^\infty s^{-n-1} |h(s)| ds \right] \tag{4.14}$$

$\forall x \in]0, \infty[$. However, this establishes the equivalences i) $\Leftrightarrow$ iii) and ii) $\Leftrightarrow$ iv).

Proof of d). We have proved the result for positive h. For an arbitrary h satisfying i)–iv), the result follows by considering its positive and negative parts h^+ and h^-, respectively.

Proof of e). Define $\bar{w}_p$ by

$$\bar{w}_p(x) := \frac{1}{\sigma^2(n-m)}\left[x^m \int_0^x s^{-m-1}|h(s)|ds + x^n \int_x^\infty s^{-n-1}|h(s)|ds\right],$$

and note that if h satisfies iii), then the same is true for $|h|$. Now, since $\bar{w}_p$ satisfies the ODE (4.2) with h replaced by $|h|$ (because of b)),

$$\int_0^t e^{-rs}|h(X_s)|ds = \bar{w}_p(x) - e^{-rt}\bar{w}_p(X_t) + \sqrt{2}\sigma \int_0^t e^{-rs} X_s \bar{w}_p'(X_s) dW_s \ .$$

Taking expectations and noting that the stochastic integral has expectation zero because of c) and Lemma 3.4, we obtain

$$e^{-rt} E\bar{w}_p(X_t) = \bar{w}_p(x) - E\int_0^t e^{-rs}|h(X_s)|ds \ .$$

However, (4.14) and the monotone convergence theorem imply that the right-hand side of this equation converges to zero as $t \to \infty$, and so, $\lim_{t\to\infty} e^{-rt} E\bar{w}_p(X_t) = 0$. Now, e) follows from the fact that $|w_p| \le \bar{w}_p$.

Proof of a). Suppose that $\liminf_{x\to\infty} x^{-n}|h(x)| > 0$, and let $\epsilon > 0$ and y be such that $x^{-n}|h(x)| \ge \epsilon \ \forall x \ge y$. Then,

$$\int_y^\infty x^{-n-1}|h(x)|dx \ge \epsilon \int_y^\infty x^{-1}dx = \infty \ ,$$

and so, (iii) does not hold. Similarly, we prove that $\liminf_{x\downarrow 0} x^{-m}|h(x)| = 0$. □

Note that, at this generality, the necessary conditions a) of the preceding proposition cannot be strengthened.

Example 4.2. Assume that the function h is defined by

$$h(x) = \begin{cases} x^{n+1} & \text{if } x \in [k, k+k^{-2}] \ , \ k \in \mathbb{N}, \\ 0 & \text{otherwise.} \end{cases}$$

We can calculate

$$w_p(1) = \frac{1}{\sigma^2(n-m)} \sum_{k=1}^\infty \frac{1}{k^2} = \frac{\pi^2}{6\sigma^2(n-m)} < \infty \ ,$$

and so, w_p satisfies statement iv) of the preceding proposition. However,

$$\limsup_{x\to\infty} x^{-n} h(x) = \infty. \qquad \square$$

Remark 4.3. Using (3.8) we find that, given any constant λ,

$$\begin{aligned} E\int_0^\infty e^{-rt} X_t^\lambda dt &= \int_0^\infty e^{-rt} E[X_t^\lambda] dt \\ &= x^\lambda \int_0^\infty \exp\left\{\left[\sigma^2\lambda^2 + (b-\sigma^2)\lambda - r\right] t\right\} dt, \end{aligned}$$

which implies that

$$E\int_0^\infty e^{-rt} X_t^\lambda dt < \infty \quad \Leftrightarrow \quad \lambda \in \,]m, n[\ .$$

As a consequence, if there exist constants k, l, and C such that $m < k < l < n$ and

$$|h(x)| \leq C(x^k + x^l) \qquad \forall x \in]0, \infty[\ ,$$

then statements i)–iv) of Proposition 4.1 are satisfied. □

Remark 4.4. At this point, note that since

$$w_p''(x) = \frac{1}{\sigma^2(n-m)} \left(m(m-1)x^{m-2} \int_0^x s^{-m-1} h(s) ds \right.$$
$$\left. + n(n-1)x^{n-2} \int_x^\infty s^{-n-1} h(s) ds \right) - \frac{1}{\sigma^2} x^{-2} h(x) \ ,$$

the bounds given by (4.12) imply that if $h \in L^p_{loc}(]0, \infty[)$, then $w_p \in W^{2,p}_{loc}(]0, \infty[)$. In particular, if h is a monotone function, then it is bounded on compact subsets of $]0, \infty[$ and therefore $w_p \in W^{2,\infty}_{loc}(]0, \infty[)$, whereas if h is continuous, then $w_p \in C^2(]0, \infty[)$. □

We will also need the following lemma.

LEMMA 4.5. *Consider any function h satisfying the requirements of Assumption* A4.b. *If w_p is the function defined by* (4.6)*, then w_p is increasing and* $\lim_{x\to\infty} w_p(x) = \infty$. *Moreover,*

$$n \int_x^\infty s^{-n-1} h(s) ds = x^{-n} h(x) + \int_x^\infty s^{-n} dh(s) \ . \tag{4.15}$$

Proof. Let any $x < y$, and denote by X^x and X^y the solutions of (2.1) with initial conditions $X_0 = x$ and $X_0 = y$, respectively. In view of (3.7), $X^x_t < X^y_t$ $\forall t$, P-a.s., and therefore (4.7) and the fact that h is increasing imply that $w_p(x) < w_p(y)$, which proves that w_p is increasing.

Now, let any y such that $h(y) > 0$; such a y exists because we have assumed that $\lim_{x\to\infty} h(x) = \infty$. For $x > y$,

$$w_p(x) \geq \frac{1}{\sigma^2(n-m)} \left[x^m \int_0^y s^{-m-1} h(s) ds + x^n \int_x^\infty s^{-n-1} h(s) ds \right]$$
$$\geq \frac{1}{\sigma^2(n-m)} \left[x^m \int_0^y s^{-m-1} h(s) ds + \frac{1}{n} h(x) \right] \ .$$

Since $m < 0 < n$ and $\lim_{x\to\infty} h(x) = \infty$, the last term in this expression tends to ∞ as $x \to \infty$, and therefore $\lim_{x\to\infty} w_p(x) = \infty$.

Finally, with reference to Proposition 4.1a), let (a_k) be any sequence converging to ∞ such that $\lim_{k\to\infty} a_k^{-n} h(a_k) = 0$. Using the dominated and the monotone convergence theorems and the integration by parts formula (see, for example, Revuz and Yor [22, Proposition 0.4.5]), we calculate

$$\int_x^\infty s^{-n-1} h(s) ds = \lim_{k\to\infty} \int_x^{a_k} s^{-n-1} h(s) ds$$
$$= \lim_{k\to\infty} \left(-\frac{1}{n} \int_x^{a_k} h(s) ds^{-n} \right)$$
$$= \lim_{k\to\infty} \left(\frac{1}{n} x^{-n} h(x) - \frac{1}{n} a_k^{-n} h(a_k) + \frac{1}{n} \int_x^{a_k} s^{-n} dh(s) \right)$$
$$= \frac{1}{n} x^{-n} h(x) + \frac{1}{n} \int_x^\infty s^{-n} dh(s) \ ,$$

which proves (4.15). □

Remark 4.6. Note that the calculations used to establish (4.15) remain true if h satisfies conditions i)–iv) of Proposition 4.1 and is decreasing instead of increasing. Also, using similar arguments we can show that if h is a monotone function satisfying i)–iv) of Proposition 4.1, then

$$m\int_0^x s^{-m-1}h(s)ds = -x^{-m}h(x) + \int_0^x s^{-m}dh(s) \ .$$

Moreover, we can use this identity, (4.15), and a simple limiting argument to show that

$$\lim_{x\to\infty} x^{-n}h(x) = \lim_{x\downarrow 0} x^{-m}h(x) = 0 \ ,$$

which is stronger than Proposition 4.1a). □

5. The solution of the control problem. Consider the control problem described in section 2 and assume A4. One possibility is that abandonment is never optimal. Given an initial condition $x > 0$, never abandoning yields a payoff equal to $w_p(x)$ (see Proposition 4.1d), whereas abandoning straightaway yields a payoff equal to K. As a consequence, we should expect that abandonment is never optimal if $w_p(x) \geq K \ \forall x > 0$, in which case the value function is equal to w_p. If abandonment is ever optimal, it should occur whenever the commodity price is sufficiently small. In this case, we should expect that the optimal policy consists of producing optimally as long as the commodity price is larger than a certain value y and abandoning as soon as the commodity price falls below y. If this policy is optimal, we should find a solution w of the HJB equation (4.1) such that $w(x) = K \ \forall x \in]0, y[$ and, in view of (4.3) and Proposition 4.1b),

$$w(x) = Ax^m + Bx^n + w_p(x) \qquad \forall x \in]y, \infty[\ ,$$

where w_p is given by (4.6) and A, B are parameters to be specified. Note that, given any $A, B \in \mathbb{R}$, this candidate solution of (4.1) can be expressed as

$$w(x) = g(x) + Bx^n + w_p(x)\mathrm{I}_{\{x\geq y\}} \qquad \forall x \in]0, \infty[\ ,$$

where g is a bounded function. Therefore, in view of (3.8) and Proposition 4.1e), we must have $B = 0$ because otherwise (3.3) cannot hold. The remaining two parameters A and y should be specified by the requirement that w is C^1 at y (the "smooth pasting condition" of optimal stopping). The next lemma is concerned with this issue.

LEMMA 5.1. *The system of equations*

$$Ay^m + w_p(y) = K \ , \tag{5.1}$$

$$mAy^{m-1} + w_p'(y) = 0 \tag{5.2}$$

for the parameters A and y has a solution with $y > 0$ if and only if

$$\inf_{x\in]0,\infty[} w_p(x) < K \ , \tag{5.3}$$

or equivalently, if and only if

$$\inf_{x\in]0,\infty[} h(x) < rK \ . \tag{5.4}$$

In this case, $y > 0$ is the unique solution of

$$\int_y^\infty s^{-n-1}[h(s) - rK]ds = 0 \tag{5.5}$$

whereas

$$A = -\frac{1}{m} y^{-m+1} w_p'(y) \ . \tag{5.6}$$

Proof. The proof is organized as follows. We first show that the system of equations (5.1), (5.2) is equivalent to (5.5), (5.6). We then show that if (5.5) has a solution $y > 0$, then (5.3) is true. Next, we prove that (5.3) implies (5.4), and finally, we prove that if (5.4) is true, then (5.5) has a unique solution $y > 0$.

It is straightforward to verify that the system of equations (5.1) and (5.2) is equivalent to the system of equations consisting of (5.6) and

$$f(y) := mK - mw_p(y) + yw_p'(y) = 0 \ . \tag{5.7}$$

Using the fact that $\sigma^2 mn = -r$, we can see that

$$f(y) = \frac{1}{\sigma^2} y^n \int_y^\infty s^{-n-1}[h(s) - rK]ds \ , \tag{5.8}$$

which proves that the system (5.1), (5.2) is equivalent to (5.5), (5.6).

Since $yw_p'(y) > 0 \ \forall y > 0$ and w_p increases to ∞ as $y \to \infty$ (see Lemma 4.5), it is clear that (5.7) can have a solution only if (5.3) is true. Therefore existence of solution of (5.5) implies (5.3).

Now, assume that (5.3) is true, and let any $z > 0$ such that

$$w_p(z) - K < 0 \ .$$

Using the identity $\sigma^2 mn = -r$, we can calculate that this is equivalent to

$$z^m \int_0^z s^{-m-1}[h(s) - rK]ds + z^n \int_z^\infty s^{-n-1}[h(s) - rK]ds < 0 \ ,$$

which implies (5.4), because otherwise, both integrands would be nonnegative functions and the inequality would not hold.

Finally, suppose that (5.4) is true, and let any $z > 0$ such that $h(z) < rK$. Since h is increasing, given $y < z$,

$$\begin{aligned} f(y) &\leq \frac{1}{\sigma^2} y^n \left[\int_y^z s^{-n-1}[h(z) - rK]ds + \int_z^\infty s^{-n-1}[h(s) - rK]ds \right] \\ &= \frac{h(z) - rK}{\sigma^2 n} + y^n \left[-\frac{[h(z) - rK]z^{-n}}{\sigma^2 n} + \frac{1}{\sigma^2} \int_z^\infty s^{-n-1}[h(s) - rK]ds \right] \\ &\underset{y \downarrow 0}{\longrightarrow} \frac{h(z) - rK}{\sigma^2 n} < 0 \ , \end{aligned}$$

which proves that $f(y)$ is negative for sufficiently small y. On the other hand, f increases to ∞ as $y \to \infty$ (see Lemma 4.5). As a consequence, (5.5) has at least one solution. However, this solution is unique because f is strictly increasing as follows:

$$
\begin{aligned}
f'(y) &= -\frac{1}{\sigma^2}y^{-1}[h(y)-rK] + \frac{n}{\sigma^2}y^{n-1}\int_y^\infty s^{-n-1}[h(s)-rK]ds \\
&= \frac{1}{\sigma^2}y^{n-1}\left[-y^{-n}[h(y)-rK] + n\int_y^\infty s^{-n-1}[h(s)-rK]ds\right] \\
&= \frac{1}{\sigma^2}y^{n-1}\int_y^\infty s^{-n}dh(s) \;>\; 0\ ,
\end{aligned}
$$

where we have used (4.15) with $h(\cdot)-rK$ in place of h. □

We can now prove the main result of this section.

THEOREM 5.2. *Consider the control problem described in section* 2*, and assume that* A4 *holds. We have the following two cases:*

a) *If* $\inf_{x\in]0,\infty[} h(x) \geq rK$*, then the value function* v *is equal to the function* w_p *defined by* (4.6)*.*

b) *If* $\inf_{x\in]0,\infty[} h(x) < rK$*, then the value function* v *is given by*

$$
v(x) = \begin{cases} K & \text{if} \quad x \in [0,y], \\ Ax^m + w_p(x) & \text{if} \quad x \in [y,\infty[, \end{cases} \tag{5.9}
$$

where w_p *is given by* (4.6)*,* y *is the unique solution of* (5.5)*, and the parameter* A *is given by* (5.6)*.*

In both cases, the optimal production process $\tilde{U}$ *is given by* (3.4)*. In the first case, the optimal abandonment time is* $\tilde{\tau} = \infty$*, whereas in the second case, the optimal abandonment time is given by* $\tilde{\tau} = \inf\{t \geq 0 :\ X_t \in [0,y]\}$*.*

Proof. First, note that, in both cases, the candidate value functions belong to $W^{2,\infty}_{loc}(]0,\infty[)$ (see Remark 4.4; also note that, in case b, the function v defined by (5.9) is C^1 at y, by construction).

We now prove that, in both cases, the candidate value functions are nondecreasing functions. In the first case, this has been proved in Lemma 4.5. In the second case, it follows from the fact that, for $x > y$,

$$
v'(x) = x^{m-1}\left[x^{-m+1}w_p'(x) - y^{-m+1}w_p'(y)\right] > 0
$$

because

$$
\begin{aligned}
\frac{d}{dx}\left(x^{-m+1}w_p'(x)\right) &= \frac{1}{\sigma^2}x^{n-m-1}\left[-x^{-n}h(x) + n\int_x^\infty s^{-n-1}h(s)ds\right] \\
&\overset{(4.15)}{=} \frac{1}{\sigma^2}x^{n-m-1}\int_x^\infty s^{-n}dh(s) \;>\; 0\ ,
\end{aligned}
$$

which proves that the function $x \to x^{-m+1}w_p'(x)$ is increasing.

If (5.3) is not true, then it is clear that w_p satisfies the HJB equation (4.1), that it satisfies (3.3) (because of Proposition 4.1e), and that the process M defined as in (3.2) is a square integrable martingale if stopped at any constant time (because of Lemma 3.4 and Proposition 4.1c). Therefore it satisfies the assumptions of the Verification Theorem 3.2, and hence the statements concerning part a) of the theorem follow.

If (5.3) is true, then the function v defined by (5.9) will satisfy the HJB equation (4.1) if

$$
-rK + h(x) \leq 0 \qquad \forall x \in [0,y[\tag{5.10}
$$

and

$$K \leq v(x) \qquad \forall x \in]y, \infty[\ . \tag{5.11}$$

(5.11) follows trivially from the fact that v is nondecreasing. On the other hand, in view of (5.5), if $z := \inf\{s > 0 : \ h(s) > rK\}$, then $y < z$. However, this proves (5.10) because h is increasing. We conclude that w satisfies the HJB equation (4.1) as well as (3.3) (because of Proposition 4.1e) and that the process M defined as in (3.2) is a square integrable martingale if stopped at any constant time (because of Lemma 3.4 and Proposition 4.1c); i.e., it satisfies the requirements of Theorem 3.2, and the proof is complete. □

6. Closed form solution for a specific payoff function. In this section we study the special case of the problem solved in section 5, which arises when the control set is $\mathcal{U} = [0, c]$ for some constant $c > 0$, and the running payoff function $\overline{h}$ is given by

$$\overline{h}(x, u) = [\alpha x - \beta]u - \gamma \ . \tag{6.1}$$

Here, α, β, γ are some given positive constants. With reference to the natural resource industry, β is the extraction cost per unit of the resource, $1 - \alpha$ is proportional to the "royalties," and γ represents the running cost; for further information concerning the choice of these parameters, the reader may consult Brennan and Schwartz [7]. A standing assumption in this section is that

$$r > b \ . \tag{6.2}$$

In the absence of (6.2), it is easy to show that the policy consisting of producing at any constant positive capacity has infinite payoff, and so, Assumptions A2 and A4.b do not hold.

The functions h, u defined by (2.4), (2.7) are now given by

$$h(x) = [\alpha x - \beta]^+ c - \gamma \ , \qquad u(x) = c \, \mathrm{I}_{\{x \geq \beta/\alpha\}} \ ,$$

respectively, and $\overline{h}$ clearly satisfies A4.b. Noting that (6.2) implies that $n > 1$ and using the fact that $mn = -r/\sigma^2$ as well as the identities

$$\sigma^2 n(n-1) = r - bn \ , \qquad \sigma^2 m(m-1) = r - bm \ , \tag{6.3}$$

it is a matter of simple calculations to verify that, in the case that we consider here, the function w_p defined by (4.6) is given by

$$w_p(x) = \frac{\beta c}{(n-m)(r-bn)} \left(\frac{\alpha}{\beta}\right)^n x^n - \frac{\gamma}{r} \ , \tag{6.4}$$

if $x \leq \beta/\alpha$, and by

$$w_p(x) = \frac{\beta c}{(n-m)(r-bm)} \left(\frac{\alpha}{\beta}\right)^m x^m + \frac{\alpha c}{r-b} x - \frac{\beta c + \gamma}{r} \ , \tag{6.5}$$

if $x \geq \beta/\alpha$. Also, the function g which identifies with the left-hand side of (5.8), the solution of which determines the "stopping barrier" y, is given by

$$f(x) = \frac{\beta c}{n(n-1)} x^{-n} \left[\left(\frac{\alpha x}{\beta}\right)^n - \frac{(\gamma + rK)(n-1)}{\beta c} \right] \ ,$$

if $x \leq \beta/\alpha$, and by

$$f(x) = \frac{\alpha c}{n-1} x^{-n} \left[x - \frac{(\beta c + \gamma + rK)(n-1)}{\alpha c n} \right] ,$$

if $x \geq \beta/\alpha$.

With reference to Theorem 5.2, depending on whether abandonment is part of the optimal policy and, if yes, on whether the unique solution y of $f(y) = 0$ is larger than or less than β/α, the optimal policy can take three qualitatively different forms, depending on parameter values. In the first case (the *PW-case*), it is optimal either to produce at full capacity or not to produce at all (i.e., wait); abandonment is never optimal. In the second case (the *PS-case*), at each time instant, it is optimal either to produce at full capacity or to abandon. In the third case (the *PWS-case*), the scenario of optimal actions consists of producing at full capacity, not producing at all, and abandoning. In view of Theorem 5.2 and the preceding calculations, we can now analyze the three cases.

The PW-case. The strategy defined by

$$\tilde{U}_t = c \, \mathrm{I}_{\{X_t \geq \beta/\alpha\}} \quad \forall t \geq 0 \quad \text{and} \quad \tilde{\tau} = \infty$$

is optimal if and only if

$$\inf_{x \in]0,\infty[} h(x) \equiv -\gamma \geq rK \ . \tag{6.6}$$

The value function v coincides with the function w_p defined piecewisely by (6.4) and (6.5).

The PS-case. The strategy defined by

$$\tilde{U}_t = c \ \ \forall t \geq 0 \quad \text{and} \quad \tilde{\tau} = \inf\{t \geq 0 : \ X_t \in]0, y]\} \ ,$$

where

$$y = \frac{(\beta c + \gamma + rK)(n-1)}{\alpha c n} \ ,$$

is optimal if and only if (6.6) is not true and $y \geq \beta/\alpha$, which is equivalent to

$$(\gamma + rK)(n-1) \geq \beta c \ . \tag{6.7}$$

The value function is given by $v(x) = K$ if $x \leq y$ and by

$$v(x) = -\frac{\alpha c}{m(r-b)} y^{-m+1} x^m + \frac{\alpha c}{r-b} x - \frac{\beta c + \gamma}{r}$$

if $x \geq y$.

The PWS-case. The strategy defined by

$$\tilde{U}_t = c \, \mathrm{I}_{\{X_t \geq \beta/\alpha\}} \quad \forall t \geq 0 \quad \text{and} \quad \tilde{\tau} = \inf\{t \geq 0 : \ X_t \in]0, y]\},$$

where

$$y = \frac{\beta}{\alpha} \left[\frac{(\gamma + rK)(n-1)}{\beta c} \right]^{1/n} \ , \tag{6.8}$$

is optimal if and only if (6.6) is not true and $y \leq \beta/\alpha$, or equivalently, if and only if (6.6) and (6.7) do not hold. In this case, the value function v is given by $v(x) = K$ if $x \leq y$, by

$$v(x) = \frac{\beta c}{(n-m)(r-bn)} \left(\frac{\alpha}{\beta}\right)^n \left[x^n - \frac{n}{m} y^{n-m} x^m\right] - \frac{\gamma}{r} \tag{6.9}$$

if $x \in [y, \beta/\alpha]$, and by

$$v(x) = \frac{\beta c}{n-m} \left[\frac{1}{r-bm} \left(\frac{\alpha}{\beta}\right)^m - \frac{n}{m(r-bn)} \left(\frac{\alpha}{\beta}\right)^n y^{n-m}\right] x^m + \frac{\alpha c}{r-b} x - \frac{\beta c + \gamma}{r} \tag{6.10}$$

if $x \in [\beta/\alpha, \infty[$. It is worth noting that if we use (6.8) and the identities

$$\frac{1}{r-bn} = -\frac{m}{r(n-1)}\ , \qquad \frac{1}{r-bm} = -\frac{n}{r(m-1)}\ ,$$

which follow directly from (6.3) and the fact that $mn = -r/\sigma^2$, we can show that (6.9) is equivalent to

$$v(x) = \frac{\gamma + rK}{r(n-m)} \left[n \left(\frac{x}{y}\right)^m - m \left(\frac{x}{y}\right)^n\right] - \frac{\gamma}{r}\ ,$$

whereas (6.10) is equivalent to

$$v(x) = \frac{\beta c n}{r(n-m)} \left[\frac{1}{n-1} \left(\frac{\alpha y}{\beta}\right)^n - \frac{1}{m-1} \left(\frac{\alpha y}{\beta}\right)^m\right] \left(\frac{x}{y}\right)^m + \frac{\alpha c}{r-b} x - \frac{\beta c + \gamma}{r}\ .$$

7. Options on stocks in the natural resource industry. In this section, we consider a company whose total asset value conforms to the model analyzed in sections 5 and 6. We assume that the commodity price is modeled by the geometric Brownian motion defined by (2.1) and A4.a and the running payoff function is given by (6.1) in the preceding section. Moreover, we make the assumption that $K < 0$, which conforms to the idea that $-K$ is the investment's abandonment cost, and we suppose that, at abandonment, the investment is scrapped and therefore becomes totally worthless.

Now, given a function $f \in W^{2,p}_{loc}$, we can show using the Itô–Tanaka formula that $f \circ X$ is a geometric Brownian motion if and only if $f(x) = Cx^k$ for some $C, k \in \mathbb{R}$. Therefore, in view of Theorem 5.2, the total asset value of the company, and therefore its stock price, is not in general a geometric Brownian motion. Plainly, this means that one cannot value options on these stocks using the Black & Scholes formula. However, in the special case that we consider here, the Black & Scholes formula can be used to calculate an approximate value for a European option when the commodity price is high, as well as an upper bound for the value of such an option.

If the commodity price is high relative to the boundary of the "production region," the value function of the company can be approximated by

$$v(x) \approx \frac{\alpha c}{r-b} x - \frac{\beta c + \gamma}{r}\ , \qquad x \gg y \vee (\beta/\alpha)\ , \tag{7.1}$$

because $m < 0$. (Note that this approximation holds for any of the PW, PS, PWS cases.) Also, in the PS or PWS case, the probability that the company is abandoned

within a short time interval can be reasonably neglected. Therefore a European call option on the stock of the company with a *short* time to maturity consisting of payout $(v(X_T)\mathrm{I}_{\{T<\tilde{\tau}\}} - p)^+$ at maturity T is almost like a European call option giving the right to buy $\alpha c/(r-b)$ units of commodity at time T for the price $p + (\beta c + \gamma)/r$. Similarly, a European put option on the investment with strike price p is almost equivalent to a European put option with strike price $p + (\beta c + \gamma)/r$ on $\alpha c/(r-b)$ units of commodity. The price of such options is given by the Black & Scholes formula, so for very large commodity prices, the valuation of European options on stocks in the investment has no problems.

Now, in view of the assumption that $r > b$, the facts that $m < 0$, $1 < n$, and the identities (6.3), we can easily check that, in any of the PW, PS, PWS cases, the second derivative of the value function v is nonnegative, which implies that v is convex. Combining this with the asymptotic expression (7.1), and taking into account that $v(0) = -\gamma/r$ or $K < 0$, depending on whether the PW or the PS, PWS is the case, we can conclude that

$$v(x) \leq \frac{\alpha c}{r-b} x \qquad \forall x \geq 0 \ ,$$

which implies that

$$\big(v(X_T)\mathrm{I}_{\{T<\tilde{\tau}\}} - p\big)^+ \leq \left(\frac{\alpha c}{r-b} X_T - p\right)^+ \qquad \forall T \geq 0 \ , \qquad P\text{-a.s.}$$

As a consequence, the value of a European option on the stock of the company is bounded from above by the value of a European option on the stock of a fictitious company whose stock price is given by $\alpha c/(r-b)X$, which can be calculated by the Black & Scholes formula.

8. Conclusion. An improved model for evaluating natural resources projects and stocks in the natural resource industry has been formulated and studied. This will be useful for both investors who are interested in finding out discrepancies between quoted prices and underlying values of companies, as well as for company executives who are looking for tools to evaluate investment decisions. Here, we have adopted the viewpoint that one can identify asset value with total stock value for a company without debt or outstanding warrants. Our analysis can most easily be applied to companies who (almost) exclusively have producing fields with firmly established reserves with relatively little uncertainty attached. As Paddock, Siegel, and Smith [20] mention, there are a number of publicly quoted companies which more or less fit this profile. Also, the mathematical analysis presented analyzes a much larger class of payoffs than previously possible. With reference to the natural resource industry, it is worth repeating that our model relies on the assumption that the investment/firm under consideration has access to an infinite amount of the resource. Although such an assumption does not hold in the real world, it can be defended as a good approximation of reality by using a number of arguments. Its relaxation is the subject of current research.

We are hereby able to move beyond the natural resource industry to other sectors. One example concerns investments into microchip production facilities. The investment decisions for companies producing microchips fall squarely into the context of our model. In this problem, the X process would be the price of one particular type of microchip; the choice of a geometric Brownian motion with negative drift reflects the fact that microchips of a particular variety are on average declining in price as

their availability increases, the production processes become more widespread, and new microchip generations are introduced. The controlled process U would be the amount of microchips produced per unit of time and the payoff function could be chosen appropriately. In this model, it is plain that the assumption discussed at the end of the previous paragraph presents no difficulties at all because there is no upper limit for the amount of chips producible over an infinite time horizon.

Finally, there is a natural need for comparing our theoretical results with market data. As mentioned in [20], there is enough data available in a variety of forms to be able to conduct some reasonable analysis.

Acknowledgments. We would like to express our gratitude to Mark Davis for numerous helpful discussions during the first stages of this research. We also wish to thank Martin Clark, Lane Hughston, Vassilios Katsouros, Ralph Korn, and Nilay Shah for useful conversations and suggestions.

REFERENCES

[1] A. Bensoussan, M. Crouhy, and D. Galai, *Stochastic equity volatility related to the leverage effect* I: *Equity volatility behaviour*, Appl. Math. Finance, 1 (1994), pp. 63–85.
[2] A. Bensoussan, M. Crouhy, and D. Galai, *Stochastic Equity Volatility Related to the Leverage Effect* II*: Valuation of European Equity Options and Warrants*, preprint, 1994).
[3] A. Bensoussan and J. L. Lions, *Applications des Inéquations Variationnelles au Contrôle Stochastique*, Dunod, Paris, 1978.
[4] D. P. Bertsekas and S. E. Shreve, *Stochastic Optimal Control: The Discrete Time Case*, Academic Press, New York, 1978.
[5] J. M. Bismut and B. Skalli, *Temps d'arrêt optimal, théorie générale des processus et processus de Markov*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, 39 (1977), pp. 301–313.
[6] K. A. Brekke and B. Øksendal, *Optimal switching in an economic activity under uncertainty*, SIAM J. Control Optim., 32 (1994), pp. 1021–1036.
[7] M. J. Brennan and E. S. Schwartz, *Evaluating natural resource investments*, J. Business, 58 (1985), pp. 135–157.
[8] H. Brezis, *Analyse Fonctionnelle*, Masson, Paris, 1992.
[9] G. Cortazar and E. S. Schwartz, *A compound option model of production and intermediate inventories*, J. Business, 66 (1993), pp. 517–540.
[10] M. H. A. Davis and M. Zervos, *A problem of singular stochastic control with discretionary stopping*, Ann. Appl. Probab., 4 (1994), pp. 226–240.
[11] A. Dixit, *Entry and exit decisions under uncertainty*, Journal of Political Economy, 97 (1989), pp. 620–638.
[12] A. K. Dixit and R. S. Pindyck, *Investment under Uncertainty*, Princeton University Press, Princeton, NJ, 1994.
[13] N. El Karoui, *Les Aspects Probabilistes du Contrôle Stochastique*, Lecture Notes in Math. 876, Springer-Verlag, New York, 1981.
[14] A. G. Fakeev, *Optimal stopping rules for stochastic processes with continuous parameter*, Theory Probab. Appl., 15 (1970), pp. 324–331.
[15] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.
[16] I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus*, Springer-Verlag, New York, 1988.
[17] T. Ø. Kobila, *A class of solvable stochastic investment problems involving singular controls*, Stochastics Stochastics Rep., 43 (1993), pp. 29–63.
[18] N. V. Krylov, *Controlled Diffusion Processes*, Springer-Verlag, New York, 1980.
[19] R. L. McDonald and D. R. Siegel, *Investment and the valuation of firms when there is an option to shut down*, Internat. Econom. Rev., 26 (1985), pp. 331–349.
[20] J. L. Paddock, D. R. Siegel, and J. L. Smith, *Option valuation of claims on real assets: The case of offshore petroleum leases*, Quart. J. Econom., 8 (1988), pp. 479–508.
[21] R. S. Pindyck, *Irreversible investment, capacity choice, and the value of the firm*, Amer. Econom. Rev., 79 (1988), pp. 969–985.
[22] D. Revuz and M. Yor, *Continuous Martingales and Brownian Motion*, Springer-Verlag, Berlin, Heidelberg, 1991.

[23] H. Shirakawa, *Evaluation of Investment Opportunity under Entry and Exit Decisions*, Technical report 96-13, Department of Industrial Engineering and Management, Tokyo Institute of Technology, Tokyo, Japan, 1996.
[24] A. N. Shiryayev, *Optimal Stopping Rules*, Springer-Verlag, New York, 1978.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 6, pp. 2103–2127, November 1998

014

MATRIX RATIONAL H_2 APPROXIMATION: A GRADIENT ALGORITHM BASED ON SCHUR ANALYSIS*

PASCALE FULCHERI† AND MARTINE OLIVI†

Abstract. This paper deals with the rational approximation of specified order n to transfer functions which are assumed to be matrix-valued functions in the Hardy space for the complement of the closed unit disk endowed with the L_2-norm. An approach is developed leading to a new algorithm, the first one to our knowledge which concerns matrix-transfer functions in L_2-norm. This approach generalizes the ideas presented in [L. Baratchart, M. Cardelli, and M. Olivi, *Automatica*, 27(1991), pp. 413–418] in the scalar case but involves substantial new difficulties.

Using the Douglas–Shapiro–Shields factorization of transfer functions, the criterion for the rational approximation problem above is expressed in terms of inner matrix functions of McMillan degree n. These functions, which possess a manifold structure, are represented by means of local coordinate maps obtained in [D. Alpay, L. Baratchart, and A. Gombani, *Oper. Theory Adv. Appl.*, 73(1994), pp. 30–66] from a tangential Schur algorithm and for which the coordinates range over n copies of the unit ball. A gradient algorithm is then employed to solve the approximation problem using the coordinate maps to describe the manifold locally and changing from one coordinate map to another when required. However, while processing the gradient algorithm a boundary point can be reached. It is proved that such a point can be considered as an initial point for searching for a local minimum of lower degree while a local minimum of McMillan degree $k < n$ provides a starting point for searching for a local minimum at degree $k + 1$. The minimization process then pursues through different degrees. The convergence of this algorithm to a local minimum of appropriate degree is proved and demonstrated on a simple example.

Key words. rational approximation, identification, discrete time systems, inner matrices, gradient algorithm, Schur analysis

AMS subject classifications. Primary 41A20, 93B30; Secondary 47N70, 93B29

PII. S0363012995284230

1. Introduction.

The identification of linear time-invariant systems can be formalized as a rational approximation problem in which some criterion function is optimized over a set of systems. This approach has led to a wide variety in model structure, performance criteria, and actual methods of estimation (see [38] and the bibliography therein). Our interest is focused mainly on the particular class of discrete time, linear, time-invariant, and strictly causal systems and their strictly proper transfer functions. The *order* of such a system is defined to be its McMillan degree, that is, the dimension of the state space in its minimal realizations. The criterion which is chosen here is the L_2-norm, and our approximation problem states in the Hardy space $\bar{H}_2^{p\times m}$ of the complement of the unit disk: *given a transfer function* $F \in \bar{H}_2^{p\times m}$*, we are concerned in minimizing*

$$\|F - H\|_2^2 = \frac{1}{2\pi}\mathrm{Tr}\int_0^{2\pi} [F-H](e^{it})[F-H](e^{it})^* dt\,, \tag{1}$$

as H ranges over the set of rational stable (i.e., analytic for $|z| > 1$) functions of order at most n. Here, the symbol Tr stands for the trace and the superscript $*$ denotes transpose-conjugate. It should be noticed that in a stochastic framework, (1) is equal

*Received by the editors April 5, 1995; accepted for publication (in revised form) January 5, 1998; published electronically August 31, 1998.
http://www.siam.org/journals/sicon/36-6/28423.html
†INRIA, BP 93, 06902 Sophia-Antipolis Cedex, France (olivi@sophia.inria.fr).

to the mean square error between the output of a given system and the output of a model of fixed order when both systems have the same white noise input (see [32]).

The above problem has received attention in [41], [40], [4], and more recently [37] and [30], either in the discrete-time form studied here or in the continuous-time equivalent. Many qualitative results have been proved in [6], such as the existence of a best approximant and the property usually called normality: if F is not itself rational of degree at most n, then a best local approximant H has degree exactly n. In [8], an algorithm to find local minima in the L_2 rational approximation problem is described for scalar systems. It is the purpose of this paper to present an algorithm which enables the results of this previous paper to be extended to the multivariable case.

Let us recall the main line of our approach in the scalar case (see [8] and [13]). In this case, the minimum in (1) must be performed over the set of irreducible fractions p/q, where q is a polynomial of degree n whose roots belong to the unit disk $\mathbb{U}$. Our optimization problem being linear with respect to the parameters of the numerator p, we are led to minimize a cost function Ψ^n defined on the set $\mathcal{P}_n^1$ of monic polynomials q of degree n whose roots belong to $\mathbb{U}$. This set can be described by the coefficients of q and is open and bounded in $\mathbb{R}^n$; the function Ψ^n is smooth, so that we can use a gradient algorithm, producing a sequence of improving estimates, which either converges to a local minimum or meets the boundary of the domain at some point having some roots of modulus one. However, roots on the unit circle cancel and the cost function Ψ^n extends to a neighborhood of the closure of $\mathcal{P}_n^1$. At the boundary the extension of Ψ^n can be interpreted as the cost function of a lower-order approximation problem. Thus the search for a local minimum can continue through different orders, until such a minimum (of order $k \leq n$) is actually reached. Conversely, multiplying by $z-1$ or $z+1$ a minimum of order k provides an initial point for the optimization problem at order $k+1$: at such a point, the opposite of the gradient points inside the domain. Finally, the procedure can continue until a local minimum of order n is actually found.

Transition to the multivariable case involves substantial new difficulties, mainly due to the fact that the domain of the cost function is no longer an open subset of a Euclidean space but it does possess a manifold structure. A manifold has a covering by countably many open coordinate neighborhoods, each of these coordinate neighborhoods corresponding to an open subset of some $\mathbb{R}^d$ by a local coordinate homeomorphism (d is then the dimension of the manifold). The methods developed for the Euclidean case then apply to each of the coordinate neighborhoods separately. Over a manifold, an optimization problem can be tackled by using a search algorithm through the manifold as a whole, using the coordinate maps to describe the manifold locally and changing from one coordinate map to another when required. Such a representation of the elements of the domain has the advantage to get rid of redundancy and ensure identifiability [22]. Using state space representations, it was first established by Hazewinkel and Kalman [26] that the set of stable transfer functions of fixed degree possesses a manifold structure. Several atlases of local coordinate maps (called sets of overlapping canonical forms in system theory) have been derived from this approach ([33], [25]). However, this manifold is never compact, and convergence of a gradient algorithm to points outside can occur. To avoid this problem, a transfer function will be represented by means of the inner-unstable or Douglas–Shapiro–Shields factorization (see [15]). The elimination of the parameters in which the system is linear (namely, those of the unstable factor) allows us to perform the search for an optimum on the manifold of *inner matrix functions* of degree n. We are

then in a position to proceed to the generalization to the multivariable case of the above-mentioned procedure.

The paper is organized as follows: Section 2 states the problem within the framework of the Hardy spaces and introduces the cost function by means of the inner-unstable factorization. In section 3, we first recall some results of [1], in which the theory of reproducing kernel Hilbert spaces is used to construct local coordinates of the manifold of inner matrix functions of fixed McMillan degree: such functions are obtained by iterating a linear fractional transformation which changes an inner function into another one, the McMillan degree being increased by one. Then, a fractional representation of this transformation is given in which the numerator (a polynomial matrix) and the denominator (a polynomial) are polynomial functions in the local coordinates. This representation allows us, in section 4, to study the cost function on the boundary of the domain and to elaborate an algorithm which converges generically to a local minimum. The numerical aspects have been examined in section 5.

2. Minimizing over the set of inner matrices. The Hardy spaces H_2 and H_∞ of the unit disk are the closed subspaces of $L_2(\mathbb{T})$ and $L_\infty(\mathbb{T})$, respectively, consisting of functions whose Fourier coefficients (a_n) satisfy $a_n = 0$ when $n < 0$; while the Hardy space $\bar{H}_{2,0}$ consists of functions for which $a_n = 0$ when $n \geq 0$. Note the orthogonal decomposition

$$L_2(\mathbb{T}) = H_2 \oplus \bar{H}_{2,0}.$$

It is well known (see, e.g., [27]) that members of H_2 are the nontangential limits on $\mathbb{T}$ of analytic functions f in the unit disk for which the functions $f_r(t) = f(re^{it}),\ r < 1$, are bounded in L_2-norm as $r \to 1$. Members of H_∞ correspond to bounded holomorphic functions in this process. Similarly, members of $\bar{H}_{2,0}$ correspond to analytic functions f in the complement of the unit disk vanishing at infinity and satisfying an analogous growth condition for $r > 1$. Thus, f belongs to H_2 (resp., to $\bar{H}_{2,0}$) if and only if it can be written as

$$f(z) = \sum_{k\geq 0} a_k\ z^k \qquad \left(\text{resp.,}\quad f(z) = \sum_{k>0} a_k\ z^{-k} \right), \qquad \sum |a_k|^2\ <\ \infty. \tag{2}$$

Note that (2) is the Taylor expansion at 0 (resp., at ∞) and at the same time the Fourier expansion if we substitute $z = e^{i\theta}$.

The space $L_2^{p\times m}(\mathbb{T})$ of $(p \times m)$-matrices whose entries belong to $L_2(\mathbb{T})$ becomes a Hilbert space when endowed with the scalar product

$$\langle F, G\rangle = \frac{1}{2\pi}\mathrm{Tr}\int_0^{2\pi} F(e^{it})G(e^{it})^*\, dt. \tag{3}$$

The corresponding norm will also be given, for $F = (f_{ij})$, by $\|F\|_2^2 = \sum_{i,j} \|f_{ij}\|_2^2$, and the orthogonal decomposition

$$L_2^{p\times m}(\mathbb{T}) = H_2^{p\times m} \oplus \bar{H}_{2,0}^{p\times m} \tag{4}$$

is still valid. Taking into account the fact that $\bar{z} = z^{-1}$ on $\mathbb{T}$, and using the notation

$$G^\sharp(z) = G(1/\bar{z})^*,$$

(3) may be converted into the line integral

$$\langle F, G\rangle = \frac{1}{2i\pi}\mathrm{Tr}\int_{\mathbb{T}} G^\sharp(z)F(z)\frac{dz}{z}.$$

The Banach space $L_\infty^{p\times m}(\mathbb{T})$ is endowed with the norm

$$\|F\|_\infty = \sup_\theta ||F(e^{i\theta})||,$$

where $||\ .\ ||$ denotes the operator norm $\mathbb{C}^m \to \mathbb{C}^p$. The prefix $\mathcal{R}$ in front of the name of some set ($\mathcal{R}\bar{H}_{2,0}^{p\times m}$, $\mathcal{R}H_2^{p\times m}$, etc.) will indicate that we consider the *real* subspace of functions whose Fourier coefficients are real. Such functions are relevant in most applications. However, the natural framework for our study is the complex case which plainly includes the real case by restriction. When necessary, the results will be stated for real transfer functions.

The normality result mentioned in the introduction allows us to state the rational approximation problem in degree n as follows: *Given $F \in \bar{H}_{2,0}^{p\times m}$, minimize* (1) *over the set $\Sigma_{p,m}^-(n)$ of rational stable functions of McMillan degree exactly n.* It is well known that $\Sigma_{p,m}^-(n)$ possesses the structure of a real analytic manifold of dimension $2n(m+p)$ (see, e.g., [24]). We shall now give a description of this set which suits our purpose by using the inner-unstable or Douglas–Shapiro–Shields factorization (see [15] and [11]).

Recall that a $\mathbb{C}^{p\times p}$-valued analytic function Q in the unit disk is called *inner* if it is analytic in $\mathbb{U}$ and takes unitary values on the unit circle $\mathbb{T}$:

$$Q(e^{it})\ Q(e^{it})^* \ = \ Q(e^{it})^*\ Q(e^{it}) \ = \ I_p, \tag{5}$$

where I_p is the identity matrix of size p. This equality implies that the inverse of a rational inner functions agrees with $Q^\sharp$ and thus is analytic outside the unit disk. Naturally associated with Q is the space $QH_2^p \subset H_2^p$ which is invariant by the shift operator (i.e., multiplication by z), and its orthogonal complement $\mathcal{H}(Q)$. Note that $\mathcal{H}(Q)$ consists of vectors $v \in H_2^p$ of the form Qu for some u in $\bar{H}_{2,0}^p$. These spaces and the inner-unstable factorization are closely related to the shift realization (see [19]). Observe that the McMillan degree of a rational matrix may be defined even if this matrix fails to be analytic at infinity, using, for instance, Smith–McMillan forms (see [28]). Furthermore, the McMillan degrees of Q and Q^{-1} agree.

Proposition 1 (inner-unstable factorization). *Any rational function H in $\bar{H}_{2,0}^{p\times m}$ can be written*

$$H \ = \ Q^{-1}\ C, \tag{6}$$

where Q is a $(p\times p)$-rational inner function and C a $(p\times m)$-rational matrix whose columns belong to $\mathcal{H}(Q)$. The matrices Q and C may be chosen left co-prime. With this condition, the factorization is unique up to a common left unitary factor and Q and H have same McMillan degree.

The matrix Q is called the *left inner factor* of H and the matrix Q^{-1} is usually named in system theory an *all-pass stable transfer function.* To ensure uniqueness in the inner-unstable factorization, we shall require that Q satisfies the condition

$$Q(1) = I_p. \tag{7}$$

The set of $\mathbb{C}^{p\times p}$-valued rational inner functions of degree n will be denoted by $\mathcal{I}_n^p$, and by $\mathcal{I}_n^p(1)$ we denote the subset of functions satisfying the extra condition (7). As

previously mentioned, $\mathcal{RI}_n^p$ and $\mathcal{RI}_n^p(1)$ will denote the corresponding sets of *real* inner functions. It is proved in [1] that $\mathcal{I}_n^p$ and $\mathcal{I}_n^p(1)$ are smooth manifolds of dimension $2np + p^2$ and $2np$, respectively (embedded in $H_\infty^{p\times p}$), while $\mathcal{RI}_n^p$ and $\mathcal{RI}_n^p(1)$ have dimension $np + p(p-1)/2$ and np, respectively. Moreover, the set $\Sigma_{p,m}^-(n)$ is a vector bundle whose base space is $\mathcal{I}_n^p(1)$ and whose fiber above Q is the vector space $\mathcal{F}_Q$ of matrices C whose columns belong to $\mathcal{H}(Q)$ (see [12]).

Now, we can write our approximation problem as

$$\min_{Q,C} \|F - Q^{-1}C\|_2^2,$$

where $Q \in \mathcal{I}_n^p(1)$ and $C \in \mathcal{F}_Q$. Observe that for fixed Q, the minimum is obtained when C is the projection of QF onto $\mathcal{F}_Q$. Since $F \in \bar{H}_{2,0}^{p\times m}$, C is also the projection of QF onto $H_2^{p\times m}$ that we shall denote by $L(Q)$. Therefore, minimizing (1) is equivalent to minimizing the function

$$\begin{array}{rccl} \Psi^n : & \mathcal{I}_n^p(1) & \to & \mathbb{R}, \\ & Q & \to & \|F - Q^{-1}L(Q)\|_2^2, \end{array} \tag{8}$$

which is going to be the main purpose of the remainder of this paper. The first step consists of studying the domain of this function and will be the content of the next section.

First of all, we give a fractional representation of an inner matrix which will be useful in the sequel. If q is a polynomial of degree n, we define its *reciprocal polynomial* as being

$$\widetilde{q}(z) = z^n\ q^\sharp(z), \tag{9}$$

and if D is a polynomial matrix whose degree does not exceed n, we also put

$$\widetilde{D}(z)\ =\ z^n\ D^\sharp(z). \tag{10}$$

Recall that the degree of a polynomial matrix is defined to be the degree of its highest degree entry. While both this degree and the McMillan degree are used in this work, there should be no confusion from the context which is used.

PROPOSITION 2. *An inner matrix $Q \in \mathcal{I}_n^p$ has a representation of the form $Q = D/\widetilde{q}$ by means of a polynomial matrix D whose degree does not exceed n and a polynomial q of exact degree n whose roots belong to the open unit disk, satisfying $D\widetilde{D} = q\widetilde{q}I_p$ and $\det D = \epsilon q\widetilde{q}^{p-1}$, ϵ being a complex number of modulus one. Conversely, these conditions are sufficient for the rational matrix $D/\widetilde{q}$ to belong to $\mathcal{I}_n^p$.*

Proof. Since Q^{-1} is analytic outside the unit disk, it has a representation of the form $\widetilde{D}/q$, where q is, up to a constant factor, its *polynomial of poles* (see [28]). Condition (5) yields an analogous representation for Q, i.e., $Q = D/\widetilde{q}$, so that $D\widetilde{D} = q\widetilde{q}I_p$. It also implies that $\det Q$ is an inner scalar rational function, that is to say a Blaschke product, and the number of zeros of $\det Q$ within $\mathbb{U}$ determines the McMillan degree of Q, by the Potapov decomposition [35]. ∎

3. Parametrization of inner matrices. We describe here a parametrization of the set of inner functions obtained in [1] from a matrix version of the classical Schur algorithm that we now explain; in a fundamental paper [36], Schur proved that every function $f \in \mathcal{I}_n^1$ can be uniquely parameterized by a sequence y_j, $j = n, \ldots, 1$, of

complex numbers with $|y_j| < 1$. Moreover, Schur gave an algorithm for computing these parameters:

$$y_j = f_j(0),$$

where $f_n = f$ and

$$f_{j-1}(z) = \frac{f_j(z) - f_j(0)}{(1 - \overline{f_j(0)} f_j(z))\, z}, \quad j = n, \dots, 1. \tag{11}$$

Since f_j is an inner function it follows from the maximum modulus principle that $|y_j| < 1$, and f_j has degree j, since a zero is eliminated at each step. Since f has degree n, f_0 is equal to a constant of modulus one. Other sequences of inner functions of decreasing degree may be constructed from f in a similar way. The most general recursion formula is the following (see [21]):

$$\frac{f_{j-1}(z) + \mu_j}{1 + \bar{\mu}_j f_{j-1}(z)} = \frac{f_j(z) - y_j}{1 - \overline{y_j} f_j(z)} \frac{1 - \bar{w}_j z}{z - w_j}, \quad j = n, \dots, 1,$$

where the w_j's are the interpolation points, $y_j = f_j(w_j)$ and the μ_j's belong to $\mathbb{U}$. The w_j's and the μ_j's being given, the sequence of numbers y_j completely characterizes the function f, which can recover inductively by the linear fractional transformations:

$$f_j(z) = \frac{[(z - w_j) + \bar{\mu}_j y_j (1 - \bar{w}_j z)]\, f_{j-1}(z) + [\mu_j (z - w_j) + y_j (1 - \bar{w}_j z)]}{[\bar{y}_j (z - w_j) + \bar{\mu}_j (1 - \bar{w}_j z)]\, f_{j-1}(z) + [\mu_j \bar{y}_j (z - w_j) + (1 - \bar{w}_j z)]}. \tag{12}$$

The map $(y_1, \dots, y_n, f_0) \to f$ is a diffeomorphism from the product of n copies of the open unit disk and of a copy of the unit circle onto $\mathcal{I}_n^1$.

This Schur algorithm is related to the classical interpolation problems of Nevanlinna–Pick and Carathéodory–Fejér (see [39]), which have a remarkable diversity of applications in systems engineering (see [5], [29]). Several approaches allow the extension of these problems to matrix-valued analytic functions (see [42] and the bibliography therein); however, the operator-theoretic one, involving reproducing kernel Hilbert spaces, clarifies the connections between interpolation and realization theory and gives a unified presentation of these problems (see, e.g., [17], [16], [3]). Another fundamental treatment can be found in [18], which emphasizes the relevance of the commutant lifting theorem in these interpolations issues and also presents several applications to engineering problems.

3.1. Reproducing kernel Hilbert spaces. For the convenience of the reader, we shall recall some basic facts about reproducing kernel Hilbert spaces which may be found in [16]. A complex Hilbert space $\mathcal{H}$ of $\mathbb{C}^p$-valued functions defined on some Ω open in $\mathbb{C}$ is called a *reproducing kernel Hilbert space* (RKHS) if there exists a $\mathbb{C}^{p\times p}$-valued function $K(z, w)$ defined on $\Omega \times \Omega$ such that for every choice of $w \in \Omega$, $c \in \mathbb{C}^p$ and $f \in \mathcal{H}$:

(i) $K(., w)c \in \mathcal{H}$,

(ii) $\langle f, K(., w)c\rangle_{\mathcal{H}} = c^* f(w)$.

The function K is called the *reproducing kernel,* and the main facts are that it is unique and it is a *positive function* in the following sense:

$$\sum_{i,j=1}^{r} c_j^* K(w_j, w_i) c_i \geq 0 \tag{13}$$

for every choice of points $w_1, w_2, \ldots, w_r \in \Omega$, and vectors $c_1, c_2, \ldots, c_r \in \mathbb{C}^p$.

The Hardy space H_2^p is clearly a reproducing kernel Hilbert space whose kernel is

$$\frac{I_p}{1-\bar{w}z}, \quad w \in \mathbb{U}, \quad z \in \mathbb{U},$$

and property (ii) is just the Cauchy formula. Finite dimensional Hilbert spaces of $\mathbb{C}^p$-valued functions are also reproducing kernel Hilbert spaces. Let $(f_1, f_2, \ldots, f_N)$ be some base of a finite dimensional Hilbert space. Then its reproducing kernel is easily computed to be

$$K(z,w) = (f_1(z), f_2(z), \ldots, f_N(z))P^{-1}(f_1(w), f_2(w), \ldots, f_N(w))^*,$$

where $P = (P_{ij})$ is the Gram matrix with entries $P_{ij} = \langle f_j, f_i \rangle$. The space $\mathcal{H}(Q)$ introduced in the previous section as being the orthogonal complement of QH_2^p in H_2^p is a reproducing kernel Hilbert space with reproducing kernel

$$K_Q(z,w) = \frac{I_p - Q(z)Q(w)^*}{1-\bar{w}z}, \tag{14}$$

which is the projection onto $\mathcal{H}(Q)^p$ of the reproducing kernel of H_2^p. This is readily seen with the help of the evaluation

$$\pi^+\left(Q(z)^{-1}\frac{I_p c}{1-\bar{w}z}\right) = \frac{Q(w)^* c}{1-\bar{w}z}, \tag{15}$$

where π^+ denotes the orthogonal projection onto H_2^p.

More generally, a RKHS is attached to every J-inner function. The study of these spaces, which play a central role in the theory of realization and interpolation, originates with de Branges and Rovnyak (see [14]). Put

$$J = \begin{pmatrix} I_p & 0 \\ 0 & -I_p \end{pmatrix}.$$

A $\mathbb{C}^{2p\times 2p}$-valued rational function Θ is called J-inner if at every point of analyticity of Θ in $\mathbb{U}$, $J - \Theta(z)J\Theta(z)^*$ is positive semidefinite:

$$\Theta(z)J\Theta(z)^* \leq J, \tag{16}$$

and equality holds for z point of analyticity on $\mathbb{T}$. Consider the space H_2^{2p} endowed with the sesquilinear Hermitian form, $\langle f, g \rangle_J = \langle f, Jg \rangle$. This form is not positive definite but it is nondegenerate. Hence, the space ΘH_2^{2p} has an orthogonal complement in H_2^{2p}, which we call $\mathcal{H}(\Theta)$. Restricted to $\mathcal{H}(\Theta)$, the form $\langle ., . \rangle_J$ is positive definite, so that $\mathcal{H}(\Theta)$ is a Hilbert space. Moreover, it is a reproducing kernel Hilbert space with reproducing kernel

$$K_\Theta(z,w) = \frac{J - \Theta(z)J\Theta(w)^*}{1-\bar{w}z} \tag{17}$$

and the dimension of $\mathcal{H}(\Theta)$ agrees with the McMillan degree of Θ. In the next section, we shall make an intensive use of one-dimensional $\mathcal{H}(\Theta)$ spaces; let f be the function defined by

$$f(z) = \frac{\begin{pmatrix} u \\ v \end{pmatrix}}{1-\bar{w}z},$$

where $w \in \mathbb{U}$, $u \in \mathbb{C}^p$ with $\|u\| = 1$, and $v \in \mathbb{C}^p$; let $\mathcal{M}$ be the linear span of f endowed with the form $\langle\ ,\ \rangle_J$. If $\|v\| \leq 1$, then the Gram matrix $P = \langle f, f\rangle_J$ is positive and $\mathcal{M}$ is of the form $\mathcal{H}(\Theta)$, where Θ is unique up to a J-unitary constant multiplier on the right. It can be specified by the formula

$$\Theta(z) \ = \ I_{2p} - (1 - \bar{\xi}z) f(z) P^{-1} f(\xi)^* J$$

for any point $\xi \in \mathbb{T}$. In the sequel, we shall work with the J-inner function associated with $\mathcal{M}$ which satisfies the condition $\Theta(1) = I_{2p}$. It is given by

$$\Theta(w,u,v)(z) \ = \ I_{2p} - (1-z)\frac{1-|w|^2}{1-\|v\|^2}\frac{\begin{pmatrix} u \\ v \end{pmatrix}}{1-\bar{w}z}\frac{\begin{pmatrix} u^* & v^* \end{pmatrix}}{1-w} J. \tag{18}$$

3.2. The linear fractional transformation associated with a *J*-inner function. In this section we introduce the linear fractional transformation T_Θ, which generalizes (12) to the matrix case (for a precise comparison see the remark after Theorem 7). The statements and the proofs of this section and the following are adapted from [1].

LEMMA 3. *Let*

$$\Theta \ = \ \begin{pmatrix} \Theta_{11} & \Theta_{12} \\ \Theta_{21} & \Theta_{22} \end{pmatrix} \tag{19}$$

be a $(2p \times 2p)$-rational J-inner function analytic in $\mathbb{U}$ and let A be a $(p \times p)$-rational inner function. Then $(\Theta_{21}A + \Theta_{22})$ is invertible in $\mathbb{U}$ and

$$T_\Theta(A) \ = \ (\Theta_{11}A + \Theta_{12})\ (\Theta_{21}A + \Theta_{22})^{-1} \tag{20}$$

is inner. Note that if $\Theta(1) = I_{2p}$ and $A(1) = I_p$, then $[T_\Theta(A)](1) = I_p$, and if A and Θ have real coefficients, then $T_\Theta(A)$ also has real coefficients.

Proof. First, let us show that $(\Theta_{21}A + \Theta_{22})$ is invertible at every point of $\mathbb{U}$. Indeed, condition (16) implies

$$\Theta_{22}\Theta_{22}^* \ \geq \ I_p + \Theta_{21}\Theta_{21}^* \quad \text{in } \mathbb{U},$$

so that $\Theta_{22}\Theta_{22}^*$ is positive definite, and Θ_{22} is invertible at any point of $\mathbb{U}$. Now, we have

$$I_p \ \geq \ \Theta_{22}^{-1}(\Theta_{22}^{-1})^* + (\Theta_{22}^{-1}\Theta_{21})(\Theta_{22}^{-1}\Theta_{21})^* \quad \text{in } \mathbb{U},$$

and thus $\|\Theta_{22}(z)^{-1}\Theta_{21}(z)\| \ < \ 1,\ \forall z \in \mathbb{U}$. The matrix A, being inner, is contractive in $\mathbb{U}$: $\|A(z)\| \ \leq \ 1,\ \forall z \in \mathbb{U}$, so that $\|\Theta_{22}(z)^{-1}\Theta_{21}(z)A(z)\| \ < \ 1,\ \forall z \in \mathbb{U}$. Finally, $(\Theta_{21}A + \Theta_{22}) \ = \ \Theta_{22}\ (I_p + \Theta_{22}^{-1}\Theta_{21}A)$ is invertible at any point of $\mathbb{U}$, and thus $B = T_\Theta(A)$ is analytic in $\mathbb{U}$. Then, condition (5) for B can be written

$$B^*B - I_p \ = \ \begin{pmatrix} B^* & I_p \end{pmatrix} J \begin{pmatrix} B \\ I_p \end{pmatrix} \ = \ 0 \quad \text{on } \mathbb{T}.$$

Using the relation

$$\begin{pmatrix} B \\ I_p \end{pmatrix} \ = \ \Theta \begin{pmatrix} A \\ I_p \end{pmatrix} (\Theta_{21}A + \Theta_{22})^{-1}, \tag{21}$$

we obtain

$$B^*B - I_p \ = \ ((\Theta_{21}A + \Theta_{22})^{-1})^* \ \begin{pmatrix} A^* & I_p \end{pmatrix} \ \Theta^* J \Theta \ \begin{pmatrix} A \\ I_p \end{pmatrix} (\Theta_{21}A + \Theta_{22})^{-1},$$

and since condition (5) is satisfied for A, it will be satisfied for B as well. ∎

LEMMA 4. *The matrix* $B = T_{\Theta(w,u,v)}(A)$, *where* $\Theta(w,u,v)$ *is given by* (18), *satisfies the interpolation condition*

$$B(w)^* u = v. \tag{22}$$

Proof. Indeed, it can be verified that $\Theta(w,u,v)$ satisfies the equation

$$\begin{pmatrix} u^* & -v^* \end{pmatrix} \Theta(w) = 0.$$

Thus

$$\begin{pmatrix} u^* & -v^* \end{pmatrix} \Theta(w) \begin{pmatrix} A(w) \\ I_p \end{pmatrix} = 0,$$

and together with (21) this implies our interpolation condition. ∎

Now, the question is the converse: let B be some rational inner matrix, and Θ J-inner analytic in $\mathbb{U}$. Can we write B in the form $B = T_\Theta(A)$ for some inner matrix A? First, note that if $B = T_\Theta(A)$, then A is the rational function given by

$$A = (\Theta_{11} - B\Theta_{21})^{-1}(B\Theta_{22} - \Theta_{12}), \tag{23}$$

unless $\det(\Theta_{11} - B\Theta_{21})$ vanishes identically. This may not happen since condition (16) for Θ implies $\Theta_{11}^*\Theta_{11} - \Theta_{21}^*\Theta_{21} = I_p$ on $\mathbb{T}$. So, $\Theta_{11} - B\Theta_{21}$ is invertible at any point of $\mathbb{T}$. However, it may fail to be invertible at some point of $\mathbb{U}$, so that A may not be analytic in $\mathbb{U}$. To ensure analyticity, we must make an additional assumption.

THEOREM 5. *Let* B *be a rational inner function, and let* $\Theta(w,u,v)$ *be the* J*-inner function* (18). *There exists an inner function* A *such that* $B = T_\Theta(A)$ *if and only if the interpolation condition* $B(w)^* u = v$ *is satisfied. We then have* $\deg B = \deg A + 1$.

Proof. This is a special case of a more general result proved in [1, Thm. 3.3] which is based on the links between $\mathcal{H}(\Theta)$ and $\mathcal{H}(Q)$ spaces. For more details on these problems we refer the reader to [2]. The content of the result of [1] is the following: let $B \in \mathcal{I}_n^p$ and Θ be a J-inner $(2p \times 2p)$-rational function; consider the map

$$\begin{array}{rccl} \tau : & \mathcal{H}(\Theta) & \to & H_2^p, \\ & \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} & \to & f_1 - Bf_2. \end{array}$$

Then there exists an inner function A such that $B = T_\Theta(A)$ if and only if τ is an isometry from $\mathcal{H}(\Theta)$ to $\mathcal{H}(B)$. Moreover, $\deg B = \deg\Theta + \deg A$. We shall admit this result.

If $\Theta(w,u,v)$ is given by (18), the conditions:

(i) τ is an isometry from $\mathcal{H}(\Theta)$ to $\mathcal{H}(B)$,

(ii) $B(w)^* u = v$

are equivalent. Indeed, τ sends the generator of $\mathcal{H}(\Theta)$ as follows:

$$\tau : \frac{\begin{pmatrix} u \\ v \end{pmatrix}}{1 - \bar{w}z} \to \frac{u - B(z)\, v}{1 - \bar{w}z}.$$

With the help of the evaluation (15), it is readily proved that $\frac{u - B(z)\, v}{1 - \bar{w}z} \in \mathcal{H}(B)$ if and only if condition (ii) holds. In this case,

$$\left\langle \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}, \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} \right\rangle_J = \|f_1\|_2^2 - \|f_2\|_2^2 = \|f_1\|_2^2 - \|Bf_2\|_2^2 = \|f_1 - Bf_2\|_2^2,$$

and τ is an isometry from $\mathcal{H}(\Theta)$ (endowed with the scalar product $\langle\ ,\ \rangle_J$) to $\mathcal{H}(B)$. This proves the theorem. ∎

3.3. Description of the charts. If Θ is of the form (18), then $T_\Theta(A)$ and A have the same value at $z = 1$. We can construct from the identity matrix I_p, using n linear fractional transformations of this type, an inner matrix of degree n which belongs to $\mathcal{I}_n^p(1)$. Conversely, any matrix of $\mathcal{I}_n^p(1)$ can be obtained in this way. This is the content of the tangential Schur algorithm for which we need the following lemma.

LEMMA 6. *Let B be an inner function and $w \in \mathbb{U}$. Then, there exists $u \in \mathbb{C}^p$, $\|u\| = 1$, such that*

$$\|B(w)^* u\| < 1.$$

Proof. Suppose that for all unit vector $u \in \mathbb{C}^p$, $\|B(w)^* u\| = 1$. Then, since

$$\|K_B(.,w)u\|_2^2 \;=\; u^* K_B(w,w)u \;=\; \frac{1 - \|B(w)^* u\|}{1 - \bar{w}w},$$

for all $u \in \mathbb{C}^p$, $K_B(.,w)u \;=\; 0$. So, $K_B(.,w)$ is identically 0 and the matrix B must be constant. But this contradicts the fact that B has McMillan degree n. ∎

THEOREM 7 (tangential Schur algorithm). *Let $Q \in \mathcal{I}_n^p(1)$, and $w_k \in \mathbb{U}$, $k = n, \ldots, 1$. Then, for $k = n, \ldots, 1$ there exist unit vectors $u_k \in \mathbb{C}^p$ such that the vectors $y_k \in \mathbb{C}^p$ given by*

$$y_k = Q^{(k)}(w_k)^* u_k \tag{24}$$

satisfy $\|y_k\| < 1$, where $Q^{(n)} = Q$,

$$Q^{(k)} \;=\; T_{\Theta_k}(Q^{(k-1)}), \tag{25}$$

and $\Theta_k = \Theta(w_k, u_k, y_k)$ is the J-inner matrix given by (18). *Then*

$$Q \;=\; T_{\Theta_n}(T_{\Theta_{n-1}} \ldots T_{\Theta_1}(I_p)) \ldots \;=\; T_{\Theta_n \ldots \Theta_1}(I_p).$$

Proof. This is an obvious consequence of Theorem 5 and Lemma 6. ∎

Let $\mathbf{w} = (w_1, w_2, \ldots, w_n)$ and $\mathbf{u} = (u_1, u_2, \ldots, u_n)$. Define the subset $\mathcal{V}_{(\mathbf{w},\mathbf{u})}$ of $\mathcal{I}_n^p(1)$ by

$$\mathcal{V}_{(\mathbf{w},\mathbf{u})} \;=\; \{Q \in \mathcal{I}_n^p(1) \;/\; \|Q^{(k)}(w_k)^* u_k\| < 1\},$$

and the function $\varphi_{(\mathbf{w},\mathbf{u})}$ by

$$\begin{array}{rccl} \varphi_{(\mathbf{w},\mathbf{u})} \;:\; & \mathcal{V}_{(\mathbf{w},\mathbf{u})} & \to & \mathcal{B}_p^n, \\ & Q & \to & (y_1, y_2, \ldots, y_n), \end{array}$$

where the $Q^{(k)}$'s and the *Schur parameters* y_k's are defined recursively by (24) and (25), and $\mathcal{B}_p^n$ denotes the product of n copies of the unit ball of $\mathbb{C}^p$.

Remark. Note that in the scalar case and for $w_j = 0$, the transformation $Q^{(j)} = T_{\Theta_j}(Q^{(j-1)})$ is given by

$$Q^{(j)}(z) = \frac{(z - |y_j|^2)\, Q^{(j-1)}(z) + (1 - z)u_j \bar{y}_j}{(z-1)\bar{u}_j y_j\, Q^{(j-1)}(z) + (1 - |y_j|^2 z)}.$$

This formula is exactly (12) in which y_j has been replaced by $u_j\bar{y}_j$, since the interpolation condition is now $Q^{(j)}(0) = u_j\bar{y}_j$, and μ_j is chosen to be $-u_j\bar{y}_j$. The general formula with an arbitrary μ_j can also be obtained by a T_Θ transformation where Θ is of the form (18) multiplied by an adequate constant J-inner function. In this case the normalization (7) is not conserved.

THEOREM 8. *The family $(\mathcal{V}, \varphi)$ defines a C^∞ atlas on $\mathcal{I}_n^p(1)$, which is compatible with its natural structure of embedded submanifold of $H_2^{p\times p}$.*

Proof. It follows from Lemma 6 that the collection of sets $\mathcal{V}_{(\mathbf{w},\mathbf{u})}$ covers $\mathcal{I}_n^p(1)$. It remains to prove that the map $\varphi_{(\mathbf{w},\mathbf{u})}$ is an homeomorphism and that the change of chart $\varphi_{(\mathbf{w},\mathbf{u})} \circ \varphi_{(\mathbf{w}',\mathbf{u}')}$ is smooth. The map $\varphi_{(\mathbf{w},\mathbf{u})}$ is one-to-one and onto by construction. The coefficients of $Q^{(k)}$ depend continuously on that of $Q^{(k-1)}$ and on $y_k, \dots, y_1$, so that the coefficients of Q depend continuously on the Schur parameters. Since the matrix Q is inner and thus bounded in the unit disk, $\|Q(z)\| \le 1$, $\forall z \in \mathbb{U}$, Lebesgue's theorem finally implies that $\varphi^{-1}_{(\mathbf{w},\mathbf{u})}$ is continuous. Conversely, note that the evaluation map $Q \to Q(w_n)^* u_n$ is continuous, so that $Q \to y_n$ is continuous. The coefficients of $Q^{(n-1)}$ depend continuously on that of Q and on y_n, and, if two normalized rational functions of bounded degree are closed in $H_2^{p\times p}$, then their coefficients must also be closed; then, the map $Q \to Q^{(n-1)}$ from $\mathcal{I}_n^p$ to $\mathcal{I}_{n-1}^p$, both endowed with the H_2-topology, is continuous. We thus prove inductively that $\varphi_{(\mathbf{w},\mathbf{u})}$ is continuous and consequently is an homeomorphism. Furthermore, the map $\varphi_{(\mathbf{w},\mathbf{u})} \circ \varphi_{(\mathbf{w}',\mathbf{u}')} : \mathcal{B}_p^n \to \mathcal{B}_p^n$ is C^∞, as a bounded rational function. ∎

In any chart of this atlas, *the local coordinates are the $2np$ real and imaginary parts of the components of the Schur parameters.* Note that an atlas for $\mathcal{RI}_n^p(1)$ can be obtained in a similar way, for which the w_i's lie in $(-1, 1)$, the u_i's and the y_i's have real components; indeed, in Lemma 6 u can be chosen real, and if A and Θ have real coefficients, $T_\Theta(A)$ also has real coefficients. The range of the charts is thus the product of n copies of the unit ball of $\mathbb{R}^p$.

3.4. Fractional representation in the local coordinates. In this section, we give a fractional representation of the form $D^{(k)}/\widetilde{q}^{(k)}$ for the matrix $Q^{(k)}$ (see Proposition 2). We introduce the map $\mathcal{S}_{(w,u)} : (A, y) \to T_{\Theta(w,u,y)}(A)$, so that the inner matrix $Q = \varphi^{-1}_{(\mathbf{w},\mathbf{u})}(y)$ is computed by the iterative process:

$$I_p \to Q^{(1)} \to \cdots \to Q^{(k)} = \mathcal{S}_{(w_k,u_k)}(Q^{(k-1)}, y_k) \to \cdots \to Q^{(n)} = Q.$$

LEMMA 9. *For any $A \in \mathcal{I}_k^p$, $w \in \mathbb{U}$, $u \in \mathbb{C}^p$, $\|u\| = 1$, and $v \in \mathbb{C}^p$, $\|v\| < 1$, we have*

$$S_{(w,u)}(A, y) = A + \frac{1-\beta_w}{1 - u^*Ay - \beta_w(y^*y - u^*Ay)}(u - Ay)(y^* - u^*A), \tag{26}$$

with $\beta_w = b_w/\widetilde{b}_w$, where $b_w(z) = (z - w)(1 - \bar{w})$.

Proof. Using (18) yields

$$T_{\Theta(w,u,y)}(A) = \left(A + (1-\beta_w)\frac{u(y^* - u^*A)}{1 - y^*y}\right)\left(I_p + (1-\beta_w)\frac{y(y^* - u^*A)}{1 - y^*y}\right)^{-1}.$$

A classical formula (see [28, Appendix A.20]) allows us to compute the inverse as follows

$$\left(I_p + (1-\beta_w)\frac{y(y^* - u^*A)}{1 - y^*y}\right)^{-1} = I_p - (1-\beta_w)\frac{y(y^* - u^*A)}{1 - u^*Ay - \beta_w(y^*y - u^*Ay)},$$

from which we deduce (26) by expanding the product. ∎

PROPOSITION 10. *A fractional representation $D^{(k)}/\widetilde{q}^{(k)}$ of the inner matrix $Q^{(k)} = T_{\Theta_k \dots \Theta_1}(I_p)$ can be computed by the recursion formulas: $D^{(0)} = I_p$, $\widetilde{q}^{(0)} = 1$, and for $k = 1, \dots, n$,*

$$
\begin{aligned}
D^{(k)} = {} & (\widetilde{b}_{w_k} - y_k^* y_k \, b_{w_k}) D^{(k-1)} - (\widetilde{b}_{w_k} - \ b_{w_k}) \Big\{ u_k u_k^* D^{(k-1)} + D^{(k-1)} y_k y_k^* \\
& - \widetilde{q}^{(k-1)} \ u_k y_k^* + \frac{u_k^* D^{(k-1)} y_k \ D^{(k-1)} - D^{(k-1)} y_k u_k^* D^{(k-1)}}{\widetilde{q}^{(k-1)}} \Big\},
\end{aligned} \tag{27}
$$

$$
\widetilde{q}^{(k)} = (\widetilde{b}_{w_k} - y_k^* y_k \, b_{w_k}) \, \widetilde{q}^{(k-1)} - (\widetilde{b}_{w_k} - b_{w_k}) \, u_k^* D^{(k-1)} y_k, \tag{28}
$$

where $b_{w_k} = (1 - \bar{w}_k)(z - w_k)$. The stable polynomial $q^{(k)}$ has degree k, and the coefficients of the polynomials $\widetilde{q}^{(k)}$ and $d_{ij}^{(k)}$ (the entries of $D^{(k)}$) are polynomial functions in the local coordinates.

Proof. Assume that such a fractional representation has been obtained for $Q^{(k-1)}$. Replacing $Q^{(k-1)}$ by $D^{(k-1)}/\widetilde{q}^{(k-1)}$ in $S_{(w_k,u_k)}(Q^{(k-1)}, y_k)$ given by (26) yields a fractional representation for $Q^{(k)}$. Note that (27) actually defines a polynomial matrix, since $\widetilde{q}^{(k-1)}$ does indeed divide $u_k^* D^{(k-1)} y_k \ D^{(k-1)} - D^{(k-1)} y_k \ u_k^* D^{(k-1)}$. In order to prove this, put $u_k^* = (\bar{u}_1^k, \dots, \bar{u}_p^k)$, $y_k^* = (\bar{y}_1^k, \dots, \bar{y}_p^k)$, and $D^{(k-1)} = (d_{ij}^{(k-1)})$. A straightforward computation shows that

$$
\begin{aligned}
& \left(u_k^* D^{(k-1)} y_k \, D^{(k-1)} - D^{(k-1)} y_k \, u_k^* D^{(k-1)} \right)_{ij} \\
& \qquad = \sum_{l,m} \left(d_{lm}^{(k-1)} d_{ij}^{(k-1)} - d_{im}^{(k-1)} d_{lj}^{(k-1)} \right) \bar{u}_l^k y_m^k,
\end{aligned}
$$

where $d_{lm}^{(k-1)} d_{ij}^{(k-1)} - d_{im}^{(k-1)} d_{lj}^{(k-1)}$ is a minor of order 2 of $D^{(k-1)}$. But in the fractional representation $\widetilde{q}^{(k-1)}$ is, up to a constant factor, the polynomial of poles of $Q^{(k-1)}$, which is the least common denominator of all the minors of $Q^{(k-1)}$ (see [28]). Thus, it must divide all the minors of order 2 of $D^{(k-1)}$.

Now, let us prove by induction that, for $k = 1, \dots, n$, the coefficients of $d_{ij}^{(k)}$ and $\widetilde{q}^{(k)}$ are polynomial functions in the local coordinates. This is true for $d_{ij}^{(1)}$ and $\widetilde{q}^{(1)}$ and we shall assume that this is also true for $d_{ij}^{(k-1)}$ and $\widetilde{q}^{(k-1)}$: for $l = 1, \dots, n$, put

$$
y_j^l = \xi_j^l + i \, \eta_j^l,
$$

where y_j^l is the jth component of y_l; then the coefficients of $d_{ij}^{(k-1)}$ and $\widetilde{q}^{(k-1)}$ belong to the ring $\mathcal{P}_{k-1}$ of complex polynomials in the $2(k-1)p$ variables ξ_j^l and η_j^l, $l = 1, \dots, k-1$, $j = 1, \dots, p$. In order to prove our assumption at order k, we must verify that $\widetilde{q}^{(k-1)}$ divides all the minors of order 2 of $D^{(k-1)}$ in the ring $\mathcal{P}_{k-1}[z]$. Let

$$
D^{(k-1)} \begin{pmatrix} i_1 & \cdots & i_l \\ j_1 & \cdots & j_l \end{pmatrix}
$$

be the minor of $D^{(k-1)}$ computed from the lines $i_1, \dots, i_l$ and the columns $j_1, \dots, j_l$. Since the matrix $D/\widetilde{q}$ is the inverse of $\widetilde{D}/q$, the minors of order 2 of $D^{(k-1)}$ are related

to those of order $p-2$ of $\widetilde{D}^{(k-1)}$ by the formula (see [20]):

$$\begin{aligned}&\{q^{(k-1)}\}^{p-3}\, D^{(k-1)}\begin{pmatrix} i_1 & i_2 \\ j_1 & j_2 \end{pmatrix}\\ &\qquad = (-1)^{\{i_1+i_2+j_1+j_2\}}\, \widetilde{D}^{(k-1)}\begin{pmatrix} i'_1 & \cdots & i'_{p-2} \\ j'_1 & \cdots & j'_{p-2} \end{pmatrix} \widetilde{q}^{(k-1)},\end{aligned} \tag{29}$$

where $\{i_1, i_2, i'_1, \ldots, i'_{p-2}\}$ and $\{j_1, j_2, j'_1, \ldots, j'_{p-2}\}$ form complete sets of rows and columns. If $\widetilde{q}^{(k-1)}$ is irreducible we are done. We shall prove this still by induction. First, $\widetilde{q}^{(0)}$ is irreducible. Then, assume that $\widetilde{q}^{(l-1)}$ is irreducible while $\widetilde{q}^{(l)}$ can be factored as

$$\widetilde{q}^{(l)} = \alpha\beta, \quad \alpha \in \mathcal{P}_l[z], \quad \beta \in \mathcal{P}_l[z].$$

The polynomial $\widetilde{q}^{(l)}$ can be viewed as a polynomial in the $2p$ coordinates $\xi_1^l, \ldots, \xi_p^l$, $\eta_1^l, \ldots, \eta_p^l$, with coefficients in $\mathcal{P}_{l-1}[z]$:

$$\widetilde{q}^{(l)} = \left(\widetilde{b}_{w_l} - b_{w_l}\left\{\sum_{j=1}^{p}(\xi_j^l)^2 + (\eta_j^l)^2\right\}\right)\widetilde{q}^{(l-1)} - (\widetilde{b}_{w_l} - b_{w_l})\sum_{j=1}^{p}\left\{\sum_{i=1}^{p}\bar{u}_i^l d_{ij}^{(l-1)}\right\}(\xi_j^l + i\eta_j^l).$$

If α does not depend on ξ_1^l, for example, then α must divide $-b_{w_l}\widetilde{q}^{(l-1)}$ and since b_{w_l} does not divide $\widetilde{q}^{(l)}$, we must have $\alpha = \widetilde{q}^{(l-1)}$. Therefore, $\widetilde{q}^{(l-1)}$ must divide each component of $u_l^* D^{(l-1)}$ in $\mathcal{P}_{l-1}[z]$. But this is clearly impossible; indeed, since $Q^{(l-1)}(1) = I_p$ we should have $u_l^* D^{(l-1)} = \widetilde{q}^{(l-1)} u_l$ for every choice of local coordinates. Thus both α and β have degree one in each ξ_j^l and η_j^l. But then, writing $\alpha = \alpha_1\xi_1^l + \cdots + \alpha_p\xi_1^p + \alpha'_1\eta_1^l + \cdots + \alpha'_p\eta_1^p + \alpha_0$ and $\beta = \beta_1\xi_1^l + \cdots + \beta_p\xi_1^p + \beta'_1\eta_1^l + \cdots + \beta'_p\eta_1^p + \beta_0$, we can see that such a factorization is impossible, so that $\widetilde{q}^{(l)}$ is actually irreducible. ∎

Though the quotient in formula (27) is exact, we fail in searching for an explicit formula for it, and we do not know if such a formula exists.

4. A generic algorithm to find a local minimum. The closure of $\mathcal{I}_n^p(1)$ in $H_2^{p\times p}$ is a compact set, so that we can think of using a gradient algorithm to find a local minimum of the function Ψ^n defined by (8) in section 2. The elements of $\mathcal{I}_n^p(1)$ will be parameterized as explained in the previous section, the local coordinates being the real and imaginary parts of the components of the vectors $y_1, \ldots, y_n$. We shall work with the local representations of Ψ^n and denote by $\Psi^n_{(\mathbf{w},\mathbf{u})}$ the local representation associated with the chart defined by $(\mathbf{w}, \mathbf{u})$:

$$\begin{array}{rlcl}\Psi^n_{(\mathbf{w},\mathbf{u})}: & \mathcal{B}_p^n & \to & \mathbb{R}, \\ & \mathbf{y} = (y_1, \ldots y_n) & \to & \Psi^n \circ \varphi^{-1}_{(\mathbf{w},\mathbf{u})}(\mathbf{y}).\end{array}$$

4.1. Limit points in the charts. The object of this section is to study what happens when, running a gradient algorithm, the norm of some Schur parameter tends to 1. In the scalar case, the structure of $\mathcal{I}_n^1(1)$ is particularly simple, since only one coordinate map is needed; as some $|y_k|$ tends to 1, the boundary of $\mathcal{I}_n^1(1)$ is reached. This boundary has been completely studied in the case of real functions; it has been proved in [7] that the set $\mathcal{RI}_n^1(1)$ can be identified with the set $\mathcal{P}_n^1$ of monic stable polynomials of degree n and established in [10] that its closure is a

topological manifold with boundary, this boundary having corners. The smooth part of the boundary, which plays an important role in the algorithm, consists of those polynomials having exactly one irreducible factor over $\mathbb{R}$ whose roots are of modulus one. In the matrix case, as some $\|y_k\|$ tends to 1, either the chart is no more available and another one must be used, or some point of the boundary of $\mathcal{I}_n^p(1)$ is reached. Moreover, as we shall see later, the closure of $\mathcal{I}_n^p(1)$ is no more a topological manifold with boundary, and possesses some *singular* boundary points (see Proposition 13).

Proposition 11 below, describes *regular* boundary points. Observe that if $\|y\| = 1$, the J-inner function $\Theta(w, u, y)$ is no more defined; however, if u^*Ay is not identically equal to 1, the transformation $S_{(w,u)}$ keeps a sense and is given by

$$S_{(w,u)}(A, y) = A + \frac{(u - Ay)(y^*A - u^*)}{(1 - u^*Ay)}.$$

PROPOSITION 11. *Let $\mathbf{y} \in \partial\mathcal{B}_p^n$, the boundary of $\mathcal{B}_p^n$, $\mathbf{w} \in \mathbb{U}^n$, and $\mathbf{u} \in \partial\mathcal{B}_p^n$, and let $(D^{(k)}, \widetilde{q}^{(k)})$ be the sequence associated with $\mathbf{w}, \mathbf{u}, \mathbf{y}$ by the recursion formulas* (27) *and* (28). *A sequence*

$$I_p \to Q^{(1)} \to \cdots \to Q^{(k)} = \mathcal{S}_{(w_k,u_k)}(Q^{(k-1)}, y_k) \to \cdots \to Q^{(n)}$$

of inner matrices can be computed, provided that $u_k^*Q^{(k-1)}(w_k)y_k$ is not identically equal to 1 as $\|y_k\| = 1$, or equivalently, $\widetilde{q}^{(k)}$ does not vanish identically. In this case, $\mathbf{y}$ will be called a regular limit point in the chart defined by $(\mathbf{w}, \mathbf{u})$. Then $Q^{(k)} = D^{(k)}/\widetilde{q}^{(k)}$, and

(a) *$q^{(k)}$ still has degree k,*

(b) *if $\|y_k\| = 1$, then $\widetilde{q}^{(k)}$ and $D^{(k)}$ have common roots on $\mathbb{T}$ and $Q^{(k)}$ has degree less than k.*

Moreover, there exists a neighborhood $\mathcal{W}$ of $\mathbf{y}$, such that $\varphi^{-1}_{(\mathbf{w},\mathbf{u})}$ extends smoothly to $\mathcal{W}$.

Proof. Assume that these assertions have been proved until order $k-1$, and let us prove that they still hold at order k. If $\|y_k\| < 1$, there is nothing to prove. If $\|y_k\| = 1$, then

$$\widetilde{q}^{(k)} = (1 - |w_k|^2)(1 - z)(\widetilde{q}^{(k-1)} - u_k^*D^{(k-1)}y_k),$$

and since $Q^{(k-1)}$ is inner, by the maximum modulus principle, either $u_k^*Q^{(k-1)}y_k$ is identically equal to 1, and $\widetilde{q}^{(k)}$ vanishes identically, or

$$\widetilde{q}^{(k)}(0) = (1 - |w_k|^2)\widetilde{q}^{(k-1)}(0)(1 - u_k^*Q^{(k-1)}(0)y_k)$$

does not vanish and thus $q^{(k)}$ has degree k. In this case, $Q^{(k)} = \mathcal{S}_{(w_k,u_k)}(Q^{(k-1)}, y_k) = D^{(k)}/\widetilde{q}^{(k)}$ is well defined and still inner; 1, which is a root of $\widetilde{q}^{(k)}$, must also be a root of $D^{(k)}$, and the degree of $Q^{(k)}$ cannot exceed that of $Q^{(k-1)}$. More precisely,

$$\deg Q^{(k)} = \deg Q^{(k-1)} - \sharp\{\xi \in \mathbb{T}, y_k = Q^{(k-1)}(\xi)^* u_k\}.$$

By induction, the first part of the proposition is proved. Now, $\varphi^{-1}_{(\mathbf{w},\mathbf{u})}(\mathbf{y}) = D^{(n)}/\widetilde{q}^{(n)}$ and since, by Proposition 10, the coefficients of the polynomials $\widetilde{q}^{(n)}$ and $d_{ij}^{(n)}$ are polynomial functions in the local coordinates, there exists a neighborhood $\mathcal{W}$ of $\mathbf{y}$, such that $\varphi^{-1}_{(\mathbf{w},\mathbf{u})}$ extends smoothly to $\mathcal{W}$. ∎

In the next lemma, we shall prove that any inner matrix of degree strictly less than n can be viewed, up to a unitary matrix, as a boundary point of $\mathcal{I}_n^p(1)$ of this type.

LEMMA 12. *For each $Q \in \mathcal{I}_d^p(1)$ of degree d strictly less than n, there exist $\mathbf{w}' \in \mathbb{U}^n$, $\mathbf{u}' \in \partial\mathcal{B}_p^n$, $\mathbf{y}' \in \partial\mathcal{B}_p^n$, and a unitary matrix $\mathcal{U}$, such that $\mathbf{y}'$ is a regular limit point in the chart defined by $(\mathbf{w}', \mathbf{u}')$ and $\mathcal{U}Q = \varphi^{-1}_{(\mathbf{w}',\mathbf{u}')}(\mathbf{y}')$.*

Proof. Let $\mathbf{y} = (y_1, \ldots, y_d)$, $\mathbf{w} = (w_1, \ldots, w_d)$, and $\mathbf{u} = (u_1, \ldots, u_d)$ be such that $Q = \varphi^{-1}_{(\mathbf{w},\mathbf{u})}(\mathbf{y})$:

$$I_p \to Q^{(1)} = \mathcal{S}_{(w_1,u_1)}(I_p, y_1) \to Q^{(2)} \cdots \to Q = \mathcal{S}_{(w_d,u_d)}(Q^{(n-1)}, y_d).$$

The Schur transform $S_{(w,u)}$ applied to some unitary matrix $\mathcal{X}$ and some unit vector y such that $y \neq \mathcal{X}^* u$ will give another unitary matrix. Thus, we can construct a unitary matrix $\mathcal{U}$ in the following way:

$$\mathcal{U}_0 = I_p \to \mathcal{U}_1 = \mathcal{S}_{(w_1',u_1')}(I_p, y_1') \to \mathcal{U}_2 \cdots \to \mathcal{U} = \mathcal{S}_{(w_{n-d}',u_{n-d}')}(\mathcal{U}_{n-d-1}, y_{n-d}'),$$

where $w_1', \ldots, w_{n-d}'$, are chosen arbitrarily and $u_1', \ldots, u_{n-d}', y_1', \ldots, y_{n-d}'$, are unit vectors in $\mathbb{C}^p$, satisfying for $k = 1, \ldots, n-d$, ${u_k'}^* \mathcal{U}_{k-1} y_k' \neq 1$. Since we have

$$S_{(w,\,\mathcal{X}u)}(\mathcal{X}A, y) \;=\; \mathcal{X} S_{(w,\,u)}(A, y)$$

for any unitary matrix $\mathcal{X}$, $\mathcal{U}Q$ can be computed by the following iterative process:

$$I_p \to \mathcal{U}_1 = \mathcal{S}_{(w_1',u_1')}(I_p, y_1') \to \mathcal{U}_2 \cdots \to \mathcal{U} = \mathcal{S}_{(w_{n-d}',u_{n-d}')}(\mathcal{U}_{n-d-1}, y_{n-d}')$$

$$\to \mathcal{U}Q^{(1)} = \mathcal{S}_{(w_1,\mathcal{U}u_1)}(\mathcal{U}, y_1) \to \mathcal{U}Q^{(2)} \cdots \to \mathcal{U}Q = \mathcal{S}_{(w_d,\mathcal{U}u_d)}(\mathcal{U}Q^{(n-1)}, y_d).$$

Put

$$\begin{aligned}
\mathbf{w}' &= (w_1', \ldots, w_{n-d}', w_1, \ldots, w_d),\\
\mathbf{u}' &= (u_1', \ldots, u_{n-d}', \mathcal{U}u_1, \ldots, \mathcal{U}u_d),\\
\mathbf{y}' &= (y_1', \ldots, y_{n-d}', y_1, \ldots, y_d),
\end{aligned}$$

then $\mathbf{y}'$ is a regular limit point in the chart defined by $(\mathbf{w}', \mathbf{u}')$ and $\mathcal{U}Q = \varphi^{-1}_{(\mathbf{w}',\mathbf{u}')}(\mathbf{y}')$ as required. ∎

The next proposition is concerned with *singular* boundary points.

PROPOSITION 13. *Let $\eta(t) : [0,1] \to \overline{\mathcal{B}_p^n}$ be a smooth path whose terminal point $\mathbf{y} = \eta(1)$ belongs to $\partial\mathcal{B}_p^n$ and let $D(t)/\widetilde{q}(t)$ be the fractional representation of $Q(t) = \varphi^{-1}_{(\mathbf{w},\mathbf{u})}(\eta(t))$ obtained by the recursion formulas (27) and (28). Assume that $\widetilde{q}(t)$ vanishes identically as $t \to 1$. Then, $Q(t)$ converges to some inner function Q_η, depending in general on the path and whose degree may be less than or equal to n; it is given by the number of roots of $q(t)$ which converges within the unit disk.*

Proof. Since $D(t)$ and $\widetilde{q}(t)$ are polynomial functions in the local coordinates, they do converge, respectively, to a polynomial matrix D and a polynomial $\widetilde{q}$. We deal with the case where $\widetilde{q}$ is the zero polynomial. However, the quotient $D(t)/\widetilde{q}(t)$ must converge to some inner function Q_η. Let $q(t)(z) = q_n(t)z^n + \cdots + q_1(t)z + q_0(t)$. As $t \to 1$, each coefficient tends to 0 while the quotients $q_k(t)/q_n(t)$ being the well-known elementary symmetric polynomials in the roots (of modulus at most 1) are bounded by

the binomial coefficients $\binom{n}{k}$ and thus converges. The polynomial $q(t)/q_n(t)$ converges to some monic polynomial, which may have roots of modulus one, while the number of its roots within the unit circle gives the degree of Q_η.

Now, we are going to show that the limit Q_η actually depends on the path and may have degree as well less than or equal to n. Let us study the case where $n = 1$ ($\eta(t) = y_1(t)$). Formulas (27) and (28) yield

$$\widetilde{q}^{(1)}(t) = \widetilde{b}_{w_1} u_1^*(u_1 - y_1(t)) + b_{w_1}(u_1 - y_1(t))^* u_1 - b_{w_1}(u_1 - y_1(t))^*(u_1 - y_1(t)),$$

and

$$D^{(1)}(t) = \widetilde{q}^{(1)}(t) I_p - (\widetilde{b}_{w_1} - b_{w_1})(u_1 - y_1(t))(u_1 - y_1(t))^*,$$

so that

$$Q^{(1)}(t) = I_p - \frac{\widetilde{b}_{w_1} - b_{w_1}}{\widetilde{q}^{(1)}(t)}\ (u_1 - y_1(t))(u_1 - y_1(t))^*.$$

Now, as $t \to 1$, $\widetilde{q}^{(1)}(t)$ vanishes identically by assumption, and thus $y_1(t)$ must converge to u_1. Let

$$y_1(t) = u_1 - \sum_{k \geq l} (1-t)^k \xi_k, \quad \xi_l \neq 0, \quad \xi_k \in \mathbb{C}^p$$

be its expansion. Consequently, $\widetilde{q}^{(1)}(t) \sim (\widetilde{b}_{w_1} u_1^* \xi_l + b_{w_1} \xi_l^* u_1)(1-t)^l$ and $Q^{(1)}(t)$ converges to I_p, unless $u_1^* \xi_l = 0$. In this case, let s be the smallest index satisfying $s > l$ and $u_1^* \xi_s \neq 0$. Then, if $s < 2l$, $Q(t)$ still converges to I_p, while if $s \geq 2l$,

$$\widetilde{q}^{(1)}(t) \sim (\widetilde{b}_{w_1} u^* \xi_{2l} + b_{w_1} \xi_{2l}^* u_1 - \widetilde{b}_{w_1} \xi_l^* \xi_l)(1-t)^{2l},$$

and

$$Q^{(1)}(t) \to I_p - \frac{\widetilde{b}_{w_1} - b_{w_1}}{\widetilde{b}_{w_1} u_1^* \xi_{2l} + b_{w_1} \xi_{2l}^* u_1 - \widetilde{b}_{w_1} \xi_l^* \xi_l}\ \xi_l \xi_l^*,$$

which is an inner function of degree 1. In conclusion, as $y_1(t)$ converges to u_1, $Q^{(1)}(t)$ converges either to I_p or to some inner matrix of degree 1, in which case, we leave the domain of the chart while staying inside the manifold. The same situation arises at each order, though it may be more complicated if the norms of several Schur parameters go to 1. □

As an illustration, the closure of $\mathcal{RI}_1^2(1)$ can be viewed as a cone. The summit represents the identity matrix and is a singular boundary point while the base represents the circle of orthogonal matrices of determinant -1 and forms a regular boundary. Two charts are needed to describe this manifold. For example, the chart given by $w = 0$ and $u = (1, 0)^*$ parametrizes all the inner functions except for those of the form (a line of the cone) $\left(\begin{smallmatrix} 1 & 0 \\ 0 & \frac{z-a}{1-az} \end{smallmatrix}\right)$, $a \in (-1, 1)$, while the chart given by $w = 0$ and $u = (0, 1)^*$ parametrizes all the inner functions except for those of the form $\left(\begin{smallmatrix} \frac{z-a}{1-az} & 0 \\ 0 & 1 \end{smallmatrix}\right)$.

4.2. Properties of the local representations of the criterion. The object of this section is to study the behavior of the local representations of the criterion at the neighborhood of a boundary point of $\mathcal{I}_n^p(1)$. We have distinguished in the last

section two kinds of limit points, the regular and the singular ones. In both cases, if $\eta(t)$ is a path whose terminal point $\mathbf{y}$ corresponds to a boundary point of degree $d < n$, say Q_η, then it is easily proved that

$$\lim_{t\to 1} \Psi^n_{(\mathbf{w},\mathbf{u})}(\eta(t)) = \Psi^d(Q_\eta).$$

However, regular limit points play a central role in our algorithm, mainly due to the fact that the local representations of the criterion extends smoothly at the neighborhood of such points. To prove this result, we shall need the following expression for Ψ^n.

PROPOSITION 14. *Let $G(z) = F^\sharp(z)/z$ and $Q = D/\widetilde{q}$ as in Proposition 2. Let R be the remainder in the Weierstrass division in $H_2^{p\times m}$ of $G\widetilde{D}$ by q:*

$$G\widetilde{D} \;=\; V\; q \;+\; R. \tag{30}$$

Then q divides RD and if P is the matrix quotient, of degree at most $n-1$, we have that

$$\Psi^n(Q) \;=\; \|F\|_2^2 - \left\langle F, \frac{\widetilde{P}}{q} \right\rangle. \tag{31}$$

Proof. Since $Q^{-1}L(Q)$ and $F - Q^{-1}L(Q)$ are orthogonal, the cost function can be rewritten:

$$\Psi^n(Q) = \|F\|_2^2 - \langle F, Q^{-1}L(Q)\rangle.$$

The orthogonal projection $L(Q)$ of QF onto $H_2^{p\times m}$ is easily computed from the division (30) as being given by $L(Q) \;=\; \widetilde{R}/\widetilde{q}$, where $\widetilde{R} \;=\; z^{n-1}\; R^\sharp(z)$. Now, multiplying (30) by D on the right shows that q divides RD, and $Q^{-1}L(Q) = \widetilde{D}/q\;\widetilde{R}/\widetilde{q} = \widetilde{P}/q$. ∎

PROPOSITION 15. *Assume that $G(z) = F^\sharp(z)/z$ is analytic in $D_r = \{z,\ \|z\| \le r\}$ for some $r > 1$. Let $\mathbf{y} \in \partial\mathcal{B}_p^n$ be a regular limit point in some chart defined by $(\mathbf{w},\mathbf{u})$ (see Proposition 11) and let $Q = \varphi^{-1}_{(\mathbf{w},\mathbf{u})}(\mathbf{y})$ belong to $\mathcal{I}_d^p$ for some $d < n$. Then, $\Psi^n_{(\mathbf{w},\mathbf{u})}$ extends in some open neighborhood of $\mathbf{y}$ to a smooth function still denoted by $\Psi^n_{(\mathbf{w},\mathbf{u})}$. Moreover, we have*

$$\Psi^n_{(\mathbf{w},\mathbf{u})}(\mathbf{y}) \;=\; \Psi^d\left(Q(1)^{-1}Q\right).$$

Proof. Let $\mathcal{W}$ be a neighborhood of $\mathbf{y}$ on which, by Proposition 11, $\varphi^{-1}_{(\mathbf{w},\mathbf{u})}$ extends smoothly. We may assume that in $\mathcal{W}$, $|\widetilde{q}(0)| \ge \mu$, for some $\mu > 0$. In order to proceed to our extension, we shall use the expression (31) of Ψ^n, in which the polynomial matrices D, R, and P, and the polynomial q, depend on the local coordinates. A well-known integral representation for the remainder R (cf. [39]) is

$$R(z) \;=\; \frac{1}{2i\pi}\int_{\mathbb{T}} \frac{G\widetilde{D}}{q}\frac{q(\xi)-q(z)}{\xi - z}d\xi.$$

We may also restrict $\mathcal{W}$ so that the roots of q lie in a disk $D_s = \{z, |z| < s\}$ for some s, $1 < s < r$. Then, we can extend R in $\mathcal{W}$ by putting

$$R(z) \;=\; \frac{1}{2i\pi}\int_{\Gamma} \frac{G\widetilde{D}}{q}\frac{q(\xi)-q(z)}{\xi - z}d\xi, \tag{32}$$

where Γ is any contour lying in the open annulus between D_s and D_r. Indeed, the coefficient of order k of R is given by

$$R_k \;=\; \frac{1}{2i\pi}\int_\Gamma \frac{G\widetilde{D}}{q}(\xi^{n-k-1}q_n + \cdots + q_{k+1})d\xi$$

and since $|q(\xi)| > \mu\ d(\Gamma, D_s)^n$, the integrand is bounded, and its derivatives are also bounded. Finally, Lebesgue's theorem says that the integral representation (32) defines a smooth function. The extension of R is still the remainder of the division (30). In $\mathcal{W}$, q keeps on dividing RD and the quotient extends smoothly P. As for R, $\Psi^n_{(\mathbf{w},\mathbf{u})}$ extends smoothly by the integral representation

$$\Psi^n_{(\mathbf{w},\mathbf{u})}(\mathbf{y}) \;=\; \|F\|_2^2 - \frac{1}{2i\pi}\overline{\mathrm{Tr}\int_\Gamma G(z)\frac{\widetilde{P}}{q}(z)dz}. \qquad \square$$

Let us give two important consequences of Proposition 15.

LEMMA 16. *Let $Q \in \mathcal{I}_k^p(1)$ for some $k < n$ and let $\mathbf{y} = (y_1, \ldots, y_k)$ be its Schur parameters in some chart defined by $\mathbf{w} = (w_1, \ldots, w_k)$ and $\mathbf{u} = (u_1, \ldots, u_k)$. Let $w_0 \in \mathbb{U}$, u_0 and y_0 be two distinct unit vectors and put*

$$\mathcal{U} = I_p - \frac{(u_0 - y_0)(u_0 - y_0)^*}{1 - u_0^* y_0},$$

$\mathbf{w}' = (w_0, w_1, \ldots, w_k)$, $\mathbf{u}' = (u_0, \mathcal{U}u_1, \ldots, \mathcal{U}u_k)$, and $\mathbf{y}' = (y_0, y_1, \ldots, y_k)$. Then $\mathbf{y}'$ is a regular limit point in the chart defined by $(\mathbf{w}', \mathbf{u}')$ and $Q' = \mathcal{U}Q$ is given by $Q' = \varphi^{-1}_{(\mathbf{w}',\mathbf{u}')}(y')$. Moreover,

$$\psi^{k+1}_{(\mathbf{w}',\mathbf{u}')}(\mathbf{y}') = \psi^k_{(\mathbf{w},\mathbf{u})}(\mathbf{y}). \tag{33}$$

COROLLARY 17. *Suppose that $\mathbf{y}$ is a local minimum of $\psi^k_{(\mathbf{w},\mathbf{u})}(\mathbf{y})$. Then, the gradient of $\psi^{k+1}_{(\mathbf{w}',\mathbf{u}')}$ at $\mathbf{y}'$ is orthogonal to the surface $\mathcal{S} = \{(y_0, \ldots, y_n),\ \|y_0\| = 1,\ \|y_j\| < 1,\ j = 1, \ldots n\}$ and points outwards.*

Proof. From Proposition 15 we see that the projection of the gradient of $\psi^{k+1}_{(\mathbf{w}',\mathbf{u}')}$ at $\mathbf{y}'$ on $\mathcal{S}$ is just the gradient of $\psi^k_{(\mathbf{w},\mathbf{u})}$ at $\mathbf{y}$, whence orthogonality holds. Moreover, it cannot point inwards because this would imply that Q' which is rational of order k is a local minimum at order $k+1$, and this is impossible except if F itself has degree $k+1$ (cf. [6]). $\square$

4.3. The algorithm. The algorithm searching for a local minimum at order n splits into four main operations.

A. *Choosing an initial point.* This choice involves the choice of $(\mathbf{w}, \mathbf{u})$ indexing a chart. The initial point $Q_i = \varphi^{-1}_{(\mathbf{w},\mathbf{u})}(\mathbf{y}_i)$ may have degree less than or equal to the target order n.

B. *Minimizing at fixed order k.* A software is used which integrates the vector field $-\mathrm{grad}\,\Psi^k_{(\mathbf{w},\mathbf{u})}$ from an initial point $\mathbf{y}_i \in \mathcal{B}_p^k$. The cost function is computed by (31) where $q = q^{(k)}$ and $\widetilde{D} = \widetilde{D}^{(k)}$ are given by the following recursion formulas, immediately deduced from (27) and (28),

$$\widetilde{D}^{(l)} = (b_{w_l} - y_l^* y_l \widetilde{b}_{w_l})\widetilde{D}^{(l-1)} - (b_{w_l} - \widetilde{b}_{w_l}) \left\{ \widetilde{D}^{(l-1)} u_l u_l^* + y_l y_l^* \widetilde{D}^{(l-1)} \right.$$

$$\left. -q^{(l-1)}\ y_l u_l^* + \frac{y_l^* \widetilde{D}^{(l-1)} u_l\ \widetilde{D}^{(l-1)} - \widetilde{D}^{(l-1)} u_l y_l^* \widetilde{D}^{(l-1)}}{q^{(l-1)}} \right\}, \tag{34}$$

$$q^{(l)} = (b_{w_l} - y_l^* y_l \widetilde{b}_{w_l})\, q^{(l-1)} - (b_{w_l} - \widetilde{b}_{w_l})\, y_l^* \widetilde{D}^{(l-1)} u_l, \tag{35}$$

and initialized by $q^{(0)} = 1$ and $\widetilde{D}^{(0)} = 1$. Then, one of the following possibilities occurs:

(i) *a local minimum is reached.* If $k = n$, we are done, while if $k < n$, this local minimum provides an initial point for searching for a minimum of order $k + 1$, as described in point D.

(ii) *the norm of some Schur parameter tends to* 1. This situation has been studied in section 4.1; either a change of chart is necessary, or a boundary point of the manifold is reached. More precisely, if the polynomial $\widetilde{q}^{(k)}$ nearly vanishes while its roots stay far from the unit circle, then the limit point belongs to $\mathcal{I}_k^p(1)$, and the first eventuality is true. In any other case, a boundary point is reached.

C. *Meeting a boundary point.* Such a boundary point, up to an unitary matrix, is an element Q_b of $\mathcal{I}_d^p(1)$ for some $d < k$, and the criterion at order k converges to $\Psi^d(Q_b)$. Then, a minimization process at order d can restart from Q_b. If only the first Schur parameter has norm 1, we can directly deduce from Lemma 16 some chart and Schur parameters for Q_b. Otherwise, the matrix Q_b must be computed from the recursion formulas (27) and (28), eliminating the roots of modulus one. Then, an adequate chart has to be provided.

D. *Choosing an adequate coordinate chart.* Given a normalized inner matrix Q, of order k, we must find a couple $(\mathbf{w}, \mathbf{u})$ such that Q belongs to the local neighborhood $\mathcal{V}_{(\mathbf{w},\mathbf{u})}$ defined in section 3.3, or equivalently such that a sequence $Q^{(k)} = Q, Q^{(k-1)}, \ldots, Q^{(1)} = I_p$ of inner functions of decreasing degree can be constructed by the Schur algorithm. The fractional representation of $Q^{(l-1)}$ is computed from that of $Q^{(l)}$ by the recursion formulas

$$b_{w_l}\, \tilde{b}_{w_l}\, \tilde{D}^{(l-1)} = (\tilde{b}_{w_l} - y_l^* y_l\, b_{w_l})\tilde{D}^{(l)} - (\tilde{b}_{w_l} - b_{w_l}) \left(\tilde{D}^{(l)} u_l u_l^* + y_l y_l^* \tilde{D}^{(l)} - q^{(l)}\ y_l u_l^* + \frac{y_l^* \tilde{D}^{(l)} u_l\ \tilde{D}^{(l)} - \tilde{D}^{(l)} u_l y_l^* \tilde{D}^{(l)}}{q^{(l)}} \right),$$

$$b_{w_l}\, \tilde{b}_{w_l}\, q^{(l-1)} = (\tilde{b}_{w_l} - y_l^* y_l\, b_{w_l})\, q^{(l)} - (\tilde{b}_{w_l} - b_{w_l})\, y_l^* \tilde{D}^{(l)} u_l.$$

The polynomial $b_{w_l}\, \tilde{b}_{w_l}$ divides the right-hand sides, so that $\tilde{D}^{(l-1)}$ and $q^{(l-1)}$ actually are polynomial, and $Q^{(l-1)}$ has degree $l - 1$ as required.

E. *Increasing the degree.* When the minimization procedure leads to a local minimum of order $k < n$, say Q_m, then Lemma 16, for any choice of w_0, u_0, and $y_0 \neq u_0$, provides a boundary point Q'_m of $\mathcal{I}_{k+1}^p(1)$ together with a local parametrization $Q'_m = \varphi^{-1}_{(\mathbf{w}',\mathbf{u}')}(\mathbf{y}'_m)$, deduced from a local parametrization $\varphi^{-1}_{(\mathbf{w},\mathbf{u})}(\mathbf{y}_m)$ of Q_m, satisfying (33). Since by Corollary 17, $-\operatorname{grad} \Psi^{k+1}_{(\mathbf{w}',\mathbf{u}')}$ points inwards at $\mathbf{y}'$, this point can be used as an initial point for a minimization process at order $k + 1$.

The point is that the value of the criterion, where the criterion must be understood as being Ψ^k when working at order k, decreases continuously, being conserved while the order changes, so that the minimization process pursues through different orders.

To ensure the good behavior of the algorithm, we shall make two extra assumptions. First, we shall assume that grad Ψ^k *does not vanish on the boundary of* $\mathcal{I}_k^p(1)$*, for* $1 \leq k \leq n$. Second, we shall require *all the critical points of* Ψ^k *in* $\mathcal{I}_k^p(1)$ *to be nondegenerate,* i.e., to have a second derivative which is a nondegenerate quadratic form. These two properties hold generically, that is for almost every F in some sense, and we refer the reader to [8] for the first one, and [6] for the second one. They ensure in particular that critical points in $\mathcal{I}_k^p(1)$ are finite in number. Since the criterion decreases continuously, we never meet twice the same local minimum and *this ensures that the procedure eventually comes to an end.* Note that if the minimization process stops at a critical point which is not a minimum, since this point is nondegenerate, it will be unstable under small perturbations, thereby allowing us to continue the procedure.

The choice of an initial point is crucial for our purpose (see the example in the next section). In many problems, we hope that some more information or engineering judgment could help us to select an initial point which ensures rapid convergence of the procedure to the global minimum. However, it is well known that the L_2 approximation problem possesses many local minima. Since our final goal is to find the global minimum, we may think of initializing the algorithm at enough points to reach all local minima and compare between them. But we do not know what "enough" means and we do not have a bound for the number of initializing points. Consequently, more efficient strategies should be investigated. For instance, we can find all the local minima at order 1 and then, initialize our procedure at order 2, by replacing them on the boundary of $\mathcal{I}_2^p(1)$ as described in point D, choosing w_0, u_0, and y_0 in several ways, and so on, step by step, until the target order. This strategy gives rather good results.

The choice of a local chart at the neighborhood of a given point is an important and difficult task. The main purpose of using coordinates is to be able to perform calculations on a computer and as such it is desirable that the numerical conditioning of the chart is good. A criterion must be chosen to decide upon the quality of local coordinates around a point on a manifold. Moreover, a distortion occurs when mapping part of a manifold to Euclidean space, so that the sequence of improving estimates produced by an optimization algorithm is dependent on the choice of the chart, and it would be interesting to select the charts with the view to improve the convergence of the algorithm. But in this case, the selection strategy will depend upon the problem at hand and bring along a lot of "overhead costs." The present version of our algorithm uses a basic selection strategy, which minimizes the norm of the Schur parameters at each step of the Schur algorithm over a finite atlas. This point must be improved and is presently under study.

4.4. A numerical example. The sole purpose of the following example is to demonstrate the procedure of computing local minima. For more real-data simulations, we refer to the scalar case paper [8] or [9]. This example has been first considered in [31] to demonstrate the procedure of computing the minimal degree approximation in a Hankel-norm model reduction problem and refers to a fourth-order system:

$$F(z) = \begin{pmatrix} \dfrac{1+z}{z^2-z+1/4} & \dfrac{1}{z-1/2} \\ \dfrac{-z^2+z+1}{z^3+1/2z^2-1/4z-1/8} & \dfrac{z-1/4}{z^2+z+1/4} \end{pmatrix},$$

or equivalently $F = N/d$, where

$$d(z) = z^4 - 1/2z^2 + 1/16$$

and

$$N(z) = \begin{pmatrix} z^3 + 2z^2 + 5/4z + 1/4 & z^3 + 1/2z^2 - 1/4z - 1/8 \\ -z^3 + 3/2z^2 + 1/2z - 1/2 & z^3 - 5/4z^2 + 1/2z - 1/16 \end{pmatrix}.$$

The system has four poles located at $1/2$, $1/2$, $-1/2$, $-1/2$. According to the theory, if we look for a minimum of (1) with $n = 4$, we must recover the function F itself, since from consistency, the criterion has no other critical points [12]. We shall use this fact to test the procedure.

The function to be approximated is represented in the program by a great number of Fourier coefficients (computed from frequency data in practice). Thus in this example, the input of the program is not actually the function F but the 200 first Fourier coefficients of its rational entries. The software package Scilab is used for the implementation. We have run a great number of tests changing the starting point and the initial chart. We present here a case in which every step of the algorithm must be visited before, according to the theory, we finally recover the function F.

Step 1. *We integrate at order* 4 *and reach the boundary.* The initial point has parameters $\mathbf{y} = ((0.5, 0.5)^*, ((0.5, 0.5)^*, (-0.5, -0.5)^*, (0.5, 0.5)^*)$ in the chart indexed by $\mathbf{w} = (0, 0, 0, 0)$ and $\mathbf{u} = ((1, 0)^*, (1, 0)^*, (0, 1)^*, (1, 0)^*)$, and corresponds to the inner matrix $Q_i = D_i/\widetilde{q}_i$, where

$$\widetilde{D}_i(z) = \begin{pmatrix} -0.3 + 0.4z + 0.4z^2 - 0.8z^3 + 0.5z^4 & -0.5 + 0.8z - 0.4z^2 + 0.4z^3 - 0.3z^4 \\ 0.3 - 0.4z + 0.4z^2 - 0.8z^3 + 0.5z^4 & 0.5 - 0.8z + 0.4z^2 + 0.4z^3 - 0.3z^4 \end{pmatrix}, \tag{36}$$

$$q_i(z) = z^2(z^2 - 1.2z + 0.4). \tag{37}$$

Note that q_i is not exactly the stable polynomial $q^{(4)}$ computed from the recursion formulas (34) and (35) which has a leading coefficient equal to 0.3125. As we integrate the opposite of the gradient using the Scilab function "ode," the norm of the first parameter tends to 1, while $\widetilde{q}^{(1)}(0) = 0.49$ stays far from 0. Thus we have reached a *regular* boundary point Q_b of parameters

$$\mathbf{y}_b = ((0.509, 0.86)^*, (0.357, 0.55)^*, (-0.659, -0.405)^*, (0.556, 0.264)^*).$$

The criterion is equal to 3.786.

Step 2. *We integrate at order* 3 *and get a local minimum.* We put $u_0 = (1, 0)^*$ and $y_0 = (0.509, 0.86)^*$ and we compute the unitary matrix

$$\mathcal{U} = I_p - \frac{(u_0 - y_0)(u_0 - y_0)^*}{1 - u_0^* y_0}.$$

Lemma 16 implies that $Q_b = \mathcal{U}Q$, where Q is the normalized inner matrix of degree 3 of parameters $\mathbf{y} = ((0.357, 0.55)^*, (-0.66, -0.405)^*, (0.556, 0.265)^*))$ in the chart indexed by $\mathbf{w} = (0, 0, 0)$ and $\mathbf{u} = ((0.509, 0.86)^*, (0.86, -0.509)^*, (0.509, 0.86))$. According to the theory, the criterion at Q is still equal to 3.786. We restart the minimization procedure from this point and find a third degree minimum for

$$\mathbf{y}_m = ((-0.574, 0.652)^*, (0.0214, -0.433)^*, (0.205, 0.428)^*),$$

where the criterion is equal to 0.997 and the relative error to 0.05.

Step 3. *We increase the degree and get out of the domain of the chart.* This third order local minimum provides starting points for fourth order minimizations. For instance, applying Lemma 16 with $w_0 = 0$, $u_0 = (1,0)^*$, and $y_0 = (0,1)^*$, which are distinct unit vectors, yields to the initial point of parameters

$$\mathbf{y} = ((0,1)^*, (-0.574, 0.652)^*, (0.021, -0.433)^*, (0.205, 0.428)^*)$$

in the chart indexed by $\mathbf{w} = (0,0,0,0)$ and

$$\mathbf{u} = ((1,0)^*, (0.86, 0.509)^*, (-0.509, 0.86)^*, (0.86, 0.509)^*).$$

The minimization process leads us to leave the domain of the chart. Indeed, it produces a sequence of inner functions whose denominators computed by formulas (27) and (28), have leading coefficients which tends to 0 but roots which stay far from the unit circle. We stop at

$$\mathbf{y} = ((0.88, 0.096)^*, (-0.688, 0.102)^*, (0.169, 0.232)^*, (0.264, -0.027)^*)$$

at which the value of $\widetilde{q}(0)$ is about 0.125 which can produce important errors in the computation.

Step 4. *We change the chart and recover the function* F. We choose to work with a finite subset of the atlas described in section 3.3; the family $(\mathcal{V}_{(\mathbf{w},\mathbf{u})}, \varphi_{(\mathbf{w},\mathbf{u})})$ where $\mathbf{w} = (0,0,0,0)$, and $\mathbf{u}$ is composed of unit vectors either equal to $e_1 = (1,0)^*$ or to $e_2 = (0,1)^*$. This family is a covering of the manifold $\mathcal{I}_n^p(1)$. At each step of the Schur algorithm, we choose $u_k = e_j$, where e_j is the vector for which the norm of the Schur parameter $y_k = Q^{(k)}(0)^*(e_j)$ is the smallest. It may happen that this process provides Schur parameters of norm almost equal to 1. In this case we can try each chart of our finite atlas to find a better one. In our case this process gives a new chart indexed by $\mathbf{w} = (0,0,0,0)$ and $\mathbf{u} = ((1,0)^*, (1,0)^*, (0,1)^*, (1,0)^*)$. The parameters of the point are given in this chart by $\mathbf{y} = ((0.632, -.278)^*, (-0.578, -.337)^*, (0.262, 0.192)^*, (0.157, -.142)^*)$. The minimization can continue and the minimum is reached for

$$\mathbf{y}_m = ((0.495, -0.32)^*, (-0.57, -0.328)^*, (0.266, 0.202)^*, (0.146, -0.129)^*).$$

The approximant computed from these parameters agrees with F with four significant digits.

If we start in the same initial chart $\mathbf{w} = (0,0,0,0)$, and

$$\mathbf{u} = ((1,0)^*, (1,0)^*, (0,1)^*, (1,0)^*),$$

from the point

$$\mathbf{y} = ((0.5, -0.5)^*, (-0.5, -0.5)^*, (0.5, 0.5)^*, (0.5, -0.5)^*),$$

we immediately reach the minimum with a very good accuracy. This emphasizes the importance of the choice of the initial point. On the other hand, if we start from the same initial point Q_i given by (36) and (37), but in the chart indexed by $\mathbf{w} = (0,0,0,0)$ and $\mathbf{u} = ((0,1)^*, (1,0)^*, (0,1)^*, (0,1)^*)$ (the Schur parameters are given by $\mathbf{y} = ((-.338, -.444)^*, (.0476, .506)^*, (.515, .515)^*, (-.3, -.3)^*)$), then we do not meet the boundary and we again reach the minimum with a very good accuracy. This illustrates the dependence on the chart of the iterative path produced by the gradient algorithm.

5. Conclusion. A rational approximation problem in L_2-norm has been studied. A new parametrization of stable all-pass transfer functions has been used, based on Schur analysis [1]. Such an overlapping parametrization (in differential geometry an atlas of charts) has allowed us to use classical optimization procedures within a local neighborhood, changing the neighborhood when necessary, in order to solve our minimization problem. Using the state space approach, other parametrizations of stable all-pass transfer functions are available as the one obtained in [25] in continuous-time, based on the work of Ober on balanced canonical forms [33]. A link between the two approaches is probable and a better understanding of the situation seems desirable. In this connection, a state space formulation of the Schur algorithm has been described in continuous-time in [23]. A balanced canonical form for discrete time stable all-pass systems has been obtained in the SISO case [34] by requiring the realization to be balanced and such that the reachability matrix is upper triangular with positive diagonal entries. This canonical form can be parametrized by the Schur parameters obtained in the classical algorithm (11). The generalization of these results to the multivariable case is under study.

Using this parametrization, a minimization algorithm has been described and its convergence to local minima has been proved. We have implemented this algorithm using the matrix-based scientific software Scilab and demonstrated the procedure of computing a local minimum in many simple examples. Later, using this work, a software package named Hyperion has been implemented by J. Grimm to solve a problem provided by the French CNES: identify from frequency data a 2×2 hyperfrequency filter of order 8. Very good results have been obtained on this problem [9]. However, the selection strategy algorithm used in this package is still basic and must be improved. This is going to be the object of forthcoming research.

Acknowledgments. The authors would like to thank D. Alpay who introduced them to Schur analysis and L. Baratchart for providing them with several helpful suggestions during this research. We would also like to thank E. Saff, A. Gombani, and N. Duddley Ward for their very careful reading of this paper and useful criticisms.

REFERENCES

[1] D. Alpay, L. Baratchart, and A. Gombani, *On the differential structure of matrix-valued rational inner functions*, Oper. Theory Adv. Appl., 73 (1994), pp. 30–66.

[2] D. Alpay and H. Dym, *Hilbert spaces of analytic functions, inverse scattering and operator models* I, Integral Equations Operator Theory, 7 (1984), pp. 589–641.

[3] D. Alpay and H. Dym, *On application of reproducing kernel spaces to the Schur algorithm and rational J–unitary factorization*, Oper. Theory Adv. Appl., 18 (1986), pp. 89–159.

[4] J. Aplevich, *Gradient method for optimal linear system reduction*, Internat. J. Control, 18 (1973), pp. 767–772.

[5] J. Ball, I. Gohberg, and L. Rodman, *Interpolation of rational matrix functions*, Birkhäuser Verlag, Basel, 1990.

[6] L. Baratchart, *Existence and generic properties for L^2 approximants of linear systems*, IMA Journal of Math. Control and Identification, 3 (1986), pp. 89–101.

[7] L. Baratchart, *On the topological structure of inner functions and its use in identification*, in Analysis of Controlled Dynamical Systems, Lyon, France, 1990, Progress in Systems and Control Theory, Vol. 8, Birkhäuser-Verlag, Basel, 1990, pp. 51–59.

[8] L. Baratchart, M. Cardelli, and M. Olivi, *Identification and rational L^2 approximation : a gradient algorithm*, Automatica, 27(1991), pp. 413–418.

[9] L. Baratchart, J. Grimm, J. Leblond, M. Olivi, F. Seyfert, and F. Wielonsky, *Identification d'un filtre hyperfréquence*. Rapport Technique 219, INRIA, 1998.

[10] L. Baratchart and M. Olivi, *Index of critical points in L^2-approximation*, System Control Lett., 10 (1988), pp. 167–174.

[11] L. Baratchart and M. Olivi, *Inner-unstable factorization of stable rational transfer functions*, in Modeling, Estimation and Control of Systems with Uncertainty, G. D. Masi, A. Gombani, and A. Kurzhansky, eds., Vol. 10 of Progress in System and Control Theory, Birkhäuser-Verlag, Basel, 1991, pp. 22–39.

[12] L. Baratchart and M. Olivi, *Critical points and error rank in best H^2 matrix rational approximation of fixed McMillan degree*, Constructive Approximation, 14 (1998), pp. 273–300.

[13] L. Baratchart, M. Olivi, and F. Wielonsky, *On a rational approximation problem in the real Hardy space H_2*, Theoretical Computer Science, 94 (1992), pp. 175–197.

[14] L. de Branges and J. Rovnyak, *Canonical models in quantum scattering theory*, in C. Wilcox, editor, Perturbation theory and its applications in quantum mechanics, Holt, Rinehart and Winston, New York, 1966, pp. 295–392.

[15] R. Douglas, H. Shapiro, and A. Shields, *Cyclic vectors and invariant subspaces for the backward shift operator*, Annales de l'Institut Fourier (Grenoble), 20 (1970), pp. 37–76.

[16] H. Dym, *J-contractive matrix functions, reproducing kernel spaces and interpolation*, CBMS Lecture Notes, Vol. 71, American Mathematical Society, Providence, RI, 1989.

[17] H. Dym, *Shifts, realizations and interpolation, redux*, Oper. Theory Adv. Appl., 73 (1994), pp. 182–243.

[18] C. Foias and A. Frazho, *The commutant lifting approach to interpolation problems*, Oper. Theory Adv. Appl., OT44, Birkhäuser-Verlag, Basel, 1990.

[19] P. Fuhrmann, *Linear Systems and Operators in Hilbert Spaces*, McGraw–Hill, New York, 1981.

[20] F. Gantmacher, *Theorie des matrices* I: *theorie generale*, Dunod, Paris, 1966.

[21] J. Garnett, *Bounded Analytic Functions*, Academic Press, New York, London, 1981.

[22] K. Glover and J. Willems, *Parametrization of linear dynamical systems: canonical forms and identifiability*, IEEE Trans. Automat. Control, 19 (1974), pp. 640–646.

[23] A. Gombani and M. Olivi, *A new parametrization of rational inner functions of fixed degree: Schur parameters and realizations*, Math. Control Signals Systems, submitted.

[24] B. Hanzon, *On the differentiable manifold of fixed order stable linear systems*, Systems Control Lett., 13 (1989), pp. 345–352.

[25] B. Hanzon, *A new balanced canonical form for stable multivariable systems*, IEEE Trans. Automat. Control, 40 (1995), pp. 374–378.

[26] M. Hazewinkel and R. Kalman, *Moduli and canonical forms for linear systems*, Tech. report, Economic Institute, Erasmus University, Rotterdam, 1974.

[27] K. Hoffman, *Banach Spaces of Analytic Functions*, Dover, New York, 1988.

[28] T. Kailath, *Linear Systems*, Prentice–Hall, Englewood Cliffs, NJ, 1980.

[29] T. Kailath, *Signal processing applications of some moment problems*, in Moments in Mathematics, Vol. 37, Proc. Symposia in Applied Mathematics, American Mathematical Society, Providence, RI, 1987, pp. 71–109.

[30] W. Krajewski, A. Lepschy, M. Redivo-Zaglia, and U. Viaro, *A program for solving the L_2 reduced-order model problem with fixed denominator degree*, Numer. Algorithms, 9(1995), pp. 355–377.

[31] S. Kung and D. Lin, *Optimal Hankel norm model reduction: Multivariable systems*, IEEE Trans. Automat. Control, 26 (1981), pp. 832–854.

[32] L. Meier and D. Luenberger, *Approximation of linear constant system*, in IEEE Trans. Automat. Control, 12 (1967), pp. 585–588.

[33] R. Ober, *Balanced realizations: Canonical form, parametrization, model reduction*, Internat. J. Control, 46 (1987), pp. 643–670.

[34] R. Peeters and B. Hanzon, *A balanced canonical form for discrete-time stable all-pass systems*, Systems and Networks: Mathematical Theory and Applications, Vol. 2, Akademie-Verlag, Berlin, 1994, pp. 417–420.

[35] V. Potapov, *The multiplicative structure of J-contractive matrix functions*, Amer. Math. Soc. Transl., 15 (1960), pp. 131–243.

[36] I. Schur, *On power series which are bounded in the interior of the unit circle*, I. Schur. methods in operator theory and signal processing, Oper. Theory Adv. Appl., 18 (1986). Translation from J. Reine Angew. Math. 147, 205-232 (1917).

[37] J. Spanos, M. Milman, and D. Mingori, *A new algorithm for L_2 optimal model reduction*, Automatica, 28 (1992), pp. 897–909.

[38] B. Wahlberg, *On System Identification and Model Reduction*, Report LiTH-ISY-I-0847, University of Linköping, Sweden, 1987.

[39] J. L. Walsh, *Interpolation and approximation by rational functions in the complex domain*, Amer. Math. Soc. Publ., 1962.
[40] D. Wilson, *Model reduction for multivariables systems*, Internat. J. Control, 20 (1974), pp. 57–64.
[41] D. A. Wilson, *Optimum solution of model-reduction problem*, Proc. IEEE, 117 (1970), pp. 1161–1165.
[42] N. Young, *The Nevanlinna–Pick problem for matrix-valued functions*, J. Operator Theory, 15 (1986), pp. 239–265.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 6, pp. 2128–2147, November 1998

015

APPROXIMATE CONTROLLABILITY OF A SEMILINEAR HEAT EQUATION IN $\mathbb{R}^{N*}$

LUZ DE TERESA†

Abstract. We prove the approximate controllability of the semilinear heat equation in $\mathbb{R}^N$. We introduce the weighted Sobolev spaces of Escobedo and Kavian and in that functional setting we adapt the technique introduced by Fabre, Puel, and Zuazua for the problem in bounded domains. That is, we first prove the approximate controllability of the linear equation and by a fixed-point method obtain the main result.

Key words. approximate controllability, unbounded domains, weighted Sobolev spaces

AMS subject classifications. 93B05, 35K05, 35K55

PII. S0036012997322042

1. Introduction. This paper is concerned with the approximate controllability of the semilinear heat equation in unbounded domains Ω when the control acts on the interior of Ω.

The general statement of the problem is as follows: Let ω be an open and nonempty subset of Ω. We consider the following semilinear heat equation:

$$\begin{cases} y_t - \Delta y + f(y) = h\chi_\omega \ \text{ in } \ Q = \Omega \times (0,T), \\ y = 0 \ \text{ on } \ \Sigma = \partial\Omega \times (0,T), \\ y(x,0) = y^0(x) \ \text{ in } \ \Omega, \end{cases} \tag{1.1}$$

where $h = h(x,t) \in L^2(\omega \times (0,T))$, χ_ω is the characteristic function of ω, and $y^0 \in L^2(\Omega)$. We shall assume that f is a real and globally Lipschitz function such that $f(0) = 0$. Let M be its Lipschitz constant, i.e.,

$$|f(s) - f(\sigma)| \le M|s - \sigma|, \qquad \forall s, \forall \sigma \in \mathbb{R}. \tag{1.2}$$

We say that the system (1.1) is approximately controllable in $L^2(\Omega)$ at time $T > 0$ if the following holds: "For every $y^0 \in L^2(\Omega)$, the set of reachable states at time $T > 0$,

$$E(T) = \{y(x,T),\ y \ \text{ solution of (1.1) with } \ h \in L^2(Q)\}$$

is dense in $L^2(\Omega)$."

When Ω is a bounded set, Fabre, Puel, and Zuazua [6] proved the approximate controllability of (1.1) in $L^p(\Omega), 1 \le p < \infty$. Their proof is divided in two parts: a) approximate controllability of the linearized system; b) fixed-point technique.

This technique cannot be applied when Ω is an unbounded set since the compactness of Sobolev's embeddings is one of the main ingredients used in b). In [14], we proved the approximate controllability of the semilinear heat equation in unbounded domains by an approximation method (used also in [13] for an insensitizing control problem). We considered the control problem in bounded domains of the form $\Omega_r = \Omega \cap B_r$, where B_r denotes the ball centered in zero of radius r. We showed

*Received by the editors May 28, 1997; accepted for publication September 26, 1997; published electronically August 31, 1998.

http://www.siam.org/journals/sicon/36-6/32204.html

†Instituto de Matemáticas, Universidad Nacional Autonoma de México, Circuito Exterior, C.U. 04510, D.F. México (deteresa@matem.unam.mx).

that the controls proposed in [6], for the problem restricted to Ω_r, converge weakly in $L^p(\omega \times (0,T))$ as $r \to \infty$ to an approximate control for our problem in the whole domain Ω. Nevertheless, this proof is indirect and rather technical and it does not provide a method to effectively compute numerically the control.

The aim of this paper is to adapt the techniques introduced in [6] to unbounded domains by introducing the weighted Sobolev spaces of Escobedo and Kavian [4] that guarantee the compactness of the Sobolev's embeddings. The use of these spaces is interesting since it allows us to prove our approximate controllability result in a "direct" way. It could also be interesting for numerical purposes.

Even if it is a known technique in the study of the asymptotic behavior of some systems, it has not been used for controllability problems, as far as we know.

Nevertheless, this proof is valid only when $\Omega = \mathbb{R}^N$ or a cone-like domain, since it is necessary to make a change of variables. For simplicity we limit ourselves to the case $\Omega = \mathbb{R}^N$ and $p = 2$.

Under these conditions, the semilinear heat equation given in (1.1) reads as follows:

$$\begin{cases} u_t - \Delta u + f(u) = h\chi_\omega \quad \text{in} \quad \mathbb{R}^N \times (0,T), \\ u(x,0) = u^0(x) \quad \text{in} \quad \mathbb{R}^N. \end{cases} \tag{1.3}$$

The main result of the paper follows.

PROPOSITION 1.1. *If f is globally Lipschitz and $f(0) = 0$, system (1.3) is approximately controllable in $L^2(\mathbb{R}^N)$ at any time $T > 0$. Furthermore, we can reach a dense set of final states by using controls of the form*

$$h(x,t) = (t+1)^{-N/2-1} K^{-1}\left(\frac{x}{\sqrt{1+t}}\right) \|\phi\|_{L^1(q')} \tilde{\lambda}\chi_\omega,$$

where $\tilde{\lambda} \in sgn\tilde{\phi}$ and $\tilde{\phi}(x,t) = \phi(\frac{x}{\sqrt{1+t}}, \log(1+\mathrm{t}))$ with ϕ solution of a suitable heat equation, $q' = \{(y,s), s \in (0, \log(T+1)), y = e^{-s/2}x, x \in \omega\}$ and K a weight.

In order to implement the fixed point argument introduced by Fabre, Puel, and Zuazua in [6], we must first study the approximate controllability problem associated with the linear system with potential:

$$\begin{cases} u_t - \Delta u + a(x,t)u = h\chi_\omega \quad \text{in} \quad \mathbb{R}^N \times (0,T), \\ u(x,0) = u^0(x) \quad \text{in} \quad \mathbb{R}^N \end{cases} \tag{1.4}$$

with $a(x,t) \in L^\infty((0,T) \times \mathbb{R}^N)$.

To avoid the problems related to the noncompactness of the Sobolev's embeddings in $\mathbb{R}^N$, we consider the operator

$$Lf = -\Delta f - \frac{y \cdot \nabla f}{2} = -\frac{1}{K}\mathrm{div}(K\nabla f),$$

$$K(y) = \exp\left(\frac{|y|^2}{4}\right),$$

$$D(L) \subset L^2(K) = \left\{ f : \int_{\mathbb{R}}^N K(y)|f(y)|^2 dy < \infty \right\}$$

and the evolution equation

$$\begin{cases} v_s + Lv + A(s,y)v = \dfrac{N}{2}v + H(y,s)\chi_{\omega'(s)} \quad \text{in} \quad \mathbb{R}^N \times (0,S), \\ v(y,0) = v^0(y) \quad \text{in} \quad \mathbb{R}^N \end{cases} \tag{1.5}$$

with $v^0(y) \in L^2(K)$, ω' a suitable set in $\mathbb{R}^N \times (0,S)$, and $S = \log(\mathrm{T}+1)$. A change of variables transforms (1.4) in (1.5). In fact, if we define for $s \geq 0,\ y \in \mathbb{R}^N$

$$\begin{aligned} v(y,s) &= e^{\frac{sN}{2}} u(e^{s/2}y, e^s - 1), \\ A(y,s) &= e^s a(e^{s/2}y, e^s - 1), \\ H(y,s) &= e^{\frac{s(N+2)}{2}} h(e^{s/2}y, e^s - 1); \end{aligned} \tag{1.6}$$

then, for $u^0 \in L^2(K)$ and u solution of (1.4), v verifies (1.5) with $v^0 = u^0$, $S = \log(T+1)$, and $\omega'(s) = e^{-s/2}\omega$.

Reciprocally, if we know a solution v of (1.5), and we define for $t \geq 0$, $x \in \mathbb{R}^N$

$$u(x,t) = (1+t)^{-N/2} v\left(\frac{x}{\sqrt{1+t}}, \log(1+t)\right), \tag{1.7}$$

it is not difficult to see that u satisfies (1.4) with $u^0 = v^0$, $T = e^S - 1$, and

$$\begin{aligned} a(x,t) &= (1+t)^{-1} A\left(\frac{x}{\sqrt{1+t}}, \log(1+t)\right), \\ h(x,t) &= (1+t)^{-N/2-1} H\left(\frac{x}{\sqrt{1+t}}, \log(1+t)\right). \end{aligned}$$

That change of variables is interesting because the operator L defined above has compact inverse in $L^2(K)$ and then the equation (1.5) can be studied in the same manner as the heat equation in a bounded region Ω of $\mathbb{R}^N$.

The paper is organized as follows. In section 2 we introduce the weighted Sobolev spaces of Escobedo and Kavian and give some a priori estimates of the norm of the solution in these spaces. In section 3 we state some preliminary results concerning the existence and properties of the minima of a functional arising in the approximate controllability of the linear case. Section 4 is devoted to proving the approximate controllability of the linear case. We conclude with section 5 by proving Proposition 1.1 by a fixed-point method. Finally, in section 6 we discuss some extensions of the methods used in this paper.

2. Weighted Sobolev spaces. In this section we introduce the weighted Sobolev spaces of Escobedo and Kavian. We are going to see that system (1.5) is well posed in these spaces. We give some a priori estimates on the norm of the solution.

DEFINITION 2.1. *Let $K(y) = \exp(|y|^2/4)$. We define*

$$L^2(K) = \left\{ f; \int_{\mathbb{R}^N} K(y)|f(y)|^2 dy < \infty \right\},$$

$$H^s(K) = \{ f \in L^2(K); D^\alpha f \in (L^2(K))^N, \forall \alpha : |\alpha| \leq s \}.$$

We endow this space with the inner products $(f,g)_{L^2(K)} = \int_{\mathbb{R}^N} Kfg dy$, $(f,g)_{H^1(K)} = (f,g)_{L^2(K)} + (\nabla f, \nabla g)_{(L^2(K))^N}$, $(f,g)_{H^2(K)} = (f,g)_{H^1(K)} + (\Delta f, \Delta g)_{L^2(K)}$ with, respectively, the associated norm $\| \ \|_{L^2(K)}$, $\| \ \|_{H^1(K)}$, $\| \ \|_{H^2(K)}$.

We recall the following result due to Escobedo and Kavian (see [4], [7]).

LEMMA 2.2.

i) *There exists* $C > 0$ *such that for every* $v \in H^1(K)$

$$\int_{\mathbb{R}^N} K(y)|v(y)|^2|y|^2dy \le C\int_{\mathbb{R}^N} K(y)|\nabla v(y)|^2dy.$$

ii) *The imbedding* $H^1(K) \hookrightarrow L^2(K)$ *is compact.*

iii) $\forall v \in H^1(K)$, $\frac{N}{2}\int_{\mathbb{R}^N} K(y)|v|^2dy \le \int_{\mathbb{R}^N} K(y)|\nabla v|^2dy$.

iv) $v \in H^1(K) \iff K^{1/2}v \in H^1(\mathbb{R}^N)$.

v) $\forall f \in L^2(K)$ *there exists a unique* $u \in H^2(K)$ *such that* $Lu = f$, *i.e.,* $D(L) = H^2(K)$.

vi) $\varphi_1 = \exp(-|y|^2/4)$ *is an eigenfunction of* L *corresponding to* $\lambda_1 = N/2$, *the minimum eigenvalue of* L, *i.e.,* $L\varphi_1 = N/2\varphi_1$.

vii) L *is a positive operator, self-adjoint in* $L^2(K)$ *with compact inverse.*

viii)

$$\begin{array}{l} \textit{If} \quad N = 1,\ v \in H^1(K), \quad \textit{then} \quad K^{1/2}v \in L^\infty(\mathbb{R}). \\ \textit{If} \quad N = 2,\ H^1(K) \subset L^q(K) \quad \forall q \ge 2, \quad \textit{and} \quad q < \infty. \\ \textit{If} \quad N \ge 3,\ H^1(K) \subset L^{2^*}(K) \ \textit{with}\ 2^* = \frac{2N}{N-2}. \end{array}$$

Remark 1. Moreover, we have $L^2(K) \subset L^1(\mathbb{R}^N)$ with continuous embedding. If $v \in L^2(K)$, then

$$\int_{\mathbb{R}^N} |v| = \int_{\mathbb{R}^N} \frac{1}{K^{1/2}} K^{1/2}|v| \le \left(\int_{\mathbb{R}^N} K|v|^2\right)^{1/2} \left(\int_{\mathbb{R}^N} \frac{1}{K}\right)^{1/2} < \infty.$$

The results of Lemma 2.2 allow us to prove (see, e.g., [9, Theorem 4.1, p. 257]), that if $H \in L^\infty(0, S; L^2(K))$, $A(s, y) \in L^\infty(\mathbb{R}^N \times (0, S))$, and $u_0 \in L^2(K)$, then (1.5) has a unique solution:

$$\begin{array}{l} v \in C([0, S]; L^2(K)) \cap C((0, S]; H^2(K)), \\ v_s \in L^\infty((0, S]; L^2(K)). \end{array} \tag{2.1}$$

Moreover, v is given by the variation of constants formula. That is,

$$v(s) = S_*(s)v^0 + \int_0^s S_*(s-\sigma)(A(\sigma)v(\sigma) + H(\sigma))d\sigma, \tag{2.2}$$

where S_* is the analytic semigroup generated by $L-(N/2)I$ in $L^2(K)$. This semigroup is given in the following way (cf. Kavian [7], Escobedo–Zuazua [5]): For every $g \in L^2(K)$, for every $s > 0$, and for every $y \in \mathbb{R}^N$

$$(S_*(s)g)(y) = e^{sN/2}(G(e^s - 1) * g)(e^{s/2}y),$$

where G is the heat kernel, i.e., $G(x, t) = (4\pi t)^{-N/2}\exp(\frac{-|x|^2}{4t})$.

PROPOSITION 2.3. *Let* v *be the solution of* (1.5) *given in* (2.1). *Then, there exists a constant* $C = C(A, S) > 0$ *depending only on* A, S *such that*

i)

$$\|v\|_{\infty,K} \le C(\|v^0\|_{L^2(K)} + \|H\|_{\infty,K}), \tag{2.3}$$

ii)

$$\|v(s)\|_{H^1(K)} \le C(1 + s^{-1/2})(\|v^0\|_{L^2(K)} + \|H\|_{\infty,K}) \ \forall s > 0, \tag{2.4}$$

iii)

$$\|v_s(s)\|_{(H^1(K))'} \leq C(1+s^{-1/2})\left(\|v^0\|_{L^2(K)} + \|H\|_{\infty,K}\right) \ \forall s>0, \tag{2.5}$$

where $\|\ \|_{\infty,K}$ *denotes the norm in* $L^\infty((0,S);L^2(K))$ *and* $(H^1(K))'$ *is the dual space of* $H^1(K)$.

Proof. The proofs of i) and ii) are classical. We give a sketch of the proof. For proving i) it is enough to multiply (1.5) by v (in $L^2(K)$) and then to apply Hölder's, Schwarz's, and Gronwall's inequalities.

For proving ii) we express v by the variation of constants formula (2.2) and take $H^1(K)$ norms (see, e.g., [3] or [12]). To prove iii) we observe that (1.5) is satisfied in $(H^1(K))'$, that is,

$$v_s = -Lv + A(s,y)v + \frac{N}{2}v + H(y,s)\chi_{\omega'(s)} \ \text{ in } \ (H^1(K))'.$$

Let $\varphi \in H^1(K)$; then

$$\begin{aligned}\langle v_s(s),\varphi\rangle_1 = \ & -\int_{\mathbb{R}^N} K(Lv(s))\varphi - \int_{\mathbb{R}^N} KA(s)v(s)\varphi + \int_{\mathbb{R}^N}\frac{N}{2}Kv(s)\varphi \\ & + \int_{\mathbb{R}^N} KH(s)\varphi,\end{aligned}$$

where $\langle\ ,\ \rangle_1$ denotes the duality pairing $((H^1(K))', H^1(K))$. We then have that

$$\|v_s(s)\|_{(H^1(K))'} \leq C\{\|v(s)\|_{H^1(K)} + \|H(s)\|_{L^2(K)}\},$$

and by ii) we conclude

$$\|v_s(s)\|_{(H^1(K))'} \leq C(1+s^{-1/2})\left(\|v^0\|_{L^2(K)} + \|H(s)\|_{\infty,K}\right) \ \forall s>0,$$

where the constant C may vary from line to line. ∎

3. Study of a functional arising in the controllability of linear systems. In this section we study the existence and properties of the minima of a functional arising in the linear control problem. Let $\phi^0 \in L^2(K)$, $v^1 \in L^2(K)$, $A(y,s) \in L^\infty(\mathbb{R}^N \times (0,S))$, and $\alpha > 0$. We introduce the functional

$$J(\phi^0; A, v^1) = \frac{1}{2}\left(\int_{q'} |\phi(x,s)|dsdx\right)^2 + \alpha\|\phi^0\|_{L^2(K)} - \int_{\mathbb{R}^N} Kv^1\phi^0, \tag{3.1}$$

where $q' = \{(y,s); s \in (0,T), y = e^{-s/2}x, x \in \omega\}$ and ϕ denotes the solution of the transposed problem:

$$\begin{cases} -\phi_s + L\phi + A(y,s)\phi = \dfrac{N}{2}\phi \text{ in } \ \mathbb{R}^N \times (0,S), \\ \phi(S) = \phi^0 \ \text{ in } \ \mathbb{R}^N. \end{cases} \tag{3.2}$$

The main result of this section is the following.

PROPOSITION 3.1. *For every* $\alpha > 0, v^1 \in L^2(K)$ *and* $A \in L^\infty(\mathbb{R}^N \times (0,S))$, $J(.;A,v^1)$ *is a real strictly convex continuous function in* $L^2(K)$ *and satisfies*

$$\liminf_{\|\phi^0\|_{L^2(K)}\to\infty} \frac{J(\phi^0;A,v^1)}{\|\phi^0\|_{L^2(K)}} \geq \alpha. \tag{3.3}$$

The functional $J(.;A,v^1)$ *achieves its minimum at a unique point* $\hat{\phi}^0 \in L^2(K)$. *Furthermore,*

$$\hat{\phi}^0 = 0 \Longleftrightarrow \alpha \geq \|v^1\|_{L^2(K)}.$$

The proof of Proposition 3.1 is based on the following two results. The first one is a unique continuation property that is a direct consequence of a result due to Saut and Scheurer [10, Th. 1.1]. The second one is a classical compactness result (see, e.g., [11, Th. 5, p. 84]). We state it for completeness.

PROPOSITION 3.2. *Let* ω *be an open an nonempty set,* $\omega'(s) = e^{-s/2}\omega$ *and* $q' = \{(y,s), s \in (0,T), y \in \omega'(s)\}$. *Assume that* $A(y,s) \in L^\infty(\mathbb{R}^N \times (0,S))$. *Let* $\phi \in L^2((0,S);H^2(K))$ *be such that*

$$\begin{cases} -\phi_s + L\phi + A(y,s)\phi = \dfrac{N}{2}\phi \text{ in } \mathbb{R}^N \times (0,S), \\ \phi = 0 \text{ in } q'. \end{cases} \tag{3.4}$$

Then $\phi \equiv 0$.

THEOREM 3.3. *Let* X, B, Y *be Banach spaces such that* $X \subset B \subset Y$ *with continuous embeddings, the embedding* $X \subset B$ *being compact. Let* $1 \leq p \leq \infty$. *If* $\mathcal{F}$ *is a bounded subset of* $L^p(0,T;X)$ *and*

$$\|\tau_h f - f\|_{L^p(0,T-h;Y)} \to 0 \text{ as } h \to 0 \text{ uniformly for } f \in \mathcal{F},$$

where $\tau_h f(t) = f(t+h)$, *then* $\mathcal{F}$ *is relatively compact in* $L^p(0,T;B)$ *(in* $C([0,T];B)$ *if* $p = \infty$*).*

Proof of Proposition 3.1. The continuity and strict convexity are immediate. In order to prove (3.3) we proceed by contradiction. That is, suppose that there exists a sequence $\{\phi_n^0\} \subset L^2(K)$ such that

$$\|\phi_n^0\|_{L^2(K)} \to \infty, \qquad n \to \infty, \tag{3.5}$$

$$\liminf_{n\to\infty} \frac{J(\phi_n^0;A,v^1)}{\|\phi_n^0\|_{L^2(K)}} < \alpha. \tag{3.6}$$

We put $\psi_n^0 = \frac{\phi_n^0}{\|\phi_n^0\|_{L^2(K)}}$. Then,

$$\frac{J(\phi_n^0;A,v^1)}{\|\phi_n^0\|_{L^2(K)}} = \frac{1}{2}\|\phi_n^0\|_{L^2(K)}\left(\int_{q'}|\psi_n|\right)^2 + \alpha - \int_{\mathbb{R}^N} Kv^1\psi_n^0, \tag{3.7}$$

where ψ_n denotes the solution of (3.2) with data $\psi_n(S) = \psi_n^0$.

Since $\|\psi_n^0\|_{L^2(K)} = 1$, we can extract a subsequence (still denoted ψ_n^0) such that $\psi_n^0 \rightharpoonup \psi^0$ weakly in $L^2(K)$. On the other hand, by Proposition 2.3, $\|\psi_n\|_{L^2((0,S);L^2(K))} \leq C$, and, therefore, we can extract a subsequence (of the previous one) such that

$$\psi_n \rightharpoonup \psi \quad \text{weakly in } L^2((0,S);L^2(K)). \tag{3.8}$$

But (2.4) and (2.5) in Proposition 2.3 and Theorem 3.3 imply that for every $T \geq \varepsilon > 0$

$$\psi_n \to \psi \quad \text{strongly in } C([0,S-\varepsilon];L^2(K)). \tag{3.9}$$

Multiplying the equation satisfied by ψ_n by suitable test functions, and passing to the limit along the sequence, we obtain that $\psi \in L^2(0,S;H^2(K))$ satisfies

$$-\psi_s + L\psi + A\psi = \frac{N}{2}\psi \quad \text{in} \quad \mathbb{R}^N \times (0,S), \qquad \psi(S) = \psi^0.$$

Let us see that (3.5) and (3.6) imply $\psi \equiv 0$. Observe that (3.8) and (3.9) imply that $\psi_n \chi_{\omega'(s)} \to \psi \chi_{\omega'(s)}$ in $L^1(\mathbb{R}^N \times (0,S))$. In fact,

$$\int_{q'} |\psi - \psi_n| \le \int_0^{S-\varepsilon} \int_{\mathbb{R}^N} |\psi - \psi_n| + \int_{S-\varepsilon}^{S} \int_{\mathbb{R}^N} |\psi - \psi_n|.$$

In view of (3.9), the first term in the right-hand side converges to zero as $n \to \infty$. The second term is, by (3.8), bounded by εC with C independent of n. Therefore,

$$\int_{q'} |\psi| = \lim_{n\to\infty} \int_{q'} |\psi_n|.$$

We observe that $\int_{q'} |\psi| = 0$. Otherwise, (3.5) and (3.7) contradict (3.6). This proves that $\psi \equiv 0$ in q'. Hence, by Proposition 3.2, $\psi \equiv 0$ in $\mathbb{R}^N \times (0,S)$ and $\psi^0 = 0$. But

$$J(\phi_n^0; A, v^1) \ge \|\phi_n^0\|_{L^2(K)} \left(\alpha - \int_{\mathbb{R}^N} K v^1 \psi_n^0 dy \right).$$

In view of (3.5) and the convergence $\psi_n^0 \rightharpoonup 0$, we obtain a contradiction with (3.6) and we prove (3.3).

Now, if $\alpha \ge \|v^1\|_{L^2(K)}$, we have $J(\phi^0; A, v^1) \ge 0$ for every ϕ^0 and, hence, $\hat{\phi}^0 = 0$. Suppose now $\hat{\phi}^0 = 0$; then $J(\phi^0; A, v^1) \ge 0$ for all $\phi^0 \in L^2(K)$. In particular,

$$\lim_{t\to 0^+} \frac{J(t\phi^0; A, v^1)}{t} \ge 0 \quad \forall \phi^0 \in L^2(K),$$

and, therefore,

$$\alpha \|\phi^0\|_{L^2(K)} \ge \int_{\mathbb{R}^N} K v^1 \phi^0 \quad \forall \phi^0 \in L^2(K).$$

In particular, for $\phi^0 = v^1 \in L^2(K)$ we get $\alpha \ge \|v^1\|_{L^2(K)}$. □

In order to study the nonlinear case, we need to make precise the dependence of the minima with respect to the potential. This is gathered in the following proposition.

PROPOSITION 3.4.

(i) *If we denote by M the mapping*

$$M : L^\infty(\mathbb{R}^N \times (0,S)) \times L^2(K) \to L^2(K); \; M(A, v^1) = \hat{\phi}^0,$$

and if W is a compact subset of $L^2(K)$ and B a bounded subset of $L^\infty(\mathbb{R}^N \times (0,S))$, then $M(B \times W)$ is a bounded subset of $L^2(K)$.

(ii) *Moreover, if $A_n \rightharpoonup A$ weakly* in $L^\infty((0,S) \times \mathbb{R}^N)$ and $v_n^1 \to v^1$ strongly in $L^2(K)$, then $\hat{\phi}_n^0$ converges strongly in $L^2(K)$ to $\hat{\phi}^0$.*

Proof. In order to get (i), we first prove that (3.3) is uniform for (A, v^1) in $B \times W$. We again argue by contradiction and follow the same argument used in the proof of (3.3): suppose that there exist sequences $\{A_n\} \subset B; \{v_n^1\} \subset W$ such that the

corresponding sequence of minimizers $\{\hat{\phi}_n^0\} \subset L^2(K)$ verify (3.5). Since B is bounded and W is compact, there exist $A \in L^\infty(\mathbb{R}^N \times (0,S))$, $v^1 \in L^2(K)$, and a subsequence (still denoted by n) such that

$$A_n \rightharpoonup A \quad \text{weakly* in} \quad L^\infty(\mathbb{R}^N \times (0,S)), \tag{3.10}$$

and

$$v_n^1 \to v^1 \quad \text{strongly in} \quad L^2(K). \tag{3.11}$$

Suppose that

$$\liminf_{n\to\infty} \frac{J(\hat{\phi}_n^0; A_n, v_n^1)}{\|\hat{\phi}_n^0\|_{L^2(K)}} < \alpha. \tag{3.12}$$

As above, we denote by $\psi_n^0 = \frac{\phi_n^0}{\|\phi_n^0\|_{L^2(K)}}$. Since $\|\psi_n^0\|_{L^2(K)} = 1$ and A_n is bounded in $L^\infty(\mathbb{R}^N \times (0,S))$, the solution ψ_n of (3.2) corresponding to A_n with data $\psi_n(S) = \psi_n^0$, is also bounded. We can repeat the arguments of the previous proof to obtain

$$\liminf_{n\to\infty} \frac{J(\hat{\phi}_n^0; A_n, v_n^1)}{\|\hat{\phi}_n^0\|_{L^2(K)}} \geq \alpha, \tag{3.13}$$

which contradicts (3.12). Now, if we suppose that the range of M is not bounded, we can construct a sequence $\{A_n, v_n^1, \hat{\phi}_n^0\} \subset B \times W \times L^2(K)$ satisfying (3.10), (3.11), and $\|\hat{\phi}_n^0\|_{L^2(K)} \to \infty$. But for every n,

$$J(\hat{\phi}_n^0; A_n, v_n^1) \leq J(0; A_n, v_n^1) = 0, \tag{3.14}$$

which contradicts (3.13) and proves (i).

We now prove (ii). From (i), we know that the minimizers $\hat{\phi}_n^0$ are bounded in $L^2(K)$. Hence, they weakly converge to an element $\tilde{\phi}^0 \in L^2(K)$ and $\hat{\phi}_n$ weakly converge in $L^2(0,S;L^2(K))$ to the solution $\tilde{\phi}$ of (3.2) corresponding to A and $\tilde{\phi}(S) = \tilde{\phi}^0$. We can argue, as in the proof of Proposition 3.1, to show that

$$\int_{q'} |\tilde{\phi}| = \lim_{n\to\infty} \int_{q'} |\hat{\phi}_n|,$$

and then

$$J(\tilde{\phi}^0; A, v^1) \leq \liminf_{n\to\infty} J(\hat{\phi}_n^0; A_n, v_n^1). \tag{3.15}$$

Let us prove that for every $\phi^0 \in L^2(K)$,

$$\lim_{n\to\infty} J(\phi^0; A_n, v_n^1) = J(\phi^0; A, v^1). \tag{3.16}$$

We denote by ϕ_n the solution of (3.2) with potential A_n and $\phi_n(S) = \phi^0$. Since $\{A_n\}$ is a uniformly bounded sequence, Proposition 2.3 and Theorem 3.3 give us the strong convergence of ϕ_n in $L^2(0,S;L^2(K))$ to the solution ϕ of (3.2) with potential A and $\phi(S) = \phi^0$. Passing to the limit in each term of the functional, we obtain (3.16).

Moreover, we have that

$$\forall \phi^0 \in L^2(K), \quad J(\hat{\phi}_n^0; A_n, v_n^1) \leq J(\phi^0; A_n, v_n^1). \tag{3.17}$$

Combining (3.15) and (3.17), we obtain

$$J(\tilde{\phi}^0; A, v^1) \leq \liminf_{n\to\infty} J(\hat{\phi}_n^0; A_n, v_n^1) \leq \limsup_{n\to\infty} J(\hat{\phi}_n^0; A_n, v_n^1)$$
$$\leq \lim_{n\to\infty} J(\phi^0; A_n, v_n^1) \quad \forall \phi^0 \in L^2(K); \tag{3.18}$$

hence, we obtain

$$\forall \phi^0 \in L^2(K), \quad J(\tilde{\phi}^0; A, v^1) \leq J(\phi^0; A, v^1). \tag{3.19}$$

The strict convexity of $J(.; A, v^1)$ and (3.19) imply that $\tilde{\phi}^0 = \hat{\phi}^0$. But, with this equality, (3.18) gives

$$J(\hat{\phi}^0; A, v^1) = \lim_{n\to\infty} J(\hat{\phi}_n^0; A_n, v_n^1). \tag{3.20}$$

On the other hand, we have

$$\lim_{n\to\infty} \int_{\mathbb{R}^N} K v_n^1 \hat{\phi}_n^0 dy = \int_{\mathbb{R}^N} K v^1 \hat{\phi}^0 dy,$$
$$\int_{q'} |\hat{\phi}| dy ds = \lim_{n\to\infty} \int_{q'} |\hat{\phi}_n| dy ds,$$
$$\|\hat{\phi}^0\|_{L^2(K)} \leq \liminf_{n\to\infty} \|\hat{\phi}_n^0\|_{L^2(K)}.$$

Equation (3.20) then implies that

$$\lim_{n\to\infty} \|\hat{\phi}_n^0\|_{L^2(K)} = \|\hat{\phi}^0\|_{L^2(K)}.$$

Since $\hat{\phi}_n^0$ converges weakly to $\hat{\phi}^0$ in $L^2(K)$ and $L^2(K)$ is uniformly convex, we deduce that the convergence is strong in $L^2(K)$, which ends the proof of Proposition 3.4. □

We are now going to interpret the results of Proposition 3.1: Since the functional $J(.; A, v^1)$ is convex continuous with real values, it possesses a subdifferential at every point of $L^2(K)$ (e.g., Aubin [1, p. 187]). At its minimum, we have $0 \in \partial J(\hat{\phi}^0; A, v^1)$.

Let us now prove the following proposition.

PROPOSITION 3.5. *For every $\phi^0 \in L^2(K)$, $\phi^0 \neq 0$, denoting by ϕ the solution of* (3.2) *with $\phi(S) = \phi^0$, we have*

$$\partial J(\phi^0; A, v^1) = \Big\{ \xi \in L^2(K); \exists \lambda \in sgn(\phi)\chi_{q'} \quad \textit{satisfying}$$
$$\int_{\mathbb{R}^N} K\xi(y)\theta^0(y)dy = \left(\int_{q'} |\phi| dy\ ds\right)\left(\int_{q'} \lambda\theta dy\ ds\right)$$
$$+\alpha \int_{\mathbb{R}^N} \frac{K\phi^0(y)}{\|\phi^0\|_{L^2(K)}}\theta^0(y)dy - \int_{\mathbb{R}^N} K v^1(y)\theta^0(y)dy \textit{ for every } \theta^0 \in L^2(K),$$
$$\textit{where } \theta \textit{ is the solution of } (3.2) \textit{ with } \theta(S) = \theta^0 \Big\}.$$

Proof. Since the functions A and v^1 are fixed, we write $J(\phi^0)$. We have $J(\phi^0) = \jmath_1(\phi^0) + \jmath_2(\phi^0)$, where

$$\jmath_1 = \frac{1}{2}\left(\int_{q'} |\phi|\right)^2; \quad \jmath_2 = \alpha\|\phi^0\|_{L^2(K)} - \int_{q'} K v^1 \phi^0.$$

Since $\jmath_2$ is Gateaux differentiable at every $\phi^0 \neq 0$ in $L^2(K)$, and $\jmath_1$ is subdifferentiable at every $\phi^0 \neq 0$ in $L^2(K)$, we have for every $\phi^0 \neq 0$ in $L^2(K)$: $\partial J(\phi^0) = \partial \jmath_1(\phi^0) + \partial \jmath_2(\phi^0)$.

We now determine the set $\partial \jmath_1(\phi^0)$. Let $\xi \in \partial \jmath_1(\phi^0)$. By definition, we have

$$\forall \theta^0 \in L^2(K) \quad (\xi, \theta^0)_{L^2(K)} \leq \lim_{t \to 0^+} \frac{\jmath_1(\phi^0 + t\theta^0) - \jmath_1(\phi^0)}{t}. \tag{3.21}$$

We can prove that

$$\lim_{t \to 0^+} \frac{\jmath_1(\phi^0 + t\theta^0) - \jmath_1(\phi^0)}{t} = \left(\int_{q'} |\phi| \right) \left(\int_{q'/\mathcal{A}} (sgn\phi)\theta + \int_{\mathcal{A}} |\theta| \right),$$

where $\mathcal{A} = \{(y,s) | \phi(y,s) = 0\} \cap q'$. In view of (3.21), that implies

$$(\xi, \theta^0)_{L^2(K)} \leq \left(\int_{q'} |\phi| \right) \left(\int_{q/\mathcal{A}} (sgn\phi)\theta + \int_{\mathcal{A}} |\theta| \right). \tag{3.22}$$

Let us call F the mapping from $L^2(K)$ to $L^1(q')$, $F(\theta^0) = \theta$, with θ the solution of (3.2) corresponding to θ^0. Then, the mapping $F(\theta^0) \to (\xi, \theta^0)_{L^2(K)}$ is a linear form on $F(L^2(K)) \subset L^1(q')$ and applying the Hahn–Banach theorem, there exists a linear form $\mathcal{V}$ on $L^1(q')$, such that $\forall \theta^0 \in L^2(K)$, $(\xi, \theta^0)_{L^2(K)} = \mathcal{V}(\theta)$, and for every $\vartheta \in L^1(q')$

$$\mathcal{V}(\vartheta) \leq \left(\int_{q'} |\phi| \right) \left(\int_{\mathcal{A}} |\vartheta| + \int_{q'/\mathcal{A}} (sgn\phi)\vartheta \right). \tag{3.23}$$

From (3.23), we deduce that $\mathcal{V}$ is linear and continuous on $L^1(q')$ and, hence, $\mathcal{V} \in L^\infty(q')$. Therefore, for every $\vartheta \in L^1(q')$, we have

$$\left| \int_{q'} \mathcal{V}(y,s)\vartheta(y,s)dyds - \left(\int_{q'} |\phi| \right) \left(\int_{q'/\mathcal{A}} (sgn\phi)\vartheta \right) \right| \leq \left(\int_{q'} |\phi| \right) \left(\int_{\mathcal{A}} |\vartheta| \right). \tag{3.24}$$

Take first $\vartheta \in L^1(q')$, whose support is contained in $q'/\mathcal{A}$ to obtain $\mathcal{V}(y,s) = (\int_{q'} |\phi|) \frac{\phi(y,s)}{|\phi(y,s)|}$ for almost every $(y,s) \in q'/\mathcal{A}$. Now take ϑ whose support is contained in $\mathcal{A}$. We obtain that $|\mathcal{V}(y,s)| \leq (\int_{q'} |\phi|)$ almost everywhere on $\mathcal{A}$. This proves that $\mathcal{V} = (\int_{q'} |\phi|)\lambda$ with $\lambda \in sgn\phi\chi_{q'}$.

Reciprocally, let $\mathcal{V} \in (\int_{q'} |\phi|) sgn\phi\chi_{q'}$. Then, if θ is the solution of (3.2) with $\theta(S) = \theta^0$, the mapping $\theta^0 \to \int_{q'} \mathcal{V}\theta dyds$ is linear and continuous on $L^2(K)$. Thus, there exists a unique $\xi \in L^2(K)$ such that

$$(\xi, \theta^0)_{L^2(K)} = \int_{q'} \mathcal{V}\theta(y,s)dyds. \tag{3.25}$$

One can easily prove that ξ satisfies (3.22) and, hence, $\xi \in \partial \jmath_1$ concluding the proof of Proposition 3.5. □

4. The linear case. In this section, we prove the approximate controllability in $L^2(\mathbb{R}^N)$ of the linear heat equation with potential. To this aim we need the following approximate controllability result in $L^2(K)$.

PROPOSITION 4.1. *Let $v^1 \in L^2(K)$ with $\|v^1\|_{L^2(K)} > \alpha$, and $\hat{\phi}$ the solution of (3.2) with $\hat{\phi}(S) = \hat{\phi}^0$ the minimizer of $J(.; A, v^1)$. Then, there exists $\lambda \in sgn(\hat{\phi})\chi_{q'}$ such that the solution v of*

$$\begin{cases} v_s + Lv + A(y,s)v = \dfrac{N}{2}v + \|\hat{\phi}\|_{L^1(q')}K^{-1}\lambda\chi_{\omega'(s)} \text{ in } \mathbb{R}^N \times (0,S), \\ v(0) = 0 \quad in \quad \mathbb{R}^N \end{cases} \tag{4.1}$$

satisfies

$$v(S) = v^1 - \alpha\frac{\hat{\phi}^0}{\|\hat{\phi}^0\|_{L^2(K)}}.$$

Therefore,

$$\|v(S) - v^1\|_{L^2(K)} = \alpha.$$

Remark 2. If $\|v^1\|_{L^2(K)} \leq \alpha$, we can take $v = 0$ to obtain $\|v(S) - v^1\|_{L^2(K)} \leq \alpha$ with control zero.

Proof of Proposition 4.1. Since $\|v^1\|_{L^2(K)} > \alpha, \hat{\phi}^0$ minimizes $J(.; A, v^1)$ and J is subdifferentiable in $\hat{\phi}^0 \neq 0$, we have that $0 \in \partial J(\hat{\phi}^0)$. In view of Proposition 3.5, there exists $\lambda \in sgn(\phi)\chi_q$ such that for every $\theta^0 \in L^2(K)$

$$0 = \left(\int_{q'} |\hat{\phi}|\right)\left(\int_{q'} \lambda\theta\right) + \alpha\int_{\mathbb{R}^N} K\frac{\hat{\phi}^0}{\|\hat{\phi}^0\|_{L^2(K)}}\theta^0 - \int_{\mathbb{R}^N} Kv^1\theta^0, \tag{4.2}$$

where θ is the solution of (3.2) with $\theta(S) = \theta^0$.

Now, if we multiply (4.1) by θ in $L^2(K)$, we obtain

$$\begin{aligned} &\int_0^S\int_{\mathbb{R}^N} Kv_s\theta - \int_0^S\int_{\mathbb{R}^N} \theta\mathrm{div}(K\nabla v) + \int_0^S\int_{\mathbb{R}^N} KA\theta v \\ &= \frac{N}{2}\int_0^S\int_{\mathbb{R}^N} K\theta v + \|\hat{\phi}\|_{L^1(q')}\int_0^S\int_{\mathbb{R}^N} \lambda\theta\chi_{\omega'}. \end{aligned} \tag{4.3}$$

Therefore,

$$\int_{\mathbb{R}^N} Kv(S)\theta^0 = \|\hat{\phi}\|_{L^1(q')}\int_{q'} \lambda\theta. \tag{4.4}$$

By (4.2), we see that (4.4) is equivalent to

$$(v(S), \theta^0)_{L^2(K)} = \left(-\alpha\frac{\hat{\phi}^0}{|\hat{\phi}^0|_{L^2(K)}} + v^1, \theta^0\right)_{L^2(K)},$$

which concludes the proof. □

We can now prove the main result in this section

PROPOSITION 4.2. *Let $u^0, u^1 \in L^2(\mathbb{R}^N)$, $\omega \subset \mathbb{R}^N$ an open and nonempty set. Then, for every $\alpha > 0$, there exists $h \in L^\infty(\mathbb{R}^N \times (0,T))$ such that the solution u of*

$$\begin{cases} u_t - \Delta u + a(x,t)u = h\chi_\omega & \text{in } \mathbb{R}^N \times (0,T), \\ u(x,0) = u_0(x) & \text{in } \mathbb{R}^N \end{cases} \tag{4.5}$$

satisfies

$$\|u(T) - u^1\|_{L^2(\mathbb{R}^N)} \leq \alpha.$$

Proof. We divide the proof into several steps.

Step 1. The case $u^0 = 0$ and $u^1 \in L^2(K)$.

We define A like in (1.6) and $v^1(y) = (T+1)^{N/2}u^1((T+1)^{1/2}y)$. We have that $v^1 \in L^2(K)$. If $\|v^1\|_{L^2(K)} > \alpha$ we construct H like in Proposition 4.1; otherwise $H \equiv 0$. Then v solution of (4.1) satisfies $\|v(S) - v^1\|_{L^2(K)} \leq \alpha$ with $S = \log(T+1)$.

Therefore, $u(x,t) = (1+t)^{-N/2}v(\frac{x}{\sqrt{1+t}}, \log(1+t))$ satisfies (4.5) with $h(x,t)\chi_\omega = (1+t)^{-N/2-1}e^{\frac{-x^2}{1+t}}\|\phi\|_{L^1(q')}\lambda$, $\lambda \in sgn\phi$, and

$$\begin{aligned} \|u(T) - u^1\|^2_{L^2(\mathbb{R}^N)} \leq & \int_{\mathbb{R}^N} e^{\frac{|x|^2}{4(1+T)}}(1+T)^{-N}|v\left(\frac{x}{\sqrt{1+T}}, S\right) - v^1\left(\frac{x}{\sqrt{1+T}}\right)|^2 dx \\ & \leq \int_{\mathbb{R}^N} K(1+T)^{-N/2}|v(S,y) - v^1(y)|^2 dy \\ & \leq (1+T)^{-N/2}\|v(S) - v^1\|^2_{L^2(K)} \leq \alpha^2. \end{aligned}$$

Step 2. The case $u^0 = 0$ and $u^1 \in L^2(\mathbb{R}^N)$.

Since $L^2(K) \subset L^2(\mathbb{R}^N)$ with dense inclusion, we know that there exists a sequence $u^1_n \subset L^2(K)$ and $\tilde{N} > 0$ such that $n > \tilde{N}$ implies

$$\|u^1_n - u^1\|_{L^2(\mathbb{R}^N)} < \frac{\alpha}{2}.$$

We just proved that there exists $h_n \in L^\infty((0,T); L^2(\mathbb{R}^N))$ such that the solution u of (4.5) satisfies

$$\|u(T) - u^1_n\|_{L^2(\mathbb{R}^N)} \leq \frac{\alpha}{2}.$$

Therefore, h_n with $n > \tilde{N}$ is an approximate control of our problem.

Step 3. The general case, i.e., u^0, $u^1 \in L^2(\mathbb{R}^N)$ arbitrary.

We write $u = y + Y$, where y is the solution of

$$\begin{cases} y_t - \Delta y + a(x,t)y = 0 & \text{in } \mathbb{R}^N \times (0,T), \\ y(x,0) = u_0(x) & \text{in } \mathbb{R}^N. \end{cases} \tag{4.6}$$

Then $y(T) \in L^2(\mathbb{R}^N)$. We construct $h(u^1 - y(T))$ such that the solution Y of

$$\begin{cases} Y_t - \Delta Y + a(x,t)Y = h\chi_\omega & \text{in } \mathbb{R}^N \times (0,T), \\ Y(x,0) = 0 & \text{in } \mathbb{R}^N \end{cases} \tag{4.7}$$

satisfies

$$\|Y(T) - (u^1 - y(T))\|_{L^2(\mathbb{R}^N)} \leq \alpha. \tag{4.8}$$

Therefore, in (4.5) it is enough to chose $h = h(u^1 - y(T))$, since the unique solution is $u = y + Y$ and in view of (4.8) we conclude the proof. ∎

5. The nonlinear case. We first analyze the case where f is of class C^1, $f(0) = 0$ and verifies (1.2). We define the continuous function

$$g(z) = \begin{cases} \dfrac{f(z)}{z} \text{ if } z \neq 0, \\ f'(0) \text{ if } z = 0. \end{cases}$$

For $\tilde{\nu} \in L^2(0, S; L^2(K))$, $u_0 \in L^2(K)$ we consider the solution u of

$$\begin{cases} u_t - \Delta u + g(\tilde{\nu})u = h\chi_\omega \text{ in } \mathbb{R}^N \times (0, T), \\ u(0) = u_0 \text{ in } \mathbb{R}^N. \end{cases} \tag{5.1}$$

Remark 3. The definition of g indicates that if $\tilde{\nu} = u$, then u is the solution of the semilinear equation (1.3).

We define v and H like in (1.6), $\nu(y, s) = e^{sN/2}\tilde{\nu}(e^{s/2}y, e^s - 1)$ and $G(\nu, s) = e^s g(e^{-sN/2}\nu)$, $\omega'(s) = e^{-s/2}\omega$. We see that if u is the solution of (5.1), then v satisfies

$$\begin{cases} v_s + Lv + G(s, \nu)v = \dfrac{N}{2}v + H\chi_{\omega'(s)} \quad \text{in} \quad \mathbb{R}^N \times (0, S), \\ v(0) = u^0 \quad \text{in} \quad \mathbb{R}^N \end{cases} \tag{5.2}$$

We are going to work on this equation to apply the fixed-point method introduced in [6] to obtain the approximate control of (1.3).

We write $v = w + V$, where w is the solution of

$$\begin{cases} w_s + Lw + G(s, \nu)w = \dfrac{N}{2}w \quad \text{in} \quad \mathbb{R}^N \times (0, S), \\ w(y, 0) = u^0(y) \quad \text{in} \quad \mathbb{R}^N. \end{cases} \tag{5.3}$$

For every $\nu \in L^2(K)$, $G(s, \nu) \in L^\infty((0, S) \times \mathbb{R}^N)$ and, therefore, $w \in C([0, S]; L^2(K))$. We can apply Theorem 3.3 to obtain that

$$\{v^1 - w(S), \quad \text{when} \quad \nu \in L^2(0, S; L^2(K))\} \tag{5.4}$$

is a compact subset of $L^2(K)$.

In view of Proposition 4.1 we know that there exists $\lambda(\nu, v^0, v^1) \in sgn(\hat{\phi}(\nu, v^0, v^1))\chi_{q'}$ with $\hat{\phi}$ minimizing $J(\phi^0; G(s, \nu), v^1 - w(S))$ and such that the solution V of

$$\begin{cases} V_s + LV + G(s, \nu)V = \dfrac{N}{2}V + \|\hat{\phi}\|_{L^1(q')}K^{-1}\lambda\chi_{\omega'} \quad \text{in} \quad \mathbb{R}^N \times (0, S), \\ V(y, 0) = 0 \quad \text{in} \quad \mathbb{R}^N \end{cases} \tag{5.5}$$

satisfies

$$\|V(S) - v^1 + w(S)\|_{L^2(K)} \leq \alpha.$$

We deduce that $v = w + V$ satisfies $\|v(S) - v^1\|_{L^2(K)} \leq \alpha$.

For $\lambda \in sgn(\hat{\phi}(\nu, v^0, v^1))\chi_{q'}$, we denote by $v(\lambda)$ the solution of

$$\begin{cases} v_s + Lv + G(s, \nu)v = \dfrac{N}{2}v + \|\hat{\phi}\|_{L^1(q')}K^{-1}\lambda\chi_{\omega'} \quad \text{in} \quad \mathbb{R}^N \times (0, S), \\ v(y, 0) = v^0 \quad \text{in} \quad \mathbb{R}^N, \end{cases} \tag{5.6}$$

and consider the following set valued mapping:

$\Lambda : L^2((0,S);L^2(K)) \to \mathcal{P}\left(L^2((0,S);L^2(K))\right)$ with

$$\Lambda(\nu) = \{v(\lambda), \lambda \in sgn(\hat{\phi}(\nu, v^0, v^1))\chi_{q'} \quad \text{and} \quad \|v(S) - v^1\|_{L^2(K)} \leq \alpha\}. \tag{5.7}$$

We just proved that $\Lambda(\nu)$ is always a nonempty subset of $L^2((0,S);L^2(K))$ and we have the following result.

PROPOSITION 5.1. *If f is of class C^1 in $\mathbb{R}$, $f(0) = 0$ and satisfies (1.2), then:*

(i) *There exists a compact subset X of $L^2(0,S;L^2(K))$ such that for every $\nu \in L^2(0,S;L^2(K))$, $\Lambda(\nu) \subset X$.*

(ii) *For all $\nu \in L^2(0,S;L^2(K))$, $\Lambda(\nu)$ is a nonempty, convex, and compact subset of $L^2(0,S;L^2(K))$.*

(iii) *Λ is upper hemicontinuous on $L^2(0,S;L^2(K))$.*

Proof. (i) Since $G \in L^\infty((0,S)\times\mathbb{R}^N)$ and from Proposition 3.4, the solutions of

$$\begin{cases} -\phi + L\phi + G(t,\nu)\phi = \dfrac{N}{2}\phi \quad \text{in} \quad \mathbb{R}^N \times (0,S), \\ \phi(S) = \hat{\phi}^0, \end{cases} \tag{5.8}$$

where $\hat{\phi}^0$ is the minimizer of $J(.;G(t,\nu),v^1)$, are bounded in $L^\infty(0,S;L^2(K))$. Therefore, there exists a bounded set X in $L^2(0,S;L^2(K))$ such that for every $\nu \in L^2(0,S;L^2(K))$, $\Lambda(\nu) \subset X$. Let us now prove that we can choose X being compact in $L^2(0,S;L^2(K))$. For this, it is sufficient to prove that the set $\mathcal{V} = \{v(\lambda), \lambda \in sgn(\hat{\phi}(\nu,v^0,v^1)), \nu \in L^2(0,S;L^2(K))\}$ is relatively compact in $L^2(0,S;L^2(K))$.

If $v = v(\lambda) \in \mathcal{V}$, then there exist $\nu \in L^2(0,S;L^2(K))$ and $\lambda \in sgn(\hat{\phi}(\nu,v^0,v^1))$ such that we can write $v = v^* + \hat{v} + V(\lambda)$, where $v^*, \hat{v}$, and $V = V(\lambda)$ are defined by the following equations:

$$\begin{cases} v^*_s + Lv^* = \dfrac{N}{2}v^* \quad \text{in} \quad \mathbb{R}^N \times (0,S), \\ v^*(0) = u^0 \quad \text{in} \quad \mathbb{R}^N, \end{cases}$$

$$\begin{cases} \hat{v}_s + L\hat{v} + G(s,\nu)(v^* + \hat{v}) = \dfrac{N}{2}\hat{v} \quad \text{in} \quad \mathbb{R}^N \times (0,S), \\ \hat{v}(0) = 0 \quad \text{in} \quad \mathbb{R}^N, \end{cases}$$

$$\begin{cases} V_s + LV + G(s,\nu)V = \dfrac{N}{2}V + \|\hat{\phi}(\nu,v^0,v^1)\|_{L^1(q')}K^{-1}\lambda\chi_\omega \quad \text{in} \quad (0,S)\times\mathbb{R}^N, \\ V(0) = 0 \quad \text{in} \quad \mathbb{R}^N. \end{cases}$$

The function v^* is fixed in $L^2(0,S;L^2(K))$. When varying ν in $L^2(0,S;L^2(K))$, $G(s,\nu)v^*$ describes a bounded set in $L^2(0,S;L^2(K))$. From (2.3), (2.4), (2.5), and applying again Theorem 3.3, we see that $\hat{v}$ describes (when varying ν) a relatively compact set $B_1 \subset L^2(0,S;L^2(K))$.

Since the functions $\hat{\phi}(\nu,v^0,v^1)$ are bounded in $L^2(0,S;L^2(K))$, we have

$$\|\hat{\phi}\|_{L^1(q')}\|K^{-1/2}\lambda\|_{L^\infty(\mathbb{R}^N\times(0,T))} \leq C$$

for some constant $C > 0$.

Therefore, using again (2.3), (2.4), (2.5), and Theorem 3.3, we see that $V(\lambda)$ describes a relatively compact set B_2 in $L^2(0,S;L^2(K))$. That proves that $\mathcal{V} \subset v^* + B_1 + B_2$ is relatively compact in $L^2(0,S;L^2(K))$. We choose for X the closure of $\mathcal{V}$ in $L^2(0,S;L^2(K))$ to obtain (i).

(ii) We have already seen that for all $\nu \in L^2(0,S;L^2(K))$, $\Lambda(\nu)$ is a nonempty set of $L^2(0,S;L^2(K))$. Since the sets $sgn(\hat{\phi})$ and $B(v^1,\alpha)$ are convex, it is easy to prove that $\Lambda(\nu)$ is convex so we are just going to prove that it is compact. As we already have $\Lambda(\nu) \subset X$ with X compact, we only have to prove that it is closed. Let $\{v_n\}_n$ be a sequence of elements of $\Lambda(\nu)$, which converges in $L^2(0,S;L^2(K))$ to an element $v \in X$. Let us prove that $v \in \Lambda(\nu)$.

There exist functions $\lambda_n \in sgn(\hat{\phi})$ such that

$$(5.9)\begin{cases} v_n' + Lv_n + G(s,\nu)v_n = \dfrac{N}{2}v_n + \|\hat{\phi}\|_{L^1(q')}K^{-1}\lambda_n\chi_{\omega'(s)} & \text{in } (0,S)\times\mathbb{R}^N, \\ v_n(0) = v^0 \quad \text{in } \mathbb{R}^N, \\ \|v_n(S) - v^1\|_{L^2(K)} \le \alpha. \end{cases}$$

Since $\|\lambda_n\|_\infty \le 1$, there exists a subsequence λ_n converging weakly* in $L^\infty(q')$ to an element $\lambda \in L^\infty(q')$. Furthermore, as $\lambda_n \in sgn(\hat{\phi})$ for every n, we have that $\|\lambda\|_\infty \le 1$ and $\lambda = sgn_0(\hat{\phi})$ in $\{\hat{\phi}(\nu,v^0,v^1) \neq 0\} \cap q'$. This proves that $\lambda \in sgn\hat{\phi}$.

Now, by Proposition 2.3 and passing to the limit in (5.9), we obtain

$$\begin{cases} v_s + Lv + G(s,\nu)v = \dfrac{N}{2}v + \|\hat{\phi}\|_{L^1(q')}K^{-1}\lambda\chi_{\omega'(s)} & \text{in } (0,S)\times\mathbb{R}^N, \\ v(0) = v^0 \quad \text{in } \mathbb{R}^N. \end{cases}$$

Due to the smoothing effects, $v_n(S)$ converges in $L^2(K)$ to $v(S)$ and, therefore, $\|v(S) - v^1\|_{L^2(K)} \le \alpha$. This proves that $v \in \Lambda(\nu)$ and concludes the proof of (ii).

(iii) We first recall that Λ is hemicontinuous at $\nu_0 \in L^2(0,S;L^2(K))$ if

$$\nu \to \sigma(\Lambda(\nu),\varphi) = \sup_{v\in\Lambda(\nu)} \int_0^S \int_{\mathbb{R}^N} K\varphi v dy ds$$

is upper semicontinuous at ν_0 for every $\varphi \in L^2(0,S;L^2(K))$.

We then have to show that

$$\forall \nu_0 \in L^2(0,S;L^2(K)), \quad \limsup_{\nu_n\to\nu_0} \sigma(\Lambda(\nu_n),\varphi) \le \sigma(\Lambda(\nu_0),\varphi).$$

We denote by $\langle u,v\rangle$ the integral $\int_0^S\int_{\mathbb{R}^N} Kuv\,dyds$. From (ii), we know that $\Lambda(\nu)$ is compact in $L^2(0,S;L^2(K))$. Then, for every $n \in \mathbb{N}$, there exists $v_n \in \Lambda(\nu_n)$ such that

$$\sigma(\Lambda(\nu_n),\varphi) = \langle v_n,\varphi\rangle.$$

From (i), $(v_n)_n \subset X$ and, therefore, $v_n \to v$ in $L^2(0,S;L^2(K))$. Let us prove that $v \in \Lambda(\nu_0)$. We write $\phi_n = \hat{\phi}(\nu_n,v^0,v^1)$. There exists $\lambda_n \in sgn(\phi_n)$ such that v_n satisfies

$$\begin{cases} v_{n,s} + Lv_n + G(s,\nu_n)v_n = \dfrac{N}{2}v_n + \|\phi_n\|_{L^1(q')}K^{-1}\lambda_n\chi_\omega & \text{in } \mathbb{R}^N\times(0,S), \\ v(y,0) = v^0 \quad \text{in } \mathbb{R}^N, \\ \|v_n(S) - v^1\|_{L^2(K)} \le \alpha. \end{cases}$$

We will need the following lemma.

LEMMA 5.2. *If $\nu_n \to \nu_0$ strongly in $L^2(0,S;L^2(K))$, then $\hat{\phi}(\nu_n, v^0, v^1) \to \hat{\phi}(\nu_0, v^0, v^1)$ strongly in $L^2(K)$.*

Proof of Lemma 5.2. The functions $G(s,\nu_n)$ are bounded in $L^\infty((0,S)\times\mathbb{R}^N)$. Hence, for a subsequence, it converges in the weak* topology to an element $\tilde{G}$ in $L^\infty((0,S)\times\mathbb{R}^N)$. But G is continuous and that means that $\tilde{G} = G(s,\nu_0)$. Moreover, from (5.4) we know the existence of a subsequence $w_n(S)$ (where w_n is the solution of (5.3) corresponding to $G(s,\nu_n)$) such that $w_n(S) \to w(S)$ strongly in $L^2(K)$, where w is the solution of (5.3) corresponding to $G(s,\nu_0)$. In view of Proposition 3.4 (ii), we have that

$$\hat{\phi}_n^0 \to \hat{\phi}^0 \quad \text{strongly in } L^2(K),$$

where $\hat{\phi}^0$ is the minimizer of $J(.; G(s,\nu_0), v^1 - w(S))$. □

From Lemma 5.2 and Proposition 2.3, $\hat{\phi}_n$ converges strongly to $\hat{\phi}(\nu_0, v^0, v^1)$ in $L^2(0,S;L^2(K))$. Hence, $\|\hat{\phi}_n\|_{L^1(q')} \to \|\hat{\phi}\|_{L^1(q')}$ and

$$\hat{\phi}_n(s,y) \to \hat{\phi}(s,y) \quad \text{almost everywhere in } q'.$$

Since $\lambda_n \in sgn(\hat{\phi}_n)$, we can extract a subsequence converging weakly* in $L^\infty(q')$ to $\lambda \in sgn(\hat{\phi})\chi_{\omega'(s)}$. We then deduce that v is the solution of

$$\begin{cases} v_s + Lv + G(s,\nu_0)v = \dfrac{N}{2}v + \|\hat{\phi}\|_{L^1(q')}K^{-1}\lambda\chi_{\omega'(s)} & \text{in } \mathbb{R}^N\times(0,S), \\ v(y,0) = v^0 & \text{in } \mathbb{R}^N. \end{cases}$$

Moreover, due to the smoothing effects we keep the condition $v(S) \in B(v^1,\alpha)$; thus $v \in \Lambda(\nu_0)$.

Now, we have $\langle w, v_n\rangle \to \langle w, v\rangle$ with $v \in \Lambda(\nu_0)$, which proves that Λ is upper hemicontinuous in ν_0. □

PROPOSITION 5.3. *If f is of class C^1, $f(0) = 0$, and satisfies* (1.2)*, then there exists $v \in L^2(0,S;L^2(K))$ such that $v \in \Lambda(v)$.*

Proof The restriction of Λ to the convex hull of $X, conv(X)$ (that is compact in $L^2(0,S;L^2(K))$) satisfies the hypothesis of Kakutani's theorem, (see, e.g., Aubin [1, p. 344]). We deduce that Λ has a fixed point v. We then have the existence of $\phi^0 \in L^2(K)$ and $\lambda \in sgn(\phi)\chi_{q'}$ such that

$$(5.10)\quad \begin{cases} -\phi_s + L\phi + G(s,v)\phi = \dfrac{N}{2}\phi & \text{in } \mathbb{R}^N\times(0,S), \\ \phi(S) = \phi^0 \text{ in } \mathbb{R}^N, \\ v_s + Lv + G(s,v)v = \dfrac{N}{2}v + \|\phi\|_{L^1(q')}K^{-1}\lambda\chi_{\omega'(s)} & \text{in } \mathbb{R}^N\times(0,S), \\ v(y,0) = v^0(y) & \text{in } \mathbb{R}^N, \\ \|v(S) - v^1\|_{L^2(K)} \le \alpha. \end{cases}$$ □

As a direct consequence of Proposition 5.3, the general case for globally Lipschitz nonlinearities is easy to obtain.

PROPOSITION 5.4. *Let f be a globally Lipschitz function with $f(0) = 0$. There exists $A > 0$ and $\{f_n\} \in C^1(\mathbb{R}^N)$, $f_n(0) = 0$ and such that*

$$(5.11)\qquad \forall n \in \mathbb{N}, \qquad \forall \sigma \in \mathbb{R}, \qquad \left|\frac{f_n(\sigma)}{\sigma}\right| \le A,$$

$$\lim_{n\to\infty} f_n = f \quad \textit{locally uniformly.}$$

For each $n \in \mathbb{N}$, if we denote by $\varphi_n, \lambda_n \in sgn(\varphi_n)$ and v_n the solution of (5.10) *associated to f_n, there exists $\tilde{G} \in L^\infty(\mathbb{R}^N \times (0,T))$ such that φ_n^0 converges strongly in $L^2(K)$ to the minimum φ^0 of $J(\cdot;\tilde{G}, v^1 - w(S))$ and (φ_n, v_n) converge strongly in $L^2(0,T;L^2(K)) \times L^2(0,T;L^2(K))$ to the solution of*

$$(5.12)\quad \begin{cases} -\phi_s + L\phi + \tilde{G}\phi = \dfrac{N}{2}\phi, \\ \phi(S) = \phi^0, \\ v_s + Lv + G(v,s)v = \dfrac{N}{2}v + \|\phi\|_{L^1(q')}K^{-1}\lambda\chi_{\omega'(s)} \quad \textit{in} \ \ \mathbb{R}^N \times (0,S), \\ v(y,0) = v^0(y) \quad \textit{in} \quad \mathbb{R}^N, \\ \|v(S) - v^1\|_{L^2(K)} \le \alpha, \end{cases}$$

where $\lambda \in sgn(\varphi)\chi_{q'}$. Furthermore, $\tilde{G}(y,s) = G(v(y,s),s)$ on the set $v(y,s) \neq 0$.

Since the proof of this proposition is a straightforward adaptation of Proposition 3.4 in [6], we refer to it to avoid technical details. We only mention that, in fact, once the sequence satisfying (5.11) is obtained, we can obtain (5.12) as a direct consequence of Propositions 5.3 and 3.4.

We conclude with the proof of Proposition 1.1.

Proof of Proposition 1.1. As in the proof of Proposition 4.2, we divide the proof into steps.

Step 1. u^0, u^1 in $L^2(K)$.

We make the change of variables $v^1(y) = (T+1)^{N/2}u^1((T+1)^{1/2}y)$ and $S = \log(1+\mathrm{T})$. From Proposition 5.4 we know the existence of $\phi \in L^2(0,S;L^2(K))$, $\lambda \in sgn(\phi)\chi_{q'}$, and v solution of

$$\begin{cases} v_s + Lv + G(v,s)v = \dfrac{N}{2}v + \|\phi\|_{L^1(q')}K^{-1}\lambda\chi_{\omega'(s)} \ \text{ in } \mathbb{R}^N \times (0,S), \\ v(y,0) = u^0(y) \quad \text{in} \quad \mathbb{R}^N, \\ \|v(S) - u^1\|_{L^2(K)} \le \alpha. \end{cases}$$

We define $u(x,t)$ as in (1.7). By construction of G, u is the solution of (1.3) with

$$h(x,t) = (t+1)^{-N/2-1}K^{-1}\left(\frac{x}{\sqrt{1+t}}\right)\|\phi\|_{L^1(q')}\tilde{\lambda}\chi_\omega,$$

where $\tilde{\lambda} \in sgn\tilde{\phi}$ and $\tilde{\phi}(x,t) = \phi(\frac{x}{\sqrt{1+t}}, \log(1+t))$. Moreover, we observe that $\tilde{\phi}$ is a solution of

$$\begin{cases} -\tilde{\phi}_s - \Delta\tilde{\phi} + g(u)\tilde{\phi} = 0 \text{ in } \mathbb{R}^N \times (0,T), \\ \tilde{\phi}(T) = (1+T)^{-N/2}\phi^0\left(\dfrac{x}{\sqrt{1+T}}, S\right) \text{ in } \mathbb{R}^N. \end{cases}$$

The end of the proof follows as in Step 1 in the proof of Proposition 4.2 obtaining

$$\|u(T) - u^1\|_{L^2(\mathbb{R}^N)} \le \alpha.$$

Step 2. $u^0 \in L^2(K)$, $u^1 \in L^2(\mathbb{R}^N)$.

Since $L^2(K) \subset L^2(\mathbb{R}^N)$ with dense inclusion, there exists a sequence $\{u_n^1\} \subset L^2(K)$ such that $u_n^1 \to u^1$ strongly in $L^2(\mathbb{R}^N)$. From the first step, we know the existence of controls h_n such that u_n, the solution of (1.3) with $h = h_n$, satisfies

$$\|u_n(T) - u_n^1\|_{L^2(\mathbb{R}^N)} \leq \frac{\alpha}{2}.$$

Let $\tilde{N}$ be such that for every $n > \tilde{N}$,

$$\|u^1 - u_n^1\|_{L^2(\mathbb{R}^N)} \leq \frac{\alpha}{2}.$$

Then u, the solution of (1.3) with $h = h_{\tilde{N}}$, satisfies $\|u(T) - u^1\|_{L^2(\mathbb{R}^N)} \leq \alpha$.

Step 3. $u_0 \in L^2(\mathbb{R}^N)$.

Then there exists a sequence u_n^0 such that $u_n^0 \to u^0$ strongly in $L^2(\mathbb{R}^N)$. Moreover, as we saw previously, there exists a sequence of controls h_n such that u_n, the solution of (1.3) corresponding to u_n^0 and h_n, satisfies

$$\|u_n(T) - u^1\|_{L^2(\mathbb{R}^N)} \leq \frac{\alpha}{2}.$$

Let $\tilde{N} > 0$ be such that

$$\|u_{\tilde{N}}^0 - u^0\|_{L^2(\mathbb{R}^N)} \leq e^{\frac{-MT}{2}} \frac{\alpha}{2},$$

where M is the constant given in (1.2).

Consider u the solution of (1.3) corresponding to u^0 and $h = h_{\tilde{N}}$. Let $z = u - u_{\tilde{N}}$. Then z satisfies

$$\begin{cases} z_t - \Delta z + f(u) - f(u_{\tilde{N}}) = 0 & \text{in } \mathbb{R}^N \times (0,T), \\ z(x,0) = u^0(x) - u_{\tilde{N}}^0(x) & \text{in } \mathbb{R}^N. \end{cases} \tag{5.13}$$

We multiply (5.13) by z and integrate over $\mathbb{R}^N$. Since f satisfies (1.2), we obtain

$$\|z(T)\|_{L^2(\mathbb{R}^N)} \leq e^{\frac{MT}{2}} \|z(0)\|_{L^2(\mathbb{R}^N)} \leq \frac{\alpha}{2},$$

and, therefore,

$$\|u(T) - u^1\|_{L^2(\mathbb{R}^N)} \leq \|z(T)\| + \|u_{\tilde{N}}(T) - u^1\| \leq \alpha. \qquad \square$$

6. Further results. In this section we mention some possible extensions of the results and techniques of this article.

6.1. Cone-like domains. All the results of this paper hold for the semilinear heat equation (1.1) when Ω is a cone-like domain. In fact, consider a domain Ω satisfying

$$0 \in \bar{\Omega}, \qquad \forall \lambda > 0, \qquad \forall x \in \Omega, \lambda x \in \Omega.$$

Then we can study the controllability of the linear equation

$$\begin{cases} u_t - \Delta u + a(x,t)u = h\chi_\omega & \text{in } \Omega \times (0,T), \\ u = 0 \quad \text{on } \partial\Omega \times (0,T), \\ u(x,0) = u^0 \quad \text{in } \Omega \end{cases} \tag{6.1}$$

in the same way as the case of the whole space $\mathbb{R}^N$.

Observe that defining v, H, A as in (1.6) we obtain that v satisfies

$$\begin{cases} v_t + Lv + A(y,t)v = \dfrac{N}{2}v + H\chi_{\omega'} & \text{in } \Omega \times (0,S), \\ v = 0 \qquad \text{on } \partial\Omega \times (0,S), \\ v(y,0) = u^0 \qquad \text{in } \Omega. \end{cases} \tag{6.2}$$

In this case the functional setting will be

$$H_0^1(K,\Omega) = \left\{ v; \int_\Omega (|v|^2 + |\nabla v|^2) K(y) dy < \infty,\ v|_{\partial\Omega} = 0 \right\}.$$

6.2. Other functional settings. In this article we have chosen to work in the $L^2(\mathbb{R}^N)$ functional setting. The same type of results can be proved in $L^p(\mathbb{R}^N)$ spaces for $1 \le p < \infty$. In this aim it is necessary to introduce the corresponding weighted Sobolev spaces, i.e., $L^p(K) = \{u : \mathbb{R}^N \to \mathbb{R} : \int_{\mathbb{R}^N} |u|^p K(y) dy < \infty\}$ and $W^{1,p}(K) = \{u \in L^p(K); \nabla u \in L^p(K)\}$.

The only essential change that has to be done in the proof is that we have to minimize a functional of the form (3.1) in the dual space to the one we have chosen to prove controllability. We refer to [6] in the case where the domain is bounded.

6.3. Insensitizing controls. Suppose that in (1.3) the initial data are partially known, i.e., $u^0 = \hat{u}^0 + \tau \bar{u}^0$, where $\bar{u}^0 \in L^2(\mathbb{R}^N)$ is unknown, $\|\bar{u}^0\|_{L^2(\mathbb{R}^N)} = 1$ and $\tau \in \mathbb{R}$ is unknown and small enough. The insensitizing problem consists of finding a control function such that some functional of the state is locally insensitive to the perturbations of these initial data. We say that the control h, ε-insensitizes $\Phi(u)$ if

$$\left| \frac{\partial \Phi(u(x,t;h,\tau))}{\partial \tau}|_{\tau=0} \right| \le \varepsilon. \tag{6.3}$$

Let $\theta \subset \mathbb{R}^N$ be an open "observation" subset. When

$$\Phi(u) = \frac{1}{2} \int_0^T \int_\Theta u^2(x,t) dx dt, \tag{6.4}$$

the condition of ε-insensitivity is equivalent to an approximate control problem. Let y and q be the solutions of the following cascade system:

$$\begin{cases} y_t - \Delta y + f(y) = h\chi_\omega \ \text{ in } \mathbb{R}^N \times (0,T), \\ y(.,0) = u^0 \ \text{ in } \ \mathbb{R}^N, \end{cases} \tag{6.5}$$

$$\begin{cases} -q_t - \Delta q + f'(y)q = y\chi_\Theta \ \text{ in } \ \mathbb{R}^N \times (0,T), \\ q(.,T) = 0 \ \text{ in } \ \mathbb{R}^N, \end{cases} \tag{6.6}$$

where χ_Θ is the characteristic function of the observation subset Θ.

Then the condition of ε-insensitivity is equivalent to

$$\|q(.,0)\|_{L^2(\mathbb{R}^N)} \le \varepsilon. \tag{6.7}$$

The techniques of this paper may be adapted to prove that when $\omega \cap \theta \neq \emptyset$, f is of class C^1 and globally Lipschitz, then there exists an ε-insensitizing control for the functional (6.4). For insensitizing controls in bounded domains we refer to [2]. The case of insensitizing controls in unbounded domains is treated in [13] by using the approximation technique mentioned in the introduction.

REFERENCES

[1] J.P. Aubin, *L'analyse non linéaire et ses motivations économiques,* Masson, Paris, 1987.

[2] O. Bodart and C. Fabre, *Controls insensitizing the norm of the solution of a semilinear heat equation,* J. Math. Anal. Appl., 195 (1995), pp. 658–683.

[3] T. Cazenave and A. Haraux, *Introduction aux problèmes d'évolution sémi-linéaires,* Collection S.M.A.I. Mathématiques et applications, Ellipses, Paris, 1980.

[4] M. Escobedo and O. Kavian, *Variational problems related to self-similar solutions of heat equation,* Nonlinear Anal., 11 (1987), pp. 1103–1133.

[5] M. Escobedo and E. Zuazua, *Large time behaviour for convection-diffusion equations in* $\mathbb{R}^N$, J. Funct. Anal., 100 (1991), pp. 119–161.

[6] C. Fabre, J.P. Puel, and E. Zuazua, *Approximate controllability of the semilinear heat equation,* Proc. Roy. Soc. Edinburgh Sect. A, 125 (1995), pp.31–61.

[7] O. Kavian, *Remarks on large time behaviour of a nonlinear diffusion equation,* Ann. Inst. H. Poincaré Anal. Non Linéaire, 4 (1987), pp. 423–452.

[8] O.A. Ladyzenskaja, V.A. Solonnikov, and N.N. Ural'ceva, *Linear and Quasilinear Equations of Parabolic Type,* AMS, Providence, RI, 1968.

[9] J.L. Lions and E. Magenes, *Problèmes aux limites non homogènes et applications,* Vol. I & II, Dunod, Paris, 1968.

[10] J.C. Saut and B. Scheurer, *Unique continuation for some evolution equations,* J. Differential Equations, 66 (1987), pp.118–139.

[11] J. Simon, *Compact sets in the space* $L^p(0,T;B)$, Ann. Mat. Pura Appl., CXLVI (1987), pp. 1173–1191.

[12] H. Tanabe, *Equations of Evolution,* Pitman, Bath, 1979.

[13] L. de Teresa, *Controls insensitizing the norm of the solution of a semilinear heat equation in unbounded domains,* ESAIM: Control, Optimization and Calculus of Variations, 2 (1997), pp.125–149.

[14] L. de Teresa and E. Zuazua, *Approximate controllability of a semilinear heat equation in unbounded domains,* Nonlinear Anal., to appear.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 6, pp. 2148–2175, November 1998

016

A STRUCTURE THEORY FOR LINEAR DYNAMIC ERRORS-IN-VARIABLES MODELS*

W. SCHERRER† AND M. DEISTLER†

Abstract. We deal with problems connected with the identification of linear dynamic systems in situations when inputs and outputs may be contaminated by noise. The case of uncorrelated noise components and the bounded noise case is considered. If also the inputs may be contaminated by noise, a number of additional complications in identification arise, in particular the underlying system is not uniquely determined from the population second moments of the observations. A description of classes of observationally equivalent systems is given, continuity properties of mappings relating classes of observationally equivalent systems to the spectral densities of the observations are derived and the classes of spectral densities corresponding to a given maximum number of outputs are studied.

Key words. errors-in-variables, factor analysis, system identification

AMS subject classifications. 93B30, 93B15, 62H25

PII. S0363012994262464

1. Introduction.

In the "main stream approach" to linear systems identification (see Deistler [7]) one of the basic assumptions is that all noise is added to the outputs (or to the equations, which is the same for our purpose). The noise thereby is assumed to be orthogonal to the inputs. In econometrics this is called the *errors-in-equations* approach. Here we are concerned with a different and, in principle, more general approach to noise modeling, where all variables may be contaminated by noise. Models of this kind are called *errors-in-variables* (EV) or *latent variables models*, or in a different but equivalent formulation, *factor models*. For the case of static systems, such models have been analyzed and used for a long time in statistics, science (in particular, chemistry), psychometrics, and econometrics (see, e.g., Adcock [1], Spearman [26], Gini [14], Frisch [13]). In the last two decades there has been a resurging interest in such models (see, e.g., Aigner et al. [2], Anderson [5]). Recently—mainly triggered by Kalman's work [18, 19, 20]—EV models have also been analyzed in systems engineering. The dynamic case has been treated, e.g., in Anderson and Deistler [3] and Deistler and Anderson [9].

The traditional errors-in-equations approach is justified in a great number of applications dealing, for instance, with prediction. On the other hand, in a number of cases the asymmetry in errors-in-equations modeling cannot be justified and may lead to "prejudiced" results (Kalman [20]). For example, in sonar array processing, when an array of n sensors is assumed to receive noisy signals from $n - m$ sources (Haykin [16]), EV models arise in a natural way. More generally, we can distinguish the following three main areas for EV modeling:

1. If we are interested in the "true system" underlying the data (rather than, for instance, in prediction) and if we cannot be sure a priori that the inputs have been observed free of noise. This is the "classical" motivation for EV models, for example, in econometrics.

*Received by the editors January 31, 1994; accepted for publication (in revised form) October 10, 1997; published electronically September 3, 1998. This research was supported by the Austrian "Fonds zur Förderung der wissenschaftlichen Forschung" Projekt P11213-MAT.

http://www.siam.org/journals/sicon/36-6/26246.html

†Institut für Ökonometrie, Operations Research und Systemtheorie, Technische Universität Wien, Argentinierstr. 8, A-1040 Vienna, Austria (W.Scherrer@tuwien.ac.at, M.Deistler@tuwien.ac.at).

2. If we want to approximate a high dimensional data vector by a small number of factors. This is the "classical" motivation for factor analysis (e.g., in psychometrics, where an example would be determining the intelligence factors underlying the test scores). A related issue is that EV modeling may considerably reduce the dimension of parameter spaces in comparison with multivariate AR or ARMA models.

3. In a number of cases, no sufficient a priori information about the number of equations and/or about the classification of the variables into inputs and outputs is available. Then, one has to use a *symmetric system model* which in turn demands a *symmetric noise model*. This point has been emphasized in particular by Kalman [18].

The systems considered are of the form

$$w(\mathrm{z})\hat{x}_t = 0, \tag{1.1}$$

where $\hat{x}_t$ is an n-dimensional vector of *latent* (i.e., not necessarily observed) real valued random variables, z is used for the backward-shift on the integers $\mathbb{Z}$ (i.e., $\mathrm{z}\,(\hat{x}_t|t \in \mathbb{Z}) = (\hat{x}_{t-1}|t \in \mathbb{Z})$) as well as for a complex variable, and where

$$w(\mathrm{z}) = \sum_{j=-\infty}^{\infty} W_j \mathrm{z}^j; \qquad W_j \in \mathbb{R}^{m\times n} \text{ and } \sum_{j=-\infty}^{\infty} \|W_j\| < \infty. \tag{1.2}$$

We will call $w(\mathrm{z})$ the *relation function*; it represents an exact (i.e., deterministic) linear system of a very general form. Clearly, systems of the form (1.1) are symmetric in the sense that no a priori classification of the variables $\hat{x}_t$ as inputs and outputs and no a priori information about causality are needed. Here also the number of equations, m, in (1.1) is not assumed to be known a priori. Without restriction of generality, we will assume that $1 \le m \le n$ holds and that $w(\mathrm{z})$ contains no linearly dependent rows.

The observed variables x_t are of the form

$$x_t = \hat{x}_t + u_t, \tag{1.3}$$

where u_t is the n-dimensional noise vector.

Throughout the paper we will assume the following:

(a.1) The processes (x_t), $(\hat{x}_t)$, and (u_t) are (wide sense) stationary with absolutely summable autocovariance functions. Thus, in particular, the spectral densities Σ, $\hat{\Sigma}$, and $\tilde{\Sigma}$ of (x_t), $(\hat{x}_t)$, and (u_t), respectively, exist and are bounded continuous functions. (In addition, limits of random variables are understood in the sense of mean square convergence.)

(a.2) $\mathrm{E}\hat{x}_t = 0$ and $\mathrm{E}u_t = 0$.

(a.3) $\mathrm{E}\hat{x}_t u_s' = 0$.

(a.4) Unless the contrary is stated explicitly, we assume that $\Sigma > 0$ holds.

Assumptions (a.1)–(a.4) are not severe restrictions of generality. They are either natural or are imposed to avoid technical problems.

Due to (1.1), the spectral density $\hat{\Sigma}$ is singular and

$$w(e^{-i\lambda})\hat{\Sigma}(\lambda) = 0 \tag{1.4}$$

holds. Note that for a given process $(\hat{x}_t)$, a relation function w satisfies (1.1) iff w satisfies (1.4) for the spectral density $\hat{\Sigma}$ of $\hat{x}_t$; in other words, there is no loss of information concerning w in going from the process $(\hat{x}_t)$ to its spectral density.

Assume that the spectral density $\hat{\Sigma}$ of corank m is given and that the $m\times n$ matrix w satisfies $w\hat{\Sigma} = 0$, where w has rank m. Then we may select m independent columns

from w and regroup the columns of w such that these columns appear in the first m positions, which gives a partitioning of w as (w_1, w_2). By a conformal rearrangement of the components of $\hat{x}_t$ and a corresponding partitioning of $\hat{x}_t = ((\hat{x}_t^1)', (\hat{x}_t^2)')'$, we obtain from (1.1) that $w_1\hat{x}_t^1 + w_2\hat{x}_t^2 = 0$. Now assume that w_1^{-1} has an absolutely summable Laurent series expansion in an annulus containing the unit circle; then by premultiplying w with w_1^{-1} we obtain

$$\hat{x}_t^1 = \underbrace{-w_1^{-1}(\mathrm{z})w_2(\mathrm{z})}_{k(\mathrm{z})}\hat{x}_t^2$$

which describes the input-output behavior of the system (1.1). In general this choice of outputs $\hat{x}_t^1$ is not unique. Note that for a nonsingular $m \times m$ transfer function $t(\mathrm{z})$, which satisfies additional conditions (e.g., that both t and t^{-1} have an absolutely summable Laurent series expansion in an annulus containing the unit circle) the relation functions $w(\mathrm{z})$ and $t(\mathrm{z})w(\mathrm{z})$ are equivalent in the sense that, for a given choice of outputs, they represent the same input-output behavior $k(\mathrm{z})$.

Note that there is a close relation to the behavioral approach developed by Willems [29]; also see Heij, Scherrer, and Deistler [17]. The main differences of the behavioral approach to the setup of this paper are that here we impose stationarity and that we do not require the system to be finite dimensional.

Under rather general conditions, the restriction that $(\hat{x}_t)$ is contained in the kernel of $w(\mathrm{z})$ (see (1.1)) can be replaced by the restriction that $(\hat{x}_t)$ is contained in the image of a suitably chosen $n \times (n-m)$ transfer function $\Lambda(\mathrm{z})$, i.e.,

$$\hat{x}_t = \Lambda(\mathrm{z})\epsilon_t, \tag{1.5}$$

where, in particular, (1.5) can be chosen to be the Wold representation of the process $(\hat{x}_t)$. This gives rise to the linear dynamic factor model

$$x_t = \Lambda(\mathrm{z})\epsilon_t + u_t, \tag{1.6}$$

where (ϵ_t) is interpreted as the $(n-m)$ dimensional factor process, which by assumption is white noise. $\Lambda(\mathrm{z})$ is called the matrix of factor loadings.

We commence from the equation

$$\Sigma = \hat{\Sigma} + \tilde{\Sigma} \tag{1.7}$$

for the spectral densities. For given Σ, the matrix $\hat{\Sigma}$ is called *compatible* (with Σ) if (1.7) is satisfied, where $\hat{\Sigma}$ and $\tilde{\Sigma}$ are positive semidefinite and where, in addition, $\hat{\Sigma}$ is singular and typically, $\tilde{\Sigma}$ satisfies further assumptions such as (a.5) or (a.6) below. Instead of *compatible*, we also use the term *observationally equivalent*. A relation function $w(\mathrm{z})$ is called *compatible* (with Σ) if there exists a compatible $\hat{\Sigma}$, such that $w\hat{\Sigma} = 0$ holds.

Without imposing additional a priori assumptions such as (a.5) or (a.6) below, the problem considered is not sufficiently structured. In particular, without such assumptions, *every* relation function w would be compatible with a given $\Sigma > 0$. This is an easy consequence of the fact that, for every singular $n \times n$ spectral density $\hat{\Sigma}$, a constant $c > 0$ exists such that $\hat{\Sigma} = c\hat{\Sigma}$ is compatible with a given $\Sigma > 0$. Thus, some additional structure has to be imposed, which can be justified in a sufficiently large number of cases. In this paper, the two following alternative assumptions are considered:

(a.5) $\tilde{\Sigma}$ is diagonal.

This case will be called the Frisch case. The idea behind this assumption is to provide a decoupling of common and individual effects between the variables. The common effects are attributed to the system and the individual effects to the noise. Another motivation for this assumption relates to the case where all noise is measurement noise and the measurement devices for each channel are independent. Note that the Frisch case, in particular for the static case, has a long tradition in econometrics and psychometrics; see, e.g., Aigner et al. [2], Anderson [5], Anderson and Rubin [6], Gini [14], Ledermann [22], and Spearman [26].

An alternative assumption is that the noise level is bounded. This will be expressed as follows:

(a.6) $\lambda_n(\tilde{\Sigma}(\lambda)) \leq \epsilon$.

This case will be called the bounded noise case. Here λ_n denotes the maximum eigenvalue of $\tilde{\Sigma}(\lambda)$ and ϵ is an a priori given bound. This assumption is justified, for instance, if all noise is measurement noise and the magnitude of the error of the measurement devices (in terms of the noise spectrum) is known a priori. Additional information about the noise spectra may be taken into account by appropriate prefiltering of the data. This, in particular, relates to scale transformations and weightings of frequency bands.

Identification of errors-in-variables models is considerably more complicated than identification of errors-in-equations models. The purpose of this paper is twofold: first, to add a further step towards a theory of identification for this general case and second, to illustrate the additional complications arising with the departure from errors-in-equations models. We restrict ourselves to structure theory, i.e., we commence from population second moments rather than from real data. The main problem considered in this paper is obtaining the underlying systems from the population second moments of the observations (x_t) given by the spectral density Σ. One of the main complications of the errors-in-variables problem is that, in general, the underlying system is not uniquely determined from Σ. This is a major difference to the errors-in-equations approach, where the underlying transfer function is uniquely determined from the second moments of the observations under a so-called *persistent excitation* condition. This nonuniqueness in the EV case is caused by a lack of a priori knowledge concerning the noise structure. This is an uncertainty about the underlying system which has nothing to do with sampling variation; it remains even in the case of an "infinite" sample.

Here our basic philosophy is not to impose additional conditions which guarantee identifiability. Such conditions, in many cases, are not justified by a priori knowledge and thus may lead to prejudiced results. For this reason, the aim considered here is to obtain *classes* of observationally equivalent systems from the second moments of the observations rather than a *single* system. Since, in general, an exact description of such equivalence classes has not yet been obtained, we will give a qualitative description in terms of topological and geometrical properties. These results may be helpful for the development of numerical procedures to compute the equivalence classes. In addition, they give an illustration of the uncertainty about the underlying model due to the lack of knowledge about the error structure.

The structure theory presented in this paper, in our opinion, is of central importance for the more general problem of identification in a linear dynamic errors-in-variables setting, where we commence from data $(x_1, \ldots, x_T)$ rather than from the second order population moments. We will give a brief sketch of this more general identification problem in order to motivate the results obtained in this paper.

In linear system identification, in many cases in a first step the data are compressed in an estimate of the second moments of the observations, in our case in an estimate Σ_T of the spectrum Σ.

Let $\Sigma_T(\lambda)$ denote an estimate of $\Sigma(\lambda)$, where T denotes the sample size. Then the class of observationally equivalent systems corresponding to Σ_T is an estimate for the class corresponding to Σ. Therefore any (numerical) procedure which constructs the equivalence class to a given Σ gives an identification procedure. Such a procedure is not only reasonable, but it seems to be the obvious one.

Under general conditions, the spectral density $\Sigma(\lambda)$ of (x_t) can be consistently estimated. If the mapping attaching to Σ the corresponding class of observationally equivalent systems is continuous, then the above estimate is consistent. Therefore, in addition to describing equivalence classes for a given Σ, the continuity of the mapping described above will be considered in this paper. Since the statistical analysis of spectral estimates is well known for a number of decades, the two problems of structure theory addressed in this paper are a major and perhaps the most important module for a general theory of identification of EV models.

The main approaches to spectral estimation are as follows: on the one hand nonparametric spectral estimation, where the spectrum is estimated at a finite number of frequencies (here the number of frequencies may increase with the sample size T); on the other hand the spectrum may be estimated by fitting AR or ARMA models.

For the rest of this paper the spectral densities Σ, $\hat{\Sigma}$, and $\tilde{\Sigma}$ as well as the relation function $w(e^{-i\lambda})$ are considered for *arbitrary* but *fixed* frequency λ. If we commence from a nonparametric spectral estimate, and if no additional a priori assumptions on the order of the relation functions $w(z)$ are imposed, then our results, obtained for an arbitrary but fixed frequency, can be applied immediately.

However, our results can also be applied for varying frequencies by putting them together pointwise, e.g., a relation function w is compatible with Σ if and only if $w(e^{-i\lambda})$ is compatible with $\Sigma(\lambda)$ for every frequency λ. In particular for instance, we can check whether a given relation function w is compatible. However, we do not analyze, e.g., the additional restrictions on the equivalence classes coming from rational parametrization for $\Sigma, \hat{\Sigma}, \tilde{\Sigma}$, and w with bounded order. Such an approach seems to be very complicated; see Stemmer [27] and also some remarks in section 5.

Clearly, the results obtained for fixed frequency also apply to situations where only a narrow frequency band is considered.

No information from the observations, besides the second moments, Σ is used. This is a reasonable limitation; however, it should be mentioned that in the non-Gaussian case, higher order moments may provide important information to identify the system (see, e.g., Deistler [8], Tugnait [28]). In this respect, EV models are different from errors-in-equations models; however, we will not comment further on this issue here.

The paper is organized as follows. In section 2 we present the basic notations and definitions as well as a short description of the main results of the paper. The main results are contained in section 3 (for the Frisch case) and section 4 (for the bounded noise case). As an illustration, in section 5 the bivariate case is studied.

2. Problem statement. Remember that from now on we only consider the case of fixed frequency. For fixed frequency $\Sigma, \hat{\Sigma}$, and $\tilde{\Sigma}$ are constant positive semidefinite matrices with complex entries, rather than functions of the frequency λ. The relation (function) is a constant matrix $w \in \mathbb{C}^{m \times n}$ and we will assume that it is of full rank m. In order to be completely precise, we partly repeat definitions which have been given before now for the case of fixed frequency.

From equations (1.7) and (1.4) and our assumptions we have, for the case of fixed frequency,

$$\begin{gathered}\Sigma, \hat{\Sigma}, \tilde{\Sigma} \in \mathbb{C}^{n\times n},\\ \Sigma = \hat{\Sigma} + \tilde{\Sigma}; \qquad \hat{\Sigma} \geq 0, \tilde{\Sigma} \geq 0,\\ \hat{\Sigma} \text{ is singular.}\end{gathered} \tag{2.1}$$

For a given Σ, a matrix $\hat{\Sigma}$ is called *compatible* with Σ if $\hat{\Sigma}$ and $\tilde{\Sigma} = \Sigma - \hat{\Sigma}$ satisfy (2.1) and where, typically, $\tilde{\Sigma}$ satisfies further assumptions such as (a.5) or (a.6). Analogously then, $\tilde{\Sigma}$ and the decomposition (2.1) are called *compatible* with Σ. A relation w, i.e., a full rank matrix $w \in \mathbb{C}^{m\times n}$, is called *compatible* with Σ if there exists a compatible $\hat{\Sigma}$ such that $w\hat{\Sigma} = 0$.

The set of all compatible relations corresponding to Σ with m rows is called the *m-relation set* $\underline{\mathcal{R}}_m$ (of Σ). Sometimes we use the notation $\underline{\mathcal{R}}_m(\Sigma)$. For many purposes it is appropriate to describe the system in terms of the linear m-dimensional subspace of $\mathbb{C}^n$ generated by the rows of the relation w. By $\mathcal{R}_m(\Sigma)$ we denote the set of all such subspaces corresponding to $\underline{\mathcal{R}}_m(\Sigma)$. By system we mean either a relation w or the subspace generated by the rows of w. Thus $\underline{\mathcal{R}}_m(\Sigma)$ and $\mathcal{R}_m(\Sigma)$ in this sense are the sets of all systems with m outputs compatible with Σ.

An important integer is the maximum corank of $\hat{\Sigma}$, denoted by $\mathrm{mc}(\Sigma)$, among the set of all $\hat{\Sigma}$ which are compatible with a given Σ. At the same time, $\mathrm{mc}(\Sigma)$ is the maximum number of equations and $(n - \mathrm{mc}(\Sigma))$ is the minimum number of factors. The subclass $\mathcal{R}_m$ corresponding to $m = \mathrm{mc}(\Sigma)$ is of special interest, since in many cases we want to explain as much as possible by the system.

We define $\mathcal{S}$ as the set of all spectral densities Σ and $\mathcal{S}_m$ as the subset of $\mathcal{S}$ where $\mathrm{mc}(\Sigma) = m$ holds, i.e.,

$$\mathcal{S} = \{\Sigma | \Sigma > 0\}; \quad \mathcal{S}_m = \{\Sigma | \Sigma > 0, \mathrm{mc}(\Sigma) = m\}.$$

Note that the sets $\underline{\mathcal{R}}_m(\Sigma)$, $\mathcal{R}_m(\Sigma)$, and $\mathcal{S}_m$ depend on the particular assumption (a.i); i=5,6 imposed. We will not introduce distinct notation for each assumption since it will become clear from the context which assumption is considered.

For the Frisch case it is convenient to consider the set of all compatible noise spectral densities for given Σ, $\mathcal{E}(\Sigma)$ say, and the subsets $\mathcal{E}_m(\Sigma)$ of $\mathcal{E}(\Sigma)$ corresponding to a given corank m of $\hat{\Sigma} = \Sigma - \tilde{\Sigma}$. Of course $\mathcal{E}(\Sigma) = \mathcal{E}_1(\Sigma) \cup \cdots \cup \mathcal{E}_n(\Sigma)$, and $\mathcal{E}_m(\Sigma)$ is empty for all $m > \mathrm{mc}(\Sigma)$. Since $\tilde{\Sigma}$ is diagonal with real elements, the sets $\mathcal{E}(\Sigma)$ and $\mathcal{E}_m(\Sigma)$ can be considered as subsets of $\mathbb{R}^n$.

The following three structural problems are analyzed in detail in the paper:

1. As has been said already, our basic philosophy is not to obtain identifiability by imposing additional restrictions (which in many cases would be a prejudice). The ultimate aim is to estimate classes of observationally equivalent systems. Thus one important structural problem is to describe classes of observationally equivalent systems, i.e. sets of systems which are compatible with given Σ. The main results concerning the description of classes of compatible systems are as follows:

   For the Frisch case, section 3 contains a number of results concerning observationally equivalent systems in terms of the sets $\mathcal{E}$ and $\mathcal{E}_m$. An integer of central importance for the Frisch case is the so-called Ledermann bound $\mathrm{m_L} = \sqrt{n}$.

   The structure of the set $\mathcal{E}(\Sigma)$, as well as of the sets $\mathcal{E}_m(\Sigma)$, is analyzed in Propositions 3.1 and 3.3, respectively. A central result is contained in

Proposition 3.5, namely, that "typically" for $m \leq \mathrm{m_L}$ the sets $\mathcal{E}_m(\Sigma)$ are differentiable submanifolds of $\mathbb{R}^n$ with boundaries of dimension $n - m^2$. In addition by Proposition 3.6, for $m = \mathrm{mc}(\Sigma)$, the set of systems $\mathcal{R}_m(\Sigma)$ is a differentiable submanifold of the set of all m-dimensional subspaces of $\mathbb{C}^n$.

In the bounded noise case, things are different and in a certain sense easier. Note that in the Frisch case all off-diagonal elements of $\tilde{\Sigma}$ have to be zero, which means that we have $n(n-1)$ real equality constraints on $\tilde{\Sigma}$, whereas in the bounded noise case there is only one inequality constraint on $\tilde{\Sigma}$. This is an intuitive explanation of why, in the bounded noise case, the set of systems $\mathcal{R}_m(\Sigma)$ are typically "thick" for $m \leq \mathrm{mc}(\Sigma)$, in the sense that they contain a nonvoid open subset of the set of all m-dimensional subspaces of $\mathbb{C}^n$. See Proposition 4.3.

2. As has been stated already, from the point of view of identification the continuity of the mapping attaching classes of observationally equivalent systems to Σ is important. This relates to consistency of estimation of these equivalence classes, as explained above.

   For the Frisch case we show, in particular, that the mapping attaching $\mathcal{E}(\Sigma)$ to Σ is continuous (Proposition 3.7) and that on a generic subset of $\mathcal{S}$ the mapping attaching $\mathcal{R}_m(\Sigma)$ to Σ is continuous (Proposition 3.9). For the bounded noise case, by Proposition 4.5 the mapping attaching $\mathcal{R}_m(\Sigma)$ to Σ is continuous on a generic subset of $\mathcal{S}$.

3. Another important problem is estimation of $\mathrm{mc}(\Sigma)$. For this purpose some properties of the sets $\mathcal{S}_m$ of spectral densities Σ such that $\mathrm{mc}(\Sigma) = m$ holds are analyzed.

   In the Frisch case, for $m \leq \mathrm{m_L}$, all sets $\mathcal{S}_m$ of spectral densities with $\mathrm{mc}(\Sigma) = m$ are "thick" in the sense that they contain a nonvoid open subset of $\mathcal{S}$. On the other hand, for $m > \mathrm{m_L}$, the sets $\mathcal{S}_m$ are "thin" in the sense that they have Lebesgue measure zero. In this sense, spectral densities which allow for a system having more than $\mathrm{m_L}$ outputs are a priori unlikely.

   Let Σ_T denote an unrestricted and consistent estimate of Σ and let $\mathrm{mc}(\Sigma) \leq \mathrm{m_L}$; then by the results of Proposition 3.5, generically, $\mathrm{mc}(\Sigma_T)$ will be equal to $\mathrm{mc}(\Sigma)$ from a certain T onwards. On the other hand, for $\mathrm{mc}(\Sigma) > \mathrm{m_L}$, typically $\mathrm{mc}(\Sigma_T) \leq \mathrm{m_L}$ will hold. Thus in the first case $\mathrm{mc}(\Sigma)$ can be directly determined from $\mathrm{mc}(\Sigma_T)$, whereas in the second case the distance of Σ_T to the set $\mathcal{S}_m$ has to be taken into account in order to decide whether or not $\mathrm{mc}(\Sigma) = m$ holds. Such a decision can be based on a test or an information criterion.

   Again, for the bounded noise case things are simpler. By Proposition 4.4 all sets $\mathcal{S}_m$ are "thick."

Now let us introduce some notation. For a matrix A, say, we use the corresponding lowercase letter a_{ij} to denote its i,jth entry. The (left) kernel of a matrix A is denoted by $\ker(A)$; $\mathrm{rank}(A)$ and $\mathrm{corank}(A)$, respectively, denote the rank and corank, respectively, of A. If $A \in \mathbb{C}^{n\times n}$ is a Hermitian matrix, then $\lambda_1(A) \leq \lambda_2(A) \leq \cdots \leq \lambda_n(A)$ denote its eigenvalues. For a vector v, $\mathrm{diag}(v)$ denotes the (square) diagonal matrix whose diagonal elements are the corresponding entries of v. For a complex matrix $A \in \mathbb{C}^{m\times n}$, the matrix A^* is the complex conjugate transposed matrix. For a subset $\mathcal{A}$, say, of a topological space, $\mathcal{A}^o$ denotes the interior of $\mathcal{A}$ and $\overline{\mathcal{A}}$ denotes the closure of $\mathcal{A}$.

In the remaining part of this section we will describe topologies and geometrical structures used in this paper.

For many purposes, it is appropriate to describe the system in terms of the linear m-dimensional subspace (of $\mathbb{C}^n$) generated by the rows of the relation (function) w. Let $\mathcal{G}(m,n)$ denote the Grassmannian of all complex subspaces of dimension m of $\mathbb{C}^n$. Note that these subspaces can be identified with the equivalence classes $\{tw | t \in \mathbb{C}^{m\times m}, \det(t) \neq 0\}$, $w \in \mathbb{C}^{m\times n}$, $\operatorname{rank}(w) = m$. The topology of the Grassmannian is the quotient topology and $\mathcal{G}(m,n)$ is a differentiable manifold of real dimension $2m(n-m)$. Clearly $\mathcal{R}_m(\Sigma)$ is a subset of the corresponding Grassmannian $\mathcal{G}(m,n)$.

We always identify the set of all Hermitian $\mathbb{C}^{n\times n}$ matrices with $\mathbb{R}^{n^2}$ in an obvious way. (Note that a Hermitian $n\times n$ matrix is given by its n real diagonal elements and by its $n(n-1)/2$ complex upper diagonal elements.) The set of all strictly positive matrices $\Sigma > 0$ (of all positive semidefinite matrices $\Sigma \geq 0$) is denoted by $\mathcal{S}$ (and $\mathcal{M}$, respectively). Note that $\mathcal{S}$ is an open subset of $\mathbb{R}^{n^2}$ and $\mathcal{M}$ is the closure of $\mathcal{S}$ in $\mathbb{R}^{n^2}$, i.e., $\overline{\mathcal{S}} = \mathcal{M} \subseteq \mathbb{R}^{n^2}$.

By $\mathcal{M}_m \subset \mathcal{M}$ we denote the set of all nonnegative definite matrices of corank m. $\mathcal{M}_m$ is a differentiable submanifold of $\mathbb{R}^{n^2}$ of dimension $n^2 - m^2$. By rearranging the rows and corresponding columns, we may partition $\hat{\Sigma} \in \mathcal{M}_m$ as

$$\hat{\Sigma} = \begin{array}{c} \left(\begin{array}{cc} \hat{\Sigma}_{11} & \hat{\Sigma}_{12} \\ \hat{\Sigma}_{12}^* & \hat{\Sigma}_{22} \end{array} \right) \\ \underbrace{\phantom{\hat{\Sigma}_{11}}}_{m} \quad \underbrace{\phantom{\hat{\Sigma}_{22}}}_{n-m} \end{array} \begin{array}{l} \} \, m \\ \} \, n-m \end{array},$$

where $\hat{\Sigma}_{22} > 0$ holds. For $\hat{\Sigma}_{22} > 0$ the statements $\hat{\Sigma} \geq 0$, $\operatorname{corank}(\hat{\Sigma}) = m$, and

$$\mathrm{g}_{n,m}(\hat{\Sigma}) = (\hat{\Sigma}_{11} - \hat{\Sigma}_{12}\hat{\Sigma}_{22}^{-1}\hat{\Sigma}_{12}^*) = 0$$

are equivalent. Since $\mathrm{g}_{n,m}(\hat{\Sigma})$ is a Hermitian $m\times m$ matrix, we can interpret $\mathrm{g}_{n,m}$ as a function defined on an open subset of $\mathbb{R}^{n^2}$ mapping to $\mathbb{R}^{m^2}$. As $\mathrm{g}_{n,m}$ is infinitely often differentiable and has full rank m^2 everywhere, $\mathrm{g}_{n,m}(\hat{\Sigma}) = 0$ is a local equation system for $\mathcal{M}_m$ and

$$\left(\begin{array}{cc} \hat{\Sigma}_{11} & \hat{\Sigma}_{12} \\ \hat{\Sigma}_{12}^* & \hat{\Sigma}_{22} \end{array} \right) \longmapsto (\hat{\Sigma}_{12}, \hat{\Sigma}_{22})$$

is a local coordinate system for $\mathcal{M}_m$.

In the following, we will often use partitionings analogous to the partitioning above, without further explaining the notation used.

Let us define the set of all diagonal covariance matrices

$$\mathcal{D} = \{\tilde{\Sigma} \geq 0 \,|\, \tilde{\Sigma} \text{ is diagonal } \} \subseteq \mathbb{R}^n,$$

and for an index set $\mathcal{I} \subseteq \{1,\ldots,n\}$ we define

$$\mathcal{D}_{\mathcal{I}} = \{\tilde{\Sigma} \in \mathcal{D} \,|\, \tilde{\sigma}_{ii} > 0 \text{ for } i \in \mathcal{I} \text{ and } \tilde{\sigma}_{ii} = 0 \text{ otherwise}\}.$$

$|\mathcal{I}|$ denotes the number of elements of the index set $\mathcal{I}$.

Let $\mathcal{A}$ be a metric space endowed with the metric $\mathrm{d}(x,y)$. Then for two compact subsets $\mathcal{U}, \mathcal{V} \subseteq \mathcal{A}$, the Haussdorff distance $\mathrm{d_H}(\mathcal{U},\mathcal{V})$ is defined by

$$\mathrm{d_H}(\mathcal{U},\mathcal{V}) = \min(\rho(\mathcal{U},\mathcal{V}), \rho(\mathcal{V},\mathcal{U})),$$

where

$$\rho(\mathcal{U},\mathcal{V}) = \sup_{x\in\mathcal{U}} \inf_{y\in\mathcal{V}} \mathrm{d}(x,y).$$

If $\mathcal{C}(\mathcal{A})$ denotes the set of all compact subsets of $\mathcal{A}$, then $\mathrm{d_H}$ is a metric defined on $\mathcal{C}(\mathcal{A})$. We will consider the following three cases:

1. $\mathcal{A} = \mathbb{R}^n$ and $\mathrm{d}(x,y) = \|x-y\|$ is the usual Euclidean distance. The Haussdorff distance is then defined on $\mathcal{C}(\mathbb{R}^n)$, the set of all compact subsets of $\mathbb{R}^n$.

2. Using the Haussdorff distance we can define a metric on the Grassmannian $\mathcal{G}(m,n)$: let $\mathfrak{x},\mathfrak{y} \in \mathcal{G}(m,n)$; then we define

$$\mathrm{d}_{\mathcal{G}}(\mathfrak{x},\mathfrak{y}) = \mathrm{d_H}(\{x\in\mathfrak{x}, \|x\|\le 1\}, \{y\in\mathfrak{y}, \|y\|\le 1\}).$$

In other words, the distance of two m-dimensional subspaces of $\mathbb{C}^n$ is defined as the Haussdorff distance of the intersections of these subspaces with the unit ball. This distance has a close connection to the canonical correlations of the spaces $\mathfrak{x}$ and $\mathfrak{y}$.

3. $\mathcal{A} = \mathcal{G}(m,n)$ and $\mathrm{d}(\mathfrak{x},\mathfrak{y}) = \mathrm{d}_{\mathcal{G}}(\mathfrak{x},\mathfrak{y})$. The Haussdorff distance is then defined on $\mathcal{C}(\mathcal{G}(m,n))$, the set of all compact subsets of $\mathcal{G}(m,n)$.

3. The Frisch case. Here condition (a.5), i.e., that $\tilde{\Sigma}$ is diagonal, is imposed throughout. For given Σ then, a decomposition (2.1) is called a Frisch decomposition.

3.1. The set of all observationally equivalent systems. This subsection is concerned with the description of sets of observationally equivalent (i.e., compatible) spectra $\hat{\Sigma}$ of the latent variables $(\hat{x}_t)$ and of observationally equivalent systems. For convenience, in this section we will consider sets of observationally equivalent noise spectral densities $\tilde{\Sigma}$. Since for given Σ the matrices $\hat{\Sigma}$ and $\tilde{\Sigma}$ are in an obvious one-to-one relation, this also gives a description of the set of all observationally equivalent $\hat{\Sigma}$. Replacing $\hat{\Sigma}$ by $\tilde{\Sigma}$ is only done since sets of $\tilde{\Sigma}$'s can be embedded in $\mathbb{R}^n$. Whether or not sets of observationally equivalent $\hat{\Sigma}$'s or of observationally equivalent systems are of primary interest depends on the particular application.

The main results obtained in this subsection are as follows: In Proposition 3.1 we give a topological description of the set $\mathcal{E}(\Sigma)$ of all compatible noise spectra: $\mathcal{E}(\Sigma)$ is shown to be topologically equivalent to the intersection of the unit sphere with the first orthant in $\mathbb{R}^n$. This result is important for illustrating the nonuniqueness inherent in the Frisch case. In Proposition 3.3 we consider subsets of $\mathcal{E}$ corresponding to different numbers of outputs. In particular we see that the set $\mathcal{E}_1$ corresponding to the single output case is generic and that the closures of these sets are nested in the sense that $\overline{\mathcal{E}_1} \supseteq \cdots \supseteq \overline{\mathcal{E}_n}$ holds.

Proposition 3.5 contains one of the most important results of the paper. It is shown that, for a generic set of spectra Σ, the sets $\mathcal{E}_m(\Sigma)$ are either empty (for $m > \mathrm{mc}(\Sigma)$) or differentiable manifolds of dimension $n-m^2$ with boundaries. In particular this generic set contains only spectra with a Frisch corank $\mathrm{mc}(\Sigma) \le \sqrt{n}$. As has been mentioned already, the case $m = \mathrm{mc}(\Sigma)$ is of particular interest. In many cases only systems with the maximum number of outputs are considered. In Proposition 3.6 we show that in this case, the set of observationally equivalent $\tilde{\Sigma}$ is homeomorphic to the set of all observationally equivalent systems $\mathcal{R}_m(\Sigma)$. In addition, in this case the set of observationally equivalent systems is generically a differentiable manifold of dimension $n-m^2$ with boundaries.

We start with the following description of the set $\mathcal{E} \subseteq \mathbb{R}^n$ of all observationally equivalent $\tilde{\Sigma}$. For the next Proposition see Deistler and Scherrer [11].

PROPOSITION 3.1. *$\mathcal{E}$ is homeomorphic to $\mathcal{K}^+ = \{d \in \mathbb{R}^n \mid d_i \geq 0, \|d\| = 1\}$. Thus $\mathcal{E}$ is compact and is a topological manifold with boundaries of real dimension $n-1$.*

Proof. For each $d \in \mathcal{K}^+$, the matrix $(\Sigma - \lambda \operatorname{diag}(d))$ is positive semidefinite and singular (and thus compatible with Σ) iff λ is the smallest real number, λ_0 say, for which $(\Sigma - \lambda \operatorname{diag}(d))$ is singular. Clearly then, $1/\lambda_0$ is the largest eigenvalue of $\Sigma^{-1/2} \operatorname{diag}(d) \Sigma^{-*/2}$. Now we define a function on $\mathcal{K}^+$ by $d \mapsto \lambda_0 \operatorname{diag}(d)$. This function is continuous, because the largest eigenvalue is a continuous function of the matrix elements. (See, e.g., Golub and van Loan [15].) Since Σ is nonsingular, the inverse mapping defined by $\tilde{\Sigma} \mapsto (\tilde{\sigma}_{11}, \ldots, \tilde{\sigma}_{nn})/\|(\tilde{\sigma}_{11}, \ldots, \tilde{\sigma}_{nn})\|$ is well defined and continuous, too.

The second statement of the proposition is an immediate consequence of the first. □

It is a trivial consequence of the proposition above that the Frisch decompositions always exist and that $\tilde{\Sigma}$ is never unique without imposing further restrictions. The intersection of $\mathcal{E}(\Sigma)$ with a coordinate axis corresponds to the regression of one component of x_t on all other components, i.e., to the case where only one component of x_t is corrupted by noise. It has been shown in Schachermayer and Deistler [24] that $\mathcal{E}$ is smooth exactly at the points where $\operatorname{corank}(\Sigma - \tilde{\Sigma}) = 1$ holds.

In the next step we analyze the partitioning of the set $\mathcal{E}$ as $\mathcal{E} = \mathcal{E}_1 \cup \cdots \cup \mathcal{E}_n$, i.e., according to different numbers of outputs. The first problem in this context is to determine $\operatorname{mc}(\Sigma)$. We have the following result (compare Deistler and Anderson [10]).

PROPOSITION 3.2. *$m = \operatorname{mc}(\Sigma)$ if and only if $\underline{\mathcal{R}}_m(\Sigma) \neq \emptyset$ and $\underline{\mathcal{R}}_m(\Sigma)$ contains no w with a column equal to zero.*

Proof. As can be easily seen, $m > \operatorname{mc}(\Sigma)$ if and only if $\underline{\mathcal{R}}_m(\Sigma) = \emptyset$. Let $w \in \underline{\mathcal{R}}_{\operatorname{mc}(\Sigma)}(\Sigma)$. Clearly, by the appropriate choice of a nonsingular transformation t, a column of w can be made equal to the first unit vector. By omitting the first $m - \operatorname{mc}(\Sigma)$ rows from w, we obtain an element of $\underline{\mathcal{R}}_m(\Sigma)$ with a zero column.

Conversely, suppose that $w \in \underline{\mathcal{R}}_m(\Sigma)$ contains a zero column, e.g., the first one. We can interpret $(\Sigma - \tilde{\Sigma})$ as a variance-covariance matrix of a certain vector of complex valued random variables. The regression of the first component of this random vector on the remaining components gives a relation. By adding this relation as a row to w, we get a compatible relation for Σ with rank $m+1$. □

Note that Proposition 3.2 does not provide us with a criterion of great practical use. Practically useful criteria are available for the case $\operatorname{mc}(\Sigma) \geq m_u = (n+1)/2$ and for the static case $\operatorname{mc}(\Sigma) = 1$. See Anderson and Rubin [6], Anderson and Deistler [4], and Deistler and Scherrer [11] for the first case and see Frisch [13], Kalman [18], and Klepper and Leamer [21] for the static case.

PROPOSITION 3.3. *For every $1 \leq m \leq n$, the set $\mathcal{E}_m$ is open and dense in $\mathcal{E}_m \cup \cdots \cup \mathcal{E}_n$.*

Proof. Let $\tilde{\Sigma}^k \in \mathcal{E}_{m+1} \cup \cdots \cup \mathcal{E}_n$ be a convergent sequence with $\tilde{\Sigma}^k \to \tilde{\Sigma}^0$. Since $\mathcal{E}$ is compact, $\tilde{\Sigma}^0 \in \mathcal{E}$ holds. Note that the corank is an upper semicontinuous function of the matrix elements, since the determinant is a continuous function. Therefore $\operatorname{corank}(\Sigma - \tilde{\Sigma}^0) \geq m+1$ holds and thus $\mathcal{E}_{m+1} \cup \cdots \cup \mathcal{E}_n$ is closed.

We now show that $\mathcal{E}_m$ is dense in $\mathcal{E}_m \cup \cdots \cup \mathcal{E}_n$ by showing that every neighborhood of a $\tilde{\Sigma}^0 \in \mathcal{E}_{k+1}$ contains a $\tilde{\Sigma}$ in $\mathcal{E}_k$. Let w be a basis of $\ker(\Sigma - \tilde{\Sigma}^0)$. From $w\Sigma = w\tilde{\Sigma}^0 \neq 0$, we see that at least one column of w and the corresponding diagonal element of $\tilde{\Sigma}^0$ are unequal to zero. Then, after rearrangement of variables in x_t and by a suitable transformation t, we may write $w = (I, w_2)$ and $\tilde{\sigma}^0_{11} > 0$ holds. Now $\tilde{\Sigma} = \tilde{\Sigma}^0 - \epsilon \operatorname{diag}(1, 0, \ldots, 0)$ for $0 < \epsilon < \tilde{\sigma}^0_{11}$ is still compatible and the corank of $\Sigma - \tilde{\Sigma} = \Sigma - \tilde{\Sigma}^0 + \epsilon \operatorname{diag}(1, 0, \ldots, 0)$ has been decreased by one. Thus $\tilde{\Sigma} \in \mathcal{E}_k$. □

It is an immediate consequence of the above proposition that $\overline{\mathcal{E}_m(\Sigma)} = \mathcal{E}_m(\Sigma) \cup \cdots \cup \mathcal{E}_n(\Sigma) = \mathcal{E}_m(\Sigma) \cup \cdots \cup \mathcal{E}_{\mathrm{mc}(\Sigma)}(\Sigma)$, and in addition by the inequality $\tilde{\Sigma} \leq \Sigma$, it follows that $\overline{\mathcal{E}_m(\Sigma)}$ and in particular $\mathcal{E}_{\mathrm{mc}(\Sigma)}(\Sigma)$ are compact subsets of $\mathbb{R}^n$.

The next proposition gives some basic results for the sets $\mathcal{R}_m(\Sigma)$ and $\underline{\mathcal{R}}_m(\Sigma)$, respectively. Note that for any compatible m-relation there exists a compatible $\tilde{\Sigma} \in \overline{\mathcal{E}_m(\Sigma)} = \mathcal{E}_m(\Sigma) \cup \cdots \cup \mathcal{E}_n(\Sigma)$. Thus the sets $\underline{\mathcal{R}}_m(\Sigma)$ and $\mathcal{R}_m(\Sigma)$ are related to $\overline{\mathcal{E}_m(\Sigma)}$ rather than to $\mathcal{E}_m(\Sigma)$. Furthermore there is, in general, no one-to-one relation between $\tilde{\Sigma}$ and the system, since the mapping attaching to $\tilde{\Sigma}$ the kernel of $(\Sigma - \tilde{\Sigma})$ is not injective in general. This can be seen from $w\Sigma = w\tilde{\Sigma}$ by considering the case where w has a zero column.

PROPOSITION 3.4. *$\mathcal{R}_m(\Sigma)$ is a compact subset of $\mathcal{G}(m,n)$.*

Proof. This is an immediate consequence of the fact that $\mathcal{E}_m(\Sigma) \cup \cdots \cup \mathcal{E}_n(\Sigma)$ is compact. □

Now let us consider the mapping

$$\begin{array}{rccc} \mathrm{f}_m : & \mathcal{M}_m \times \mathcal{D} : & \longrightarrow & \mathcal{M} \subseteq \mathbb{R}^{n^2}, \\ & (\hat{\Sigma}, \tilde{\Sigma}) & \longmapsto & \Sigma = (\hat{\Sigma} + \tilde{\Sigma}), \end{array}$$

and the restrictions $\mathrm{f}_{m,\mathcal{I}}$ of f_m on $\mathcal{M}_m \times \mathcal{D}_{\mathcal{I}}$ attaching to every $\hat{\Sigma}$ with corank m and every $\tilde{\Sigma} \in \mathcal{D}_{\mathcal{I}}$ the corresponding Σ.

A heuristic motivation for the dimension of $\mathcal{E}_m$ can be given as follows. The dimension of the domain of definition of f_m is $n^2 - m^2 + n$; the dimension of $\mathcal{M}$ is n^2. Thus by comparing the dimensions, one may expect that for $m < \mathrm{m_L}$, where

$$\mathrm{m_L} = \sqrt{n}, \tag{3.1}$$

the set $\mathcal{E}_m$ is either empty or has dimension $n - m^2$. In the following we give a precise formulation of this statement. On the other hand, for $m \geq \mathrm{m_L}$ the set $\mathcal{E}_m$ can be expected to either be empty or to consist of a finite number of points. The number $\mathrm{m_L}$ is called the Ledermann bound. (Strictly speaking this term has been used only for the static case where the Ledermann bound is $\mathrm{m_L} = (-1 + \sqrt{1 + 8n})/2$. See, e.g., Ledermann [22].)

Let $\mathcal{S}^r \subseteq \mathcal{S}$ be the set of all Σ, which are regular points of all the mappings $\mathrm{f}_{m,\mathcal{I}}$, $1 \leq m \leq n$ and $\mathcal{I} \subseteq \{1, \ldots, n\}$. Note that $(\hat{\Sigma}, \tilde{\Sigma}) \in (\mathcal{M}_m \times \mathcal{D}_{\mathcal{I}})$ is called a regular point of $\mathrm{f}_{m,\mathcal{I}}$ if the derivative has full rank n^2 in $(\hat{\Sigma}, \tilde{\Sigma})$. A point $\Sigma \in \mathbb{R}^{n^2}$ is called a regular point of $\mathrm{f}_{m,\mathcal{I}}$ if all decompositions $(\hat{\Sigma}, \tilde{\Sigma}) \in \mathrm{f}_{m,\mathcal{I}}^{-1}(\Sigma)$ are regular points of $\mathrm{f}_{m,\mathcal{I}}$. (Note that, therefore, all points Σ for which $\mathrm{f}_{m,\mathcal{I}}^{-1}$ is empty are regular.) A point Σ is an "irregular" point of $\mathrm{f}_{m,\mathcal{I}}$ iff there exists a decomposition $\Sigma = \hat{\Sigma} + \tilde{\Sigma}$, $(\hat{\Sigma}, \tilde{\Sigma}) \in (\mathcal{M}_m \times \mathcal{D}_{\mathcal{I}})$, where the derivative of $\mathrm{f}_{m,\mathcal{I}}$ has rank less than n^2.

For this set $\mathcal{S}^r$ of *regular* points we have the following. (For parts of this proposition compare Dufour [12], Deistler and Scherrer [11], and Scherrer [25].)

PROPOSITION 3.5.

1. *The complement of $\mathcal{S}^r$, i.e., $\mathcal{S} \setminus \mathcal{S}^r$, is a set of Lebesgue measure zero (in this sense $\mathcal{S}^r$ is generic in $\mathcal{S}$).*
2. *$\mathcal{S}_m \cap \mathcal{S}^r = \emptyset$ for all $m > \mathrm{m_L}$.*
3. *$\mathcal{S}_m \cap \mathcal{S}^r$ is open and nonvoid in $\mathcal{S}$, and dense in $\mathcal{S}_m$, for all $m \leq \mathrm{m_L}$.*
4. *For $\Sigma \in \mathcal{S}^r$ and $m < \mathrm{m_L}$, the set $\mathcal{E}_m(\Sigma)$ is either empty or a differentiable submanifold of $\mathbb{R}^n$ with boundaries of dimension $n - m^2$. For $\Sigma \in \mathcal{S}^r$ and $m = \mathrm{m_L}$, the set $\mathcal{E}_m(\Sigma)$ contains only a finite number of points.*

Proof. 1. By the theorem of Sard (see Milnor [23]) we know that the set of "irregular" points of $\mathrm{f}_{m,\mathcal{I}}$ is a set of Lebesgue measure zero. Therefore $\mathcal{S} \setminus \mathcal{S}^r$ has Lebesgue measure zero, since it is a finite union of sets of measure zero.

2. The domain of definition of $\mathrm{f}_{m,\mathcal{I}}$ has dimension $n^2 - m^2 + |\mathcal{I}|$, which is strictly smaller than n^2 for $m > \mathrm{m_L}$.

3. We first prove that $\mathcal{S}^r$ is open in $\mathcal{S}$. Suppose that $\mathcal{S}^r$ is not open. Then there exists a sequence $\Sigma^k \in \mathcal{S} \setminus \mathcal{S}^r$ with $\Sigma^k \to \Sigma^0 \in \mathcal{S}^r$. For each Σ^k there exists a Frisch decomposition $(\hat{\Sigma}^k, \tilde{\Sigma}^k) \in (\mathcal{M}_{m(k)}, \mathcal{D}_{\mathcal{I}(k)})$, which is an "irregular" point of $\mathrm{f}_{m(k),\mathcal{I}(k)}$. There is only a finite number of possible combinations $(m(k), \mathcal{I}(k))$; thus at least one of them, $(s, \mathcal{J})$ say, must occur infinitely often. Since $(\hat{\Sigma}^k, \tilde{\Sigma}^k)$ are bounded, there exists a convergent subsequence. Putting this together we have $(\hat{\Sigma}^k, \tilde{\Sigma}^k) \in (\mathcal{M}_s, \mathcal{D}_{\mathcal{J}}) \to (\hat{\Sigma}^0, \tilde{\Sigma}^0)$. Clearly we have $\Sigma^0 = \hat{\Sigma}^0 + \tilde{\Sigma}^0$ and $\hat{\Sigma}^0 \in \mathcal{M}_m$, $m \geq s$, $\tilde{\Sigma}^0 \in \mathcal{D}_{\mathcal{I}}$, $\mathcal{I} \subseteq \mathcal{J}$. Since $(\hat{\Sigma}^0, \tilde{\Sigma}^0)$ is a regular point of $\mathrm{f}_{m,\mathcal{I}}$, we may construct a neighborhood $\mathcal{U}$ of $(\Sigma^0, \tilde{\Sigma}^0)$ as indicated in Lemma A.2. Now by assumption, $(\Sigma^k, \tilde{\Sigma}^k)$ must be an element of $\mathcal{U}$ for all k large enough; thus $(\hat{\Sigma}^k, \tilde{\Sigma}^k)$ is a regular point of $\mathrm{f}_{s,\mathcal{J}}$ for all such k, in contradiction to our assumptions.

In Proposition 3.12 we will show, that $\mathcal{S}_m$ is nonvoid and that $\mathcal{S}_m \cup \cdots \cup \mathcal{S}_1$ is open. Therefore, for $\Sigma^0 \in \mathcal{S}_m$, there exists an open neighborhood $\mathcal{U} \subseteq \mathcal{S}_m \cup \cdots \cup \mathcal{S}_1$. By Lemma A.3 and by the continuity of $\mathrm{f}_{m,\mathcal{I}}$, we may find a $\Sigma = \hat{\Sigma} + \tilde{\Sigma} \in \mathcal{U}$, where $(\hat{\Sigma}, \tilde{\Sigma})$ is a regular point of $\mathrm{f}_{m,\mathcal{I}}$. Since $\mathrm{f}_{m,\mathcal{I}}$ is locally surjective in $(\hat{\Sigma}, \tilde{\Sigma})$, we may find an open neighborhood $\mathcal{V} \subseteq \mathcal{S}_n \cup \cdots \cup \mathcal{S}_m$ of Σ. Thus $\mathcal{V} \cap \mathcal{U} \subseteq \mathcal{S}_m$ is a nonvoid open neighborhood of Σ. Since $\mathcal{S} \setminus \mathcal{S}^r$ has Lebesgue measure zero also $\mathcal{S}^r \cap \mathcal{V} \cap \mathcal{U} \subseteq \mathcal{S}^r \cap \mathcal{S}_m$ is open and nonvoid, which proves that $\mathcal{S}_m \cap \mathcal{S}^r$ is nonvoid. For each $\Sigma \in \mathcal{S}_m \cap \mathcal{S}^r$ we may construct neighborhoods $\mathcal{U}$, $\mathcal{V}$ as above. Then $\mathcal{U} \cap \mathcal{V} \cap \mathcal{S}^r \subseteq \mathcal{S}_m \cap \mathcal{S}^r$ is an open neighborhood of Σ.

Let $\Sigma^0 \in \mathcal{S}_m$. By the same arguments as above, we see that in any neighborhood of Σ^0 there exists an $\Sigma \in \mathcal{S}_m$, which has an open neighborhood $\mathcal{U} \subseteq \mathcal{S}_m$, i.e., $\Sigma \in \mathcal{S}_m^o$. Since $\mathcal{S} \setminus \mathcal{S}^r$ has Lebesgue measure zero, we can find in any neighborhood of Σ an element in $\mathcal{S}^r$ and thus in $\mathcal{S}^r \cap \mathcal{S}_m$.

4. Let $\Sigma^0 \in \mathcal{S}^r$, $\mathrm{mc}(\Sigma^0) \geq m$, and $\mathrm{m_L} > m$ hold. Thus there exists a decomposition $\Sigma^0 = \hat{\Sigma}^0 + \tilde{\Sigma}^0$, $\hat{\Sigma}^0 \in \mathcal{M}_m$, and $\tilde{\Sigma}^0 \in \mathcal{D}_{\mathcal{I}}$ for some index set $\mathcal{I}$. Since $\Sigma^0 > 0$, we may arrange the variables such that $\mathcal{I} = \{1, \ldots, l\}$ and that $\hat{\Sigma}^0_{22} > 0$ holds. Since $(\hat{\Sigma}^0, \tilde{\Sigma}^0)$ is a regular point of $\mathrm{f}_{m,\mathcal{I}}$, we may construct a neighborhood $\mathcal{U} \subseteq \mathbb{R}^n$ as in Lemma A.1.

Let $\mathcal{J} \subseteq \mathcal{I}$, $|\mathcal{J}| = m^2$ be such that the derivative of $\mathrm{g}_{n,m}(\Sigma^0 - \tilde{\Sigma})$ with respect to $d_{\mathcal{J}}(\tilde{\Sigma})$ has full rank m^2. If $\mathcal{K} = \{1, \ldots, n\} \setminus \mathcal{J}$, then the mapping

$$\begin{array}{rccc} \mathrm{h}: & \mathcal{U} & \to & \mathbb{R}^n, \\ & \tilde{\Sigma} & \mapsto & (\mathrm{g}_{n,m}(\Sigma^0 - \tilde{\Sigma}), d_{\mathcal{K}}(\tilde{\Sigma})) \end{array}$$

has full rank n in the point $\tilde{\Sigma}^0$. Thus we may further restrict $\mathcal{U}$ in a way such that $\mathrm{h}: \mathcal{U} \subseteq \mathbb{R}^n \to \mathcal{V} = \mathrm{h}(\mathcal{U}) \subseteq \mathbb{R}^n$ is a local diffeomorphism of $\mathbb{R}^n$. By this construction we have found a coordinate system h of $\mathbb{R}^n$, such that

$$\mathrm{h}(\mathcal{E}_m(\Sigma^0) \cap \mathcal{U}) = (\{0\} \times (\mathbb{R}^+)^{n-m^2}) \cap \mathcal{V}$$

holds.

Now we consider the case $m = \mathrm{m_L}$. Note that $\Sigma^0 \in \mathcal{S}^r$ implies that $\mathrm{mc}(\Sigma^0) \leq \mathrm{m_L}$ and thus $m = \mathrm{m_L} = \mathrm{mc}(\Sigma^0)$ or $\mathcal{E}_m(\Sigma^0)$ is empty. If $\mathcal{E}_m(\Sigma^0)$ contains infinitely many points, then $\mathcal{E}_m(\Sigma^0)$ must contain a limiting point, $\tilde{\Sigma}^0$ say, since $\mathcal{E}_{\mathrm{mc}}(\Sigma^0)$ is closed and bounded. On the other hand, since all points $\tilde{\Sigma} \in \mathcal{E}_m$ give regular points $(\Sigma^0 - \tilde{\Sigma}, \tilde{\Sigma})$ of $\mathrm{f}_{m,\mathcal{I}}$ (and in this case $\mathrm{f}_{m,\mathcal{I}}$ is locally injective) all points of $\mathcal{E}_m$ must be isolated. □

From the proposition above we see that $\mathcal{S}^r$ is a set of "well-behaved" spectral densities in the sense that, for every $\Sigma \in \mathcal{S}^r$, the sets $\mathcal{E}_m(\Sigma)$ are nice mathematical

objects, i.e., they are differentiable manifolds. In addition, $\mathrm{mc}(\Sigma)$ is a continuous function on $\mathcal{S}^r$, since $\mathcal{S} \setminus \mathcal{S}^r$ contains all points of discontinuity of $\mathrm{mc}(\Sigma)$.

Note that, in addition, $\mathcal{S}_1$ is fully contained in $\mathcal{S}^r$, since for all $\Sigma \in \mathcal{S}_1$ and $\tilde{\Sigma}^0 \in \mathcal{E}(\Sigma)$,

$$\mathrm{g}_{n,1}(\Sigma - \tilde{\Sigma}) = \sigma_{11} - \tilde{\sigma}_{11} - \Sigma_{12}(\Sigma_{22} - \tilde{\Sigma}_{22})^{-1}\Sigma_{12}^*$$

is defined in a neighborhood of $\tilde{\Sigma}^0$ and has full rank 1. Thus for all $\Sigma \in \mathcal{S}_1$, the set $\mathcal{E}(\Sigma)$ is differentiable submanifold of $\mathbb{R}^n$ with boundaries of dimension $n-1$.

The set $\mathcal{S} \setminus \mathcal{S}^r$ in particular contains all spectral densities with Frisch corank larger than the Ledermann bound. Thus the latter case is highly nongeneric.

For $\Sigma \in \mathcal{S} \setminus \mathcal{S}^r$ the following cases may occur (see Scherrer [25]):

1. $\mathcal{E}_m(\Sigma)$ is not a differentiable submanifold of $\mathbb{R}^n$.
2. $\mathcal{E}_m(\Sigma)$ is a differentiable submanifold of $\mathbb{R}^n$, but the dimension of $\mathcal{E}_m(\Sigma)$ is either larger or smaller than $n - m^2$.

As has been stated already, in many cases the class of observationally equivalent systems with a maximum number of equations is of particular interest. In this case we have the following.

PROPOSITION 3.6. *For $m = \mathrm{mc}(\Sigma)$, the sets $\mathcal{E}_m(\Sigma)$ and $\mathcal{R}_m(\Sigma)$ are homeomorphic. If, in addition, $\Sigma \in \mathcal{S}^r$ and $\mathrm{mc}(\Sigma) < \mathrm{m_L}$ hold, then $\mathcal{R}_m(\Sigma)$ is a differentiable submanifold of $\mathcal{G}(m,n)$ with boundaries of dimension $n - m^2$ and $\mathcal{E}_m(\Sigma)$ and $\mathcal{R}_m(\Sigma)$ are diffeomorphic.*

Proof. First we note that, for $m = \mathrm{mc}(\Sigma)$, the mapping

$$\begin{array}{rccc} \kappa_m : & \mathcal{E}_m(\Sigma) & \longrightarrow & \mathcal{R}_m(\Sigma), \\ & \tilde{\Sigma} & \longmapsto & \ker(\Sigma - \tilde{\Sigma}) \end{array}$$

is bijective by Proposition 3.2; this is true because we can compute $\tilde{\Sigma}$ from the equation $w\Sigma = w\tilde{\Sigma}$, since $w \in \underline{\mathcal{R}}_m(\Sigma)$ contains no zero column for $m = \mathrm{mc}(\Sigma)$. (Note that w and tw give the same $\tilde{\Sigma}$.)

Now let $\tilde{\Sigma}^0 \in \mathcal{E}_m(\Sigma)$, where w.l.o.g. we assume that $(\Sigma_{22} - \tilde{\Sigma}^0_{22}) > 0$ holds. Therefore the matrix $w = (I, -\Sigma_{12}(\Sigma_{22} - \tilde{\Sigma}_{22})^{-1})$ is a basis for $\ker(\Sigma - \tilde{\Sigma})$, and this is true in a neighborhood $\mathcal{U} \subseteq \mathcal{E}_m(\Sigma)$ of $\tilde{\Sigma}^0$. In addition, in a neighborhood $\mathcal{V} \subseteq \mathcal{G}(m,n)$ of $\ker(\Sigma - \tilde{\Sigma}^0)$, we can use the coordinate mapping

$$\begin{array}{rccc} \mathrm{k} : & \mathcal{G}(m,n) \supseteq \mathcal{V} & \longrightarrow & \mathcal{V}' \subseteq \mathbb{R}^{2m(n-m)}, \\ & \{t(w_1, w_2) | \det(t) \neq 0\} & \longmapsto & w_1^{-1} w_2 \end{array}$$

for $\mathcal{G}(m,n)$. Therefore $\kappa_m(\tilde{\Sigma}) = \mathrm{k}^{-1}(-\Sigma_{12}(\Sigma_{22} - \tilde{\Sigma}_{22})^{-1})$ for all $\tilde{\Sigma} \in \mathcal{U}$ and $\kappa_m(\cdot)$ is continuous, since k is a homeomorphism.

Next, define the mapping

$$\mathrm{h} : w \in \mathbb{C}^{m \times n} \longmapsto \mathrm{diag}((w_j^* w \Sigma_j)/(w_j^* w_j))_{j=1,\ldots,n},$$

where w_j denotes the jth column of w and Σ_j denotes the jth column of Σ. The mapping h is defined and continuous for all $w \in \mathbb{C}^{m \times n}$, which have no zero column. Then we have $\kappa_m^{-1}(\mathfrak{x}) = \mathrm{h}(I, \mathrm{k}(\mathfrak{x}))$ for all $\mathfrak{x} \in \mathcal{V} \cap \mathcal{R}_m(\Sigma)$, which proves that κ_m^{-1} is continuous.

Now we consider the case $\Sigma \in \mathcal{S}^r$ and $m = \mathrm{mc}(\Sigma) < \mathrm{m_L}$. As is shown in the proof of Proposition 3.5, we can use the mapping

$$\begin{array}{rccc} \mathrm{h} : & \mathcal{U} \subseteq \mathcal{E}_m(\Sigma) & \longrightarrow & \mathcal{U}' \cap (\mathbb{R}^+)^{n-m^2}, \\ & \tilde{\Sigma} & \longmapsto & x = d_{\mathcal{K}}(\tilde{\Sigma}) \end{array}$$

as a coordinate mapping for $\mathcal{E}_m(\Sigma)$, where $d_{\mathcal{K}}(\tilde{\Sigma})$ denotes a (suitable) selection of $n-m^2$ diagonal entries of $\tilde{\Sigma}$. (Note that $\mathcal{U}' \subseteq \mathbb{R}^{n-m^2}$ is open in $\mathbb{R}^{n-m^2}$.) Since $\mathcal{E}_m(\Sigma)$ and $\mathcal{R}_m(\Sigma)$ are homeomorphic by the above considerations, we can choose $\mathcal{U},\mathcal{U}'$ and $\mathcal{V},\mathcal{V}'$ in a way such that $\kappa_m(\mathcal{U}) = \mathcal{R}_m(\Sigma) \cap \mathcal{V}$ holds.

Next we consider the mapping

$$\mathrm{k} \circ \kappa_m \circ \mathrm{h}^{-1} : x \mapsto \tilde{\Sigma} \mapsto \ker(\Sigma - \tilde{\Sigma}) \mapsto -\Sigma_{12}(\Sigma_{22} - \tilde{\Sigma}_{22})^{-1}.$$

Note that (by choosing $\mathcal{U}'$ small enough) $\mathrm{k} \circ \kappa_m \circ \mathrm{h}^{-1}$ is defined and differentiable on $\mathcal{U}'$. (Thus we allow for the moment also small negative entries in $\tilde{\Sigma}$.) The derivative is given by

$$\delta x \mapsto \delta\tilde{\Sigma} \mapsto \delta\ker(\Sigma - \tilde{\Sigma}) \mapsto \underbrace{-\Sigma_{12}(\Sigma_{22} - \tilde{\Sigma}_{22})^{-1}}_{w_2} \delta\tilde{\Sigma}_{22}(\Sigma_{22} - \tilde{\Sigma}_{22})^{-1},$$

which is zero iff $\delta\tilde{\Sigma}_{22} = 0$, since w_2 contains no zero column. Since $\mathrm{g}_{n,m}(\Sigma - \tilde{\Sigma}) \equiv 0$, each direction $\delta\tilde{\Sigma}$ in the tangent space of $\mathcal{E}_m(\Sigma)$ must fulfill the equation

$$-\delta\tilde{\Sigma}_{11} - w_2\delta\tilde{\Sigma}_{22}w_2^* = 0.$$

Thus $\delta\tilde{\Sigma}_{22} = 0$ implies $\delta\tilde{\Sigma} = 0$, and $\mathrm{k} \circ \kappa_m \circ \mathrm{h}^{-1}$ has full rank $n-m^2$ in all points of $\mathcal{U}' \cap (\mathbb{R}^+)^{n-m^2}$. Therefore we can find a diffeomorphism $\mathrm{k}' : \mathcal{V}' \subseteq \mathbb{R}^{2m(n-m)} \to \mathcal{V}'' \subseteq \mathbb{R}^{2m(n-m)}$ such that $\mathrm{k}' \circ \mathrm{k} \circ \kappa_m \circ \mathrm{h}^{-1}$ takes the form

$$(x_1, \ldots, x_{n-m^2}) \mapsto (x_1, \ldots, x_{n-m^2}, 0, \ldots, 0).$$

(Again we have to shrink the neighborhoods $\mathcal{U},\mathcal{U}'$ and $\mathcal{V},\mathcal{V}',\mathcal{V}''$ suitably.) Putting this together we have found a coordinate mapping $\mathrm{k}'' = \mathrm{k}' \circ \mathrm{k} : \mathcal{V} \to \mathcal{V}''$ of $\mathcal{G}(m,n)$ such that

$$\mathrm{k}''(\mathcal{R}_m(\Sigma) \cap \mathcal{V}) = ((\mathbb{R}^+)^{n-m^2} \times \{0\}) \cap \mathcal{V}''$$

holds. □

3.2. Continuity results. If we commence from real data, Σ is not known a priori but has to be estimated. As is well known under general assumptions on (x_t), e.g., the usual nonparametric estimates Σ_T of the spectral density of Σ are consistent. Then, e.g., $\mathcal{R}_m(\Sigma_T)$ is an estimate for $\mathcal{R}_m(\Sigma)$. In this context the question arises whether the estimate $\mathcal{R}_m(\Sigma_T)$ is close to the "true" set of systems $\mathcal{R}_m(\Sigma)$, if Σ_T is close to Σ; in other words, whether the mapping $\Sigma \mapsto \mathcal{R}_m(\Sigma)$ is continuous. Then if $\Sigma_T \to \Sigma$, this implies $\mathcal{R}_m(\Sigma_T) \to \mathcal{R}_m(\Sigma)$, i.e., $\mathcal{R}_m(\Sigma_T)$ are consistent estimates.

First we consider the continuity of the mappings $\Sigma \mapsto \mathcal{E}(\Sigma)$ and $\Sigma \mapsto \mathcal{E}_m(\Sigma)$. Proposition 3.8 shows that the mapping $\Sigma \mapsto \overline{\mathcal{E}_m(\Sigma)}$ is continuous for $\Sigma \in \mathcal{S}^r$ and Proposition 3.7 shows that $\Sigma \mapsto \mathcal{E}(\Sigma)$ is globally continuous. Proposition 3.9 shows that the mapping $\Sigma \mapsto \mathcal{R}_m(\Sigma)$ is continuous for $\Sigma \in \mathcal{S}^r$.

PROPOSITION 3.7. *The mapping $\mathcal{S} \to \mathcal{C}(\mathbb{R}^n) : \Sigma \to \mathcal{E}(\Sigma)$ is continuous (with respect to the Haussdorff distance).*

Proof. We consider a sequence $(\Sigma^k \in \mathcal{S})$ which converges to $\Sigma^0 \in \mathcal{S}$, i.e., $\Sigma^k \to \Sigma^0$. In order to prove that $\mathrm{d_H}(\mathcal{E}(\Sigma^0), \mathcal{E}(\Sigma^k)) \to 0$ holds, by Lemmas A.4 and A.5, we have to construct, for every $\tilde{\Sigma}^0 \in \mathcal{E}(\Sigma^0)$, a sequence $(\tilde{\Sigma}^k \in \mathcal{E}(\Sigma^k))$ with $\tilde{\Sigma}^k \to \tilde{\Sigma}^0$. It is easy to see that $\tilde{\Sigma}^k = \lambda\tilde{\Sigma}^0$, where λ is the reciprocal of the largest eigenvalue of the matrix $(\Sigma^k)^{-1/2}\tilde{\Sigma}^0(\Sigma^k)^{-*/2}$, gives such a sequence, since $\lambda \to 1$. Note that the largest

eigenvalue of $(\Sigma^0)^{-1}\tilde{\Sigma}^0(\Sigma^0)^{-1}$ is equal to one. (Compare the proof of Proposition 3.1.) ∎

PROPOSITION 3.8. *The mapping* $\mathcal{S}^r \to \mathcal{C}(\mathbb{R}^n)$*:* $\Sigma \to \overline{\mathcal{E}_m(\Sigma)}$ *is continuous (with respect to the Haussdorff distance).*

Proof. We consider a point $\Sigma^0 \in \mathcal{S}^r$ and a sequence $\Sigma^k \in \mathcal{S}^r$ with $\Sigma^k \to \Sigma^0$. By Proposition 3.5, w.l.o.g. we can assume that $\mathrm{mc}(\Sigma^k) = \mathrm{mc}(\Sigma^0) = m_0 \leq \mathrm{m_L}$ holds. If $m > m_0$, then $\overline{\mathcal{E}_m(\Sigma^0)} = \emptyset = \overline{\mathcal{E}_m(\Sigma^k)}$. Thus we have only to consider the case $m \leq m_0$.

Let $\tilde{\Sigma}^0 \in \overline{\mathcal{E}_m(\Sigma^0)}$; then $(\hat{\Sigma}^0, \tilde{\Sigma}^0) \in (\mathcal{M}_s \times \mathcal{D}_\mathcal{I})$ is a regular point of $\mathrm{f}_{s,\mathcal{I}}$, where $\hat{\Sigma}^0 = \Sigma^0 - \tilde{\Sigma}^0$, $s \geq m$, and $\mathcal{I} \subseteq \{1, \ldots, n\}$. For each neighborhood $\mathcal{U} \subseteq (\mathcal{M}_s \times \mathcal{D}_\mathcal{I})$ of $(\hat{\Sigma}^0, \tilde{\Sigma}^0)$, the image $\mathrm{f}_{s,\mathcal{I}}(\mathcal{U}) \subseteq \mathbb{R}^{n^2}$ is a neighborhood of Σ^0 in $\mathbb{R}^{n^2}$, since $\mathrm{f}_{s,\mathcal{I}}$ is locally surjective. Since $\Sigma^k \to \Sigma^0$, we have $\Sigma^k \in \mathrm{f}_{s,\mathcal{I}}(\mathcal{U})$ for all k large enough and we can find a decomposition $\Sigma^k = \hat{\Sigma}^k + \tilde{\Sigma}^k$, $(\hat{\Sigma}^k, \tilde{\Sigma}^k) \in \mathcal{U}$ (and thus $\tilde{\Sigma}^k \in \overline{\mathcal{E}_m(\Sigma^k)}$) for all such k's. Now, by considering a sequence of "shrinking" neighborhoods of $(\hat{\Sigma}^0, \tilde{\Sigma}^0)$, we can construct a sequence $\tilde{\Sigma}^k \in \overline{\mathcal{E}_m(\Sigma^k)}$ with $\tilde{\Sigma}^k \to \tilde{\Sigma}^0$. By Lemmata A.4 and A.5 we therefore have $\mathrm{d_H}(\overline{\mathcal{E}_m(\Sigma^0)}, \overline{\mathcal{E}_m(\Sigma^k)}) \to 0$. ∎

In the following proposition we describe the continuity results for the relation sets $\mathcal{R}_m(\Sigma)$. Note that by Proposition 3.4, the relation sets $\mathcal{R}_m(\Sigma)$ are compact subsets of the Grassmannian $\mathcal{G}(m, n)$.

PROPOSITION 3.9. *The function* $\mathcal{S}^r \to \mathcal{C}(\mathcal{G}(m,n))$*:* $\Sigma \mapsto \mathcal{R}_m(\Sigma)$ *is continuous (with respect to the Haussdorff distance).*

Proof. This result immediately follows from Proposition 3.8 and Lemma A.7. ∎

Note that for $\mathrm{mc}(\Sigma) > \sqrt{n}$, typically the set $\mathcal{E}_m(\Sigma)$ for $m = \mathrm{mc}(\Sigma)$ will be a singleton, i.e., the system corresponding to the maximum number of outputs will be unique; see Scherrer [25]. If $m = \mathrm{mc}(\Sigma)$ is known, then estimates for Σ taking into this restriction may be used. For this case the following result holds.

PROPOSITION 3.10. *For a sequence* $\Sigma^k \in \mathcal{S}$*,* $\mathrm{mc}(\Sigma^k) \geq m$*,* $\Sigma^k \to \Sigma^0 \in \mathcal{S}$*,* $\mathcal{E}_m(\Sigma^0) = \{\tilde{\Sigma}^0\}$*, and* $\mathcal{R}_m(\Sigma^0) = \{\mathfrak{r}^0\}$*, we have*

$$\mathrm{d_H}(\mathcal{E}_m(\Sigma^k), \mathcal{E}_m(\Sigma^0)) \to 0 \quad \text{and} \quad \mathrm{d_H}(\mathcal{R}_m(\Sigma^k), \mathcal{R}_m(\Sigma^0)) \to 0.$$

Proof. Note that for $m < \mathrm{mc}(\Sigma^0)$, the set $\mathcal{E}_m(\Sigma^0)$ contains infinitely many elements. Thus the assumption $\mathcal{E}_m(\Sigma^0) = \{\tilde{\Sigma}^0\}$ implies $\mathrm{mc}(\Sigma^0) = m$. Since $\mathcal{S}_1 \cup \cdots \cup \mathcal{S}_m$ is open (see Proposition 3.12), our assumptions imply that $\mathrm{mc}(\Sigma^k) = m$ for all k large enough. Thus $\mathcal{E}_m(\Sigma^0)$ and $\mathcal{E}_m(\Sigma^k)$ are compact subsets of $\mathbb{R}^n$ for all such k's.

Now we define a sequence $(\tilde{\Sigma}^k \in \mathcal{E}_m(\Sigma^k))$ by

$$\|\tilde{\Sigma}^0 - \tilde{\Sigma}^k\| = \inf_{\tilde{\Sigma} \in \mathcal{E}_m(\Sigma^k)} \|\tilde{\Sigma}^0 - \tilde{\Sigma}\|.$$

Since the $(\tilde{\Sigma}^k)$'s are bounded, there exist a limiting point $\tilde{\Sigma}^l$, say. As in the proof of Lemma A.5, we can see that $\tilde{\Sigma}^l \in \mathcal{E}_m(\Sigma^0)$ and thus $\tilde{\Sigma}^l = \tilde{\Sigma}^0$. Therefore Lemmata A.4 and A.5 imply the convergence of the sets $\mathcal{E}_m(\Sigma^k) \to \mathcal{E}_m(\Sigma^0)$.

The convergence of the sets $\mathcal{R}_m(\Sigma^k) \to \mathcal{R}_m(\Sigma^0)$ follows from an analogous reasoning. ∎

In the next proposition, continuity results are considered when the true system is observed with small noise satisfying (a.5). Then the set of observationally equivalent systems is small and close to the true system.

PROPOSITION 3.11 (low noise). *Let* $\hat{\Sigma}^0 \in \mathcal{M}_m$ *and* $w \in \mathbb{C}^{m \times n}$ *be a basis for the left kernel of* $\hat{\Sigma}^0$ *such that* w *contains no zero column. If* $\Sigma^k \in \mathcal{S} \to \hat{\Sigma}^0$*, then* $\mathcal{E}(\Sigma^k) \to$

$\{0 \in \mathbb{C}^{n\times n}\}$ *holds. If we have, in addition,* $\mathrm{mc}(\Sigma^k) \geq m$, *then* $\mathcal{E}_m(\Sigma^k) \to \{0 \in \mathbb{C}^{n\times n}\}$ *and* $\mathcal{R}_m(\Sigma^k) \to \{\ker(\hat{\Sigma}^0)\}$ *hold.*

Proof. By assumption there is a vector $v \in \ker(\hat{\Sigma}^0)$ such that all components v_i of v are nonzero. Let us define $\epsilon^k = v(\Sigma^k - \hat{\Sigma}^0)v^*$. Then we have

$$0 \leq v(\Sigma^k - \tilde{\Sigma})v^* = \underbrace{v\hat{\Sigma}^0 v^*}_{=0} + \underbrace{v(\Sigma^k - \hat{\Sigma}^0)v^*}_{=\epsilon^k} - v\tilde{\Sigma}v^*,$$

and thus

$$\tilde{\sigma}_{ii} \leq \epsilon^k / v_i^2 \to 0$$

for all $\tilde{\Sigma} \in \mathcal{E}(\Sigma^k)$. The second statement can be proved analogously to the proof of Proposition 3.10. ∎

3.3. Some further properties of $\mathcal{S}_m$. In applications one might be interested in determining $\mathrm{mc}(\Sigma)$ from data. For this purpose one may use, for instance, a sequence of likelihood ratio tests. In order to derive the properties of such a procedure, the properties of the sets $\mathcal{S}_m$ have to be investigated. In addition to Proposition 3.5, we have the following properties of these sets: from Propositions 3.5 and 3.12 we see that $\mathcal{S}_1$ is a nonvoid open subset of $\mathcal{S}$, $\mathcal{S}_2, \ldots, \mathcal{S}_m$, $m \leq \mathrm{m_L}$ contain nonvoid open subsets, and $\mathcal{S}_m$, $m > \mathrm{m_L}$ are thin subsets of $\mathcal{S}$. The closures of the sets $\mathcal{S}_m$, $m \geq \mathrm{m_L}$ are nested. It should be noted, however, that for the derivation of the statistical properties of likelihood ratio test, in addition a manifold structure usually is required.

PROPOSITION 3.12.

1. $\mathcal{S}_m$ *is nonvoid for all* $1 \leq m \leq n$.
2. $\mathcal{S}_m \cup \cdots \cup \mathcal{S}_1$ *is open in* $\mathcal{S}$ *for all* $1 \leq m \leq n$.
3. $\overline{\mathcal{S}_m} = \overline{\mathcal{S}_m \cup \cdots \cup \mathcal{S}_n}$ *for* $m \geq \mathrm{m_L}$.

Proof. 1. Let $w = (1, \ldots, 1)$ and $O \in \mathbb{C}^{n\times(n-m)}$ with $wO = 0$ and $O^*O = I \in \mathbb{C}^{(n-m)\times(n-m)}$. We now prove that $\Sigma = OO^* + \epsilon I$ is an element of $\mathcal{S}_m$ for all $0 < \epsilon < 1/(n-1)$. Clearly $\mathrm{mc}(\Sigma) \geq m$. Suppose that $\Sigma = \hat{\Sigma} + \tilde{\Sigma}$ is a Frisch decomposition of Σ. Now $0 \leq w\hat{\Sigma}w^* = wOO^*w^* + \epsilon ww^* - w\tilde{\Sigma}w^*$ implies that $\tilde{\sigma}_{ii} \leq n\epsilon$ for all i. Thus we have for all $v = \lambda O^*$, $\lambda \in \mathbb{C}^{1\times n-m}$, $\lambda\lambda^* = 1 = vv^*$,

$$v\hat{\Sigma}v^* = vOO^*v^* + \epsilon vv^* - v\tilde{\Sigma}v^* = 1 + \epsilon - v\tilde{\Sigma}v^* \geq (1+\epsilon) - n\epsilon = 1 - (n-1)\epsilon > 0.$$

Therefore $O^*\hat{\Sigma}O > 0$, and thus $\mathrm{corank}(\hat{\Sigma}) \leq m$ must hold.

2. Consider a sequence $\Sigma^k \in \mathcal{S}_n \cup \cdots \cup \mathcal{S}_{m+1}$, $\Sigma^k \to \Sigma^0 \in \mathcal{S}$. For each k there exists a Frisch decomposition $\Sigma = \hat{\Sigma}^k + \tilde{\Sigma}^k$ with $\mathrm{corank}(\hat{\Sigma}^k) \geq m+1$. Since the $(\hat{\Sigma}^k, \tilde{\Sigma}^k)$ are bounded there exists a limiting point $(\hat{\Sigma}^0, \tilde{\Sigma}^0)$. For this point we have $\Sigma^0 = \hat{\Sigma}^0 + \tilde{\Sigma}^0$, $\hat{\Sigma}^0 \geq 0$, $\mathrm{corank}(\hat{\Sigma}^0) \geq m+1$, and $\tilde{\Sigma}^0 \geq 0$ which implies $\Sigma^0 \in \mathcal{S}_n \cup \cdots \cup \mathcal{S}_{m+1}$.

3. Let $\Sigma^0 \in \mathcal{S}_{m+1}$. We consider all matrices Σ of the form $\Sigma = \Sigma^0 + \Lambda\Lambda^*$, $\Lambda \in \mathbb{C}^{n\times n}$, where the largest eigenvalue of $\Lambda\Lambda^*$ is smaller than some $\epsilon > 0$. Clearly this set contains an open and nonvoid subset of $\mathcal{S}$. Since $\mathcal{S}^r$ is dense in $\mathcal{S}$, there exists a Λ such that $\Sigma^0 + \Lambda\Lambda^* \in \mathcal{S}^r$, and therefore $\mathrm{mc}(\Sigma^0 + \Lambda\Lambda^*) \leq \mathrm{m_L}$. We now define n matrices $\Sigma^i = \Sigma^0 + \sum_{j=1}^{i} \lambda_j\lambda_j^*$, where λ_j denotes the jth column of Λ. It is trivial to see that $\mathrm{mc}(\Sigma^i) \geq \mathrm{mc}(\Sigma^{i-1}) - 1$ holds. Thus we must have $\mathrm{mc}(\Sigma^i) = m$ for some index i. In this way we may find a matrix $\Sigma \in \mathcal{S}_m$ in any neighborhood of $\Sigma^0 \in \mathcal{S}_{m+1}$. ∎

4. The bounded noise case. In a number of applications, the assumption (a.5) that $\tilde{\Sigma}$ is diagonal, i.e., that the noise components are mutually uncorrelated, is not appropriate. Here we consider the alternative assumption that the noise level is bounded. This assumption is expressed as

(a.6) $\lambda_n(\tilde{\Sigma}(\lambda)) \leq \epsilon$.

The idea behind this assumption is that the order of magnitude of the noise is known a priori and nothing else. We will adhere to assumption (a.6) throughout this section. Clearly, terms such as *compatible* or *m-relation set* in this section relate to assumption (a.6) and not to (a.5).

If we replace assumption (a.6) by

$$\mathrm{tr}\tilde{\Sigma} \leq \epsilon,$$

most results can be shown analogously. Note that in the static case, $\mathrm{tr}(\tilde{\Sigma}) = \mathrm{E}u'u$ is the mean square error.

In Proposition 4.1, for a given relation w a corresponding minimal noise spectrum $\tilde{\Sigma}$ is derived. In Proposition 4.2, a characterization of compatible relations is given and it is shown that $\mathrm{mc}(\Sigma)$ can be easily determined from the eigenvalues of Σ. From Proposition 4.3, we see that for $\lambda_m(\Sigma) < \epsilon$ the sets $\mathcal{R}_m(\Sigma)$, in a certain sense, are of the same topological dimension as $\mathcal{G}(m,n)$, namely, $2m(n-m)$. For the Frisch case, on the other hand for the special case $m = \mathrm{mc}(\Sigma) < \mathrm{m_L}$, we see from Propositions 3.5 and 3.6 that $\mathcal{R}_m(\Sigma)$ has generically dimension $n-m^2$ which is smaller than $2m(n-m)$. This result is not implausible, since (a.5) imposes more restrictions than (a.6). Some topological properties of the sets $\mathcal{S}_m$ are considered in Proposition 4.4; it is shown that all sets $\mathcal{S}_m$ are "thick" in the sense that they contain an open nonvoid subset of $\mathcal{S}$. Finally, Proposition 4.5 shows that the mapping $\Sigma \mapsto \mathcal{R}_m(\Sigma)$ is continuous on a generic subset of $\mathcal{S}$.

For the next proposition see Kalman [20].

PROPOSITION 4.1. *Let Σ be given, let w be an arbitrary but fixed relation, and consider the set of all $\tilde{\Sigma}$ satisfying* (1.7) *and* (1.4), *and $\hat{\Sigma}, \tilde{\Sigma} \geq 0$. With respect to the semi-ordering given by semipositivity of matrices, this set has a unique minimal element*

$$\tilde{\Sigma}_w = \Sigma w^*(w\Sigma w^*)^{-1} w\Sigma. \tag{4.1}$$

Proof. Since w has full rank, there is "coordinate transformation" $x_t \to t(\mathrm{z})x_t$, such that $\bar{w} = wt^{-1} = (I, 0)$. If $\bar{\Sigma}$, $\hat{\bar{\Sigma}}$, and $\tilde{\bar{\Sigma}}$ denote the corresponding transformed spectral densities Σ, $\hat{\Sigma}$, and $\tilde{\Sigma}$, respectively, then in an obvious partitioning,

$$\begin{pmatrix} \bar{\Sigma}_{11} & \bar{\Sigma}_{12} \\ \bar{\Sigma}_{21} & \bar{\Sigma}_{22} \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & \hat{\bar{\Sigma}}_{22} \end{pmatrix} + \begin{pmatrix} \tilde{\bar{\Sigma}}_{11} & \tilde{\bar{\Sigma}}_{12} \\ \tilde{\bar{\Sigma}}_{21} & \tilde{\bar{\Sigma}}_{22} \end{pmatrix}$$

holds. Since $\tilde{\bar{\Sigma}}$ must be positive semidefinite, the above equation implies $\tilde{\bar{\Sigma}}_{22} \geq \bar{\Sigma}_{21}\bar{\Sigma}_{11}^{-1}\bar{\Sigma}_{12}$. This inequality now gives

$$\tilde{\bar{\Sigma}} \geq \begin{pmatrix} \bar{\Sigma}_{11} & \bar{\Sigma}_{12} \\ \bar{\Sigma}_{21} & \bar{\Sigma}_{21}\bar{\Sigma}_{11}^{-1}\bar{\Sigma}_{12} \end{pmatrix} = \bar{\Sigma}\bar{w}^*(\bar{w}\bar{\Sigma}\bar{w}^*)^{-1}\bar{w}\bar{\Sigma},$$

and thus by backsubstituting, $\tilde{\Sigma} = t^{-1}\tilde{\bar{\Sigma}}t^{-*} \geq \Sigma w^*(w\Sigma w^*)^{-1}w\Sigma$. ∎

An (almost) immediate consequence of the above proposition is the following characterization of compatible relations, and of the maximum corank, for a given Σ.

PROPOSITION 4.2.

1. *$w \in \mathbb{C}^{m\times n}$, $\mathrm{rank}(w) = m$ is compatible with Σ iff $wSw^* \geq 0$ holds, where $S = \epsilon\Sigma - \Sigma\Sigma$.*

2. $\mathrm{mc}(\Sigma) = m$ *iff* $\lambda_1(\Sigma) \leq \lambda_2(\Sigma) \leq \cdots \leq \lambda_m(\Sigma) \leq \epsilon < \lambda_{m+1}(\Sigma) \leq \cdots \leq \lambda_n(\Sigma)$ *holds.*

Proof. By Proposition 4.1, we know that w is compatible iff

$$\tilde{\Sigma}_w = \Sigma w^*(w\Sigma w^*)^{-1} w\Sigma \leq \epsilon I$$

holds. As can be easily seen, this equivalent to

$$\begin{pmatrix} \epsilon I & \Sigma w^* \\ w\Sigma & w\Sigma w^* \end{pmatrix} \geq 0,$$

which in turn is equivalent to $w\Sigma w^* \geq \frac{1}{\epsilon} w\Sigma\Sigma w^*$.

Now $\underline{\mathcal{R}}_m(\Sigma)$ is not empty iff $S = \epsilon\Sigma - \Sigma\Sigma$ has at least m eigenvalues greater than or equal to zero. Thus item 2 follows immediately. □

Note that in this case, as opposed to the Frisch case, the characterization of $\mathrm{mc}(\Sigma)$ from Σ is easy. Note also that in the bounded noise case considered here, all m-relation sets may be empty (if ϵ is too small) and thus $\mathrm{mc}(\Sigma) = 0$ may hold.

In the next proposition, some topological and geometric properties of the m-relation set $\mathcal{R}_m(\Sigma)$ are given.

PROPOSITION 4.3.

1. $\mathcal{R}_m(\Sigma)$ *is a compact subset of* $\mathcal{G}(m,n)$.

2. *For* $\lambda_m(\Sigma) < \epsilon$, *the set* $\mathcal{R}_m(\Sigma)$ *contains an open (and nonvoid) subset of* $\mathcal{G}(m,n)$ *and* $\mathcal{R}_m(\Sigma) = \overline{\mathcal{R}_m(\Sigma)^o}$.

Proof. 1. Since $\mathcal{G}(m,n)$ is compact, it remains to prove that $\mathcal{R}_m(\Sigma)$ is closed. Let $\mathfrak{x}_k \in \mathcal{R}_m(\Sigma)$ be a sequence converging to $\mathfrak{x}_0 \in \mathcal{G}(m,n)$. Without loss of generality, let $w_0 = (I, w_2^0)$ be a basis for $\mathfrak{x}_0$ and let $w_k = (I, w_2^k)$ be the corresponding basis for $\mathfrak{x}_k$. Then clearly $w_k \to w_0$ and $\tilde{\Sigma}_{w_k} \to \tilde{\Sigma}_{w_0}$, where $\tilde{\Sigma}_w$ denotes the least squares error covariance defined by (4.1). Clearly $\tilde{\Sigma}_{w_0}$ satisfies (a.6) and thus $\mathfrak{x}_0 \in \mathcal{R}_m(\Sigma)$ holds.

2. By assumption, the matrix $S = \epsilon\Sigma - \Sigma\Sigma$ has at least m eigenvalues strictly larger than zero. If $w \in \mathbb{C}^{m\times n}$ spans the corresponding eigenspace of S, then $wSw^* > 0$ holds. Thus $\underline{\mathcal{R}}_m(\Sigma)$ and therefore, also, $\mathcal{R}_m(\Sigma)$ contains an open and nonvoid subset.

Clearly $\mathcal{R}_m^o \subseteq \mathcal{R}_m$ and thus $\overline{\mathcal{R}_m^o} \subseteq \overline{\mathcal{R}_m} = \mathcal{R}_m$.

Since unitary coordinate transformations do not change the topological structure of $\mathcal{R}_m(\Sigma)$, we may, w.l.o.g., assume that Σ is a diagonal matrix and that the diagonal elements of Σ are ordered increasing in size. Let $S = \epsilon\Sigma - \Sigma\Sigma$ be partitioned as

$$S = \begin{pmatrix} \Theta_{11} & 0 \\ 0 & \Theta_{22} \end{pmatrix},$$

where Θ_{11} is an $s \times s$, $s \geq m$ matrix and $\Theta_{11} \geq 0$ and $\Theta_{22} < 0$ hold. If $\mathfrak{x} \in \mathcal{R}_m(\Sigma)$ has a basis $w = (w_1, w_2)$, which is partitioned conformingly, then it is easy to see that w_1 must have full rank m. If $tw_1 = 0$ holds for some vector $t \in \mathbb{C}^{1\times m}$, then $0 \leq twSw^*t^* = tw_2\Theta_{22}w_2^*t^*$ implies $tw_2 = 0$, since $\Theta_{22} < 0$ holds. Since by assumption Θ_{11} has at least rank m, it follows that in any neighborhood of w_1 there exists a full rank matrix $\bar{w}_1$, such that $\bar{w}_1\Theta_{11}\bar{w}_1^* > w_1\Theta_{11}w_1^*$ holds. Let $\bar{w} = (\bar{w}_1, w_2)$; then clearly $\bar{w}S\bar{w}^* > 0$ holds, i.e., the corresponding subspace $\bar{\mathfrak{x}}$ is an element of $\mathcal{R}_m^o$. □

Item 2 of the above proposition shows that the boundary of $\mathcal{R}_m(\Sigma)$ has a simple structure. Next we consider some topological properties of the sets $\mathcal{S}_m$ of all spectral densities Σ with $\mathrm{mc}(\Sigma) = m$.

PROPOSITION 4.4. $\mathcal{S}_m$ *contains an open nonvoid subset and* $\mathcal{S}_n \cup \cdots \cup \mathcal{S}_m$ *is closed in* $\mathcal{S}$.

Proof. The result is a straightforward consequence of the fact that, for two Hermitian matrices A, E, the following relations for the eigenvalues holds:

$$\lambda_k(A) + \lambda_1(E) \leq \lambda_k(A+E) \leq \lambda_k(A) + \lambda_n(E).$$

(See, e.g., Golub and van Loan [15].) □

The above propositions state that all sets $\mathcal{S}_m$ are "thick subsets" of the set of all spectral densities. Note that in the Frisch scheme, on the contrary, the set of all spectral densities corresponding to a corank $m > \sqrt{n}$ are "thin subsets."

With the same motivation as in the Frisch case, we now consider the continuity of the mapping $\Sigma \mapsto \mathcal{R}_m(\Sigma)$.

PROPOSITION 4.5. *The mapping*

$$\begin{array}{rcl} \mathcal{S}(\epsilon) & \longrightarrow & \mathcal{C}, \\ \Sigma & \longmapsto & \mathcal{R}_m(\Sigma), \end{array}$$

is continuous for all $1 \leq m \leq n$. *Here* $\mathcal{C}$ *denotes the set of compact subspaces of* $\mathcal{G}(m,n)$ *endowed with the Haussdorff metric and* $\mathcal{S}(\epsilon) \subseteq \mathcal{S}$ *is the set of spectral densities, which have no eigenvalue equal to* ϵ.

Proof. This follows immediately from Lemma A.8. □

For the low noise case, a continuity result similar to Proposition 3.11 holds. However, in this case ϵ has to go to zero at a suitable rate.

5. The bivariate case. As an important special case and as a further explanation of some of the preceding results, we now discuss the case $n = 2$. This is done for the Frisch case and for the bounded noise case.

First we consider the Frisch case. In this case $\mathrm{mc}(\Sigma) = 2$ holds iff Σ is diagonal. This case is not really interesting since we may then choose $\hat{\Sigma} = 0$ and $\tilde{\Sigma} = \Sigma$ and thus every $w \in \mathbb{C}^{1\times 2}$ is compatible.

For $\mathrm{mc}(\Sigma) = 1$, by Proposition 3.2, w.l.o.g., we may choose a normalization $w = (1, -k)$ for the relation functions. Then k corresponds to the transfer function of the second component of $\hat{x}$ to the first component. Clearly $\hat{\Sigma}$ is compatible iff

$$\begin{array}{c} \hat{\sigma}_{11}\hat{\sigma}_{22} - |\sigma_{12}|^2 = 0, \\ 0 \leq \hat{\sigma}_{11} \leq \sigma_{11}, \\ 0 \leq \hat{\sigma}_{22} \leq \sigma_{22}, \end{array}$$

and thus k is compatible iff

$$k = \frac{\sigma_{12}}{\hat{\sigma}_{22}} \text{ where } \hat{\sigma}_{22} \in \left[\frac{|\sigma_{21}|^2}{\sigma_{11}}, \sigma_{22} \right].$$

(See Anderson and Deistler [3].) In particular here, the phase is uniquely determined and the gain is in an interval whose boundaries correspond to the Wiener filter formula, where all noise is added either to the second or first component.

Next we consider the bounded noise case. Again we introduce the normalization $w = (1, -k)$. This normalization excludes relation functions of the form $w = (0, k)$ which by Proposition 4.2 are compatible iff $s_{22} \geq 0$ holds. Using the normalization above, w is a compatible relation iff

$$wSw^* = ks_{22}k^* - ks_{21} - s_{12}k^* + s_{11} \geq 0.$$

For $s_{22} \neq 0$ this is equivalent to

$$\left(k - \frac{s_{12}}{s_{22}}\right) s_{22} \left(k - \frac{s_{12}}{s_{22}}\right)^* + \frac{\det(S)}{s_{22}} \geq 0,$$

TABLE 5.1

	s_{22}	$\det(S)/s_{22}$	compatible k	mc(Σ)
$0 \le \epsilon < \lambda_1$	< 0	< 0	none	0
$\lambda_1 \le \epsilon < \epsilon_0$	< 0	≥ 0	$\lvert k - s_{12}/s_{22}\rvert \le \sqrt{-\det(S)/s_{22}^2}$	1
$\epsilon = \epsilon_0$	0	∞	$2\Re(ks_{21}) \le s_{11}$	1
$\epsilon_0 < \epsilon < \lambda_2$	> 0	< 0	$\lvert k - s_{12}/s_{22}\rvert \ge \sqrt{-\det(S)/s_{22}^2}$	1
$\lambda_2 \le \epsilon$	> 0	≥ 0	all	2

and for $s_{22} = 0$ this is equivalent to

$$2\Re(ks_{21}) \le s_{11}.$$

In the next step we analyze the variation of the 1-relation set for a fixed spectral density Σ but with a varying noise bound ϵ. To simplify the analysis we assume that Σ is not diagonal and that $0 < \lambda_1 < \lambda_2$ holds for the eigenvalues of Σ. Clearly $\det(S) \le 0$ holds for $\lambda_1 \le \epsilon \le \lambda_2$ and $\det(S) > 0$ otherwise. Let ϵ_0 denote the value of ϵ for which s_{22} becomes zero. Since $s_{22} = e_2 S e_2^*$, where $e_2 = (0, 1)$, we see that $\lambda_1 \le \epsilon_0 \le \lambda_2$ must hold. Since Σ is not diagonal and thus e_2 is not an eigenvector of S, we have $\lambda_1 < \epsilon_0 < \lambda_2$. Putting this together gives cases shown in Table 5.1.

The three cases corresponding to mc = 1 are shown in Figure 5.1 for

$$\Sigma = \begin{pmatrix} 0.5 & 0.2 + 0.5i \\ 0.2 - 0.5i & 1.5 \end{pmatrix}. \tag{5.1}$$

Note that the set of compatible transfer functions k is unbounded for $\epsilon_0 \le \epsilon < \lambda_2$. This is a consequence of our normalization $w_1 = 1$. In these cases the relation function $w = (0, k)$ is compatible!

In Figure 5.2, we show in an exemplary way how our results, which have been obtained for a fixed frequency, can be applied for the case of varying frequencies. By putting together sets of compatible systems frequency by frequency, we obtain a "tube" containing all compatible transfer functions. To be more precise, the set of all compatible transfer functions is the set of all transfer functions contained in this tube. Note, however, that the tube may contain functions which are not transfer functions, e.g., if k is not continuous (compare (1.2)). The figures are given for the spectrum

$$A = \begin{pmatrix} 0.25 & 1 \\ 0.2 & 0.55 \end{pmatrix} \quad \Sigma(\lambda) = \left(I - Ae^{-i\lambda}\right)^{-1} \left(I - A'e^{i\lambda}\right)^{-1}. \tag{5.2}$$

Rationality of transfer functions, bounding of order, and causality impose additional restrictions on the set of compatible transfer functions; see, e.g., [9] for the Frisch case. An extreme example can be easily seen from Figure 5.2, where both equivalence classes do not contain a static system (i.e., a system of order 0).

6. Conclusion. In identification of errors-in-variables models—where also inputs may be contaminated by noise—often there is not sufficient a priori knowledge about the noise available in order to obtain unique models. Imposing additional assumptions, which are not justified by a priori knowledge, in order to get uniqueness may give prejudiced results. Therefore our basic philosophy is to attach to the data a set of observationally equivalent models, rather than a single model. Clearly such an approach faces additional complications. These complications, treated in an idealized

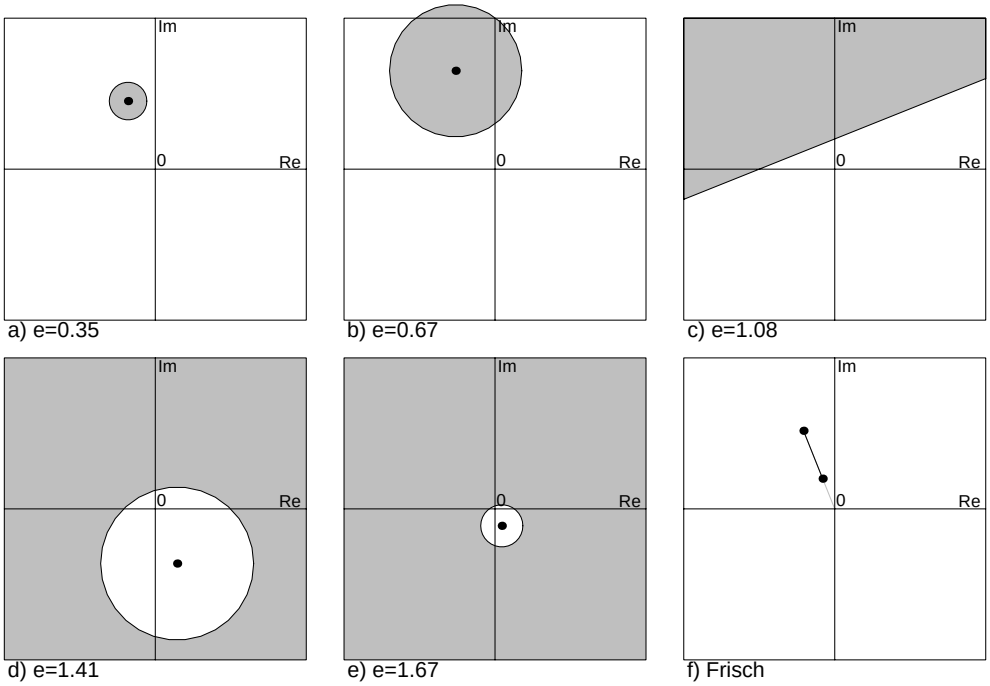


FIG. 5.1. *Here, for* Σ *given in* (5.1), *sets of compatible transfer functions* k, *evaluated at a fixed frequency, are represented in the complex plane marked by dark areas. Figures* (a)–(e) *show the bounded noise case for a growing noise bound* ϵ (*see also Table* 5.1). *Figure* (f) *shows the Frisch case. Note that the sets have dimension* 2 *in the bounded noise case and dimension* 1 *in the Frisch case.*

setting, where the relation between second moments of the observations (rather than data) and models is analyzed, are studied here.

In particular the following problems are addressed:

1. The description of the classes of observationally equivalent models. In many cases a simple analytic description is not available; for this reason we focus on topological and geometric properties of these classes.

2. For the well-posedness of the identification problem, the continuity of the mapping attaching equivalence classes to second moments of the observations is important.

3. The class of observationally equivalent models may contain systems with a different number of outputs. Here the maximum number of outputs is of special interest. For inference for this number (e.g., by likelihood ratio testing) some topological properties of sets of spectral densities corresponding to a given maximum number of outputs are analyzed.

These problems are investigated for two different assumptions on the noise, namely, for the Frisch case (mutually uncorrelated noise components) and for the case of bounded noise.

Appendix A. Some lemmata.

Lemmas A.1–A.7 deal with the Frisch case and Lemma A.8 deals with the bounded noise case.

LEMMA A.1. *Let* $\hat{\Sigma}^0 \in \mathcal{M}_m$, $\tilde{\Sigma}^0 \in \mathcal{D}_{\mathcal{I}}$, $\Sigma^0 = \hat{\Sigma}^0 + \tilde{\Sigma}^0 > 0$, *where* $\mathcal{I} = \{1, \ldots, l\}$

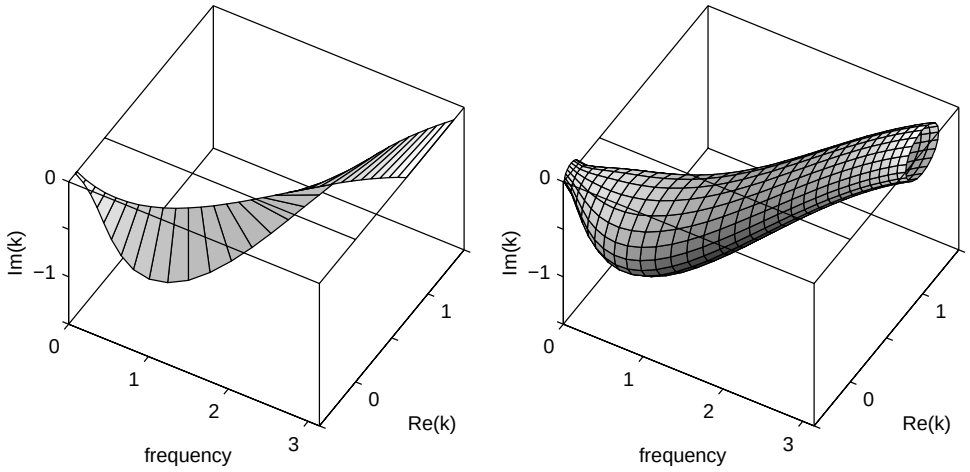


FIG. 5.2. *Here, for the spectrum* Σ *given in* (5.2), *the set of compatible transfer functions* k *is shown. The left-hand figure shows the Frisch case and the right-hand figure shows the bounded noise case. The above figures are obtained by putting together the sets shown in Figure* 5.1, *frequency by frequency. (Of course, here only one noise bound* ϵ *was used.) Note that the Frisch case gives a "band," whereas the bounded noise case gives a "tube" of compatible* k*'s.*

and $\hat{\Sigma}^0_{22} > 0$ *holds. Then the following statements hold:*

1. *We can find an open neighborhood* $\mathcal{U} \subseteq \mathbb{R}^n$ *of* $\tilde{\Sigma}^0$ *such that* $(\Sigma^0_{22} - \tilde{\Sigma}_{22}) > 0$ *and* $\tilde{\sigma}_{ii} > 0$ *for* $i \in \mathcal{I}$ *hold for all* $\tilde{\Sigma} \in \mathcal{U}$. *In this neighborhood the mapping* $\tilde{\Sigma} \mapsto \mathrm{g}_{n,m}(\Sigma^0 - \tilde{\Sigma})$ *is defined and differentiable, and* $\tilde{\Sigma} \in \mathcal{E}_m(\Sigma^0)$ *holds iff* $\mathrm{g}_{n,m}(\Sigma^0 - \tilde{\Sigma}) = 0$ *and* $\tilde{\Sigma} \geq 0$ *holds.*

2. *The mapping* $\mathrm{f}_{m,\mathcal{I}}$ *has full rank in* $(\hat{\Sigma}^0, \tilde{\Sigma}^0)$ *iff the derivative of* $\mathrm{g}_{n,m}(\Sigma^0 - \tilde{\Sigma})$, *with respect to* $d_{\mathcal{I}}(\tilde{\Sigma})$, *has full rank in* $\tilde{\Sigma}^0$. *Here* $d_{\mathcal{I}}(\tilde{\Sigma})$ *denotes the vector of the diagonal elements of* $\tilde{\Sigma}$ *corresponding to the index set* $\mathcal{I}$.

Proof. The first statement is evident. In order to prove the second statement note that the derivative of $\mathrm{f}_{m,\mathcal{I}}$ is given by

$$\begin{pmatrix} I & 0 & 0 & 0 & 0 & 0 \\ 0 & I & 0 & 0 & 0 & 0 \\ 0 & 0 & I & 0 & 0 & 0 \\ X & X & H_2 & H_1 & 0 & 0 \\ 0 & 0 & 0 & I & I & 0 \\ X & X & X & X & 0 & I \end{pmatrix} \begin{matrix} \} \partial v(\Sigma_{12}) \\ \} \partial o(\Sigma_{22}) \\ \} \partial d_2(\Sigma_{22}) \\ \} \partial o(\Sigma_{11}) \\ \} \partial d_1(\Sigma_{22}) \\ \} \partial d(\Sigma_{11}) \end{matrix} .$$

$$\underbrace{}_{\partial v(\hat{\Sigma}_{12})} \; \underbrace{}_{\partial o(\hat{\Sigma}_{22})} \; \underbrace{}_{\partial d_2(\hat{\Sigma}_{22})} \; \underbrace{}_{\partial d_1(\hat{\Sigma}_{22})} \; \underbrace{}_{\partial d_1(\tilde{\Sigma}_{22})} \; \underbrace{}_{\partial d(\tilde{\Sigma}_{11})}$$

Here block entries of this derivative, which do not play a role in our analysis, are simply marked with an X. Furthermore, the vector of diagonal elements of $\hat{\Sigma}_{22}$ and $\tilde{\Sigma}_{22}$ is split into two parts $d_1(\cdot)$ and $d_2(\cdot)$, such that $d_{\mathcal{I}}(\tilde{\Sigma}) = (d(\tilde{\Sigma}_{11})', d_1(\tilde{\Sigma}_{22})')'$.

On the other hand, the derivative of $\mathrm{g}_{n,m}(\Sigma^0 - \tilde{\Sigma})$ with respect to $d(\tilde{\Sigma})$ is given by

$$\begin{pmatrix} -I & X & X \\ 0 & H_1 & H_2 \end{pmatrix} \begin{matrix} \} \partial d(\mathrm{g}_{n,m}) \\ \} \partial o(\mathrm{g}_{n,m}) \end{matrix} .$$

$$\underbrace{}_{\partial d(\tilde{\Sigma}_{11})} \; \underbrace{}_{\partial d_1(\tilde{\Sigma}_{22})} \; \underbrace{}_{\partial d_2(\tilde{\Sigma}_{22})}$$

Since both matrices have full rank iff H_1 has full rank, statement 2 has been proven. ∎

LEMMA A.2. *Let $\Sigma^0 = \hat{\Sigma}^0 + \tilde{\Sigma}^0$, $(\hat{\Sigma}^0, \tilde{\Sigma}^0) \in \mathcal{M}_m \times \mathcal{D}_\mathcal{I}$ be a regular point of $\mathrm{f}_{m,\mathcal{I}}$. There exists an open neighborhood $\mathcal{U} \subseteq (\mathbb{R}^{n^2} \times \mathbb{R}^n)$, of $(\Sigma^0, \tilde{\Sigma}^0)$, such that for all $(\Sigma, \tilde{\Sigma}) \in \mathcal{U}$ we have that: if $\hat{\Sigma} = (\Sigma - \tilde{\Sigma}) \in \mathcal{M}_s$ and $\tilde{\Sigma} \geq 0$ hold, then*

1. *$s \leq m$,*
2. *$\tilde{\Sigma} \in \mathcal{D}_\mathcal{J}$, $\mathcal{J} \supseteq \mathcal{I}$, and*
3. *$(\hat{\Sigma}, \tilde{\Sigma})$ is a regular point of $\mathrm{f}_{s,\mathcal{J}}$.*

Proof. By a rearrangement of variables in x_t we may achieve that $\hat{\Sigma}^0_{22} > 0$ and $\mathcal{I} = \{1, \ldots, l\}$ hold. We then can choose $\mathcal{U}$ in such a way that $(\Sigma_{22} - \tilde{\Sigma}_{22}) > 0$ and $\tilde{\sigma}_{ii} > 0$ for all $i \in \mathcal{I}$ hold for all $(\Sigma, \tilde{\Sigma}) \in \mathcal{U}$. Thus we have shown 1 and 2.

By Lemma A.1 we know that the derivative of $\mathrm{g}_{n,m}(\Sigma - \tilde{\Sigma})$ with respect to $d_\mathcal{I}(\tilde{\Sigma})$ evaluated at $(\Sigma^0, \tilde{\Sigma}^0)$ has full rank m^2, since $(\hat{\Sigma}^0, \tilde{\Sigma}^0)$ is a regular point of $\mathrm{f}_{m,\mathcal{I}}$. Now let $\mathcal{U}$ be so small that this derivative has full rank m^2 for all points $(\Sigma, \tilde{\Sigma}) \in \mathcal{U}$. We consider a pair $(\Sigma, \tilde{\Sigma}) \in \mathcal{U}$, $\hat{\Sigma} = (\Sigma - \tilde{\Sigma}) \in \mathcal{M}_s$, $\tilde{\Sigma} \in \mathcal{D}_\mathcal{J}$. Again by rearrangement of the first m variables in x_t, we can achieve that the lower right $(n-s) \times (n-s)$ corner of $\hat{\Sigma}$ has full rank $n - s$. By the identity $\mathrm{g}_{n,s} = \mathrm{g}_{m,s} \circ \mathrm{g}_{n,m}$ and the chain rule, the derivative of $\mathrm{g}_{n,s}(\Sigma - \hat{\Sigma})$ with respect to $d_\mathcal{I}(\tilde{\Sigma})$ has full rank s^2, since the derivative of $\mathrm{g}_{m,s}$ has rank s^2 and since the derivative of $\mathrm{g}_{n,m}$ with respect to $d_\mathcal{I}(\tilde{\Sigma})$ has rank m^2. Therefore we again have by Lemma A.1 that $(\hat{\Sigma}, \tilde{\Sigma})$ is a regular point of $\mathrm{f}_{s,\mathcal{I}}$. ∎

LEMMA A.3. *The set of all pairs $(\hat{\Sigma}, \tilde{\Sigma}) \in (\mathcal{M}_m \times \mathcal{D}_\mathcal{I})$, where $\mathrm{f}_{m,\mathcal{I}}$ has full rank $\min(n^2, n^2 - m^2 + |\mathcal{I}|)$, is open and dense in $(\mathcal{M}_m \times \mathcal{D}_\mathcal{I})$.*

Proof. We first prove that the set of pairs $(\hat{\Sigma}, \tilde{\Sigma})$ with a full rank derivative is dense. Let us consider a pair $(\hat{\Sigma}^0, \tilde{\Sigma}^0)$, where, w.l.o.g., we may assume that $\hat{\Sigma}^0_{22} > 0$ and $\mathcal{I} = \{1, \ldots, l\}$ holds. By the results of Lemma A.1, it suffices to consider the derivative of $o(\hat{\Sigma}_{12}\hat{\Sigma}_{22}\hat{\Sigma}_{21})$ with respect to the first $s = l - m$ diagonal elements of $\hat{\Sigma}_{22}$. Let $A = \hat{\Sigma}_{12}\hat{\Sigma}_{22}^{-1}$; then this derivative is given by

$$\left(\begin{array}{ccccc}
\Re(a_{11}a_{21}^*) & \Re(a_{12}a_{22}^*) & \ldots & \ldots & \Re(a_{1s}a_{2s}^*) \\
\Im(a_{11}a_{21}^*) & \Im(a_{12}a_{22}^*) & \ldots & \ldots & \Im(a_{1s}a_{2s}^*) \\
\Re(a_{11}a_{31}^*) & \Re(a_{12}a_{32}^*) & \ldots & \ldots & \Re(a_{1s}a_{3s}^*) \\
\Im(a_{11}a_{31}^*) & \Im(a_{12}a_{32}^*) & \ldots & \ldots & \Im(a_{1s}a_{3s}^*) \\
\vdots & \vdots & & & \vdots \\
\Re(a_{11}a_{m1}^*) & \Re(a_{12}a_{m2}^*) & \ldots & \ldots & \Re(a_{1s}a_{ms}^*) \\
\Im(a_{11}a_{m1}^*) & \Im(a_{12}a_{m2}^*) & \ldots & \ldots & \Im(a_{1s}a_{ms}^*) \\
\Re(a_{21}a_{31}^*) & \Re(a_{22}a_{32}^*) & \ldots & \ldots & \Re(a_{2s}a_{3s}^*) \\
\Im(a_{21}a_{31}^*) & \Im(a_{22}a_{32}^*) & \ldots & \ldots & \Im(a_{2s}a_{3s}^*) \\
\vdots & \vdots & & & \vdots \\
\vdots & \vdots & & & \vdots \\
\Re(a_{m-1,1}a_{m1}^*) & \Re(a_{m-1,2}a_{m,2}^*) & \ldots & \ldots & \Re(a_{m-1,s}a_{m,s}^*) \\
\Im(a_{m-1,1}a_{m1}^*) & \Im(a_{m-1,2}a_{m,2}^*) & \ldots & \ldots & \Im(a_{m-1,s}a_{m,s}^*)
\end{array}\right)
\begin{array}{l}
\}\partial\Re((\cdot)_{12}) \\
\}\partial\Im((\cdot)_{12}) \\
\}\partial\Re((\cdot)_{13}) \\
\}\partial\Im((\cdot)_{13}) \\
\vdots \\
\}\partial\Re((\cdot)_{1s}) \\
\}\partial\Im((\cdot)_{1s}) \\
\}\partial\Re((\cdot)_{23}) \\
\}\partial\Im((\cdot)_{23}) \\
\vdots \\
\vdots \\
\}\partial\Re((\cdot)_{m-1,m}) \\
\}\partial\Im((\cdot)_{m-1,m})
\end{array}$$

In each neighborhood of $A^0 = \hat{\Sigma}^0_{12}(\hat{\Sigma}^0_{22})^{-1}$, we may find a matrix A such that all entries a_{ij} are nonzero and $\Re(a_{11}a_{21}^*) \neq 0$ holds. Now let us consider the left upper $(k+1) \times (k+1)$ block of the above matrix, which is partitioned as

$$\left(\begin{array}{cc} H_{11} & h_{12} \\ h_{21} & H_{22} \end{array}\right).$$

If H_{11} is regular, then the determinant of this block is equal to $\det(H_{11})(h_{22} - h_{21}H_{11}^{-1}h_{21})$. The element h_{22} is the real or imaginary part of a product of the form $a_{i,k+1}a_{j,k+1}^*$ for some $i \neq j$. It is easy to see that we can disturb these two entries $a_{i,k+1}$ and $a_{j,k+1}$ such that $(h_{22} - h_{21}H_{11}^{-1}h_{21})$ is nonzero. Then after a finite number of such steps, we end up with a full rank matrix H. Since $A = \hat{\Sigma}_{12}\hat{\Sigma}_{22}^{-1}$, these arbitrarily small changes in A may be achieved by arbitrarily small changes in $\hat{\Sigma}_{12}$. Thus we have proved that, in any neighborhood of $(\hat{\Sigma}^0, \tilde{\Sigma}^0)$, we may find a pair $(\hat{\Sigma}, \tilde{\Sigma})$ such that the derivative $\mathrm{f}_{m,\mathcal{I}}$ has full rank.

Since the determinant of the derivative is a continuous functions of the entries of $(\hat{\Sigma}, \tilde{\Sigma})$, it follows that the set of pairs $(\hat{\Sigma}, \tilde{\Sigma})$ with a full rank derivative is open. □

LEMMA A.4. *Let $\mathcal{A}$ be a metric space endowed with the metric $\mathrm{d}(x,y)$ and let $(\mathcal{U}_k \in \mathcal{C}(\mathcal{A}))$ be a sequence of compact subsets of $\mathcal{A}$ and $\mathcal{U}_0 \in \mathcal{C}(\mathcal{A})$. Then*

$$\rho(\mathcal{U}_0, \mathcal{U}_k) = \sup_{x \in \mathcal{U}_0} \inf_{y \in \mathcal{U}_k} \mathrm{d}(x,y) \longrightarrow 0$$

iff

$$i_k(x) = \inf_{y \in \mathcal{U}_k} \mathrm{d}(x,y) \longrightarrow 0 \quad \textit{for all } x \in \mathcal{U}_0,$$

i.e., iff for all $x \in \mathcal{U}_0$ there exists a sequence $y_k \in \mathcal{U}_k$, such that $y_k \to x$ holds.

Proof. Let us consider two points $x_1, x_2 \in \mathcal{U}_0$. Since $\mathcal{U}_k$ is compact, there are two points $y_1, y_2 \in \mathcal{U}_k$ such that $i_k(x_j) = \inf_{y \in \mathcal{U}_k} \mathrm{d}(x_j, y) = \mathrm{d}(x_j, y_j)$, $j = 1, 2$. From the triangle inequality, we get

$$\begin{aligned} i_k(x_1) &= \mathrm{d}(x_1, y_1) \leq \mathrm{d}(x_1, y_2) \leq \mathrm{d}(x_1, x_2) + \mathrm{d}(x_2, y_2) = \mathrm{d}(x_1, x_2) + i_k(x_2), \\ i_k(x_2) &= \mathrm{d}(x_2, y_2) \leq \mathrm{d}(x_2, y_1) \leq \mathrm{d}(x_2, x_1) + \mathrm{d}(x_1, y_1) = \mathrm{d}(x_1, x_2) + i_k(x_1), \end{aligned}$$

and thus

$$|i_k(x_1) - i_k(x_2)| \leq \mathrm{d}(x_1, x_2).$$

Now suppose that $i_k(x) \to 0$ holds for all $x \in \mathcal{U}_0$, but $\sup_{x \in \mathcal{U}_0} i_k(x)$ does not converge to zero. Then there exist an $\epsilon > 0$ and a sequence of points $x_k \in \mathcal{U}_0$, such that $i_k(x_k) \geq \epsilon$ for infinitely many k's. Since $\mathcal{U}_0$ is compact, there exists a limiting point $x_0 \in \mathcal{U}_0$ of these sequence of points x_k. By the above inequality, we further get

$$i_k(x_k) \leq \underbrace{|i_k(x_k) - i_k(x_0)|}_{\leq \mathrm{d}(x_k, x_0) \to 0} + \underbrace{i_k(x_0)}_{\to 0} \longrightarrow 0,$$

in contradiction to $i_k(x_k) \geq \epsilon$. Thus we have shown that $i_k(x) \to 0$, $\forall x \in \mathcal{U}_0$ implies that $\rho(\mathcal{U}_0, \mathcal{U}_k) \to 0$. The reverse statement is evident. Clearly $i_k(x) \to 0$ is equivalent to the existence of a sequence $y_k \in \mathcal{U}_k$, with $y_k \to x$. □

Now we consider the mapping $\Sigma \mapsto \overline{\mathcal{E}_m(\Sigma)}$. In the next proposition the semicontinuity of this mapping, in the sense that the limiting set of a sequence $\overline{\mathcal{E}_m(\Sigma^k)}$ is a subset of $\overline{\mathcal{E}_m(\lim \Sigma^k)}$, is stated.

LEMMA A.5. *For $\mathcal{S} \ni \Sigma^k \to \Sigma^0 \in \mathcal{S}$, we have*

$$\rho(\overline{\mathcal{E}_m(\Sigma^k)}, \overline{\mathcal{E}_m(\Sigma^0)}) = \sup_{\tilde{\Sigma}^k \in \overline{\mathcal{E}_m(\Sigma^k)}} \inf_{\tilde{\Sigma}^0 \in \overline{\mathcal{E}_m(\Sigma^0)}} \|\tilde{\Sigma}^k - \tilde{\Sigma}^0\| \to 0.$$

Proof. Note that $\overline{\mathcal{E}_m(\Sigma)}$ is closed and bounded by Propositions 3.1 and 3.3. Thus $\rho(\overline{\mathcal{E}_m(\Sigma^k)}, \overline{\mathcal{E}_m(\Sigma^0)})$ is well defined.

If $\mathrm{mc}(\Sigma^0) < m$ (and thus $\overline{\mathcal{E}_m(\Sigma^0)} = \emptyset$), then also $\mathrm{mc}(\Sigma^k) < m$ (and thus $\overline{\mathcal{E}_m(\Sigma^k)} = \emptyset$) must hold from some k onwards. (This follows from the fact that $\mathcal{S}_1 \cup \cdots \cup \mathcal{S}_{m-1}$ is open; see Proposition 3.12.) If we define $\rho(\emptyset, \emptyset) = 0$, then the proposition is proved in this case.

Suppose that there is an $\epsilon > 0$ such that $\rho(\overline{\mathcal{E}_m(\Sigma^k)}, \overline{\mathcal{E}_m(\Sigma^0)}) > \epsilon$ holds for infinitely many k's. Then for each such k there exists a $\tilde{\Sigma}^k \in \overline{\mathcal{E}_m(\Sigma^k)}$ such that $\inf_{\tilde{\Sigma} \in \overline{\mathcal{E}_m(\Sigma^0)}} \|\tilde{\Sigma}^k - \tilde{\Sigma}\| > \epsilon$ holds. Since the $\tilde{\Sigma}^k$'s are bounded, there is a limiting point $\tilde{\Sigma}^0$, say. Note that $\tilde{\Sigma}^0 \in \overline{\mathcal{E}_m(\Sigma^0)}$ holds, since $(\Sigma^k - \tilde{\Sigma}^k) \geq 0$, $\mathrm{corank}(\Sigma^k - \tilde{\Sigma}^k) \geq m$, and $\tilde{\Sigma}^k \geq 0$ hold.

Therefore we have

$$\inf_{\tilde{\Sigma} \in \overline{\mathcal{E}_m(\Sigma^0)}} \|\tilde{\Sigma}^k - \tilde{\Sigma}\| \leq \|\tilde{\Sigma}^k - \tilde{\Sigma}^0\| \to 0,$$

in contradiction to the construction of the $\tilde{\Sigma}^k$'s. ∎

LEMMA A.6. *For a convergent sequence* $(\Sigma^k \in \mathcal{S})$, $\Sigma^k \to \Sigma^0 \in \mathcal{S}$, *we have*

$$\rho(\mathcal{R}_m(\Sigma^k), \mathcal{R}_m(\Sigma^0)) \to 0.$$

Proof. In the case $\mathrm{mc}(\Sigma^0) < m$, we have $\mathrm{mc}(\Sigma^k) < m$ from some k_0 onwards. Thus $\mathcal{R}_m(\Sigma^0) = \emptyset = \mathcal{R}_m(\Sigma^k)$ for all $k \geq k_0$.

Now we consider the case where $\mathrm{mc}(\Sigma^0) \geq m$ holds. We suppose that there exist an $\epsilon > 0$ and a sequence $\mathfrak{x}^k \in \mathcal{R}_m(\Sigma^k)$, such that $\inf_{\mathfrak{y} \in \mathcal{R}_m(\Sigma^0)} \mathrm{d}_{\mathcal{G}}(\mathfrak{x}^k, \mathfrak{y}) \geq \epsilon$ holds for infinitely many k's. Since $\mathcal{G}(m,n)$ is compact, there exists a limiting point $\mathfrak{x}^0 \in \mathcal{G}(m,n)$ of these $\mathfrak{x}^k$'s.

To each $\mathfrak{x}^k$ there exists a corresponding $\tilde{\Sigma}^k \in \mathcal{E}(\Sigma^k)$, such that $\mathfrak{x}^k \subseteq \ker(\Sigma^k - \tilde{\Sigma}^k)$. Since the $\tilde{\Sigma}^k$'s are bounded, there exist a limiting point $\tilde{\Sigma}^0$, say. It is easy to see that $\tilde{\Sigma}^0 \in \mathcal{E}(\Sigma^0)$ and $\mathfrak{x}^0 \subseteq \ker(\Sigma^0 - \tilde{\Sigma}^0)$ hold. Thus $\mathfrak{x}^0 \in \mathcal{R}_m(\Sigma^0)$ holds and we have

$$\inf_{\mathfrak{y} \in \mathcal{R}_m(\Sigma^0)} \mathrm{d}_{\mathcal{G}}(\mathfrak{x}^k, \mathfrak{y}) \leq \mathrm{d}_{\mathcal{G}}(\mathfrak{x}^k, \mathfrak{x}^0) \to 0,$$

in contradiction to the construction of the sequence $\mathfrak{x}^k$. ∎

LEMMA A.7. *Consider a sequence* $\Sigma^k \in \mathcal{S}$, $\Sigma^k \to \Sigma^0 \in \mathcal{S}$, *where* $\mathrm{mc}(\Sigma^0) = m_0$, *and* $\overline{\mathcal{E}_s(\Sigma^k)} \to \overline{\mathcal{E}_s(\Sigma^0)}$ *for all* $m \leq s \leq m_0$ *holds. In this case,*

$$\mathrm{d_H}(\mathcal{R}_m(\Sigma^k), \mathcal{R}_m(\Sigma^0)) \to 0$$

holds.

Proof. For $\mathfrak{x}^0 \in \mathcal{R}_m(\Sigma^0)$ there is a $\tilde{\Sigma}^0 \in \overline{\mathcal{E}_m(\Sigma^0)}$ such that $\mathfrak{x}^0 \subseteq \ker(\Sigma^0 - \tilde{\Sigma}^0)$. In order to prove the proposition we consider the following two cases:

1. $\mathfrak{x}^0 = \ker(\Sigma^0 - \tilde{\Sigma}^0)$, i.e. $\tilde{\Sigma}^0 \in \mathcal{E}_m(\Sigma^0)$. Since $\overline{\mathcal{E}_m(\Sigma^k)} \to \overline{\mathcal{E}_m(\Sigma^0)}$, we can find a sequence $\tilde{\Sigma}^k \in \overline{\mathcal{E}_m(\Sigma^k)}$, such that $\tilde{\Sigma}^k \to \tilde{\Sigma}^0$ holds. There must be an index k_0 such that $\mathrm{corank}(\Sigma^k - \tilde{\Sigma}^k) = m$, and thus $\tilde{\Sigma}^k \in \mathcal{E}_m(\Sigma^k)$, holds for all $k \geq k_0$. Thus $\mathfrak{x}^k = \ker(\Sigma^k - \tilde{\Sigma}^k) \in \mathcal{R}_m(\Sigma^k)$ holds for all $k \geq k_0$. Since $\ker(\hat{\Sigma})$ continuously depends on $\hat{\Sigma}$, we have found a sequence $\mathfrak{x}^k \in \mathcal{R}_m(\Sigma^k) \to \mathfrak{x}^0$.

2. $\mathfrak{x}^0 \subset \ker(\Sigma^0 - \tilde{\Sigma}^0)$, i.e., $\tilde{\Sigma}^0 \in \mathcal{E}_s(\Sigma^0)$ for some $s > m$. We can construct a sequence $\mathfrak{y}^k \in \mathcal{R}_s(\Sigma^k) \to \mathfrak{y}^0 = \ker(\Sigma^0 - \tilde{\Sigma}^0)$ as described above. Without loss of generality, we assume that $w^0 = (I, w_2^0)$ is a basis for $\mathfrak{y}^0$. For all k large enough, $\mathfrak{y}^k$

has a basis $w^k = (I, w_2^k)$. Clearly $w^k \to w^0$ holds. Since $\mathfrak{x}^0 \subset \ker(\Sigma^0 - \tilde{\Sigma}^0) = \mathfrak{y}^0$, there exists a matrix $o \in \mathbb{C}^{m\times s}$ such that $ow^0 \in \mathbb{C}^{m\times n}$ is a basis for $\mathfrak{x}^0$. Then $ow^k \in \underline{\mathcal{R}}_m(\Sigma^k)$ holds and the corresponding subspace $\mathfrak{x}^k$ (spanned by the rows of ow^k) is an element of $\mathcal{R}_m(\Sigma^k)$. Thus we have found a sequence $\mathfrak{x}^k \in \mathcal{R}_m(\Sigma^k) \to \mathfrak{x}^0 \in \mathcal{R}_m(\Sigma^0)$.

Now, using Lemmata A.4 and A.6 gives us the desired result. ∎

LEMMA A.8. *Let $\Sigma^k \in \mathcal{S}$ be a convergent sequence of spectral densities with $\lim_{n\to\infty} \Sigma^k = \Sigma^0 \in \mathcal{S}$.*

1. $\rho(\mathcal{R}_m(\Sigma^k), \mathcal{R}_m(\Sigma^0)) = \sup_{\mathfrak{x}\in\mathcal{R}_m(\Sigma^k)} \inf_{\mathfrak{y}\in\mathcal{R}_m(\Sigma^0)} \mathrm{d}_{\mathcal{G}}(\mathfrak{x}, \mathfrak{y}) \longrightarrow 0$.
2. *If $\lambda_m(\Sigma^0) < \epsilon$, then* $\rho(\mathcal{R}_m(\Sigma^0), \mathcal{R}_m(\Sigma^k)) = \sup_{\mathfrak{x}\in\mathcal{R}_m(\Sigma^0)} \inf_{\mathfrak{y}\in\mathcal{R}_m(\Sigma^k)} \mathrm{d}_{\mathcal{G}}(\mathfrak{x}, \mathfrak{y}) \longrightarrow 0$

Proof. 1. First we note that, if $\mathrm{mc}(\Sigma^0) < m$ holds (and thus $\mathcal{R}_m(\Sigma^0)$ is void), then $\mathrm{mc}(\Sigma^k) < m$ (and thus $\mathcal{R}_m(\Sigma^k) = \emptyset$) must hold for all k large enough. If we define $\mathrm{d}_{\mathrm{H}}(\emptyset, \emptyset) = 0$, then the statement is true for this case.

Next we consider the case $\mathrm{mc}(\Sigma^0) \geq m$. If $\rho(\mathcal{R}_m(\Sigma^k), \mathcal{R}_m(\Sigma^0))$ does not converge to zero, then there exist a $\mu > 0$ and a sequence $\mathfrak{x}^k \in \mathcal{R}_m(\Sigma^k)$ such that $\inf_{\mathfrak{y}\in\mathcal{R}_m(\Sigma^0)} \mathrm{d}_{\mathcal{G}}(\mathfrak{x}^k, \mathfrak{y}) \geq \mu$ holds. Since $\mathcal{G}(m, n)$ is compact, the sequence $\mathfrak{x}^k$ has a convergent subsequence. To simplify the notation we use the same index k for this subsequence. We now have $\lim_k \mathfrak{x}^k = \mathfrak{x}^0$. If (after possibly reordering variables) $w^0 = (I, w_2^0)$ is a basis for the subspace $\mathfrak{x}^0$, then we can find $w^k = (I, w_2^k)$ which span the subspaces $\mathfrak{x}^k$. Since $\mathfrak{x}^k \to \mathfrak{x}^0$, we have $w^k \to w^0$ and $w^0(\epsilon\Sigma^0 - \Sigma^0\Sigma^0)(w^0)^* = \lim_k w^k(\epsilon\Sigma^k - \Sigma^k\Sigma^k)(w^k)^* \geq 0$. Thus $\mathfrak{x}^0 \in \mathcal{R}_m(\Sigma^0)$ and

$$\inf_{\mathfrak{y}\in\mathcal{R}_m(\Sigma^0)} \mathrm{d}_{\mathcal{G}}(\mathfrak{x}^k, \mathfrak{y}) \leq \mathrm{d}_{\mathcal{G}}(\mathfrak{x}^k, \mathfrak{x}^0) \to 0$$

holds, which gives the desired contradiction.

2. By Lemma A.4 it suffices to show that for each $w^0 \in \underline{\mathcal{R}}_m(\Sigma^0)$ a sequence $w^k \in \underline{\mathcal{R}}_m(\Sigma^k)$ with $w^k \to w^0$ can be constructed. Let $S^k = \epsilon\Sigma^k - \Sigma^k\Sigma^k$ and $S^0 = \epsilon\Sigma^0 - \Sigma^0\Sigma^0$. Since S^0 has at least m eigenvalues strictly larger than zero, we may find a sequence $v^l \to w^0$ such that $v^l\Sigma^0(v^l)^* > 0$ holds (see item 2 of Proposition 4.3.) Since $\Sigma^k \to \Sigma^0$, for each l there exists an integer $k(l)$ such that $v^l\Sigma^k(v^l)^* \geq 0$ holds for all $k \geq k(l)$. Now we define the sequence w^k by $w^k = v^l$ for $k(l) \leq k < k(l+1)$. ∎

Appendix B. Notation.

$\mathcal{S} \subseteq \mathbb{C}^{n\times n}$	set of all $n \times n$ complex positive definite matrices
$\mathcal{M} \subseteq \mathbb{C}^{n\times n}$	set of all $n \times n$ complex positive semidefinite matrices
$\mathcal{M}_m \subseteq \mathcal{M}$	set of all $n \times n$ complex positive semidefinite matrices of corank m
$\mathcal{S}_m \subseteq \mathcal{S}$	set of all Σ with $\mathrm{mc}(\Sigma) = m$
$\mathcal{G}(m, n)$	set of all m-dimensional subspaces of $\mathbb{C}^n$ (Grassmannian)
x_t	observations
$\hat{x}_t$	latent variables
u_t	noise
$\Sigma \in \mathcal{S}$	spectral density of (x_t) evaluated at a fixed frequency λ
$\Sigma_T \in \mathcal{S}$	an estimate of Σ based on a sample of length T
$\hat{\Sigma} \in \mathcal{M}_m$	spectral density of $(\hat{x}_t)$ evaluated at a fixed frequency λ
$\tilde{\Sigma} \in \mathcal{M}$	spectral density of (u_t) evaluated at a fixed frequency λ
$w \in \mathbb{C}^{m\times n}$	relation function evaluated at a fixed frequency λ

$\mathcal{E}(\Sigma)$	set of all compatible $\tilde{\Sigma}$
$\mathcal{E}_m(\Sigma) \subseteq \mathcal{E}(\Sigma)$	set of all compatible $\tilde{\Sigma}$, such that $\hat{\Sigma} = \Sigma - \tilde{\Sigma}$ has corank m
$\underline{\mathcal{R}}_m(\Sigma) \subseteq \mathbb{C}^{m \times n}$	set of all compatible w, m-relation set of Σ
$\mathcal{R}_m(\Sigma) \subseteq \mathcal{G}(m,n)$	set of all subspaces spanned by a compatible $w \in \underline{\mathcal{R}}_m(\Sigma)$
$\mathrm{mc}(\Sigma)$	maximum corank of all compatible $\hat{\Sigma} = \Sigma - \tilde{\Sigma}$
$\mathrm{m_L} = \sqrt{n}$	Ledermann bound
$\mathcal{D} \subseteq \mathbb{R}^n$	set of all positive semidefinite diagonal $n \times n$ matrices $\tilde{\Sigma}$
$\mathcal{D}_{\mathcal{I}} \subseteq \mathcal{D}$	set of all positive semidefinite diagonal $n \times n$ matrices $\tilde{\Sigma}$, where $\tilde{\sigma}_{ii} > 0$ for all $i \in \mathcal{I} \subseteq \{1, \ldots, n\}$ and $\tilde{\sigma}_{ii} = 0$ else
$\ker(A)$	kernel of a complex matrix A
$\mathrm{rank}(A), \mathrm{corank}(A)$	rank and corank of a complex matrix A
$\lambda_i(A)$	eigenvalues of a complex Hermitian matrix
$\mathcal{A}^o, \overline{\mathcal{A}}$	interior and closure of a set $\mathcal{A}$
$\mathrm{d_H}(\cdot,\cdot)$	Hausdorff distance
$\mathrm{d}_{\mathcal{G}}(\cdot,\cdot)$	gap metric, defined on $\mathcal{G}(m,n)$
$\Re(\mathrm{z}), \Im(\mathrm{z})$	real and imaginary part of a complex number z

REFERENCES

[1] R. J. Adcock, *A problem in least squares*, The Analyst, 5 (1878), pp. 53–54.

[2] D. J. Aigner, C. Hsiao, A. Kapteyn, and T. Wansbeck, *Latent variable models in econometrics*, in Handbook of Econometrics, Z. Griliches and M. D. Intriligator, eds., North–Holland, Amsterdam, 1984.

[3] B. D. O. Anderson and M. Deistler, *Identifiability in dynamic errors-in-variables models*, J. Time Series Anal., 5 (1984), pp. 1–13.

[4] B. D. O. Anderson and M. Deistler, *Identification of dynamic systems from noisy data: The single factor case*, Math. Control, Signals Systems, 6 (1993), pp. 61–65.

[5] T. W. Anderson, *Estimating linear statistical relationships*, Ann. Statist., 12 (1984), pp. 1–45.

[6] T. W. Anderson and H. Rubin, *Statistical inference in factor analysis*, in Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, J. Neyman, ed., vol. 5, University of California Press, Berkeley, Los Angeles, 1956.

[7] M. Deistler, *Linear system identification—a survey*, in From Data to Model, J. Willems, ed., Springer-Verlag, Berlin, 1989, pp. 1–25.

[8] M. Deistler, *Linear dynamic errors-in-variables models*, in Essays in Time Series and Allied Processing, Festschrift in Honor of E. J. Hannan, J. Gani and M. Priestly, eds., Appl. Probab. Trust, 1986, pp. 128–147.

[9] M. Deistler and B. D. O. Anderson, *Linear dynamic errors-in-variables models, some structure theory*, J. Econometrics, 41 (1989), pp. 39–63.

[10] M. Deistler and B. D. O. Anderson, *Identification of linear systems from noisy data: The case $m^* = 1$*, in Festschrift in honor of R.E. Kalman on the Occasion of his 60th birthday, Springer-Verlag, Berlin 1991, pp. 423–436.

[11] M. Deistler and W. Scherrer, *Identification of linear systems from noisy data*, in New Directions in Time-Series Analysis, Part II, IMA Volumes in Mathematics and Its Applications 46, D. Brillinger, P. Caines, J. Geweke, E. Parzen, M. Rosenblatt, and M. S. Taqqu, eds., Springer-Verlag, Berlin, 1993, pp. 21–42.

[12] J. P. Dufour, *Resultats generiques en analyse factorielle*, Universite des Sciences et Techniques du Languedoc, Institut de Mathematiques–Seminaire de Geometrie Differentielle, 1982–1983.

[13] R. Frisch, *Statistical Confluence Analysis by Means of Complete Regression Systems*, Publication No. 5, University of Oslo, Economic Institute, 1934.

[14] C. Gini, *Sull'interpolazione de una retta quando i valori della variabile indipendente sono affetti errori accidentali*, Metron, 1 (1921), pp. 63–82.

[15] G. H. Golub and C. F. van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, London, 1989.

[16] S. Haykin, ed., *Array Signal Processing*, Prentice–Hall, Englewood Cliffs, NJ, 1985.

[17] C. Heij, W. Scherrer, and M. Deistler, *System identification by dynamic factor models*, SIAM J. Control Optim., 35 (1997), pp. 1924–1951.

[18] R. E. Kalman, *System identification from noisy data*, in Dynamical Systems II, A. R. Bednarek

and L. Cesari, eds., Academic Press, New York, 1982, pp. 331–342.
[19] R. E. Kalman, *Identifiability and modeling in econometrics*, in Developments in Statistics 4, P. R. Krishnaiah, ed., Academic Press, New York, 1983.
[20] R. E. Kalman, *A theory for the identification of linear relations*, in A Collection of Papers Dedicated to Jacques-Louis Lions on the Occasion of his Sixtieth Birthday, Frontiers in Pure and Applied Mathematics, R. Dautray, ed., North–Holland, Amsterdam, 1991, pp. 117–132.
[21] S. Klepper and E. E. Leamer, *Consistent sets of estimates for regressions with errors in all variables*, Econometrica, 52 (1984), pp. 163–183.
[22] W. Ledermann, *On the rank of the reduced correlational matrix in multiple factor analysis*, Psychometrica, 2 (1938), pp. 85–93.
[23] J. Milnor, *Topology from a Differential Viewpoint*, The University Press of Virginia, Charlottesville, VA, 1965.
[24] W. Schachermayer and M. Deistler, *The set of observationally equivalent errors-in-variables models*, Systems Control Lett., 34 (1998), pp. 101–104.
[25] W. Scherrer, *Strukturtheorie von linearen dynamischen Fehler–in–den–Variablen Modellen mit diagonalem Fehlerspektrum*, Ph.D. thesis, Technical University Vienna, Austria, Sept. 1991.
[26] C. Spearman, *General intelligence, objectively determined and measured*, Amer. J. Psych., 15 (1904), pp. 201–293.
[27] D. Stemmer, *Testing for the Number of Equations in Linear Dynamic EV-models*, Ph.D. thesis, Technische Universität, Wien, 1995.
[28] J. K. Tugnait, *Stochastic system identification with noisy input using cumulant statics*, IEEE Trans. Automat. Control, 37 (1992), pp. 476–485.
[29] J. C. Willems, *From time series to linear systems*, Automatica, 22 (1986), pp. 561–580.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 6, pp. 2176–2193, November 1998

017

COMPUTATIONAL COMPLEXITY OF LYAPUNOV STABILITY ANALYSIS PROBLEMS FOR A CLASS OF NONLINEAR SYSTEMS*

MARC W. MCCONLEY†, BRENT D. APPLEBY†, MUNTHER A. DAHLEH‡, AND ERIC FERON§

Abstract. Nonlinear control systems can be stabilized by constructing control Lyapunov functions and computing the regions of state space over which such functions decrease along trajectories of the closed-loop system under an appropriate control law. This paper analyzes the computational complexity of these procedures for two classes of control Lyapunov functions. The systems considered are those which are nonlinear in only a few state variables and which may be affected by control constraints and bounded disturbances. This paper extends previous work by the authors, which develops a procedure for stability analysis for these systems whose computational complexity is exponential only in the dimension of the "nonlinear" states and polynomial in the dimension of the remaining states. The main results are illustrated by a numerical example for the case of purely quadratic control Lyapunov functions.

Key words. Lyapunov methods, nonlinear control systems, computational methods, robustness, bounded control

AMS subject classifications. 93B40, 93C10, 93C35, 93D05, 93D09

PII. S0363012997320316

1. Introduction.

Many dynamical systems can be represented by ordinary differential equations in the physical states of the system influenced by other parameters, such as disturbance and control inputs. The focus of state-feedback control theory is to design a *control law* (a function which maps measured states of the system to control inputs) which produces a desired performance for the system. Very different theories apply to this problem depending on whether the state derivatives are linear or nonlinear functions of the states in the differential equations defining the system. Many simple, straightforward techniques for robust optimal control of linear systems have been developed. Extensions of these methods to nonlinear systems are sometimes possible, but the analogous procedures which result from this exercise cannot typically be executed in a computationally tractable way. As a result, control of nonlinear systems has been a topic of intense research for some time.

Progress on the nonlinear control problem is difficult because of the inherent complexity of methods which are general enough to apply to arbitrary nonlinear systems. One method which has recently come into favor is to construct a stabilizing control law based on a known *control Lyapunov function* (CLF) for the system [2, 13, 27, 29]. A function is a CLF if a control law exists to render it a Lyapunov function for the closed-loop system. The computation of a stabilizing control law is straightforward

*Received by the editors April 23, 1997; accepted for publication (in revised form) January 20, 1998; published electronically September 3, 1998. This research was supported in part by Charles Stark Draper Laboratory Internal Research and Development, in part by Air Force Office of Scientific Research grant AFOSR F49620-95-0219, and in part by National Science Foundation grants 9157306-ECS and 9409715-ECS.

http://www.siam.org/journals/sicon/36-6/32031.html

†Charles Stark Draper Laboratory, 555 Technology Square MS 77, Cambridge, MA 02139 (mcconley@draper.com, appleby@draper.com).

‡Department of Electrical Engineering, Massachusetts Institute of Technology, Room 35-402, 77 Massachusetts Avenue, Cambridge, MA 02139 (dahleh@lids.mit.edu).

§Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Room 33-217, 77 Massachusetts Avenue, Cambridge, MA 02139 (feron@mit.edu).

from any of a number of universal formulas [13, 17, 27] based on the CLF and the system dynamics, so the control synthesis problem is reduced to constructing a CLF for the system and computing the region of state space over which a control exists to stabilize the system based on the given CLF. Recently, the authors have developed a computationally efficient procedure for solving a version of the nonlinear control problem (Problem 1 below) based on a given CLF [18, 19]. In its general form, the procedure requires one to construct a CLF and to determine the region of state space over which that CLF guarantees stability of the closed-loop system. In this paper, we analyze the computational complexity of these two problems for two important classes of Lyapunov functions.

We consider the problem of designing a control law which stabilizes a nonlinear system in the sense prescribed by Definition 1.2 below. The formal stability definition is a version of the concept of uniform asymptotic stability used in [16]. That definition is repeated below.

DEFINITION 1.1 (see [16]). *Given a system $\dot{x} = f(x, w)$ with $w(t) \in \mathcal{W} \subseteq R^l$ for all $t \geq 0$ and a compact subset $\Omega \subset R^n$, define $\|x\|_\Omega \doteq \inf\{\|x - y\|, y \in \Omega\}$. Then the system is* robustly uniformly asymptotically stable with respect to Ω, *or RUAS-Ω, if the following conditions hold:*

1. *For every $x(0) \in R^n$ and $w(t) \in \mathcal{W}$, the solution $x(t)$ is defined for all $t \geq 0$.*
2. *Uniform stability: There exists a radially unbounded, continuous, strictly increasing function $\delta(\epsilon)$, with $\delta(0) = 0$, such that, for any $\epsilon > 0$, $\|x(0)\|_\Omega \leq \delta(\epsilon)$, $t \geq 0$, and $w(t) \in \mathcal{W}$, we have $\|x(t)\|_\Omega < \epsilon$.*
3. *Uniform attraction: For any $r, \epsilon > 0$, there exists $T > 0$, such that for every $w(t) \in \mathcal{W}$, $\|x(t)\|_\Omega < \epsilon$ whenever $\|x(0)\|_\Omega < r$ and $t \geq T$.*

DEFINITION 1.2. *Given a system $\dot{x} = f(x, w)$ with $w(t) \in \mathcal{W} \subseteq R^l$ for all $t \geq 0$, a positively invariant set $\mathcal{X} \subseteq R^n$, and a compact subset $\Omega \subset \mathcal{X}$, the system is* robustly uniformly asymptotically stable over $\mathcal{X}$ with respect to Ω, *or RUAS($\mathcal{X}$,Ω), if it is RUAS-Ω whenever $x(0) \in \mathcal{X}$. We call the set $\mathcal{X}$ a* region of stability *for the system.*

In particular, we consider the following control synthesis problem.

PROBLEM 1. *Consider a continuous time, time-invariant, nonlinear system influenced by a control $u(t)$ in a closed subset $\mathcal{U} \subseteq R^m$ and a disturbance $w(t) \in \mathcal{W} \subseteq R^l$. The state vector $x(t) \in R^n$ is partitioned into $x_N \in R^k$ and $x_L \in R^{n-k}$, and the system has the form*

$$\begin{aligned} \begin{bmatrix} \dot{x}_N \\ \dot{x}_L \end{bmatrix} &= \begin{bmatrix} f_N(x_N) \\ f_L(x_N) \end{bmatrix} + \begin{bmatrix} A_N(x_N) \\ A_L(x_N) \end{bmatrix} x_L \\ &\quad + \begin{bmatrix} g_w^N(x_N) \\ g_w^L(x_N) \end{bmatrix} w + \begin{bmatrix} g_u^N(x_N) \\ g_u^L(x_N) \end{bmatrix} u, \end{aligned} \tag{1.1}$$

where all functions of x_N are C^1. Construct sets $\Omega \subset \mathcal{X} \subseteq R^n$ containing an equilibrium point at $x = 0$ and a static state-feedback control law $\mu : \mathcal{X} \to \mathcal{U}$ such that the closed-loop system with $u = \mu(x)$ is RUAS($\mathcal{X}$,Ω).

Note that Problem 1 includes the problem of analyzing robust stability for an autonomous system without control, since this is just the case $\mathcal{U} = \{0\}$.

Obviously, we would like $\mathcal{X}$ to be as large as possible and Ω as small as possible. When $\mathcal{W} = \{0\}$ and the system is locally asymptotically stabilizable, local stabilization theory yields a set $\mathcal{X}$ such that the system is RUAS($\mathcal{X}$,$\{0\}$) [30, 13]. To compute the region of stability generally requires computation times which are exponential in the state dimension n; for the system (1.1), however, the computations required to find $\mathcal{X}$ are tractable.

In many applications, the engineer knows from the physics of the problem that only a few physical quantities affect the system dynamics in a nonlinear way, so that the system can be modeled in the form (1.1). Such systems are also considered in [3], where the *output-feedback* stabilization problem (with output x_N) is solved based on an output control Lyapunov function, assuming that the solution to the state-feedback stabilization problem is already available and the output CLF can be constructed.

We now define some relevant terms pertaining to a system of the general form given below, where $f \in C^1(R^n \to R^n)$ and $g_w(x)$ and $g_u(x)$ are continuous:

$$\dot{x} = f(x) + g_w(x)w + g_u(x)u. \tag{1.2}$$

DEFINITION 1.3. *A* level set *of a proper, positive-definite function $V : R^n \to R$ is defined by real numbers $c_2 > c_1 \geq 0$ via $V^{-1}[c_1, c_2] \doteq \{x \in R^n \mid c_1 \leq V(x) \leq c_2\}$.*

DEFINITION 1.4 (see [5]). *Given a locally Lipschitz function $V : R^n \to R$ and a continuous vector field f on R^n, the* Lie derivative of $V(x)$ along $f(x)$ *is defined by*

$$L_f V(x) \doteq \limsup_{t \to 0^+} \frac{V(x + tf(x)) - V(x)}{t}.$$

If $V(x)$ is differentiable at a point $x \in R^n$, then $L_f V(x) = \frac{\partial V}{\partial x}(x) f(x)$.

DEFINITION 1.5 (see [12, 27]). *Consider a subset $\mathcal{W} \subseteq R^l$, a closed subset $\mathcal{U} \subseteq R^m$, a positive-definite function $W(x)$, and real numbers $c_2 > c_1 \geq 0$. A proper, positive-definite C^1 function $V(x)$ is a* robust control Lyapunov function (RCLF) with stability margin $W(x)$ with controls in $\mathcal{U}$ over $V^{-1}[c_1, c_2]$ *for the system* (1.2) *if there exists a control law $\mu : R^n \to \mathcal{U}$ such that*

$$\sup_{x \in V^{-1}[c_1,c_2]} \sup_{w \in \mathcal{W}} L_f V(x) + L_{g_w} V(x) w + L_{g_u} V(x) \mu(x) + W(x) \leq 0.$$

Equivalently,

$$\sup_{x \in V^{-1}[c_1,c_2]} \inf_{u \in \mathcal{U}} \sup_{w \in \mathcal{W}} L_f V(x) + L_{g_w} V(x) w + L_{g_u} V(x) u + W(x) \leq 0.$$

REMARK 1. *If $\mathcal{U} = R^m$, the condition in Definition* 1.5 *is equivalent to*

$$\sup_{x \in V^{-1}[c_1,c_2] \cap \ker(L_{g_u} V)} \sup_{w \in \mathcal{W}} L_f V(x) + L_{g_w} V(x) w + W(x) \leq 0.$$

By Definition 1.5, if $V(x)$ is an RCLF, then a static state-feedback control law exists such that the closed-loop system is robustly stable in the sense defined by Definition 1.2. We now proceed to show how to determine whether a given function $V(x)$ is an RCLF over a given level set in section 2. In section 3, we investigate the problem of finding the RCLF $V(x)$ which maximizes the volume of the region over which stability can be guaranteed. We also analyze the computational complexity of these procedures in each of these sections. A numerical example is presented in section 4 to illustrate the findings of the paper. The main results are summarized in section 5.

2. Stability analysis using a given RCLF. Given an RCLF $V(x)$ and a stability margin $W(x)$, our objective is to determine whether $V(x)$ is an RCLF with stability margin $W(x)$ with controls in $\mathcal{U}$ over some level set given by $V^{-1}[c_1, c_2]$ for

the nonlinear system (1.1). By Definition 1.5, this stability condition holds if and only if

$$\sup_{x\in V^{-1}[c_1,c_2]} \inf_{u\in\mathcal{U}} \sup_{w\in\mathcal{W}} L_f V(x) + L_{g_w} V(x) w + L_{g_u} V(x) u + W(x) \le 0. \tag{2.1}$$

Since the term $\inf_{u\in\mathcal{U}} L_{g_u} V(x) u$ is a constant over sets of the form $\{x \in R^n \mid L_{g_u} V(x) = \psi^T\}$ for $\psi \in R^m$, it is useful to parameterize sets of this form when evaluating the condition (2.1).

Under certain assumptions which are made precise in sections 2.1–2.2, the set $\{x \in R^n \mid L_{g_u} V(x) = \psi^T\}$ can be parameterized by x_N and a parameter $\lambda \in R^{n-k-m}$. With this parameterization, we can express $V(x)$ restricted to $\{x \in R^n \mid L_{g_u} V(x) = \psi^T\}$ by a function $V(x_N, \psi, \lambda)$. This parameterization can be chosen so that $V(x_N, \psi, \lambda)$ is minimized at $\lambda = 0$ for each (x_N, ψ).

To analyze stability over the level set $V^{-1}[c_1, c_2]$, we first make the following definitions:

$$\begin{aligned}
\mathcal{Y}(c_2) &\doteq \{(x_N, \psi) \in R^k \times R^m \mid V(x_N, \psi, 0) \le c_2\},\\
\mathcal{Z}(c_1, c_2, x_N, \psi) &\doteq \{\lambda \in R^{n-k-m} \mid c_1 \le V(x_N, \psi, \lambda) \le c_2\},\\
\Gamma(x_N, \psi, \lambda) &\doteq \sup_{w\in\mathcal{W}} L_f V(x) + L_{g_w} V(x) w + \left[\inf_{u\in\mathcal{U}} \psi^T u\right] + W(x),\\
\Gamma(c_1, c_2, x_N, \psi) &\doteq \max_{\lambda\in\mathcal{Z}(c_1,c_2,x_N,\psi)} \Gamma(x_N, \psi, \lambda).
\end{aligned}$$

The set $\mathcal{Y}(c_2)$ is simply the allowable range of (x_N, ψ) in the given level set, and $\mathcal{Z}(c_1, c_2, x_N, \psi)$ maps the level set to the parameter space of the variable λ.

PROPOSITION 2.1. *$V(x)$ is an RCLF with stability margin $W(x)$ with controls in $\mathcal{U}$ over $V^{-1}[c_1, c_2]$ if and only if $\Gamma(c_1, c_2, x_N, \psi) \le 0$ for all $(x_N, \psi) \in \mathcal{Y}(c_2)$.*

The condition in Proposition 2.1 implies that $\Gamma(c_1, c_2, x_N, \psi) \le 0$ for all $(x_N, \psi) \in R^k \times R^m$. Since $\mathcal{Y}(c_2)$ is compact, we check the condition in Proposition 2.1 by gridding $\mathcal{Y}(c_2)$ and solving a feasibility problem to determine whether $\Gamma(c_1, c_2, x_N, \psi) \le 0$ for all (x_N, ψ) in the grid.

When $\mathcal{U} = R^m$, we see from Remark 1 that we need only to check whether $\Gamma(c_1, c_2, x_N, \psi) \le 0$ for $\psi = 0$. In this case, it is necessary only to grid over the permissible values of x_N, which is a compact region of dimension k rather than dimension $(k + m)$. Therefore, Proposition 2.1 holds with the following modified definition:

$$\mathcal{Y}(c_2) \doteq \{(x_N, \psi) \in R^k \times R^m \mid \psi = 0, V(x_N, 0, 0) \le c_2\}.$$

REMARK 2. *The procedure just described also applies in the analysis of closed-loop robust stability under a control law of the form $\mu(x) = \mu_N(x_N) + \mu_L(x_N) x_L$. In this case, the control is already fixed, so we do not need to parameterize sets of the form $\{x \in R^n \mid L_{g_u} V(x) = \psi^T\}$. In other words, we simply parameterize the level set $V^{-1}[c_1, c_2]$ by the variable x_N. The only additional step is to check that $\mu(x) \in \mathcal{U}$ over $V^{-1}[c_1, c_2]$, but this is straightforward for certain common classes of control constraint sets* [21].

In sections 2.1–2.2, we fill in the details of this stability analysis procedure for two specific Lyapunov function classes and evaluate the computational complexity of the procedure in each case.

2.1. Quadratic RCLF with constant P matrix. In this section, we fill in the details of the stability analysis procedure for the special case of a standard quadratic RCLF and stability margin of the form

$$V(x) = x^T P x = x_N^T P_{NN} x_N + x_N^T P_{NL} x_L + x_L^T P_{LN} x_N + x_L^T P_{LL} x_L, \tag{2.2}$$

$$W(x) = x^T Q x = x_N^T Q_{NN} x_N + x_N^T Q_{NL} x_L + x_L^T Q_{LN} x_N + x_L^T Q_{LL} x_L. \tag{2.3}$$

We begin by parameterizing sets of the form $\{x \in R^n \mid L_{g_u} V(x) = \psi^T\}$ for $\psi \in R^m$. For a system of the form (1.1) and a Lyapunov function (2.2), we have

$$\begin{aligned} L_{g_u} V(x) &= 2x_N^T [P_{NN} g_u^N(x_N) + P_{NL} g_u^L(x_N)] + 2x_L^T Y(x_N), \\ Y(x_N) &\doteq P_{LN} g_u^N(x_N) + P_{LL} g_u^L(x_N). \end{aligned}$$

To simplify the algebra, we make the following assumption.

ASSUMPTION 1. *For all $x_N \in R^k$, rank $Y(x_N) = m \leq n - k$.*

Assumption 1 can be relaxed, if necessary, with some modifications to the analysis which follows. Theorem 2.2 shows how the set $\{x \in R^n \mid L_{g_u} V(x) = \psi^T\}$ can be parameterized by x_N and a parameter $\lambda \in R^{n-k-m}$.

THEOREM 2.2. *Given a function $V(x)$ of the form (2.2) and a vector $\psi \in R^m$,*

$$\begin{aligned} &\{x \in R^n \mid L_{g_u} V(x) = \psi^T\} \\ &\quad = \left\{ \begin{bmatrix} x_N \\ G(x_N)\lambda - P_{LL}^{-1} P_{LN} x_N - \xi(x_N, \psi) \end{bmatrix}, x_N \in R^k, \lambda \in R^{n-k-m} \right\}, \\ &\xi(x_N, \psi) \doteq P_{LL}^{-1} Y(x_N) [Y(x_N)^T P_{LL}^{-1} Y(x_N)]^{-1} \left[g_u^N(x_N)^T R x_N - \frac{1}{2}\psi \right], \\ &\quad R \doteq P_{NN} - P_{NL} P_{LL}^{-1} P_{LN}, \end{aligned}$$

for any matrix $G(x_N) \in R^{(n-k)\times(n-k-m)}$ of full rank such that $Y(x_N)^T G(x_N) \equiv 0$.

Proof. Note that since $G(x_N)$ is full rank, $G(x_N)\lambda$ completely characterizes the null space of $Y(x_N)^T$. Therefore, if any element in the set described above is an element of $\{x \in R^n \mid L_{g_u} V(x) = \psi^T\}$, then that set is equal to $\{x \in R^n \mid L_{g_u} V(x) = \psi^T\}$. Hence, it is sufficient to check the value of $L_{g_u} V(x)$ when $\lambda = 0$:

$$\begin{aligned} L_{g_u} V(x) &= 2x_N^T [P_{NN} g_u^N(x_N) + P_{NL} g_u^L(x_N)] + 2x_L^T Y(x_N) \\ &= 2x_N^T [P_{NN} g_u^N(x_N) + P_{NL} g_u^L(x_N)] - 2[P_{LL}^{-1} P_{LN} x_N + \xi(x_N, \psi)]^T Y(x_N) \\ &= 2x_N^T R g_u^N(x_N) - 2\left[g_u^N(x_N) R x_N - \frac{1}{2}\psi \right]^T = \psi^T. \end{aligned}$$

With the parameterization of Theorem 2.2, we can express $V(x)$ restricted to $\{x \in R^n \mid L_{g_u} V(x) = \psi^T\}$ as follows:

$$V(x_N, \psi, \lambda) = x_N^T R x_N + \xi(x_N, \psi)^T P_{LL} \xi(x_N, \psi) + \lambda^T G(x_N)^T P_{LL} G(x_N) \lambda.$$

In order to grid $\mathcal{Y}(c_2)$, we need bounds on the values of x_N and ψ which can be achieved on this set. The permissible values of x_N are those satisfying $x_N^T R x_N \leq c_2$. Now, for fixed x_N, $L_{g_u} V(x)$ is affine in x_L for systems of the form (1.1). Hence, each element of $L_{g_u} V(x)$ can be represented as $[L_{g_u} V(x)]_p = a_p(x_N) + x_L^T b_p(x_N)$, where

$$\begin{aligned} \begin{bmatrix} a_1(x_N) & \cdots & a_m(x_N) \end{bmatrix} &= 2x_N^T [P_{NN} g_u^N(x_N) + P_{NL} g_u^L(x_N)], \\ \begin{bmatrix} b_1(x_N) & \cdots & b_m(x_N) \end{bmatrix} &= 2Y(x_N). \end{aligned}$$

The bounds on each element ψ_p are therefore given as follows:

$$\min_{V(x)\le c_2} a_p(x_N) + x_L^T b_p(x_N) \le \psi_p \le \max_{V(x)\le c_2} a_p(x_N) + x_L^T b_p(x_N).$$

By the method of Lagrange multipliers [28], we find this to be equivalent to

$$\alpha_p(x_N) - \beta_p(x_N) \le \psi_p \le \alpha_p(x_N) + \beta_p(x_N), \tag{2.4}$$

where, for $x_N^T R x_N \le c_2$,

$$\alpha_p(x_N) \doteq a_p(x_N) - b_p(x_N)^T P_{LL}^{-1} P_{LN} x_N,$$
$$\beta_p(x_N) \doteq \sqrt{[c_2 - x_N^T R x_N][b_p(x_N)^T P_{LL}^{-1} b_p(x_N)]}.$$

The bounds in (2.4) give us the set of values of $L_{g_u} V(x)$ over which we must grid in order to complete the stability analysis. If $\psi = 0$ is within the allowable range for a given x_N, we should use this as one of the grid points to ensure that at least the condition for stability from Remark 1 for the case of unlimited control ($\mathcal{U} = R^m$) is satisfied.

By Theorem 2.2, we can parameterize the stability condition as follows for each $(x_N, \psi) \in \mathcal{Y}(c_2)$:

$$\begin{aligned}\Gamma(x_N, \psi, \lambda) = \sup_{w\in\mathcal{W}} & a_0(x_N, \psi) + b_0(x_N, \psi)^T \lambda + \lambda^T C_0(x_N)\lambda \\ & + w^T[s(x_N, \psi) + T(x_N)\lambda].\end{aligned} \tag{2.5}$$

The coefficients can be found by straightforward algebra. We can check the condition $\Gamma(c_1, c_2, x_N, \psi) \le 0$ by solving a linear matrix inequality (LMI) feasibility problem using the $\mathcal{S}$-procedure as discussed in [6]. Let us first consider the case of no disturbances ($\mathcal{W} = \{0\}$), for which the parameterized stability condition (2.5) gives us

$$\Gamma(c_1, c_2, x_N, \psi) = \max_{c_1 \le V(x_N, \psi, \lambda) \le c_2} a_0(x_N, \psi) + b_0(x_N, \psi)^T \lambda + \lambda^T C_0(x_N)\lambda. \tag{2.6}$$

Checking whether $\Gamma(c_1, c_2, x_N, \psi) \le 0$ is therefore a quadratic feasibility problem with quadratic constraints, which can be solved using the $\mathcal{S}$-procedure in the manner discussed in [6]. With only one quadratic constraint, the $\mathcal{S}$-procedure is nonconservative [6], but a potential problem arises in our case because there are two constraints. Fortunately, the two constraints are never simultaneously active, as shown in the following theorem.

THEOREM 2.3. *Suppose $(x_N, \psi) \in \mathcal{Y}(c_2)$, and consider the maximization problem (2.6). If $C_0(x_N) \not\prec 0$, then $\Gamma(c_1, c_2, x_N, \psi) = \Gamma(0, c_2, x_N, \psi)$. If $C_0(x_N) < 0$, one of the following applies for $\lambda^* = -\frac{1}{2} C_0(x_N)^{-1} b_0(x_N, \psi)$:*

1. *If $V(x_N, \psi, \lambda^*) < c_1$, then $\Gamma(c_1, c_2, x_N, \psi) = \Gamma(c_1, \infty, x_N, \psi)$.*
2. *If $V(x_N, \psi, \lambda^*) > c_2$, then $\Gamma(c_1, c_2, x_N, \psi) = \Gamma(0, c_2, x_N, \psi)$.*
3. *Otherwise, $\Gamma(c_1, c_2, x_N, \psi) = a_0(x_N, \psi) - \frac{1}{4} b_0(x_N, \psi)^T C_0(x_N)^{-1} b_0(x_N, \psi)$.*

Proof. See Appendix A.

Consider next the case where $\mathcal{W}$ is polytopic; for example, $\mathcal{W} = \{w \mid \|w\|_\infty \le 1\}$ belongs to this category. Since the expression in (2.5) is affine in w, robust stability can be analyzed exactly using Theorem 2.3 with each of the extreme points of $\mathcal{W}$ substituted for w. This approach gives us the value of

$$\begin{aligned}\Gamma(c_1, c_2, x_N, \psi) = \max_{w\in\mathcal{V}} \max_{c_1 \le V(x_N, \psi, \lambda) \le c_2} & [a_0(x_N, \psi) + s(x_N, \psi)^T w] \\ & + [b_0(x_N, \psi) + T(x_N)^T w]^T \lambda + \lambda^T C_0(x_N)\lambda,\end{aligned}$$

TABLE 2.1
Computation times for stability analysis using a quadratic RCLF with a constant P matrix.

Operation	Time (unbounded control)	Time (bounded control)
Solve for $L_{g_u}V(x) = \psi^T$	$\mathcal{O}(N_c^k n^3)$	$\mathcal{O}(N_c^{k+m} n^3)$
Compute $V(x_N, \psi, \lambda)$	$\mathcal{O}(N_c^k n^4)$	$\mathcal{O}(N_c^{k+m} n^4)$
Check $\Gamma(c_1, c_2, x_N, \psi) \leq 0$	$\mathcal{O}(2^l N_c^k n^3)$	$\mathcal{O}(2^l N_c^{k+m} n^3)$

where $\mathcal{V}$ is the set of extreme points of $\mathcal{W}$. Other classes of disturbance constraints can be handled in the $\mathcal{S}$-procedure framework, but we do not enumerate them here.

We now analyze the computational complexity of the stability analysis procedure outlined in this section. To verify whether a desired stability margin is achieved over a given level set, we must parameterize the set $\{x \in R^n \mid L_{g_u}V(x) = \psi^T\}$ and the level set. We then evaluate robust stability over the resulting parameterized set at various grid point values of x_N (and ψ if there are control limitations). Approximate computation times determined numerically for a system influenced by disturbances contained in $\mathcal{W} = \{w \in R^l \mid \|w\|_\infty \leq 1\}$ are listed in Table 2.1. The quantity N_c is the number of grid points in $\mathcal{Y}(c_2)$ over each dimension, so that the total number of grid points is roughly N_c^k (or N_c^{k+m} if there are control limitations).

The complexity of this procedure should be compared with the complexity of evaluating Lyapunov derivatives over a level set to determine the region of stability for a general nonlinear system. This problem is discussed in [8, 9, 10, 14, 15, 20, 24, 26]. For an arbitrary nonlinear system, it is necessary to grid the level set over all dimensions, so the computation times required to solve this problem to some desired accuracy grow exponentially with the state dimension. In the procedure developed here for the system (1.1), on the other hand, we grid only $(x_N, \psi) \in \mathcal{Y}(c_2)$ and evaluate whether $\Gamma(c_1, c_2, x_N, \psi) \leq 0$ at each grid point. By Theorem 2.3, this reduces to an LMI feasibility problem, and there are standard methods for solving this problem to a desired level of accuracy with a computation time which is polynomial in the state dimension [6, 23]. For example, the author implemented an ellipsoid algorithm [6] and found the computation time to vary roughly as $(n-k-m)^3$. Due to gridding over $\mathcal{Y}(c_2)$, the computation time is still exponential in $\dim(x_N)$ or in $\dim(x_N) + \dim(u)$ when control constraints are active. In other words, the computational complexity is exponential only in the degree of nonlinearity in the problem.

2.2. Quadratic RCLF with state-dependent *P* matrix. In this section, we fill in the details of the stability analysis procedure for the special case of an RCLF and stability margin which are arbitrarily nonlinear in x_N but quadratic in x_L. In other words, $V(x)$ and $W(x)$ have the form

$$V(x) = x^T P(x_N) x, \tag{2.7}$$

$$W(x) = x^T Q(x_N) x, \tag{2.8}$$

where $P \in C^1(R^k \to R^{n \times n})$ and $Q \in C^0(R^k \to R^{n \times n})$. We assume that $V(x)$ and $W(x)$ are positive-definite and that the matrices $P(x_N)$ and $Q(x_N)$ are partitioned in the same manner as for the P and Q matrices in section 2.1.

We begin by parameterizing sets of the form $\{x \in R^n \mid L_{g_u}V(x) = \psi^T\}$ for $\psi \in R^m$. We can simplify the treatment considerably by making the following assumption on the system (1.1).

ASSUMPTION 2. *For all $x_N \in R^k$, $g_u^N(x_N) = 0$.*

For a system of the form (1.1) satisfying Assumption 2 and a Lyapunov function of the form (2.7), we have

$$L_{g_u}V(x) = 2x_N^T P_{NL}(x_N) g_u^L(x_N) + x_L^T Y(x_N),$$
$$Y(x_N) \doteq 2P_{LL}(x_N) g_u^L(x_N).$$

To simplify the algebra, we make the following assumption.

ASSUMPTION 3. *For all* $x_N \in R^k$, rank $g_u^L(x_N) = m \le n-k$.

Note that Assumption 3 is equivalent to the condition that rank $Y(x_N) = m$ for all $x_N \in R^k$. This assumption can be relaxed, if necessary, with some modifications to the analysis which follows. Theorem 2.4 shows how the set $\{x \in R^n \mid L_{g_u}V(x) = \psi^T\}$ can be parameterized by x_N and a parameter $\lambda \in R^{n-k-m}$.

THEOREM 2.4. *Given a function* $V(x)$ *of the form* (2.7) *and a vector* $\psi \in R^m$,

$$\{x \in R^n \mid L_{g_u}V(x) = \psi^T\} = \left\{ \begin{bmatrix} x_N \\ \sigma(x_N, \psi) \end{bmatrix}, x_N \in R^k, \lambda \in R^{n-k-m} \right\},$$
$$\sigma(x_N, \psi) \doteq G(x_N)\lambda - P_{LL}(x_N)^{-1}P_{LN}(x_N)x_N + P_{LL}(x_N)^{-1}Y(x_N)[Y(x_N)^T P_{LL}(x_N)^{-1}Y(x_N)]^{-1}\psi,$$

for any matrix $G(x_N) \in R^{(n-k)\times(n-k-m)}$ *of full rank such that* $Y(x_N)^T G(x_N) \equiv 0$.

Proof. The proof of Theorem 2.4 follows the same arguments as the proof of Theorem 2.2.

With the parameterization of Theorem 2.4, we can express $V(x)$ restricted to $\{x \in R^n \mid L_{g_u}V(x) = \psi^T\}$ as follows:

$$V(x_N, \psi, \lambda) = x_N^T R(x_N) x_N + \psi^T [Y(x_N)^T P_{LL}(x_N)^{-1} Y(x_N)]^{-1}\psi + \lambda^T G(x_N)^T P_{LL}(x_N) G(x_N)\lambda,$$
$$R(x_N) \doteq P_{NN}(x_N) - P_{NL}(x_N)P_{LL}(x_N)^{-1}P_{LN}(x_N).$$

In order to grid $\mathcal{Y}(c_2)$, we need bounds on the values of x_N and ψ which can be achieved on this set. The permissible values of x_N are those satisfying $x_N^T R(x_N) x_N \le c_2$. Similarly, the permissible values of $L_{g_u}V(x) = \psi^T$ corresponding to each such value of x_N are those contained in the ellipsoid described by

$$\psi^T [Y(x_N)^T P_{LL}(x_N)^{-1} Y(x_N)]^{-1}\psi \le c_2 - x_N^T R(x_N) x_N.$$

This is the set of values of $L_{g_u}V(x)$ over which we must grid in order to complete the stability analysis. We should always use $\psi = 0$ as one of the grid points to ensure that at least the condition for stability from Remark 1 for the case of unlimited control ($\mathcal{U} = R^m$) is satisfied.

By Theorem 2.4, we can parameterize the stability condition as follows for each $(x_N, \psi) \in \mathcal{Y}(c_2)$:

$$\Gamma(x_N, \psi, \lambda) = \sup_{w \in \mathcal{W}} a_0(x_N, \psi) + b_0(x_N, \psi)^T\lambda + \lambda^T C_0(x_N, \psi)\lambda + w^T[s(x_N, \psi) + T(x_N, \psi)\lambda] + \sum_{i=1}^{k} [\lambda^T D_i(x_N)\lambda][h_i(x_N)^T\lambda + v_i(x_N)^T w]. \tag{2.9}$$

The coefficients can be found by straightforward algebra. We can rearrange the expression for $\Gamma(c_1, c_2, x_N, \psi)$ when $(x_N, \psi) \in \mathcal{Y}(c_2)$ by introducing the following scaling matrix:

$$K(x_N, \psi) \doteq \sqrt{c_2 - V(x_N, \psi, 0)}\left[G(x_N)^T P_{LL}(x_N) G(x_N)\right]^{-1/2},$$

where $M^{-1/2}$ is the positive-definite square root of M^{-1} for $M = M^T > 0$. Then we can replace λ by $K(x_N, \psi)\lambda$ and set $b_0 \leftarrow Kb_0$, $C_0 \leftarrow KC_0K$, $T \leftarrow TK$, $D_i \leftarrow KD_iK$, and $h_i \leftarrow Kh_i$ to obtain

$$\begin{aligned}\Gamma(c_1, c_2, x_N, \psi) = \max_{\beta(x_N,\psi)\le\lambda^T\lambda\le 1} \sup_{w\in\mathcal{W}} & a_0(x_N,\psi) + b_0(x_N,\psi)^T\lambda + \lambda^T C_0(x_N,\psi)\lambda \\ & + w^T[s(x_N,\psi) + T(x_N,\psi)\lambda] \\ & + \sum_{i=1}^{k}[\lambda^T D_i(x_N,\psi)\lambda][h_i(x_N,\psi)^T\lambda + v_i(x_N)^T w],\end{aligned} \tag{2.10}$$

where $\beta(x_N, \psi) = [c_1 - V(x_N, \psi, 0)]/[c_2 - V(x_N, \psi, 0)] \le 1$. This alternative form simplifies the development of a method to evaluate the stability condition.

Checking the condition $\Gamma(c_1, c_2, x_N, \psi) \le 0$ is a nonconvex constrained feasibility problem, which is known to be NP-hard [7, 22]. However, in practice, good upper and lower bounds can be computed in polynomial time by transforming the problem to a real μ analysis problem of the type developed in [31]. Let us first consider the case of no disturbances ($\mathcal{W} = \{0\}$). We approximate the parameterized stability condition (2.10) by

$$\begin{aligned}\tilde{\Gamma}(0, c_2, x_N, \psi) &\doteq \max_{\|\lambda\|_\infty\le 1} \phi_0(\lambda), \\ \phi_0(\lambda) &\doteq a_0 + b_0^T\lambda + \lambda^T C_0\lambda + \sum_{i=1}^{k}\lambda^T h_i(\lambda^T D_i\lambda),\end{aligned} \tag{2.11}$$

where we have suppressed the dependence of the coefficients in (2.11) on (x_N, ψ). By assuming $\mathcal{W} = \{0\}$, we guarantee that $c_1 = 0$. This implies that $\beta(x_N, \psi) \le 0$; in other words, the constraint set in (2.10) is given by $\|\lambda\|_2 \le 1$. It is useful to the development of the procedure which follows that this set is convex. We also replace the condition $\|\lambda\|_2 \le 1$ by $\|\lambda\|_\infty \le 1$. Since $\|\lambda\|_\infty \le \|\lambda\|_2$, this approach is conservative because $\Gamma(0, c_2, x_N, \psi) \le \tilde{\Gamma}(0, c_2, x_N, \psi)$. Nevertheless, we obtain the following test for stability, which is adapted from Proposition 2.1.

PROPOSITION 2.5. *If $\mathcal{W} = \{0\}$, then $V(x)$ is an RCLF with stability margin $W(x)$ with controls in $\mathcal{U}$ over $V^{-1}[0, c_2]$ if $\tilde{\Gamma}(0, c_2, x_N, \psi) \le 0$ for all $(x_N, \psi) \in \mathcal{Y}(c_2)$.*

We now show how Proposition 2.5 can be converted to a real μ analysis problem. Recall that in the real μ analysis problem, we are given a matrix M and an uncertainty class $\mathbf{\Delta}$, and we seek to evaluate

$$\mu(M, \mathbf{\Delta}) \doteq \begin{cases} \left[\min_{\Delta\in\mathbf{\Delta}}\{\bar{\sigma}(\Delta) \mid \det(I + M\Delta) = 0\}\right]^{-1} & \exists\Delta \in \mathbf{\Delta} \mid \det(I + M\Delta) = 0, \\ 0 & \text{otherwise.}\end{cases}$$

In other words, $\mu(M, \mathbf{\Delta})$ is the maximum singular value of the smallest perturbation $\Delta \in \mathbf{\Delta}$ which destabilizes a linear fractional interconnection between M and Δ by the small gain theorem [31]. Our objective here is to transform Proposition 2.5 to a problem of computing $\mu(M, \mathbf{\Delta})$ for some relevant M and $\mathbf{\Delta}$.

The following theorem is a general result on how to convert the problem of maximizing any rational function of a constrained variable to a real μ analysis problem.

THEOREM 2.6 (see [7]). *For any $q \ge 0$ and any expression $\phi_0(\lambda)$ which can be expressed as a linear fractional transformation of λ and λ^T, there exists a matrix M*

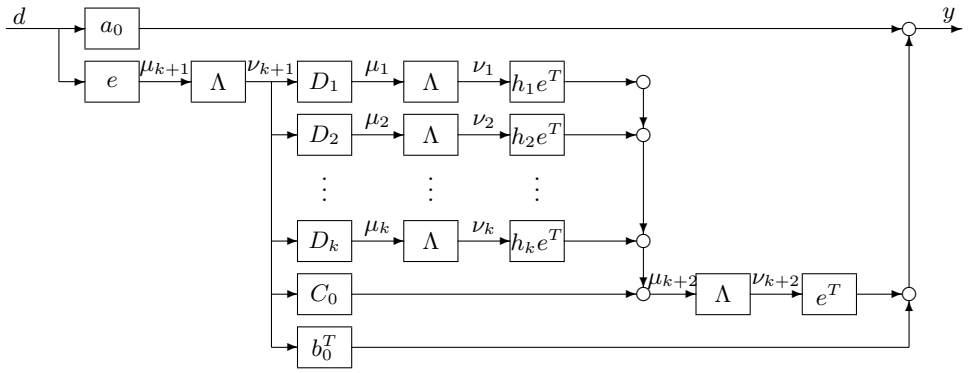


FIG. 2.1. *Block diagram representation of* $\phi_0(\lambda)$ *from* (2.11).

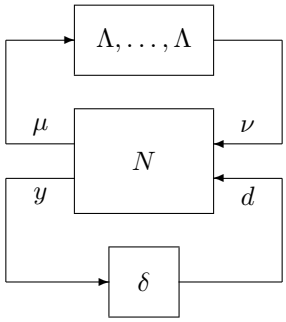


FIG. 2.2. *Block diagram from Figure* 2.1 *with uncertainty representation.*

and an uncertainty class $\mathbf{\Delta}$ *such that*

$$\max_{\|\lambda\|_\infty \le 1} |\phi_0(\lambda)| < q \Leftrightarrow \mu(M, \mathbf{\Delta}) < q. \tag{2.12}$$

A basic idea of how to compute M and $\mathbf{\Delta}$ is given in [7], where this is done (in slightly more generality) for a quadratic $\phi_0(\lambda)$. We seek to apply Theorem 2.6 when $\phi_0(\lambda)$ has the form (2.11), where $C_0 = C_0^T$ and $D_i = D_i^T$ for all $i = 1, \ldots, k$. Following [7], the first step is to generate a block diagram representation for $\phi_0(\lambda)$ from (2.11) such that $y = \phi_0(\lambda)d$ for $d, y \in R$. To this end, we define $e \doteq [1, \ldots, 1]^T$ with $\dim(e) = \dim(\lambda)$ and $\Lambda \doteq \mathrm{diag}(\lambda)$, so that $\lambda = \Lambda e$. Then $\phi_0(\lambda)$ is represented in block diagram form as in Figure 2.1.

Note that the Λ block is repeated $k+2$ times in Figure 2.1. It is useful to transform the block diagram to the alternative form shown in Figure 2.2, where

$$N = \left[\begin{array}{ccc|cc|c} 0 & \cdots & 0 & D_1 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & D_k & 0 & 0 \\ \hline 0 & \cdots & 0 & 0 & 0 & e \\ h_1e^T & \cdots & h_ke^T & C_0 & 0 & 0 \\ \hline 0 & \cdots & 0 & b_0^T & e^T & a_0 \end{array}\right], \quad \Delta = \left[\begin{array}{ccc|cc|c} \Lambda & & & & & \\ & \ddots & & & & \\ & & \Lambda & & & \\ \hline & & & \Lambda & & \\ & & & & \Lambda & \\ \hline & & & & & \delta \end{array}\right].$$

Note from (2.12) and the discussion in [7] that the condition $|\phi_0(\lambda)| < q$ is equivalent to $\mu(M, \mathbf{\Delta}) < q$ for any $q > 0$, where M is obtained by multiplying all but the last row of N by q.

At this stage we notice that $|\phi_0(\lambda)|$ is used in Theorem 2.6, but what we are really interested in is the maximum of $\phi_0(\lambda)$ itself over $\|\lambda\|_\infty \leq 1$. Fortunately, this distinction is not restrictive. Indeed, it is straightforward to find a crude lower bound L such that $\phi_0(\lambda) \geq L$ for all λ satisfying $\|\lambda\|_\infty \leq 1$. Given such a lower bound, we can define $\tilde{\phi}_0(\lambda) \doteq \phi_0(\lambda) - L$, and we have $|\tilde{\phi}_0(\lambda)| = \tilde{\phi}_0(\lambda)$ whenever $\|\lambda\|_\infty \leq 1$. By constructing $\tilde{M}$ based on $\tilde{\phi}_0(\lambda)$ rather than $\phi_0(\lambda)$, we obtain the following necessary and sufficient condition for a given bound on $\tilde{\Gamma}(0, c_2, x_N, \psi)$ to hold:

$$\mu(\tilde{M}, \mathbf{\Delta}) < q \Leftrightarrow \tilde{\Gamma}(0, c_2, x_N, \psi) < q + L.$$

In other words, the condition in Proposition 2.5 holds if $\mu(\tilde{M}, \mathbf{\Delta}) < -L$. Therefore, we can evaluate the stability condition using the standard real μ analysis procedure from [32].

Consider next the case where $\mathcal{W} = \{w \in R^l \mid \|w\|_\infty \leq 1\}$. Since the expression in (2.10) is affine in w, robust stability can be analyzed exactly by solving the nonconvex feasibility problem with each of the extreme points of $\mathcal{W}$ substituted for w. Alternatively, we could add a block $W \doteq \mathrm{diag}(w)$ to Figures 2.1 and 2.2 so that $y = \phi_0(\lambda)d$, where

$$\tilde{\Gamma}(0, c_2, x_N, \psi) \doteq \max_{\|\lambda\|_\infty \leq 1} \max_{\|w\|_\infty \leq 1} \phi_0(\lambda),$$

$$\phi_0(\lambda) \doteq a_0 + b_0^T\lambda + \lambda^T C_0 \lambda + w^T[s + T\lambda] + \sum_{i=1}^{k} (\lambda^T D_i \lambda)(h_i^T \lambda + v_i^T w).$$

Note, however, that we can only analyze stability in the special case $c_1 = 0$ using this procedure because the level set must be convex. In general, we expect the stability condition to be violated over the level set $V^{-1}[0, c_2]$ unless a matching condition on $g_w(x_N)$ and $g_u(x_N)$ holds.

We now analyze the computational complexity of the stability analysis procedure outlined in this section. The parameterizations of the set $\{x \in R^n \mid L_{g_u} V(x) = \psi^T\}$ and of the level set do not differ significantly from those used in section 2.1. Robust stability analysis can be accomplished either by general techniques for solving nonconvex constrained feasibility problems or by conversion to a real μ analysis problem. The nonconvex solution procedure is NP-hard [7, 22], but the solution (if it can be found) is exact. Alternatively, approximate bounds on $\tilde{\Gamma}(0, c_2, x_N, \psi)$ can be found using standard techniques for real μ analysis. The computational complexity of this approximation is analyzed in [32]. Approximate computation times for a system influenced by disturbances contained in $\mathcal{W} = \{w \in R^l \mid \|w\|_\infty \leq 1\}$ are listed in Table 2.2. The quantity N_c is the number of grid points in $\mathcal{Y}(c_2)$ over each dimension, so that the total number of grid points is roughly N_c^k (or N_c^{k+m} if there are control limitations). The quantity $N_\Delta \doteq (n - k - m)(k + 2) + l + 1$ is the dimension of the perturbation block Δ. The nonconvex solution procedure is at least as complex as gridding over the entire state space to evaluate the Lyapunov derivative.

3. Optimization over the RCLF. Now we turn our attention to the problem of constructing an RCLF such that the volume of the level set for guaranteed stability is maximized. This problem formulation is inspired by [10], in which a quadratic Lyapunov function is used to find the stability region of maximum volume for an

TABLE 2.2
Computation times for stability analysis using a quadratic RCLF with a state-dependent P matrix.

Operation	Time (unbounded control)	Time (bounded control)
Solve for $L_{g_u}V(x) = \psi^T$	$\mathcal{O}(N_c^k n^3)$	$\mathcal{O}(N_c^{k+m} n^3)$
Compute $V(x_N, \psi, \lambda)$	$\mathcal{O}(N_c^k n^4)$	$\mathcal{O}(N_c^{k+m} n^4)$
Nonconvex solution	$\mathcal{O}(2^l N_c^n)$	$\mathcal{O}(2^l N_c^n)$
μ upper bound	$\mathcal{O}(N_c^k N_\Delta^3)$	$\mathcal{O}(N_c^{k+m} N_\Delta^3)$

arbitrary nonlinear autonomous system of the form $\dot{x} = f(x)$. For this problem, we do not wish to impose a particular stability margin, so we consider the following modified definition for a function to be an RCLF.

DEFINITION 3.1. *Consider a subset $\mathcal{W} \subseteq R^l$, a closed subset $\mathcal{U} \subseteq R^m$, and real numbers $c_2 > c_1 > 0$. A proper, positive-definite C^1 function $V(x)$ is an RCLF with controls in $\mathcal{U}$ over $V^{-1}[c_1, c_2]$ for the system* (1.2) *if there exists a control law $\mu : R^n \to \mathcal{U}$ such that*

$$\sup_{x \in V^{-1}[c_1,c_2]} \sup_{w \in \mathcal{W}} L_f V(x) + L_{g_w} V(x) w + L_{g_u} V(x) \mu(x) < 0.$$

Equivalently,

$$\sup_{x \in V^{-1}[c_1,c_2]} \inf_{u \in \mathcal{U}} \sup_{w \in \mathcal{W}} L_f V(x) + L_{g_w} V(x) w + L_{g_u} V(x) u < 0.$$

REMARK 3. *If $\mathcal{U} = R^m$, the condition in Definition* 3.1 *is equivalent to*

$$\sup_{x \in V^{-1}[c_1,c_2] \cap \ker(L_{g_u} V)} \sup_{w \in \mathcal{W}} L_f V(x) + L_{g_w} V(x) w < 0.$$

Note that the analysis of stability over a level set in this scenario is slightly different from the procedure used in section 2 because we require the Lyapunov derivative to be strictly negative. In other words, if we define

$$\Gamma(x_N, \psi, \lambda) \doteq \sup_{w \in \mathcal{W}} L_f V(x) + L_{g_w} V(x) w + \left[\inf_{u \in \mathcal{U}} \psi^T u \right],$$
$$\Gamma(c_1, c_2, x_N, \psi) \doteq \max_{\lambda \in \mathcal{Z}(c_1,c_2,x_N,\psi)} \Gamma(x_N, \psi, \lambda),$$

then a condition for $V(x)$ to be an RCLF is given by Proposition 3.2.

PROPOSITION 3.2. *$V(x)$ is an RCLF with controls in $\mathcal{U}$ over $V^{-1}[c_1, c_2]$ if and only if $\Gamma(c_1, c_2, x_N, \psi) < 0$ for all $(x_N, \psi) \in \mathcal{Y}(c_2)$.*

Since we are optimizing over $V(x)$ at this point, we do not need to iterate over the level values to determine the largest region of stability. Therefore, we shall henceforth consider the problem of finding the function $V(x)$, which is an RCLF with controls in $\mathcal{U}$ over $V^{-1}[\gamma, 1]$ for some $\gamma > 0$ such that the volume of the set $V^{-1}[0, 1]$ is maximized. The appropriate value of γ depends on the RCLF. However, if we assume that there exists some nominal $V_0(x)$ which is an RCLF with controls in $\mathcal{U}$ over $V_0^{-1}[c_1, c_2]$, then we could fix γ to be the largest value satisfying $V^{-1}[0, \gamma] \subseteq V_0^{-1}[0, c_2]$. In this way, we can find a control law which renders the system RUAS($\mathcal{X}$,Ω) with $\mathcal{X} = V^{-1}[0, 1]$ and $\Omega = V_0^{-1}[0, c_1]$ by the method in [19]. In other words, the inner level set over which stability is guaranteed is independent of the particular RCLF used. Therefore,

a direct comparison of the volumes of sets of the form $V^{-1}[0,1]$ for different RCLFs is a meaningful comparison of the size of the stability region.

In sections 3.1–3.2, we describe the optimization procedures which might be applied to the two Lyapunov function classes of section 2 and evaluate the computational complexity of the procedure in each case.

3.1. Quadratic RCLF with constant P matrix. Our objective is to find a quadratic function $V(x)$ of the form (2.2) which is an RCLF with controls in $\mathcal{U}$ over $V^{-1}[\gamma,1]$ and which maximizes the volume of the level set $V^{-1}[0,1]$. Following [10], we can pose the following optimization problem.

PROBLEM 2. *Suppose that $V_0(x) \doteq x^T P_0 x$ is an RCLF with controls in $\mathcal{U}$ over $V_0^{-1}[c_1,c_2]$. Determine $P = P^T$ and γ from the following optimization problem:*

$$\begin{aligned} &\textit{minimize} && \det(P) \\ &\textit{subject to} && \gamma > 0, \\ &&& P/\gamma \geq P_0/c_2, \\ &&& \Gamma(\gamma, 1, x_N, \psi) < 0 \ \textit{ for all } (x_N,\psi) \in \mathcal{Y}(1). \end{aligned}$$

Problem 2 is a nonlinear, nonconvex optimization problem in the variables P and γ. From [6], we note that by casting the problem in terms of P^{-1} instead of P, we can rewrite the objective using the convex function $\log\det(P^{-1})$. However, the stability condition $\Gamma(\gamma,1,x_N,\psi) < 0$ is nonconvex in both P and P^{-1}, and there is no obvious way to reformulate Problem 2 to be convex. Consequently, Problem 2 is NP-hard. In particular, the computational complexity of this problem is comparable to the complexity of gridding over the set of symmetric positive-definite matrices P. Note that the largest value of γ which satisfies the constraint $P/\gamma \geq P_0/c_2$ can be computed exactly by $\gamma = c_2\lambda_{min}(PP_0^{-1})$.

An approximate solution to Problem 2 is developed in [11]. If the nonlinear terms in the dynamics of (1.1) are rational functions of x_N, then the system can be written in the following *linear fractional representation*:

$$\begin{aligned} \dot{x} &= Ax + B_w w + B_u u + B_p p, \\ q &= C_q x + D_{qw} w + D_{qu} u + D_{qp} p, \\ p &= \Delta(x) q, \\ \Delta(x_N) &= \operatorname{diag}(x_1 I_{r_1}, \ldots, x_k I_{r_k}). \end{aligned}$$

In other words, the nonlinear dynamics of the system are written as a linear fractional transformation (LFT) between a nominal linear time-invariant system and a perturbation block containing (possibly repeated) values of the "nonlinear" states x_N. Note that $\dim(q) = \dim(p) = \sum_{i=1}^{k} r_i$, which is the size of the perturbation block required to represent the nonlinear terms in the dynamics. This quantity could possibly be much larger than $\dim(x_N)$, depending on the types of nonlinearities present.

The control design method proposed in [11] could therefore be used to obtain an RCLF $V(x) = x^T P x$. Since the problem formulation in [11] is an LMI convex optimization problem, we can use this procedure to try to optimize any convex function of the matrix P subject to the constraint of stability over a level set. However, there are two important sources of conservatism in this procedure. The first is that a linear controller structure is assumed, and the solution to the optimization problem under this assumption may not yield the optimal Lyapunov function when nonlinear control laws are allowed. The conservatism due to this assumption could probably

TABLE 3.1
Computation times for optimization using a quadratic RCLF with a constant P matrix.

Operation	Time (unbounded control)	Time (bounded control)
Nonconvex solution	$\mathcal{O}(2^l N_c^n)$	$\mathcal{O}(2^l N_c^n)$
LFT solution	$\mathcal{O}(N_c^k[l+r]^3)$	$\mathcal{O}(N_c^{k+m}[l+r]^3)$

be reduced by assuming an LFT structure for the control as well as the plant in a manner analogous to the incorporation of time-varying parameters in the control for a linear parameter-varying (LPV) system in [25].

The second source of conservatism is probably more important. The stability constraint in [11] is derived from the small gain theorem from robust control and actually requires that the following system is stable:

$$\begin{aligned}\dot{x} &= Ax + B_w w + B_u u + B_p p,\\ q &= C_q x + D_{qw} w + D_{qu} u + D_{qp} p,\\ p &= \Delta(t) q.\end{aligned}$$

In this alternative system description, the perturbation block is allowed to be any matrix function $\Delta(t) \in \mathbf{\Delta}$ for some constraint set $\mathbf{\Delta}$ representing an upper bound on the allowable magnitude of the nonlinear states over the appropriate level set of $V(x)$. Treating nonlinearities as disturbances and applying robust control techniques is known to be an overly conservative approach to nonlinear control [1]. Consequently, there is no guarantee that the matrix P obtained using this procedure solves the original optimization problem. Nevertheless, this procedure is an alternative which may yield good results in some cases.

We now analyze the computational complexity of the optimization procedure described in this section. The optimization can be accomplished either by general nonconvex optimization techniques or by solving an equivalent real μ analysis problem applied to the LFT representation. The general nonconvex optimization problem is NP-hard [22], but the solution (if it can be found) is exact. The number of variables in this problem is equal to $n(n+1)/2$, the dimension of the set of symmetric matrices. Alternatively, approximate bounds on $\Gamma(\gamma, 1, x_N, \psi)$ can be found using standard techniques for real μ analysis [32]. Approximate computation times for a system influenced by disturbances contained in $\mathcal{W} = \{w \in R^l \mid \|w\|_\infty \leq 1\}$ are listed in Table 3.1. Note that the perturbation block in the robust stability problem with an LFT representation for the nonlinear terms has dimension $l + r$, where $r = \sum_{i=1}^k r_i$.

An alternative to the LFT representation would be to regard the system (1.1) as a quasi-LPV system and apply the procedure from [4] for LPV systems. In this procedure, an RCLF is computed by solving a family of LMI convex optimization problems parameterized over the "nonlinear" states x_N. Here again the computational complexity is exponential in k (exponential in $k+m$ in the bounded control case) and polynomial in the remaining state dimension. Note, however, that LPV systems are slightly different from the system (1.1): in (1.1) the dynamics of x_N are known and x_N is part of the state vector to be regulated to a desired equilibrium point.

3.2. Quadratic RCLF with state-dependent *P* matrix. The problem of optimizing over a function $V(x)$ of the form (2.7) is substantially more complicated than the problem in section 3.1. The optimization problem under consideration here is the following.

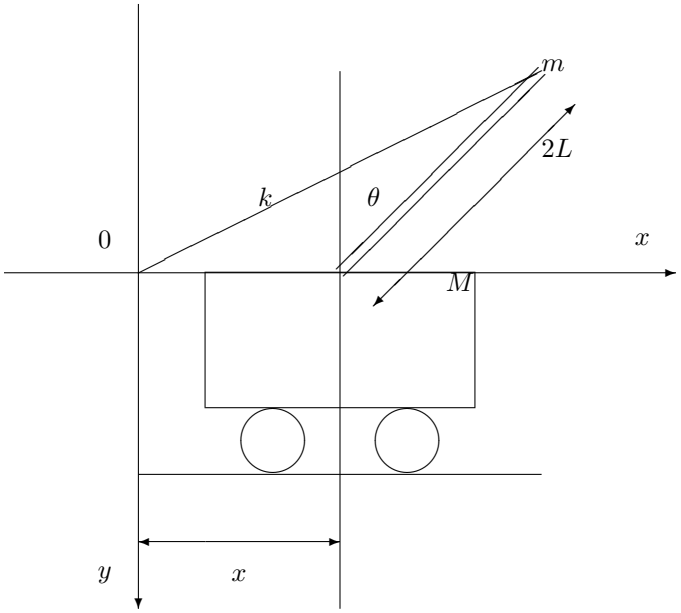


FIG. 3.1. *Cart with inverted pendulum and spring.*

PROBLEM 3. *Suppose that* $V_0(x) \doteq x^T P_0(x_N)x$ *is an RCLF with controls in* $\mathcal{U}$ *over* $V_0^{-1}[c_1, c_2]$. *Determine* $P(x_N) = P^T(x_N)$ *and* γ *from the following optimization problem:*

$$\begin{array}{ll} \textit{maximize} & \textit{vol}\, V^{-1}[0,1] \\ \textit{subject to} & \gamma > 0, \\ & V^{-1}[0,\gamma] \subseteq V_0^{-1}[0,c_1], \\ & \Gamma(\gamma, 1, x_N, \psi) < 0 \ \textit{for all}\ (x_N, \psi) \in \mathcal{Y}(1). \end{array}$$

To develop a tractable approximate solution to Problem 3, we would like to view the problem as an optimization over $P(x_N)$, which is analogous to Problem 2 but parameterized by the x_N states. We can write the objective and the second constraint in Problem 3 in terms of $P(x_N)$, although the formulas may be complicated. In the stability constraint, however, the *derivatives* of $P(x_N)$ with the x_N states appear in $\Gamma(\gamma, 1, x_N, \psi)$. Therefore, we cannot simply evaluate this constraint at a given value of x_N independent of the others; we really need to optimize over the whole function $P(x_N)$. Consequently, we cannot view Problem 3 as a parameterized collection of subproblems for fixed x_N as we would like. We do not currently have a procedure for solving Problem 3 even approximately which is not NP-hard.

4. Numerical example. In the system shown in Figure 3.1, a pole is hinged on a cart, and a spring joins the top of the pole to a fixed point on the wall behind the cart. The control is a force on the cart, which is limited by a saturation constraint. We want a control design to drive the system to the origin from an initial condition. A simplified model of the dynamics has the form

$$\begin{bmatrix} \dot{\theta} \\ \dot{x}_L \end{bmatrix} = f(\theta) + A(\theta)x_L + g_u(\theta)u,$$

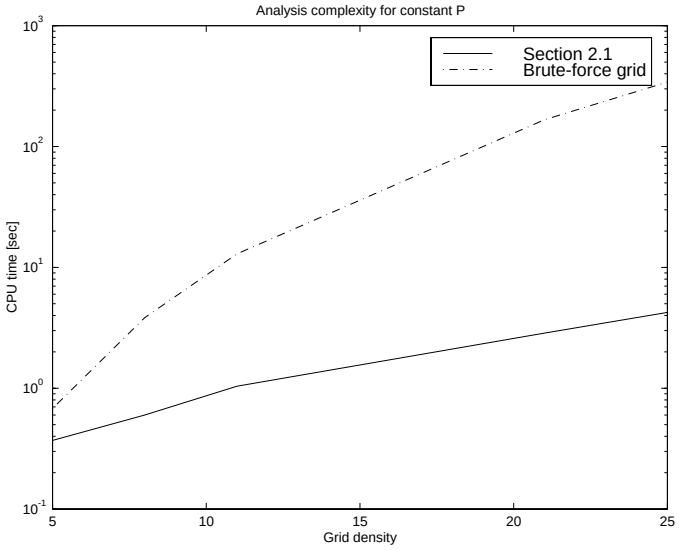


FIG. 4.1. *Computation time required for analysis with constant P using a Sparc* 20 *processor.*

where $x_N = \theta$ and $x_L = [\dot{\theta}; x; \dot{x}]$. After designing an RCLF $V(x)$ and stability margin $W(x)$, we analyze stability over $V^{-1}[0,1]$ under the saturation constraint using the method described in section 2.1. A comparison of the computational complexity of this procedure with that of a "brute-force" gridding procedure for different grid densities N_c is shown in Figure 4.1. This example shows the computational complexity increasing as a function of N_c with a higher exponent for the case of the brute-force grid than for the method of section 2.1. Specifically, for large N_c the complexity increases as N_c^4 for the brute-force grid and as N_c^2 for the method of section 2.1, as expected.

5. Conclusions. The most serious hindrance to progress on the nonlinear control problem is the inherent complexity of the class of arbitrary nonlinear systems. Therefore, it is important to identify a class of systems which is sufficiently restricted so that computations can be made tractable, yet which is general enough to apply to a wide variety of real control systems. We have shown some ways in which systems of the form (1.1) form such a class. When the dynamics are nonlinear in only a fixed number of variables x_N, the computation time required for stability analysis given a quadratic RCLF is exponential only in $\dim(x_N)$ and polynomial in the dimension of the remaining states. For an RCLF of the form $V(x) = x^T P(x_N) x$, the stability analysis problem is NP-hard. In practice, however, good bounds on the Lyapunov derivative can be found by converting the problem to a real μ analysis problem. This results in an approximate solution procedure with a computational complexity which is exponential in $\dim(x_N)$ and polynomial in the dimension of the remaining states.

Similarly, the problem of optimizing over RCLFs to get the largest possible stability region is NP-hard. We can get a computationally tractable, approximate solution procedure by finding an LFT representation for the system dynamics. The computational complexity of this problem is exponential only in $\dim(x_N)$. The computation time also depends polynomially on the dimension of the perturbation block needed to represent the nonlinear dynamics. Since this procedure relies on bounding the nonlinear terms in the dynamics and applying robust control techniques, we expect this

procedure to be overly conservative. We do not obtain a computationally tractable optimization procedure for RCLFs of the form $V(x) = x^T P(x_N)x$.

Appendix A. Proof of Theorem 2.3. Suppose that $C_0(x_N) \not< 0$. It is clear that $\Gamma(c_1, c_2, x_N, \psi) \leq \Gamma(0, c_2, x_N, \psi)$. Let $\lambda^* \in \mathcal{Z}(0, c_2, x_N, \psi)$ be such that $\Gamma(x_N, \psi, \lambda^*) = \Gamma(0, c_2, x_N, \psi)$. By the $\mathcal{S}$-procedure, this quantity is equal to the smallest r such that $\tau \geq 0$ exists to satisfy the following for *all* values of λ:

$$-\Gamma(x_N, \psi, \lambda) + r + \tau[V(x_N, \psi, \lambda) - c_2] \geq 0. \tag{A.1}$$

Therefore, $\tau[V(x_N, \psi, \lambda^*) - c_2] \geq 0$. On the other hand, the constraint and $\tau \geq 0$ combine to give $\tau[V(x_N, \psi, \lambda^*) - c_2] \leq 0$. Therefore, $\tau[V(x_N, \psi, \lambda^*) - c_2] = 0$. Since (A.1) must hold for all λ, we require $C_0(x_N) - \tau C_1(x_N) \leq 0$ for $C_1(x_N) = G(x_N)^T P_{LL} G(x_N) > 0$. Hence, if $C_0(x_N) \not\leq 0$, then we require $\tau > 0$, which implies $V(x_N, \psi, \lambda^*) = c_2$. If $C_0(x_N) \leq 0$ but $C_0(x_N) \not< 0$, then there exists $\mu \neq 0$ such that $C_0(x_N)\mu = 0$. We want to show that for any λ satisfying $V(x_N, \psi, \lambda) < c_2$, there exists $\beta \in R$ such that $\Gamma(x_N, \psi, \lambda + \beta\mu) \geq \Gamma(x_N, \psi, \lambda)$ but $V(x_N, \psi, \lambda + \beta\mu) = c_2$, and this will complete the proof. Now $V(x_N, \psi, \lambda + \beta\mu) = [\mu^T C_1(x_N)\mu]\beta^2 + [2\mu^T C_1(x_N)\lambda]\beta + V(x_N, \psi, \lambda)$. Therefore, the equation $V(x_N, \psi, \lambda + \beta\mu) = c_2$ is quadratic in β, and if $V(x_N, \psi, \lambda) < c_2$, there are two real roots: one positive and one negative. Since $\Gamma(x_N, \psi, \lambda + \beta\mu) = \Gamma(x_N, \psi, \lambda) + \beta b_0(x_N, \psi)^T \mu$, and at least one of the roots has the property $\beta b_0(x_N, \psi)^T \mu \geq 0$, we get $\Gamma(x_N, \psi, \lambda + \beta\mu) \geq \Gamma(x_N, \psi, \lambda)$ and $V(x_N, \psi, \lambda + \beta\mu) = c_2$. Therefore, there exists λ^* with the property $V(x_N, \psi, \lambda^*) = c_2$ and $\Gamma(x_N, \psi, \lambda^*) = \Gamma(0, c_2, x_N, \psi)$, and we conclude that $\Gamma(c_1, c_2, x_N, \psi) = \Gamma(0, c_2, x_N, \psi)$. Finally, suppose that $C_0(x_N) < 0$. Then $\Gamma(x_N, \psi, \lambda)$ is concave and is maximized for $\lambda^* = -\frac{1}{2} C_0(x_N)^{-1} b_0(x_N, \psi)$. If $\lambda^* \notin \mathcal{Z}(c_1, c_2, x_N, \psi)$, then $\Gamma(x_N, \psi, \lambda)$ is maximized on the boundary of $\mathcal{Z}(c_1, c_2, x_N, \psi)$ closest to λ^*. This proves the first two statements. Otherwise, $\Gamma(c_1, c_2, x_N, \psi) = \Gamma(x_N, \psi, \lambda^*) = a_0(x_N, \psi) - \frac{1}{4} b_0(x_N, \psi)^T C_0(x_N)^{-1} b_0(x_N, \psi)$, which proves the third statement.

Acknowledgment. The authors thank the anonymous reviewers for their helpful suggestions on improving this paper.

REFERENCES

[1] P. Apkarian and P. Gahinet, *A convex characterization of gain-scheduled H_∞ controllers*, IEEE Trans. Automat. Control, 40 (1995), pp. 853–864.
[2] Z. Artstein, *Stabilization with relaxed controls*, Nonlinear Anal., 7 (1983), pp. 1163–1173.
[3] S. Battilotti, *Stabilization via dynamic output feedback for systems with output nonlinearities*, Systems Control Lett., 23 (1994), pp. 411–419.
[4] G. Becker and A. Packard, *Robust performance of linear parametrically varying systems using parametrically-dependent linear feedback*, Systems Control Lett., 23 (1994), pp. 205–215.
[5] F. Blanchini, *Nonquadratic Lyapunov functions for robust control*, Automatica J. IFAC, 31 (1995), pp. 451–461.
[6] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, SIAM, Philadelphia, PA, 1994.
[7] R. P. Braatz, P. M. Young, J. C. Doyle, and M. Morari, *Computational complexity of μ calculation*, IEEE Trans. Automat. Control, 39 (1994), pp. 1000–1002.
[8] R. K. Brayton and C. H. Tong, *Constructive stability and asymptotic stability of dynamical systems*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 27 (1980), pp. 1121–1130.
[9] H.-D. Chiang and J. S. Thorp, *Stability regions of nonlinear dynamical systems: A constructive methodology*, IEEE Trans. Automat. Control, 34 (1989), pp. 1229–1241.
[10] E. J. Davison and E. M. Kurak, *A computational method for determining quadratic Lyapunov functions for non-linear systems*, Automatica J. IFAC, 7 (1971), pp. 627–636.

[11] L. El Ghaoui, *State feedback control of rational systems using linear-fractional representations and LMIs*, in Proc. American Control Conference, Baltimore, American Automatic Control Council, Evanston, IL, 1994, pp. 3563–3567.

[12] R. A. Freeman and P. V. Kokotović, *Tools and procedures for robust control of nonlinear systems*, in Proc. IEEE Conference on Decision and Control, Lake Buena Vista, FL, IEEE Computer Society Press, Los Alamitos, CA, 1994, pp. 3458–3463.

[13] R. A. Freeman and P. V. Kokotović, *Inverse optimality in robust stabilization*, SIAM J. Control Optim., 34 (1996), pp. 1365–1391.

[14] R. Genesio, M. Tartaglia, and A. Vicino, *On the estimation of asymptotic stability regions: State of the art and new proposals*, IEEE Trans. Automat. Control, 30 (1985), pp. 747–755.

[15] M. C. Lai and J. Hauser, *Computing maximal stability region using a given Lyapunov function*, in Proc. American Control Conference, San Francisco, American Automatic Control Council, Evanston, IL, 1993, pp. 1500–1502.

[16] Y. Lin, E. Sontag, and Y. Wang, *Recent results on Lyapunov-theoretic techniques for nonlinear stability*, in Proc. American Control Conference, Baltimore, 1994, pp. 1771–1775.

[17] Y. Lin and E. D. Sontag, *A universal formula for stabilization with bounded controls*, Systems Control Lett., 16 (1991), pp. 393–397.

[18] M. W. McConley, *A Computationally Efficient Lyapunov-Based Procedure for Control of Nonlinear Systems with Stability and Performance Guarantees*, Ph.D. thesis, Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, MA, 1997.

[19] M. W. McConley, B. D. Appleby, M. A. Dahleh, and E. Feron, *A control Lyapunov function approach to robust stabilization of nonlinear systems*, in Proc. American Control Conference, Albuquerque, IEEE Computer Society Press, Piscataway, NJ, 1997, pp. 329–333.

[20] A. N. Michel, N. R. Sarabudla, and R. K. Miller, *Stability analysis of complex dynamical systems*, Circuits Systems Signal Process., 1 (1982), pp. 171–202.

[21] P. Miotto, J. M. Shewchun, E. Feron, and J. D. Paduano, *High performance bounded control synthesis with application to the* F18 *HARV*, in AIAA Guidance, Navigation, and Control Conference, San Diego, American Institute of Aeronautics and Astronautics, Washington, DC, 1996.

[22] K. G. Murty and S. N. Kabadi, *Some NP-complete problems in quadratic and nonlinear programming*, Math. Programming, 39 (1987), pp. 117–129.

[23] Y. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*, SIAM, Philadelphia, PA, 1994.

[24] Y. Ohta et al., *Computer generated Lyapunov functions for a class of nonlinear systems*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 40 (1993), pp. 343–354.

[25] A. Packard, *Gain scheduling via linear fractional transformations*, Systems Control Lett., 22 (1994), pp. 79–92.

[26] D. N. Shields and C. Storey, *The behaviour of optimal Lyapunov functions*, Internat. J. Control, 21 (1975), pp. 561–573.

[27] E. D. Sontag, *A 'universal' construction of Artstein's theorem on nonlinear stabilization*, Systems Control Lett., 13 (1989), pp. 117–123.

[28] G. Strang, *Introduction to Applied Mathematics*, Wellesley-Cambridge Press, Wellesley, MA, 1986.

[29] J. Tsinias, *Versions of Sontag's "input to state stability condition" and the global stabilizability problem*, SIAM J. Control Optim., 31 (1993), pp. 928–941.

[30] M. Vidyasagar, *Nonlinear Systems Analysis*, 2nd ed., Prentice–Hall, Englewood Cliffs, NJ, 1993.

[31] P. M. Young, *Controller design with mixed uncertainties*, in Proc. American Control Conference, Baltimore, American Automatic Control Council, Evanston, IL, 1994, pp. 2333–2337.

[32] P. M. Young, M. P. Newlin, and J. C. Doyle, *Practical computation of the mixed μ problem*, in Proc. American Control Conference, Chicago, 1992, pp. 2190–2194.

SIAM J. CONTROL OPTIM.
Vol. 36, No. 6, pp. 2194–2195, November 1998

018

CORRIGENDUM: FEEDBACK STABILIZATION OVER COMMUTATIVE RINGS: THE MATRIX CASE*

V. R. SULE†

PII. S0363012997319668

In a previous paper [1] certain arguments in the proofs of Proposition 1, Proposition 4, and subsection 4.4 are erroneous. These errors are highly regretted by this author and are corrected as follows. These corrections, however, do not affect the main results of the paper.

1. In Definition 1 (of a strictly causal matrix) the second sentence should be corrected as "*A causal matrix M is called strictly causal if all entries of M are in $\mathcal{J}$.*"

2. In the sufficiency part of the proof of Proposition 1, the two sentences starting in line 3 of the paragraph "Since $I_t(N) \subseteq \mathcal{J}$, using ... units of $R^{-1}\mathcal{A}$," should be replaced by the following:

"From $NY = (I - X)d$ and the assumption that all entries of P are in $\mathcal{J}$ (and hence that of N, since d is nonzerodivisor), it follows that all entries of $I - X$ are in $\mathcal{J}$. Hence $\det X = \det(I - (I - X))$ is of the form $1 + j$ where j is in $\mathcal{J}$. This implies that $\det X$ is a unit of $R^{-1}\mathcal{A}$ as $\mathcal{J}$ is the Jacobson's radical."

3. The proof of the necessity part of Proposition 4 is erroneous. Since this result is not needed further in the paper. We replace it with the following version stating only the sufficient condition.

PROPOSITION 4. *A strictly causal $P = Nd^{-1}$ in $(\mathcal{F})_n$ is stabilizable if*

$$a + b = \mathcal{A}.$$

The proof should be read from the third paragraph of the existing proof after deleting the line "Second, the proof is by sufficiency."

4. The arguments made in subsection 4.4 are erroneous. This subsection should be replaced by the following:

"**4.4. Relaxing strict causality.** We now show that for integral domains the strict causality of P can be replaced by a more concrete and weaker condition (than strict causality), which makes Proposition 1 stronger. (Recall that the strict causality of P is required in the necessity part of the proof of Proposition 1. This condition is as follows.)

Now let $\mathcal{A}$ be an integral domain and P satisfy the condition that the equation $\det(I - PY) = 0$ has no solution Y over $\mathcal{A}$.

Since the solutions of equation (1) are invariant with respect to the choice of fractions of P, choose a pair of fractions N and $d \neq 0$. Now, as $X = I - PY$ is equivalent to $NY = (I - X)d$, it follows from the above condition that if X is a solution of (1) then $\det X \neq 0$. Thus every solution of (1), satisfies this condition. Note further that this condition is not necessary for existence of solutions X with

*Received by the editors April 11, 1997; accepted for publication (in revised form) November 4, 1997; published electronically September 3, 1998.
http://www.siam.org/journals/sicon/36-6/31966.html
†Department of Electrical Engineering, Indian Institute of Technology, Kanpur 208016, India (vrs@iitk.ernet.in).

nonzero determinants. For instance, P of example 2 does not satisfy this condition but is still stabilized by the controller $C = [0\ X_1^2+X_2^2+X_3^2]$, as can be easily checked."

Acknowledgment. The author is thankful to Kazuyoshi Mori for his query on item 2 above.

REFERENCES

[1] V. R. Sule, *Feedback stabilization over commutative rings: The matrix case*, SIAM J. Control Optim., 32(1994), pp. 1675–1695.